YU INFO 2025

Informaciono društvo Srbije

31. NACIONALNA KONFERENCIJA

u oblasti
informaciono-komunikacionih tehnologija

YU INFO 2025

ZBORNIK RADOVA

9 - 12. mart 2025. godine

Kongresni centar
hotela „GRAND“

Kopaonik, SRBIJA

decembar 2025.

ISBN: 978-86-85525-33-9

Zbornik radova 31. IKT konferencije

“YU INFO 2025”

Urednici: Prof. dr Miodrag Ivković (Informaciono društvo Srbije),
Prof. dr Dražen Drašković (Informaciono društvo Srbije,
Elektrotehnički fakultet Univerziteta u Beogradu)

Tehnička priprema: Prof. dr Dražen Drašković (Informaciono društvo Srbije,
Elektrotehnički fakultet Univerziteta u Beogradu)

Izdavač: Informaciono društvo Srbije

Datum izdanja: 1. decembar 2025.

Datum konferencije: 9-12. mart 2025. godine

Sajt konferencije: http://www.yuinfo.org

CIP – Каталогизација у публикацији
Народна библиотека Србије, Београд

004 Рачунарство. Рачунарска техника.
004.4 Рачунарски софтвер
004.8 Вештачка интелигенција
005 Менаџмент

ISBN 978-86-85525-33-9

Sadržaj

Reč dobrodošlice:
YU INFO - Ulazak u četvrtu deceniju

Poštovane koleginice i kolege,
dragi učesnici konferencije YU INFO,

sa velikim zadovoljstvom vas pozdravljamo nakon završetka 31. konferencije YU INFO, kojom simbolično ulazimo u četvrtu deceniju njenog postojanja. Nakon tri decenije kontinuiranog razvoja, rasta i prilagođavanja savremenim tehnološkim i društvenim izazovima, YU INFO danas predstavlja jedno od najznačajnijih mesta koje okupljanja akademsku zajednicu, istraživače, inženjere i stručnjake iz IKT industrije iz zemlje i regiona.

Konferencija, koja je započela svoje putovanje 1995. godine, ostala je dosledna svojoj misiji - razmeni znanja, ideja i iskustava, kao i povezivanju nauke, obrazovanja i privrede. Ulazak u četvrtu deceniju obeležavamo sa jasnom vizijom budućnosti, oslanjajući se na bogato iskustvo prethodnih godina i spremno prihvatajući nove tehnološke paradigme koje oblikuju digitalno društvo.

I ove godine nastavljen je trend velikog interesovanja autora i visokog kvaliteta prijavljenih radova, koji obuhvataju širok spektar savremenih tema iz oblasti informaciono-komunikacionih tehnologija. Radovi su organizovani u tematske oblasti koje prate aktuelne istraživačke i industrijske tokove, uključujući veštačku inteligenciju i mašinsko učenje, softversko inženjerstvo, računarske mreže i bezbednost, baze podataka i informacione sisteme, industriju 4.0, industrijsku automatizaciju i digitalnu transformaciju, kao i specifične primene IKT-a u različitim sektorima.

Zajedničko održavanje 31. konferencije YU INFO, sa jubilarnim 15. izdanjem međunarodne konferencije ICIST na Kopaoniku, predstavlja još jedan korak ka jačanju međunarodne vidljivosti i naučne relevantnosti oba skupa. Uvereni smo da će ova sinergija doprineti dodatnom unapređenju kvaliteta programa, većem broju učesnika i još sadržajnijem naučnom i društvenom iskustvu.

Zahvaljujemo se svim autorima, recenzentima, partnerima i učesnicima koji svojim doprinosom čine da YU INFO nastavi da raste i razvija se. Radujemo se zajedničkom radu, razmeni ideja i novim inicijativama i vidimo se ponovo u martu 2026. godine, na Kopaoniku.

Srdačno vas pozdravlja,

Organizacioni odbor konferencije „YU INFO 2025“

Programski i organizacioni odbor konferencije

Prof. dr Miodrag Ivković, Univerzitet u Novom Sadu

Prof. dr Dražen Drašković, Univerzitet u Beogradu

Prof. dr Milan Zdravković, Univerzitet u Nišu

Prof. dr Zora Konjović, Univerzitet “Singidunum” Beograd

Prof. dr Jelica Protić, Univerzitet u Beogradu

† Prof. dr Miroslav Trajanović, Univerzitet u Nišu

Sponzori i prijatelji konferencije

MEDIJSKI PARTNER

Program konferencije

Datum	Dan 1. 9.3.2025. nedelja	Dan 2. 10.3.2025. ponedeljak			Dan 3. 11.3.2025. utorak			Dan 4. 12.3.2025. sreda	
Satnica	Sale „Pančić 1 i 2“	Sala „Pančić 1“	Sala „Pančić 2“	Sala „Sunčani vrhovi“	Sala „Sunčani vrhovi“	Sala „Pančić 1“	Sala „Pančić 2“	Sala „Pančić 1“	Sala „Pančić 2“
9:00		YU #1: Digitalizacija i e-zdravstvo						YU #6: Vojne simulacije i primene	
9:30									
10:00									
10:30									
11:00									
11:30									
12:00									
12:30									
13:00									
13:30			ICIST - Special track: Autonomous driving						
14:00									
14:30									
15:00				YU #2: Veštačka inteligencija					
15:30	Registracija (lobi ispred Pančić 1 i 2)	ICIST - Special track: Digital water	ICIST - Generative AI and Large Language Models - Session 1		YU #4: Softverski sistemi i savremene informacione tehnologije	ICIST - Software engineering	ICIST - Special track: ICT for health, well-being and sport, Session 1		
16:00									
16:30									

nastavak satnice na narednoj stranici

<table>
<tr><td>Datum</td><td>Dan 1.
9.3.2025.
nedelja</td><td colspan="3">Dan 2.
10.3.2025.
ponedeljak</td><td colspan="3">Dan 3.
11.3.2025.
utorak</td><td colspan="2">Dan 4.
12.3.2025.
sreda</td></tr>
<tr><td>Satnica</td><td>Sale
„Pančić 1 i 2“</td><td>Sala
„Pančić 1“</td><td>Sala
„Pančić 2“</td><td>Sala
„Sunčani vrhovi“</td><td>Sala
„Sunčani vrhovi“</td><td>Sala
„Pančić 1“</td><td>Sala
„Pančić 2“</td><td>Sala
„Pančić 1“</td><td>Sala
„Pančić 2“</td></tr>
<tr><td>17:00</td><td rowspan="5">Otvaranje konferencije (pozdravne reči organizatora)

Glavna predavanja: (sala Pančić 1)</td><td rowspan="3">Panel:
Šta menadžment očekuje, a šta ima od primene veštačke inteligencije?</td><td rowspan="3">ICIST-Poster session (lobi ispred Pančić 1 i 2)</td><td></td><td rowspan="3">YU #4</td><td></td><td rowspan="3">ICIST - Generative AI and Large Language Models, Session 2</td><td></td><td></td></tr>
<tr><td>17:30</td><td></td><td rowspan="2">Glavno predavanje</td><td></td><td></td></tr>
<tr><td>18:00</td><td>AI i regulativa</td><td rowspan="3">YU #3: Mašinsko učenje i veliki jezički modeli</td><td rowspan="3">YU #5: Informacioni sistemi, računarske simulacije i edukacija</td><td rowspan="3">ICIST - AI & IoT for Smart Industry, Session 2</td><td></td><td></td></tr>
<tr><td>18:30</td><td>IBM Guardium (SBS lecture)</td><td rowspan="3">ICIST - AI & IoT for Smart Industry, Session 1</td><td rowspan="3">Special track: ICT for health, well-being and sport, Session 2</td><td></td><td></td></tr>
<tr><td>19:00</td><td></td><td></td><td></td></tr>
<tr><td>19:30</td><td>Veče dobro-došlice

WINE SESSION (sala Pančić 2)

VINARIJA Cvetković Aleksan-drovac</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>20:00</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>20:45</td><td colspan="3" rowspan="3">Konferencijska žurka - „Rok veče“
(Madicine bend @Pub & Ko Konaci)</td><td colspan="4" rowspan="3">Sponzor večeri dobrodošlice:
Serbian Business Systems

Sponzor žurke: SAGA
A NOVENTIQ Company</td></tr>
<tr><td>21:00</td></tr>
<tr><td>21:30</td></tr>
</table>

Nedelja, 9. mart 2025.

17^{00} Otvaranje konferencije sala „Pančić 1“

- **Pozdravne reči organizatora konferencija YU INFO i ICIST 2025**
- **Glavna predavanja (*Keynotes*):**
  - **Pawel Herman (Division of Computational Science and Technology, School of Electrical Engineering and Computer Science, KTH and Digital Futures)**

    Tema: Neuromorphic computing
  - **Miroslav Trajković (Zebra Technologies Corporation, USA)**

    Tema: Computer vision

Sesije / tematske oblasti konferencije “YU INFO 2025”

Sesija 1: Digitalizacija i e-zdravstvo

Sesija 2: Veštačka inteligencija

Sesija 3: Mašinsko učenje i veliki jezički modeli

Sesija 4: Softverski sistemi i savremene informacione tehnologije

Sesija 5: Informacioni sistemi, računarske simulacije i edukacija

Sesija 6: Vojne simulacije i primene

Napomena: Radovi označeni oznakom “A” u polju “Broj strana”, nisu predati u finalnu verziju Zbornika radova 31. IKT konferencije “YU INFO 2025” (kategorija radova M63), već su poslati samo kao apstrakti, koji su prezentovani bez slanja finalnog rada u celosti.

Sesija YU #1: Digitalizacija i e-zdravstvo		Br. str.
ponedeljak, 10.3.2025, 09:00-11:00, sala „Pančić 1“		14
Moderator:	**Dr Anđelka Zečević (Matematički institut SANU)**	

Sesija YU #2: Veštačka inteligencija		Br. str.
ponedeljak, 10.3.2025, 15:00-17:00, sala „Sunčani vrhovi“		38
Moderatori:	Prof. dr Boris Delibašić (Univerzitet u Beogradu, Fakultet organizacionih nauka)	

R.br.	Naslov rada	Autori	Br. strana
2.01.	Neuromorfno računarstvo i tečne mašine stanja u sistemima za otkrivanje upada nad CIC-IDS2017 skupom podataka	**Živadinović, Miloš**; Simić, Dejan	39 - 47
2.02.	Digitalna transformacija u šumarstvu – primena TinyML tehnologije i veštačke inteligencije na ivici	**Pavlović, Dejan**	48 - 52
2.03.	Sistemi za podršku odlučivanju u eri veštačke inteligencije	**Delibašić, Boris**; Radovanović, Sandro; Suknović, Milija	53 - 55
2.04.	Primjena mašinskog učenja za opisivanje slika pomoću teksta	**Mićić, Aleksije;** Đurić, Zoran	56 - 61
2.05.	Predviđanje uspešnosti memorizacije slike pomoću modela mašinskog učenja	**Matvejev, Valerijan;** Drašković, Dražen	62 - 65
2.06.	Evolucija veštačke inteligencije, izazovi i novi trendovi u zdravstvu	**Terzić, Rajko**; Majstorović, Milosav; Pantović, Vladan; Terzić, Dušan	66 - 70
2.07.	Energetska efikasnost pristupne mobilne mreže Telekom Srbija a.d.	**Aleksić, Danijela**	71 - 74

Sesija YU #3: Mašinsko učenje i veliki jezički modeli		Br. str.
ponedeljak, 10.3.2025, 18:00-20:00, sala „Sunčani vrhovi“		75
Moderatori:	**Prof. dr Dražen Drašković (Univerzitet u Beogradu, Elektrotehnički fakultet)**	

R.br.	Naslov rada	Autori	Br. strana
3.01.	Snimanje bilingvalne baze AI-SPEAK za multimodalno prepoznavanje govora	**Nosek, Tijana**; Suzić, Siniša; Stanojev, Vuk; Jakovljević, Nikša; Krstanović, Lidija; Sečujski, Milan	76 - 79
3.02.	Analiza indeksa ljudskog razvoja i njegovih komponenti korišćenjem metoda mašinskog učenja	**Radojičić, Dragana**; Stamenković, Mladen	80 - 83
3.03.	Integracija velikih jezičkih modela u obrazovanju: Percepcije studenata kroz praktične primene	**Ivankovic, Zdravko**; Pecev, Predrag; Sudar, Sasa; Damnjanovic, Jasmina; Ivankovic, Milana	84 - 88
3.04.	Uticaj augmentacija trening podataka na performanse doobučenog Whisper modela	**Suzić, Siniša**; Nosek, Tijana; Simić, Nikola; Pekar, Darko; Delić, Vlado	89 - 92
3.05.	Prikupljanje podataka i označavanje u procesu formiranja skupova podataka govora mržnje na srpskom jeziku	**Drašković, Dražen;** Radenković, Uroš; Mićović, Marko; Cincović, Jelica; Milaković, Adrian; Jocović, Vladimir;	93 - 98
3.06.	Primena veštačke inteligencije u analizi i obradi govora kroz transkripciju, verifikaciju i evaluaciju izjava	Mandić, Ana; Jelović, Marko; Bulatović, Ana; Nikolić, Filip; Mijatović, Hana; Vučićević, Nikola; Stanić, Ana; Bogdanović, Natalija; Lazović, Luka; Mihajlov, Anja; Popović, Lana; Milaković, Adrian; Jocović, Vladimir; **Drašković, Dražen**	99 - 103

Sesija YU #4: Softverski sistemi i savremene informacione tehnologije		Br. str.
utorak, 11.3.2025, 16:00-18:00, sala „Sunčani vrhovi“		104
Moderatori:	**Prof. dr Marko Mišić (Univerzitet u Beogradu, Elektrotehnički fakultet)**	

R.br.	Naslov rada	Autori	Br. strana
4.01.	Vremenska analiza procesa integracije primenom Simpsonovog pravila na procesorima sa više jezgara u Javi	Smilić, Marko; Stefanović, Časlav; Milić, Dejan; **Djurović, Sanja**; Đošić, Danijel	105 - 108
4.02.	Uporedna analiza radnih okvira Angular, React i Flutter na osnovu implementacije aplikacije za poslastičarnicu	**Tufegdžić, Janko**; Hrvačević, Luka; Potkonjak, Iva; Mutavdžić, Uroš; Cincović, Jelica; Punt, Marija	109 - 113
4.03.	Pregled pozicionih algoritama u proširenoj i virtuelnoj stvarnosti	**Ogrizović, Mihajlo**; Janković, Filip; Ilić, Aleksa; Drašković, Dražen	114 - 119
4.04.	Analiza performansi različitih algoritama broadcast komunikacije u Open MPI biblioteci	**Nastić, Miloš**; Mišić, Marko; Vuletić, Pavle; Protić, Jelica	120 - 125
4.05.	Analiza algoritama dokaza bez znanja u blokčejn tehnologiji	**Obradović, Miloš**; Vuletić, Pavle	126 - 131
4.06.	Bezbednosna kultura u razvoju novih servisa: Organizacione mere i edukacija zaposlenih	Divac, Jelena; Marenović, Neda	A
4.07.	Fiksni bežični pristup (FWA) – tehnologije, primena i perspektive	**Nemec, Dejan**	132 - 137
4.08.	Evolucija RAN arhitektura: Od distribuiranih sistema do naprednih Open RAN modela u eri 5G	**Stefanović, Katarina**	138 - 143
4.09.	Prednosti i izazovi primene prenosnih baznih stanica mobilne telefonije	**Nemec, Dejan**	144 - 149

Sesija YU #5: Informacioni sistemi, računarske simulacije i edukacija		Br. str.
utorak, 11.3.2025, 18:00-20:00, sala „Sunčani vrhovi“		150
Moderatori:	**Prof. dr Marija Punt (Univerzitet u Beogradu, Elektrotehnički fakultet)**	

Sesija YU #6: Vojne simulacije i primene		Br. str.
sreda, 12.3.2025, 9:00-11:30, sala „Pančić 1“		183
Moderatori:	**Prof. dr Miodrag Ivković** **(Univerzitet u Novom Sadu, Fakultet tehničkih nauka)**	

R.br.	Naslov rada	Autori	Br. strana
6.01.	Realizacija autonomnog kretanja na besposadnoj platformi	**Pavlovic, Rade**; Mitričević, Nina	184 - 188
6.02.	Koncept Android aplikacije za razmenu šifrovanih i digitalno potpisanih poruka elektronske pošte	Pavlovic, Filip; Dimitrijevic, Nenad; Milovanovic, Nikola; Citic, Nikola	A
6.03.	Artificial Intelligence and Military Command Information Systems: Current State and Expected Challenges	Vulić, Ivan	A
6.04.	Analiza tačnosti identifikacije parametara funkcije prenosa kanala propinjanja rakete malog dometa primenom iterativnih metoda	Stošić, Darko	A
6.05.	Merenje intenziteta svetlosti, dometa svetlosnog snopa i provera zaptivenosti ručne taktičke lampe	**Jovanović, Milena;** Tripković, Marina	189 - 191
6.06.	Termovizijsko praćenje toplotnih efekata borbene opreme pri različitim fizičkim aktivnostima korisnika	**Tripković, Marina;** Jovanović, Milena	192 - 195
6.07.	Analiza podataka o saobraćajnim nezgodama sa socio-ekonomskog aspekta korišćenjem sistema poslovne inteligencije	**Atanasijevic, Jordan**; Đukic, Dejan; Tot, Ivan	196 - 200
6.08.	Primena CMMN notacije u modelovanju dinamičnih poslovnih procesa i upravljanju slučajevima	**Atanasijevic, Jordan**; Viduka, Dejan; Tot, Ivan	201 - 206
6.09.	Interpolacija kubnim splajnom korišćenjem programskog paketa *Wolfram Mathematica*	**Atanasijevic, Jordan**; Đukic, Dejan; Tot, Ivan	207 - 212
6.10.	Predikcija efikasnog položaja radio-ometača primenom algoritama veštačke inteligencije	**Pejčić, Ognjen**; Šepa, Nemanja; Sazdić-Jotić, Boban	A
6.11.	Mogućnost primene veštačke inteligencije za poboljšanje mentalne higijene u sistemu odbrane	**Gajić, Tamara**; Stanković, Dragan; Stajković, Nenad; Belotić, Branislav; Vasić, Nikola	A

YU #1: Sesija 1

Digitalizacija i e-zdravstvo

Evaluacija statističkih metoda imputacije gena u podacima prostorne transkriptomike

Lazar Smiljković
Elektrotehnički fakultet
Beograd, Srbija
lazarsmiljkovic@etf.bg.ac.rs
0009-0004-8168-1924

Marko Mišić
Elektrotehnički fakultet
Beograd, Srbija
marko.misic@etf.bg.ac.rs
0000-0002-7369-4010

Vladimir Kovačević
BGI Research
Beograd, Srbija
vladimirkovacevic@genomics.cn
0000-0002-9843-6261

***Apstrakt* – Prostorna transkriptomika (ST) omogućava analizu ekspresije gena uz očuvanje prostorne informacije u tkivima, pružajući jedinstveni uvid u prostornu organizaciju ćelijskih tipova. Međutim, ova tehnologija često pati od problema *drop-out* efekta, gde određeni geni nisu detektovani zbog tehničkih ograničenja, a ne zato što zaista nisu prisutni u samim ćelijama. U ovom radu predstavljamo metod imputacije gena u ST podacima zasnovan na prethodno anotiranim ćelijskim tipovima pomoću CoDi ili nekog sličnog algoritma. Ovaj pristup koristi informacije iz komplementarnih *single-cell* RNA-seq (scRNA-seq) podataka za imputaciju nedostajućih vrednosti ekspresije unutar specifičnih ćelijskih tipova, očuvavajući biološku konzistentnost. Evaluacija metode na četiri različita skupa podataka pokazuje značajno poboljšanje u retenciji marker gena, sa povećanjem do 21.67% za jedinstvene marker gene i do 20.43% za top 100 marker gene. Ovi rezultati pokazuju potencijalnu efikasnost predloženog pristupa u rekonstrukciji pouzdanijih i biološki relevantnijih podataka prostorne transkriptomske i daje prostor za dalje usavršavanje metode i buduća istraživanja.**

Ključne reči – prostorna transkriptomika, imputacija gena, drop-out efekat, ćelijska anotacija, marker geni

I. Uvod

Prostorna transkriptomika predstavlja revolucionarni napredak u oblasti genomike, omogućavajući analizu ekspresije gena uz očuvanje prostornih informacija unutar tkiva [1]. Ovo otvara nove mogućnosti za razumevanje prostorne organizacije ćelijskih tipova, njihovih interakcija i diferencijalne genske ekspresije u kontekstu arhitekture tkiva [2].

I pored velikog potencijala, ST tehnologije se suočavaju sa specifičnim izazovima koji mogu ograničiti biološku analizu. Jedan od najznačajnijih problema je tzv. *drop-out* efekat [3], gde određeni geni nisu detektovani u podacima iako su zapravo ekspresovani u ćelijama. Ovo rezultira nepotpunim podacima sa visokim procentom nultih vrednosti, što može dovesti do pogrešnih bioloških zaključaka. Dodatno, većina ST tehnologija ima nižu gensku rezoluciju u poređenju sa tehnologijama sekvenciranja pojedinačnih ćelija (scRNA-seq), što dodatno otežava identifikaciju ćelijskih tipova i analizu njihovih funkcija [4].

Postojeći pristupi analizi ST podataka obično uključuju anotaciju ćelijskih tipova. Za anotaciju ćelijskih tipova razvijeno nekoliko naprednih metoda poput CoDi [5], Cell2location [6], Tangram [7], CytoSpace [8], RCTD [9] i Seurat [10]. S druge strane, postoje pokušaji poboljšavanja ST podataka imputacijom (dopunom) nedostajućih vrednosti ekspresije gena. Problem imputacije nedostajućih vrednosti ekspresije gena nije dobio preveliku pažnju istraživačke zajednice, posebno kada se uzme u obzir i očuvanje svojstava ćelijskih tipova. Ovaj rad se fokusira upravo na taj dodatan korak u obradi ST podataka, razvijajući metod za imputaciju gena koji uzima u obzir specifične karakteristike različitih ćelijskih tipova, čime značajno doprinosimo poboljšanju kvaliteta i tumačenju ST podataka.

Rad je podeljen na nekoliko poglavlja. Drugo poglavlje detaljno opisuje tehnologije prostornog sekvenciranja i *single-cell* RNA sekvenciranja, njihove prednosti i ograničenja, kao i komplementarnost ovih pristupa. U trećem poglavlju predstavljena je metoda imputacije gena zasnovana na CoDi anotaciji i statistički pristup zasnovan na klasterima, uz objašnjenje načina za očuvanje biološke tačnosti. Četvrto poglavlje opisuje korišćene skupove podataka iz različitih tehnologija sekvenciranja i metrike evaluacije zasnovane na retenciji marker gena. U petom poglavlju su prikazani rezultati evaluacije na četiri različita skupa podataka, uz detaljnu diskusiju uočenih obrazaca i poređenje sa drugim pristupima. U poslednjem, šestom poglavlju, dat je zaključak rada uz razmatranje ograničenja trenutnog pristupa i predloženim pravcima budućeg istraživanja.

II. Tehnologije sekvenciranja i izazovi

A. Pregled tehnologija prostornog sekvenciranja

Područje prostorne transkriptomike brzo se razvija, sa nekoliko ključnih tehnologija koje omogućavaju različite nivoe prostorne rezolucije .

10x Genomics Visium: Omogućava profilisanje gena sa prostornom rezolucijom od oko 55 μm, gde svaka tačka obično sadrži 1-10 ćelija. Ova tehnologija je relativno pristupačna i često se koristi, ali ima ograničenu rezoluciju na nivou pojedinačnih ćelija.

Slide-seq: Pruža veću prostornu rezoluciju (oko 10 μm) koristeći DNA-kodirane kuglice, što omogućava analizu na nivou pojedinačnih ćelija [11]. Slide-seq V2 poboljšava osetljivost u odnosu na prvu verziju.

Stereo-seq: Tehnologija sa vrlo visokom rezolucijom (subćelijska rezolucija od 500-700 nm), koja omogućava precizno mapiranje genske ekspresije na nivou pojedinačnih ćelija [12].

MERFISH i *seqFISH+*: In situ tehnologije koje direktno detektuju RNA molekule u fiksiranom tkivu koristeći fluorescentne sonde, omogućavajući subćelijsku rezoluciju.

U našem istraživanju smo koristili podatke dobijene iz nekoliko različitih tehnologija prostornog sekvenciranja, kako bismo dobili širu sliku efikasnosti naše metode

imputacije. Raznovrsnost tehnologija u analizi omogućila nam je da testiramo stabilnost naše metode imputacije u različitim uslovima, sa različitim nivoima drop-out efekta, prostornom rezolucijom i brojem detektovanih gena. Konzistentni pozitivni rezultati na svim platformama potvrđuju široku primenljivost našeg pristupa, bez obzira na specifične karakteristike tehnologije prostornog sekvenciranja.

B. Single-cell RNA sekvenciranje

Single-cell RNA sekvenciranje (scRNA-seq) omogućava merenje ekspresije gena na nivou pojedinačnih ćelija, što predstavlja značajan napredak u odnosu na tradicionalne bulk RNA-seq metode. Ova tehnologija je unapredila naše razumevanje ćelijske heterogenosti, diferencijacije i funkcije u različitim biološkim kontekstima. scRNA-seq ima nekoliko ključnih prednosti koje ga čine idealnim komplementarnim pristupom za prostornu transkriptomiku [13].

ScRNA-seq obično detektuje mnogo više gena po ćeliji, u poređenju sa prostornim tehnologijama, što omogućava detaljniju opisivanje ćelijskih tipova. Zbog direktne izolacije ćelija, scRNA-seq ima manji stepen drop-out efekta, što dovodi do pouzdanijih podataka o ekspresiji gena. Takođe, ova tehnologija pruža veću dubinu sekvenciranja, što omogućava detekciju gena sa niskom ekspresijom, koji mogu biti ključni za definisanje ćelijskih tipova. Moderne metode omogućavaju analizu velikog broja ćelija, čime se omogućava bolje razumevanje retkih ćelijskih populacija. Za razliku od nekih prostornih tehnologija, gde jedna tačka može sadržati više ćelija, scRNA-seq analizira izolovane pojedinačne ćelije, čime se eliminiše kontaminacija podataka iz drugih ćelija. Takođe, zbog dugogodišnje primene u istraživanjima, scRNA-seq ima razvijene analitičke protokole i alate.

Iako scRNA-seq nudi superiornu gensku rezoluciju i manja tehnička ograničenja, on ima jedan ključni nedostatak - gubitak prostorne informacije. Tokom pripreme uzoraka, ćelije se izdvajaju iz tkiva, što dovodi do gubitka podataka o njihovom originalnom prostornom položaju i mikrookruženju . Zbog ove razlike, scRNA-seq i prostorna transkriptomika se savršeno dopunjuju.

scRNA-seq pruža detaljan i precizan profil ekspresije gena pojedinačnih ćelija, ali bez prostornog konteksta. S druge strane, prostorna transkriptomika omogućava analizu genske ekspresije u prostornom kontekstu, ali sa nižom genskom rezolucijom i većim drop-out efektom. Integracija ovih podataka omogućava bolje razumevanje složene organizacije tkiva. Korišćenjem scRNA-seq podataka za imputaciju nedostajućih vrednosti u prostornim podacima, možemo značajno poboljšati kvalitet prostornih podataka, a da pri tome ne izgubimo ključnu prostornu informaciju.

Ova integracija predstavlja osnovu našeg pristupa imputaciji gena u prostornim podacima, gde koristimo detaljne i tačne scRNA-seq profile za rekonstrukciju nepotpunih prostornih podataka, kombinujući prednosti obe tehnologije [14].

C. Drop-out efekat i ostali izazovi

Drop-out efekat predstavlja jedan od najozbiljnijih izazova u analizi prostornih transkriptomskih podataka. Ova pojava se dešava kada gen, iako ekspresovan u ćeliji, nije detektovan zbog tehničkih ograničenja, što dovodi do netačno negativnih vrednosti u matrici ekspresije. Problem je posebno izražen u ST tehnologijama, pre svega zbog procesa obrade uzoraka koji mogu uzrokovati gubitak ili oštećenje RNK molekula. Pored toga, ST platforme imaju smanjenu osetljivost u poređenju sa tehnologijama sekvenciranja pojedinačnih ćelija (scRNA-seq), što dodatno otežava detekciju ekspresije. Još jedan faktor koji doprinosi prisustvu drop-out efekta jeste ograničen broj očitavanja po ćeliji, usled tehničkih limita dubine sekvenciranja u prostornim tehnologijama.

Analiza ST podataka dodatno je otežana zbog prisustva više ćelija u okviru jedne prostorne tačke, što može izazvati mešanje signala i otežati preciznu identifikaciju ćelijskih tipova. Tehničke varijacije između eksperimenata, usled razlika u pripremi uzoraka, uslovima sekvenciranja ili koracima obrade podataka, mogu dovesti do pojave nepravilnosti i smanjiti pouzdanost rezultata. Takođe, postoji jasan kompromis između prostorne rezolucije i dubine ekspresije – tehnologije koje omogućavaju višu prostornu preciznost često dolaze sa smanjenim brojem detektovanih gena i nižom dubinom.

Rešavanje problema drop-out efekta je ključno za dobijanje kvalitetnih i razumljivih prostornih transkriptomskih podataka. Pouzdana identifikacija ćelijskih tipova direktno zavisi od tačne ekspresije marker gena, koji često ostaju neotkriveni zbog tehničkih ograničenja, čime se narušava celokupna analiza. Nepotpuni podaci mogu sakriti važne biološke obrasce u raspodeli ekspresije gena kroz tkivo, čime se gubi prostorni kontekst koji je ključan za razumevanje funkcionalne organizacije ćelijskih populacija.

Tačne vrednosti ekspresije su osnov za niz analitičkih koraka, uključujući identifikaciju diferencijalno eksprimiranih gena i funkcionalnih puteva, čime se značajno utiče na kvalitet svih naknadnih (downstream) analiza. Dodatno, preciznija i potpunija ekspresiona matrica omogućava kvalitetniju integraciju ST podataka sa drugim vrstama omics informacija, poput prostorne proteomike ili nuklearno izolovanih transkriptomskih podataka (snRNA-seq).

Iz tog razloga, razvoj metoda za imputaciju nedostajućih vrednosti koje uspešno prevazilaze tehnička ograničenja ST tehnologija predstavlja važan korak ka unapređenju analize prostornih podataka i boljem razumevanju složenih bioloških procesa koji se odvijaju u tkivima.

III. METODOLOGIJA

Prvi korak u obradi ST podataka je anotacija ćelijskih tipova koja može da se obavi nekim od dostupnih alata. U našem istraživanju koristimo CoDi alat, ali se anotacije mogu dobiti i na drugi način. Stoga će u ovom poglavlju najpre biti opisan način za anotaciju ćelijskih tipova zasnovan na CoDi metodi, zatim predloženi statistički metod za imputaciju gena, korišćeni skupovi podataka i metrike za evaluaciju.

A. CoDi anotacija ćelijskih tipova

Naš metod kao ulaz koristi podatke koje je proizveo CoDi (*Contrastive Distance*) pristup za anotaciju ćelijskih tipova u prostornim transkriptomskim (ST) podacima, koji je razvijen

kako bi povezao ST podatke sa referentnim scRNA-seq podacima. CoDi koristi neuronsku mrežu kako bi stvorio niskodimenzionalne reprezentacije ćelija, pri čemu ćelije sličnog tipa budu bliže jedna drugoj u prostoru. Zatim koristi metriku distance između vektora genskih ekspresija u ST i scRNA-seq podacima kako bi povezao ćelije sličnih osobina.

Ovaj metod se pokazao kao veoma uspešan, čak i u složenim tkivima sa mnogo različitih tipova ćelija, pa smo ga izabrali kao polaznu osnovu za naš pristup imputaciji gena.

B. Statistički pristup imputaciji gena

Naša metoda imputacije gena oslanja se na informacije iz referentnih scRNA-seq podataka kako bi rekonstruisala nedostajuće vrednosti ekspresije u ST podacima, uzimajući u obzir razlike između ćelijskih tipova. Proces započinje anotacijom ćelijskih tipova u ST podacima pomoću CoDi metode. Umesto primene globalnog pristupa, vrši se imputacija odvojeno za svaki anotirani ćelijski tip, što omogućava veću biološku preciznost.

Za svaki gen sa nultom vrednošću (potencijalni *drop-out*) u ST podacima, prvo izračunavamo srednju vrednost i standardnu devijaciju njegove ekspresije u odgovarajućem ćelijskom tipu iz scRNA-seq podataka. Na osnovu tih parametara, po normalnoj raspodeli uzorkujemo novu vrednost, a zatim je skaliramo u skladu sa odnosom srednjih vrednosti ekspresije između ST i scRNA-seq podataka.

Da bi rezultati imputacije ostali u skladu sa biološkim obrascima, uvodimo dodatnu kontrolu. Koristimo podatke iz scRNA-seq da procenimo koliko često je određeni gen nula u nekom ćelijskom tipu. Ova verovatnoća se koristi kao kriterijum za odluku da li treba zadržati nultu vrednost u ST podacima. Na ovaj način ne popunjavamo sve vrednosti, već ostavljamo nule tamo gde je to biološki realno.

Pored statističke preciznosti, metod je optimizovan korišćenjem paralelne obrade, što ga čini pogodnim za rad sa velikim skupovima podataka. Ključna prednost ovog pristupa leži u činjenici da se imputacija vrši unutar jasno definisanih ćelijskih tipova, što omogućava očuvanje biološke specifičnosti ekspresionih profila koji opisuju različite ćelijske populacije, što je od posebnog značaja u složenim i heterogenim tkivima.

C. Skupovi podataka

Za evaluaciju naše metodologije koristili smo četiri različita skupa podataka prostorne transkriptomike, koji predstavljaju različite tehnologije sekvenciranja i tipove tkiva. Za svaki ST skup podataka koristili smo odgovarajuće scRNA-seq podatke istog tkiva kao referencu za anotaciju i imputaciju. Tabele 1 i 2 prikazuju skupove podataka.

Tabela 1 Korišćeni ST skupovi podataka

Skup podataka	Tehnologija	Broj ćelija	Broj gena	Broj klasa
Mouse Kidney	Slide-seq-V2	36299	31053	16
Mouse Brain Visium	Visium 10x	14968	31053	59
Adult Mouse Brain	Stereo-seq	45201	27998	256
Mouse Embryo Whole Brain	Stereo-seq	59704	27998	256

Tabela 2 Korišćeni scRNAseq skupovi podataka

Skup podataka	Tehnologija	Broj ćelija	Broj gena	Broj klasa
Mouse Kidney	Slide-seq-V2	43636	31053	16
Mouse Brain Visium	Visium 10x	40532	31053	59
Mouse Brain L5	Stereo-seq	160796	27998	256

D. Metrike evaluacije

Za objektivnu evaluaciju naše metode imputacije koristili smo metrike zasnovane na retenciji marker gena, što predstavlja procenat marker gena iz referentnih scRNA-seq podataka koji se zadržavaju kao marker geni u ST podacima. Marker geni su oni geni koji su karakteristični za određene ćelijske tipove i ključni su za njihovo prepoznavanje. Metrike su preuzete iz CoDi rada kako bi se održala konzistentnost u evaluaciji [5].

- *Retencija top 100 marker gena*: procenat od 100 najvažnijih marker gena (sortiranih po p-vrednosti) za svaki ćelijski tip koji se zadržavaju kao marker geni u ST podacima.

- *Retencija jedinstvenih marker gena*: procenat marker gena koji su specifični za određene ćelijske tipove (prisutni u manje od 25% svih ćelijskih tipova) koji se zadržavaju kao marker geni u ST podacima.

IV. REZULTATI I DISKUSIJA

Rezultati naše evaluacije pokazuju značajno poboljšanje retencije marker gena nakon imputacije za sve četiri skupa podataka. U Tabeli 3 i Tabeli 4 predstavljeni su rezultati retencije top 100 marker gena i jedinstvenih marker gena redom.

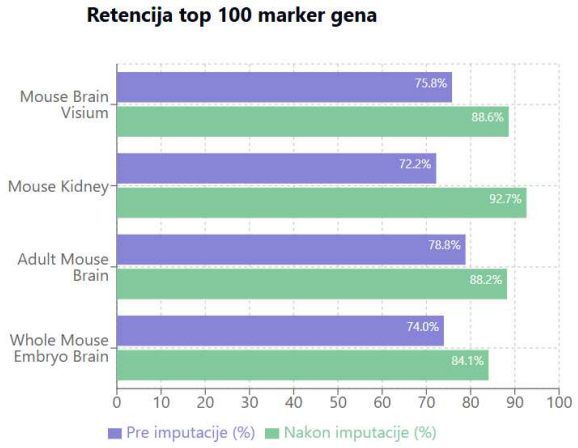
Tabela 3 Rezultati retencije top 100 marker gena

Skup podataka	Pre imputacije (%)	Nakon imputacije (%)	Poboljšanje (%)
Mouse Kidney	75.81	88.63	+12.82
Mouse Brain Visium	72.22	92.65	+20.43
Adult Mouse Brain	78.85	88.23	+9.38
Mouse Whole Brain	73.95	84.07	+10.12

Tabela 4 Rezultati retencije jedinstvenih marker gena

Skup podataka	Pre imputacije (%)	Nakon imputacije (%)	Poboljšanje (%)
Mouse Kidney	59.47	66.04	+6.57
Mouse Brain Visium	48.26	69.93	+20.43
Adult Mouse Brain	60.26	77.71	+17.45
Mouse Whole Brain	54.30	68.47	+14.17

Sl. 3 i sl. 4 grafički prikazuju poboljšanja u retenciji top 100 i jedinstvenih marker gena redom nakon imputacije.

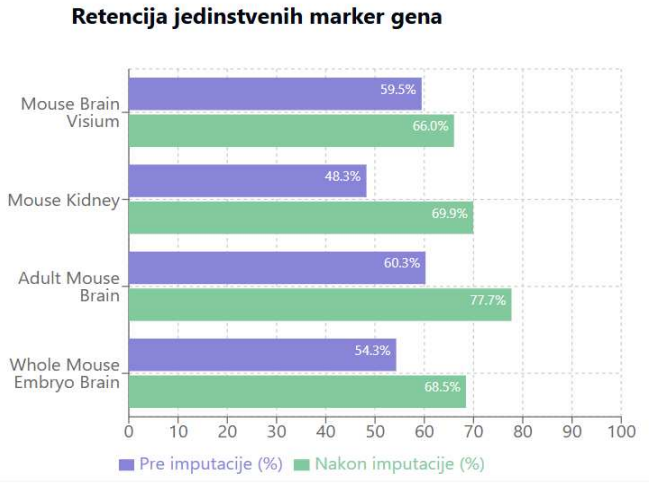


Slika 2 Grafički prikaz poboljšanja u retenciji top 100 marker gena

Najveće poboljšanje zabeleženo je za Mouse Kindey skup, sa povećanjem retencije jedinstvenih marker gena za 21.67% i top 100 marker gena za 20.43%. Značajna poboljšanja su takođe postignuta za Adult Mouse Brain (+17.45% za jedinstvene marker gene) i Whole Mouse Embryo Brain (+14.17% za jedinstvene marker gene).

Rezultati ovog istraživanja pokazuju da imputacija gena po pojedinačnim klasterima predstavlja efikasan način za popunjavanje nedostajućih vrednosti ekspresije u podacima prostorne transkriptomike. Značajno poboljšanje retencije marker gena nakon imputacije ukazuje da se biološki važna informacija uspešno očuvala.

U rezultatima se uočava nekoliko važnih obrazaca. Najveće poboljšanje zabeleženo je na skupu podataka Mouse Kidney, gde je retencija jedinstvenih marker gena porasla za 21.67%, dok je Mouse Brain Visium pokazao najskromnije poboljšanje od 6.57%. Ove razlike mogu poticati iz različitih tehnologija koje su korišćene (*Slide-seq* naspram *Visium*), specifičnih karakteristika samog tkiva ili kvaliteta ulaznih podataka. Takođe, veći efekat imputacije primećen je kod jedinstvenih marker gena, koji su specifični za pojedinačne ćelijske tipove, što je posebno važno jer upravo ti geni



Slika 1 Grafički prikaz poboljšanja u retenciji jedinstvenih marker gena

najpreciznije definišu identitet ćelija. Na kraju, sva četiri analizirana skupa pokazuju konzistentno poboljšanje nakon imputacije, što govori u prilog tome da je metodologija robusna i primenljiva na različite biološke podatke.

Ovaj pristup imputaciji gena u prostornim transkriptomskim podacima pokazao se kao efikasniji u poređenju sa postojećim metodama, pre svega zbog svog fokusa na biološku specifičnost. Za razliku od globalnih metoda koje sve ćelije tretiraju jednako, naš model primenjuje imputaciju za svaki klaster posebno, uzimajući u obzir razlike među ćelijskim tipovima.

Ovaj pristup omogućava generisanje biološki tačnijih vrednosti i očuvanje karakterističnih obrazaca ekspresije. Dodatno, kontrola verovatnoće zadržavanja nule i skaliranje na osnovu odnosa ekspresije između ST i scRNA-seq podataka doprinose očuvanju prirodnih obrazaca eskpresije. Još jedna važna prednost naše metode je skalabilnost, zahvaljujući implementaciji paralelnog procesiranja, moguće je efikasno analizirati velike skupove podataka koji sadrže stotine hiljada ćelija i desetine hiljada gena.

Povećana retencija marker gena nakon imputacije ima važan biološki značaj jer direktno utiče na kvalitet interpretacije prostornih transkriptomskih podataka. Veća retencija omogućava pouzdaniju identifikaciju i opis ćelijskih tipova u njihovom prostornom kontekstu, što je ključno za razumevanje funkcionalne organizacije tkiva. Takođe, imputacija može doprineti detekciji retkih ćelijskih populacija, čiji marker geni često ostaju neotkriveni usled izraženog *drop-out* efekta.

Precizniji podaci o ekspresiji omogućavaju i bolje mapiranje ćelijske mikrookoline i međusobnih interakcija, pružajući dublji uvid u strukturu i funkciju tkiva. Na kraju, imputirani podaci povećavaju pouzdanost niza analitičkih postupaka, uključujući identifikaciju diferencijalno eksprimiranih gena, funkcionalno obogaćivanje bioloških puteva i rekonstrukciju regulatornih mreža, što dodatno potvrđuje korisnost naše metodologije.

V. Zaključak

U ovom radu predstavljen je metod za imputaciju gena u podacima prostorne transkriptomike, koji se zasniva na imputaciji po klasteru. Ovaj metod koristi informacije iz referentnih scRNA-seq podataka kako bi rekonstruisao nedostajuće vrednosti ekspresije unutar jasno definisanih ćelijskih tipova. Na taj način se čuvaju biološka konzistentnost i specifičnost različitih ćelijskih populacija, što doprinosi tačnijoj interpretaciji prostornih transkriptomskih podataka.

Evaluacija sprovedena na četiri javno dostupna skupa podataka pokazala je značajno poboljšanje retencije marker gena nakon imputacije. Zabeleženo je povećanje retencije do 21.67% za jedinstvene marker gene i do 20.43% za top 100 marker gena, što potvrđuje efikasnost našeg pristupa u očuvanju biološki relevantnih informacija.

Uprkos postignutim rezultatima, metod ima određena ograničenja koja treba imati u vidu. Efikasnost imputacije u velikoj meri zavisi od tačnosti početne anotacije ćelijskih tipova, a uspešna primena zahteva dostupnost referentnih scRNA-seq podataka iz istog ili sličnog tkiva. Takođe, analiza velikih skupova podataka može biti tehnički zahtevna u pogledu memorijskih resursa i vremena izvršavanja, što može ograničiti primenljivost u nekim okruženjima.

Budući pravci istraživanja uključuju integraciju prostornih informacija u proces imputacije kako bi se dodatno povećala preciznost i očuvala prostorna organizacija ćelija. Planiramo proširenje metodologije na druge tipove omics podataka, kao i optimizaciju algoritma u cilju povećanja efikasnosti i skalabilnosti pri obradi izuzetno velikih skupova podataka.

Naš rad predstavlja doprinos rastućem polju prostorne transkriptomike, nudeći metod koji poboljšava kvalitet i interpretabilnost prostornih podataka. Imputacija nedostajućih vrednosti ekspresije gena omogućava prevazilaženje tehničkih ograničenja trenutnih tehnologija, što otvara mogućnosti za dublje razumevanje organizacije tkiva u prostoru [15]. Daljim razvojem i primenom ovakvih metoda možemo unaprediti istraživanja u oblastima kao što su razvoj organa [16], onkologija, neurologija i regenerativna medicina. Poboljšani prostorni podaci doprinose razvoju preciznijih modela tkiva i bolesti, što može voditi ka efikasnijim dijagnostičkim i terapijskim strategijama.

Zahvalnica

Ovaj rad je delimično finansijski podržalo Ministarstvo nauke, tehnološkog razvoja i inovacija Republike Srbije po ugovorima broj: 451-03-137/2025-03/200103 i 451-03-136/2025-03/200103 i Complete Genomics po ugovoru broj: 1847/2022-12. Autori se zahvaljuju na finansijskoj podršci.

Literatura

[1] Burgess, D. J. (2019). Spatial transcriptomics coming of age. Nature Reviews Genetics, 20(6), 317-317.

[2] Moses, L., & Pachter, L. (2022). Museum of spatial transcriptomics. Nature methods, 19(5), 534-546.

[3] Jiang, R., Sun, T., Song, D., & Li, J. J. (2022). Statistics or biology: the zero-inflation controversy about scRNA-seq data. Genome biology, 23(1), 31.

[4] Luecken, M. D., & Theis, F. J. (2019). Current best practices in single‐cell RNA‐seq analysis: a tutorial. Molecular systems biology, 15(6), e8746.

[5] Kovacevic, V., Bezulj, M., Milicevic, N., Josic, B., Fang, S., Zhang, Y., & Li, J. (2024). CoDi: Contrastive distance cell type annotation for spatially resolved transcriptomics. Preprint, available at Research Square https://doi.org/10.21203/rs.3.rs-4495419/v1.

[6] Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H. W., Li, T., ... & Bayraktar, O. A. (2022). Cell2location maps fine-grained cell types in spatial transcriptomics. Nature biotechnology, 40(5), 661-671.

[7] Biancalani, T., Scalia, G., Buffoni, L., Avasthi, R., Lu, Z., Sanger, A., ... & Regev, A. (2021). Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. Nature methods, 18(11), 1352-1362.

[8] Vahid, M. R., Brown, E. L., Steen, C. B., Zhang, W., Jeon, H. S., Kang, M., ... & Newman, A. M. (2023). High-resolution alignment of single-cell and spatial transcriptomes with CytoSPACE. Nature biotechnology, 41(11), 1543-1548.

[9] Cable, D. M., Murray, E., Zou, L. S., Goeva, A., Macosko, E. Z., Chen, F., & Irizarry, R. A. (2022). Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature biotechnology*, *40*(4), 517-526.

[10] Gribov, A., Sill, M., Lück, S., Rücker, F., Döhner, K., Bullinger, L., ... & Unwin, A. (2010). SEURAT: visual analytics for the integrated analysis of microarray data. BMC medical genomics, 3, 1-6.

[11] Stickels, R. R., Murray, E., Kumar, P., Li, J., Marshall, J. L., Di Bella, D. J., ... & Chen, F. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nature biotechnology, 39(3), 313-319.

[12] Chen, A., Liao, S., Cheng, M., Ma, K., Wu, L., Lai, Y., ... & Wang, J. (2022). Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell, 185(10), 1777-1792.

[13] Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Experimental & molecular medicine, 50(8), 1-14.

[14] Stuart, T., & Satija, R. (2019). Integrative single-cell analysis. Nature reviews genetics, 20(5), 257-272.

[15] Edsgärd, D., Johnsson, P., & Sandberg, R. (2018). Identification of spatial expression trends in single-cell gene expression data. Nature methods, 15(5), 339-342.

[16] Asp, M., Giacomello, S., Larsson, L., Wu, C., Fürth, D., Qian, X., ... & Lundeberg, J. (2019). A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. Cell, 179(7), 1647-1660.

Evaluation of Statistical Gene Imputation Methods in Spatial Transcriptomics Data

Lazar Smiljković, Marko Mišić, Vladimir Kovačević

ABSTRACT

Spatial transcriptomics (ST) enables the analysis of gene expression while preserving spatial information within tissues, offering a unique insight into the spatial organization of cell types. However, this technology often suffers from drop-out effects, where certain genes are not detected due to technical limitations rather than their true absence in cells. In this paper, we present a gene imputation method for ST data based on pre-annotated cell types using the CoDi algorithm. This approach leverages complementary single-cell RNA-seq (scRNA-seq) data to impute missing expression values within specific cell types, preserving biological consistency. Evaluation of the method on four different datasets shows significant improvement in marker gene retention, with increases of up to 21.67% for unique marker genes and up to 20.43% for the top 100 marker genes. These results demonstrate the effectiveness of the proposed approach in reconstructing more reliable and biologically meaningful spatial transcriptomic data.

Radiološki pejzaži: prepoznavanje i analiza prostornih odnosa u radiološkim izveštajima

Isidora Popović
Elektrotehnički fakultet
Univerzitet u Beogradu
pi243285m@student.etf.bg.ac.rs

Anđelka Zečević
Matematički institut
Srpska akademija nauka i umetnosti
andjelkaz@mi.sanu.ac.rs

***Apstrakt* - Prostorne anotacije su poseban vid oznaka teksta kojima se naglašavaju položaji, orijentacije i međusobni odnosi objekata. Obogaćivanje radioloških izveštaja prostornim anotacijama doprinosi dubljem domenskom razumevanju i lakšoj interpretaciji sadržaja istovremeno omogućavajući razvojnoj AI zajednici adekvatnije i pouzdanije kreiranje alata. U ovom radu će biti predstavljeni rezultati ispitivanja kapaciteta velikog jezičkog modela GPT-4 i njegovih mogućnosti da obogati radiološke izveštaje anotacijama sheme Rad-SpRL. Ovom shemom se predlaže pet ključnih elemenata za opisivanje radioloških izveštaja: *radiološki entitet* (engl. trajector), *prostorna odrednica* (engl. landmark), *prostorni indikator* (engl. spatial indicator), *nivo sigurnosti* (engl. hedge) i *dijagnoza*. Rezultati će biti predstavljeni iz perspektive *one-shot* testiranja za pojedinačne elemente ove sheme sa osvrtom na formalne metrike i kvalitativnu analizu. Sva testiranja će biti izvedena nad javno dostupnim skupom *Open-i* koji objedinjuje 2,000 deidentifikovanih radioloških izveštaja na engleskom jeziku. Posebno će biti prokomentarisana pitanja izbora pokaznih primera kao i uticaj korišćenja njihovog većeg broja.**

***Ključne reči* -** *radiološki izveštaji, prostorne anotacije, Rad-SpRL, veliki jezički modeli, GPT-4*

I. Uvod

Prema zvaničnom izveštaju Stanfordovog centra za veštačku inteligenciju okrenutu ka čoveku (Stanford University Human-Centered Artificial Intelligence, HAI) [1], radiologija je oblast medicine koja je najviše spregnuta sa progresima u svetu veštačke inteligencije. Na listi Uprave za hranu i lekove (eng. Food and Drug Administration, FDA) na koju referiše ovaj izvor [2], nalaze se 692 medicinska uređaja koja koriste veštačku inteligenciju, od čega čak 531 uređaj pronalazi svoju primenu u oblasti radiologije.

Radiologija, kao osnovna dijagnostička oblast medicine, generiše veliki broj radioloških slika i pratećih izveštaja. Dok se radiološke slike (rendgenski snimci, CT i MRI skenovi, ultrazvučne slike i drugo) obrađuju algoritmima računarskog vida[1] i razmatraju u kontekstu zadataka detekcije objekata, segmentacije ili klasifikacije [3], radiološki izveštaji se posmatraju iz ugla obrade prirodnog jezika i pojavljuju u zdacima poput ekstrakcije medicinskih informacija, klasifikacije prema stepenu ozbiljnosti i dijagnozi ili pojednostavljivanja sadržaja zarad veće čitljivosti [4]. Sa porastom broja multimodalnih modela, otvara se prostor za uparivanje radioloških slika i izveštaja i rad na zadacima poput generisanja izveštaja na osnovu zadatih radioloških snimaka [5].

Budući da radiološki izveštaji predstavljaju opis i interpretaciju dvodimenzionih ili trodimenzionih radioloških slika od strane eksperata, jezik koji ih karakteriše obiluje konstrukcijama koje opisuju prostorne odnose između medicinskih značajnih pronalazaka (lezija, tumora, imflamacija, senki i slično) i anatomskih lokaliteta (pluća, grudna kost, L1 pršljen i slično). Automatsko prepoznavanje ovih

[1] Neke kategorije koje FDA spominje među primerima softvera kao medicinskog sredstva su CADe (eng. Computer-aided Detection) tj. računarski podržana detekcija, CADt (eng. Computer-aided Triage) tj. računarski podržana trijaž, CADx (eng. Computer-aided Diagnosis) tj. računarski podržana dijagnostika i MIMPS (eng. Medical Image Management and Processing System) tj. sistem za upravljanje i obradu medicinskih slika.

lingvističkih konstrukcija bi omogućilo finiju ekstrakciju informacija iz radiološkog izveštaja, ali i potpomoglo čitanje i analizu radioloških izveštaja isticanjem posebno važnih konstrukcija ili pak njihovom vizuelizacijom na radiološkim slikama. Vođeni mogućim benefitima razumevanja jezika radiologije, u ovom radu su razmatrani kapaciteti velikog jezičkog modela GPT-4 da automatizuje izdvajanje pomenutih prostornih konstrukcija.

U sekciji koja sledi je dat ostvrt na jednu od lingvističkih shema za opisivanje radioloških prostornih anotacija, zatim je opisan skup podataka koji je korišćen u eksperimentima, a potom i dizajn samog eksperimenta i rezultati dobijeni *one-shot* testiranjem modela GPT-4.

II. Prostorna anotacija

Zadatak obogaćivanja teksta prostornim anotacijama (engl. Spatial Role Labeling, SpRL) je detaljno razmatran u kontekstu izazova kao što su navigiranje objekata glasom, generisanje naslova slika ili generisanje scene na osnovu opisa i ima svoje utemeljenje u lingvističkoj teoriji semantike okvira [6]. Svako obogaćivanje teksta anotacijama prati određenu shemu anotacije tj. skup uvedenih elemenata. Tako su, na primer, česti elementi ovih anotacija objekti koji se posmatraju, njihove dimenzije i orijentacije, lokaliteti, reperi u odnosu na koje se kreću, odrednice poput početne i krajnje tačke putanje i slično [7]. U zavisnosti od sheme i njene svrhe [8, 9, 10] može varirati nivo granularnosti i broj raspoloživih elemenata, kao i skup obaveznih i opcionih elemenata.

Jedna od shema za prostornu anotaciju radioloških izveštaja je shema Rad-SpRL [11] razvijena od strane Nacionalnog instituta za zdravstvo Amerike. Glavni elementi ove sheme su *radiološki entitet* (engl. trajector) koji predstavlja radiološki pronalazak čija se pozicija opisuje, *prostorna odrednica* (engl. landmark) koja predstavlja anatomsku lokaciju radiološkog entiteta, *dijagnoza* koja predstavlja potencijalnu dijagnozu koja se dovodi u vezu sa relacijom radiološkog entiteta i prostorne odrednice i *stepen sigurnost*i (engl. hedge) kao fraza kojom se opisuje izvesnost dijagnoze ili pronalaska. Uz ove elemente se posmatra i *prostorno obeležje* (engl. spatial indicator), najčešće predlog, koji ukazuje na postojanje prostornog opisa u tekstu. Na primer, u rečenici *Stable peripheral right lower lobe opacities seen between the anterior 7th and 8th right ribs which may represent pleural reaction or small pulmonary nodules.* fraza *Stable peripheral right lower lobe opacities* predstavlja radiološki entitet, fraza *the anterior 7th and 8th right ribs* prostornu odrednicu, fraze *pleural reaction* i *small pulmonary nodules* dve moguće dijagnoze, dok fraza *may represent* označava nivo izvesnosti pomenutih dijagnoza čija dalja precizna potvrda zavisi od dodatnih pregleda i laboratorijskih testova. Fraza *between* je u posmatranoj rečenici prostorni indikator i ukazuje na postojanje prostornih elemenata.

Moguće je da u jednoj rečenici postoji više prostornih indikatora i da svaki od njih inicira nešto drugačija obeležavanja fraza. Tako, u rečenici "*Visualized osseous structures of the thorax are without acute abnormality.*" u odnosu na prostorni indikator *of*, fraza *osseous structures* predstavlja radiološki entitet, dok je u odnosu na indikator *without* lokalitet tj. prostornu odrednicu.

U shemi Rad-SpRL elementi radiološki entitet i prostorna odrednica su obavezni elementi dok su stepen sigurnosti i moguće dijagnoze opcioni.

Slika 1. Primer prostorne anotacije

III. Skup podataka

U eksperimentima u ovom radu korišćen je javno dostupni skup radioloških izveštaja grudnog koša *Open-i* obeležen prostornim anotacijama sheme Rad-SpRL [12]. Svi izveštaji su na engleskom jeziku i prošli su proces automatske deidentifikacije kojom se čuva privatnost pacijenata i medicinskog osoblja. Na mestima potencijalno osetljivih informacija u izveštajima se nalazi token *XXXX*.

Ukupan broj izveštaja u skupu podataka iznosi 2,000. Svaki radiološki izveštaj je u formatu XML i sastoji se od pet sekcija. U uvodnoj sekciji su navedene neke opšte informacije o snimanju (datum snimanja, projekcija aparata, dodatni nalazi i slično), zatim sledi sekcija sa indikacijama, zatim sekcija sa zapažanjima, a potom i sekcija sa zaključcima. Na kraju se, opciono, nalazi sekcija sa dodatnim informacijama. Primer jednog radiološkog izveštaja ovog skupa se se može videti na Slici 2.

Ukupan broj prostornih indikatora u skupu podataka, mahom predloga, je 1,962 od kojih su najfrekventniji *of* (765 pojavljivanja), *in* (526 pojavljivanja), *without* (176 pojavljivanja), *with* (141 pojavljivanje) i *within* (102 pojavljivanja). U skupu

podataka su obeležena i 2,293 radiološka entiteta, 2,167 prostorne odrednice, 455 dijagnoza i 388 izraza izvesnosti. Primeri najčešće pojavljivanih radioloških entiteta, prostornih odrednica, dijagnoza i stepena izvesnosti su navedeni u Tabeli 1.

Najveći broj radioloških izveštaja (njih 534) ima jednu pridruženu prostornu anotaciju. Nešto manji broj izveštaja ima dve (njih 298) ili tri (njih 107) pridužene anotacije, dok su izveštaji sa većim brojem anotacija ređe prisutni[2]. Sa idejom da se lakše isprate ponašanja modela GPT-4 za svaki element sheme, u daljem radu je korišćen samo podskup izveštaja sa jednom prostornom anotacijom i to 100 nasumično izabranih izveštaja iz ove grupe. Ovakav pristup je smanjio kompleksnost obrade podataka i omogućio pokretanje većeg broja eksperimenata.

IV. Eksperiment

Kako bi se ispitala mogućnost korišćenja modela GPT-4 za prostornu anotaciju radioloških izveštaja u duhu opisane sheme Rad-SpRL, odabran je model *GPT-4o Mini*. Njegove performanse su se u inicijalnim eksperimentima kroz funkcionalnosti servisa *OpenAI playground* pokazale boljim u odnosu na starije modele i, dodatno, uklapao se u raspoložive resurse za korišćenje zvaničnog OpenAI API servisa.

Prompt za komunikaciju sa ovim servisom je dizajniran tako da prati zvanične smernice za anotaciju radioloških izveštaja i zvanične opise elemenata sheme Rad-SpRL. Kratki propratni kontekst je objašnjavao sam zadatak i formu očekivanog izlaza, dok je za primer u *one-shot* testiranjima odabran pokazni primer sheme Rad-SpRL. Na Slici 3 se može videti primer korišćenog prompta, dok će u daljem radu biti diskutovani uticaji primera kao i neka zapažanja eksperimenata *zero-shot* i *few-shot* tipa.

Svaki od radioloških izveštaja izdvojenog skupa podataka je uz opisani prompt prosleđen odabranom modelu GPT-4 kao ulaz, a potom su njegovi rezultati upoređivani sa postojećim prostornim anotacijama u skupu podataka.

Prilikom upoređivanja anotacija predloženih od strane modela GPT-4 i anotacija koje postoje u skupu podataka, praćeno je potpuno i delimično preklapanje. Potpuno preklapanje (engl. exact match) je označavalo odgovor modela GPT-4 koji se u potpunosti poklapao sa referentnom anotacijom u skupu podataka, uključujući sve karaktere i tokene. Delimično poklapanje (engl. partial match) je pratilo stepen preklapanje na nivou tokena između odgovora modela GPT-4 i referentne anotacije čime su obuhvaćeni scenariji u kojima je model GPT-4 uspeo da identifikuje relevantne segmente ali ne dovoljno precizno. Delimično preklapanje je praćeno vrednostima Žakardove sličnosti (engl. Jaccard similarity), jedne od često korišćenih mera u oblasti prirodnih jezika koja predstavlja odnos preseka tokena i ukupnog broja tokena predložene i stvarne anotacije.

Chest PA-Lat XR

Imaging Study
Xray Chest PA and Lateral

Exam:
PA and lateral views of the Chest performed XXXX/XXXX.

Comparison:
PA and lateral chest XXXX and CTA XXXX.

Indication:
XXXX year old smoking on oxygen and nasal cannula caught XXXX. XXXX to the cheek and inside of nose.

Findings:
The heart is within normal limits in size. Surgical suture material projects over the right lung apex. The lungs are hyperlucent and hyperinflated compatible with emphysema. There is left lower lobe airspace disease identified. There is moderate left pleural effusion and small right pleural effusion. No visualized pneumothorax.

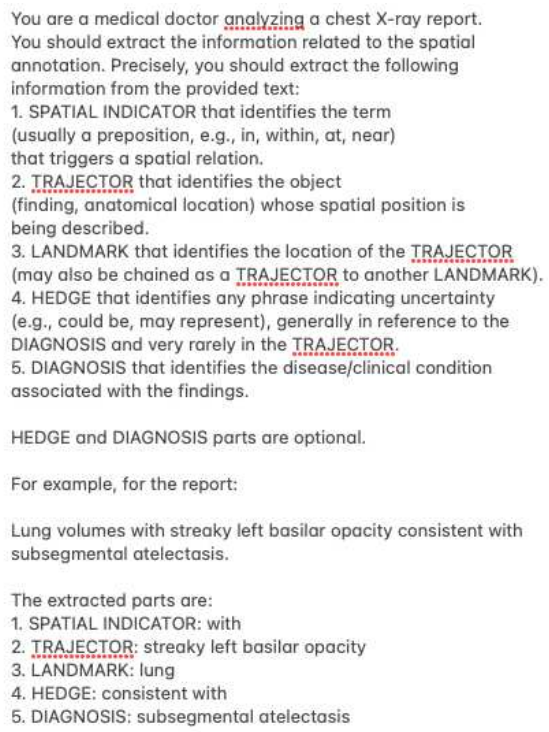
Impression:
Left lower lobe airspace disease and bilateral pleural effusions, left greater than right. This may be secondary to inhalational injury. Recommend followup to ensure complete resolution.

Slika 2. Primeri radiološkog izveštaja skupa *Open-i*

Tabela 1. Najčešće anotacije u skupu podataka *Open-i*

Radiološki entitet	*opacity, degenerative, change, pneumothorax, pleural effusion, consolidation*
Prostorna odrednica	*lung, thoracic spine, spine, left lung base, thorax*
Dijagnoza	*scarring, atelectasis, infiltrate, granuloma, emphysema*
Stepen izvesnosti	*may represent, consistent with, compatible with*

[2] U skupu se nalaze i izveštaji u kojima nema prostornih anotacija.



Slika 3. Primeri korišćenog prompta

Obe vrste preklapanja su praćene na nivou pojedinačnih elemenata sheme Rad-SpRL. Za ocenu potpunog preklapanja korišćena je preciznosti (procenat potpunih preklapanja u odnosu na ukupan broj elemenata), dok je za ocenu delimičnog preklapanja praćena prosečna vrednost dobijenih Žakardovih sličnosti.

V. Rezultati

Rezultati eskperimenta prikazani su u Tabeli 2. Kao što se može primetiti, sve vrednosti[3], bilo u terminima preciznosti i potpunog preklapanja ili u terminima Žakardove sličnosti i delimičnog preklapanja, su niske, ukazujući na poteškoće modela GPT-4 da adekvatno prepozna prostorne elemente. Manuelnom inspekcijom primera u kategoriji radioloških entiteta kod kojih nije ostvareno potpuno preklapanje, utvrđeno je da model GPT-4 ne pravi dovoljno dobru razliku klinički relevantnih informacija i da prisustvo negacije utiče na rezultat. Tako, recimo, fraza *within normal limits* često navodi model na pogrešne anotacije rezultirajući pogrešnim radiološkim entitetima (i prostornim odrednicama) u čak 29 od 100 slučajeva. U kontekstu delimičnog preklapanja, primećeno je da model GPT-4 često uz radiološke entitete veže i odrednice poput *mild* ili *moderate* (nrp. *stable calcified small granuloma* umesto *calcified small granuloma*) ili proširuje entitete (npr. *degenerative changes of the thoracic spine* umesto *degenerative changes*) pa time nudi nešto duže predloge.

Pogrešno obeležavanje prostornih odrednica je, pre svega, uslovljeno pomenutom nemogućnošću jezičkog modela da prepozna klinički relevantne scenarije. Na primer, u jednom od izveštaja su pogrešno razmatrani entiteti rečenice T*he lungs and pleural spaces show no acute abnormality.* (verovatno zbog prisustva prostorne odrednice *spaces*) umesto rečenice *Mild tortuosity of the descending thoracic aorta*. Primećeno je i da su tokeni *XXXX* nastali usled procesa deidentifikacije u jednom broju slučajeva (ukupno 10) sastavni deo prostornih odrednica, koje model nije eksplicitno izdvojio što je, dalje, rezultiralo nižim vrednostima delimičnog preklapanja.

Tabela 2. Preklapanje anotacija skupa *Open-i* i izlaza modela *GPT-4o Min*i po elementima u *one-shot* postavci eksperimenta

Element sheme Rad-SpRL	**Potpuno preklapanje**	**Delimično preklapanje**
Radiološki entitet	0.23	0.29
Prostorna odrednica	0.35	0.4
Dijagnoza	0.09	0.1
Stepen izvesnosti	0.31	0.3275

Anotiranje dijagnoza je praćeno najnižim metrikama. Analizom pojedinačnih slučajeva utvrđeno je da je model bio ekspresivniji u izražavanju situacija u kojima nije potrebno uspostaviti dijagnozu generisanjem fraza poput *No acute cardiopulmonary process*, *No acute radiographic cardiopulmonary process*, *No acute cardiopulmonary disease*, *No acute cardiopulmonary abnormalitie*s, *No acute disease, No radiographic evidence of acute cardiopulmonary disease* i slično a koje nisu netačne.

Stepen izvesnosti nije obavezan element sheme Rad-SpRL pa je u najvećem broju slučajeva njegovo anotiranje trebalo i da izostane. Utvrđeno je da je model u dodatnih 20 slučajeva dao tačne odgovore ali izražene u formi fraza *No* ili *None.* Takođe, primećeno je da je u određenom broju slučajeva (njih

[3] Vrednosti preciznosti i prosečne Žakardove sličnosti su na skali od 0 do 1 i veće vrednosti ukazuju na bolje rezultate.

6) ponovo prisutan token za deidentifikaciju u originalnim anotacijama i da ih model nije reprodukovao. Obe korekcije bi dodatno povećale metrike koje prate ovaj entitet.

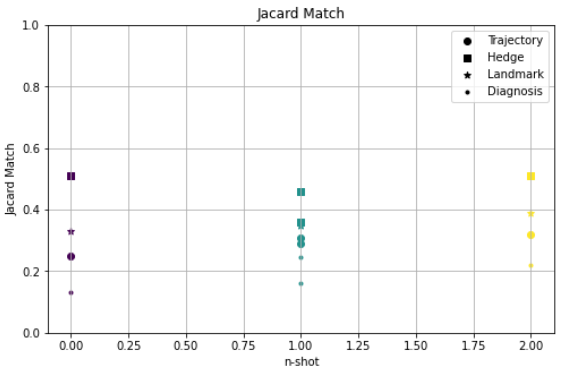
Sa idejom da se provere dobijeni zaključci koji ukazuju na ograničenja modela da adekvatno anotira prostorne elemente, ceo eksperiment je ponovljen koristeći drugi pokazni primer. On je, takođe, preuzet iz pratećih smernica za kreiranje skupa Rad-SpRL i glasi *There is airspace opacity within the left lung base, witch may represent atelectasis or infiltrate*. Prostorni indikator ovog primera je *within*, fraza *airspace opacity* predstavlja radiološki entitet, fraza *left lung base* predstavlja prostornu odrednicu, fraza *may represent* je stepen izvesnosti, dok su *atelectasis* i *infiltrate* dve moguće dijagnoze. Promena primera je rezultirala neznatnim promenama metrika. Preciznost anotiranja radioloških entiteta je porasla na 0.25 dok je prosečna Žakardova mera porasla na 0.2918. Na nivou prostornih odrednica je primećen blagi pad metrika i to preciznosti na 0.33 i prosečne Žakardove sličnosti na 0.3564. Stepen izvesnosti je ostvario najveće poboljšanje i to porast preciznosti na 0.46 i porast prosečne Žakardove sličnosti na 0.462. Anotiranje dijagnoza je i dalje bilo najizazovnije sa porastom preciznosti na svega 0.15 i prosečne Žakardove sličnosti na 0.1677. Kvalitativna analiza pojedinačnih elemenata je ukazivala na situacije primećene i prilikom korišćenja prvog pokaznog primera ostavljajući zaključak da je za potrebe automatske anotacije potrebno koristiti ili podesnije odabrane primere ili više njih.[4]

U inicijalnim eksperimentima u kojima su korišćena oba pomenuta primera (*two-shot* postavka eksperimenta) dobijene su nešto bolje metrike. Njihove vrednosti se mogu videti u Tabeli 3. Dodatne analize pojedinačnih slučajeva koji mogu biti indikativni za razumevanje ponašanja modela GPT-4 u zadacima prostorne anotacije i uočavanja razlika između prethodnih i ovako dobijenih rezultata predviđeni su za dalji rad i produblijivanje ove teme.

[4] Rezultati koji su dobijeni u *zero-shot* eksperimentima su ukazivali na lošije ponašanje modela, izuzev kod stepena izvesnosti. Vrednosti preciznosti i prosečne Žakardove mere su, redom, za radiološke entitete, prostorne odrednice, dijagnoze i stepen izvesnosti iznosile 0.23 i 0.2567, 0.31 i 0.3305, 0.12 i 0.1376, 0.51 i 0.5142. Najčešće anotirani prostorni indikatori su bili *in* (34 puta), *within* (31 put) i *of* (10 puta).

Tabela 3. Preklapanje anotacija skupa *Open-i* i izlaza modela *GPT-4o Mini* po elementima u *two-shot* postavci eksperimenta

Element sheme Rad-SpRL	**Potpuno preklapanje**	**Delimično preklapanje**
Radiološki entitet	0.28	0.3259
Prostorna odrednica	0.38	0.3929
Dijagnoza	0.21	0.2274
Stepen izvesnosti	0.51	0.5183



Slika 4. Uticaj broja primera na vrednosti Žakardove sličnosti u prostornim anotacijama

VI. Zaključak

Analiza rezultata sprovedenih eksperimenata je pokazala da model *GPT-4o Mini* ostvaruje ograničene performanse u zadacima prostorne anotacije radioloških izveštaja elementima sheme Rad-SpRL. Svaki od elemenata ove sheme nosi svoje izazove i može se anotirati sa različitom uspešnošću. Element *stepen izvesnosti* je u svim eksperimentima bio praćen najvišim metrikama (preciznost od 0.23 do 0.51 tj. vrednost prosečne Žakardove sličnosti od 0.2567 do 0.5183) dok se najizazovnijim ispostavio element *dijagnoza* (preciznost od 0.12 do 0.21 tj. prosečna Žakardova mera sličnosti od 0.1376 do 0.2274). Finiji izbor primera i povećanje njihovog broja mogu evidentno poboljšati vrednosti metrika i uticati na neka nepoželjna ponašanja modela. Međutim, oblast dijagnostičkih anotacija ostaje izazovna i zahteva dodatna istraživanja i unapređenja pristupa.

U daljem radu će biti još detaljnije ispitan prostor grešaka modela GPT-4 uz korišćenje većeg i raznovrsnijeg skupa podataka. Takođe, biće ispitano

korišćenje metrika koje mogu finije da isprate semantičku sličnost generisanih izlaza, kao i korišćenje dodatnih radioloških ontologija koje bi potencijalno mogle da unaprede trenutno najosetljiljiviji segment dijagnostike. Nakon progresa na ovim poljima, preći će se na višejezične eksperimente i prilagođavanje same sheme Rad-SpRL i modela srpskom jeziku.

Literatura

[1] AI Index Report 2024, Science and Medicine: https://hai.stanford.edu/ai-index/2024-ai-index-report/science-and-medicine, pristupljeno maj 2025. godine

[2] FAD list of AI-enabled devices: https://rad.washington.edu/news/fda-publishes-list-of-ai-enabled-medical-devices/, pristupljeno maj 2025. godine

[3] Zhang K., Khosravi B., Vahdati S., Erickson B. J. FDA *Review of Radiologic AI Algorithms: Process and Challenges*. Radiology. 2024 Jan;310(1):e230242. doi: 10.1148/radiol.230242. PMID: 38165243.

[4] Casey A., Davidson E., Poon M. et al. *A systematic review of natural language processing applied to radiology reports.* BMC Med Inform Decis Mak 21, 179 (2021). https://doi.org/10.1186/s12911-021-01533-7

[5] Shared Task on Large-Scale Radiology Report Generation @BioNLP ACL 2024: https://stanford-aimi.github.io/RRG24/

[6] Fillmore C. J. and Baker C. *Frame semantics for text understanding.* Proceedings of WordNet and Other Lexical Resources Workshop, NAACL. 2001.

[7] Kordjamshidi P., Moens M.F. and van Otterlo M., 2010, May. Spatial role labeling: Task definition and annotation scheme. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)* (pp. 413-420). European Language Resources Association (ELRA).

[8] Pustejovsky J., Kordjamshid P., Moens M., Levine A., Dworman S., and Yocum Z. 2015. SemEval-2015 Task 8: SpaceEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 884–894, Denver, Colorado. Association for Computational Linguistics.

[9] Ulinski M., Coyne B., and Hirschberg J. 2019. SpatialNet: A Declarative Resource for Spatial Relations. In *Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)*, pages 61–70, Minneapolis, Minnesota. Association for Computational Linguistics.

[10] Datta S., Ulinski M., Godfrey-Stovall J., Khanpara S., Riascos-Castaneda R. F., and Roberts K. 2020. Rad-SpatialNet: A Frame-based Resource for Fine-Grained Spatial Relations in Radiology Reports. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2251–2260, Marseille, France. European Language Resources Association.

[11] Datta S., Si Y., Rodriguez L., Shooshan S.E., Demner-Fushman D., Roberts K. *Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest X-ray reports using deep learning.* J Biomed Inform. 2020 Aug;108:103473. doi: 10.1016/j.jbi.2020.103473. Epub 2020 Jun 18. PMID: 32562898; PMCID: PMC7807990.

[12] Datta S. and Kirk R. (2020). *A dataset of chest X-ray reports annotated with Spatial Role Labeling annotations*. Mendeley Data, V1, doi: 10.17632/yhb26hfz8n.1

Radiology Landscapes: Annotation and Analysis of Spatial Relationships in Radiology Reports

Isidora Popović, Anđelka Zečević

Spatial annotations are a special type of text annotations that emphasize the positions, orientations, and mutual relations of objects. Enriching radiology reports with spatial annotations contributes to a deeper domain understanding and more straightforward interpretation while enabling the AI development community to create more adequate and reliable tools. This paper will present the results of testing the capacity of the large language model GPT-4 and its ability to enrich radiological reports with annotations according to the Rad-SpRL scheme. This scheme proposes five key elements for describing radiological reports: trajector, landmark, spatial indicator, hedge, and diagnosis. Results will be presented from the perspective of *one-shot* testing for individual elements of this scheme regarding formal metrics and qualitative analysis. All testing will be performed on the publicly available *Open-i* dataset, which aggregates 2,000 de-identified radiology reports in English. Special comments will be made on selecting demonstrative examples and using a larger number of examples.

Uticaj uvođenja ERP sistema na unutrašnju organizaciju finansijske funkcije u kompaniji

Milan Jovanović
Univerzitet u Beogradu, Fakultet organizacionih nauka
Beograd, Srbija
milan.jovanovic@fon.bg.ac.rs
0009-0004-7740-6217

Ivan Todorović
Univerzitet u Beogradu, Fakultet organizacionih nauka
Beograd, Srbija
ivan.todorovic@fon.bg.ac.rs
0000-0002-3792-582X

Ondrej Jaško
Univerzitet u Beogradu, Fakultet organizacionih nauka
Beograd, Srbija
ondrej.jasko@fon.bg.ac.rs
0000-0003-4877-0592

Petar Stanimirović
Univerzitet u Beogradu, Fakultet organizacionih nauka
Beograd, Srbija
petar.stanimirovic@fon.bg.ac.rs
0000-0002-6610-5820

Miha Marič
Univerza v Mariboru, Fakulteta za organizacijske vede
Kranj, Slovenija
miha.maric@um.si
0009-0001-6689-7944

***Apstrakt* - Finansije predstavljaju jednu od ključnih funkcija u svakoj organizaciji. Bez obzira na model organizacione strukture, finansije se javljaju kao organizaciona celina koja pruža jednu od osnovnih funkcija podrške. Prilikom uvođenja ERP rešenja finansijski modul je uvek među prvima koji se kupuju i implementiraju, neretko i jedini u inicijalnoj fazi. Precizno i brzo finansijsko izveštavanje predstavlja jedan od osnovnih motiva za nabavku ERP sistema. Pored ovih benefita, digitalizacija finansijskih procesa dovodi i do niza drugih promena, koje utiču na sve ključne elemente organizacije. U ovom radu će biti prikazano na koji način uvođenje ERP rešenja utiče na organizacioni dizajn u domenu finansijske funkcije. Rezultati su generisani na osnovu analize promena u organizaciji finansijskih poslova u više kompanija koje su uvele ERP sistem. Rezultati mogu biti iskorišćeni od strane menadžera kompanija i eksperata angažovanih na implementaciji ERP rešenja, u cilju sticanja boljeg uvida u organizacione promene koje moraju da isprate ovaj postupak, kao i u potencijalne dodatne indirektne benefite, koji se reflektuju kroz unapređenje organizacije, a ne kroz direktno poboljšanje finansijskog kontrolinga, analitike i izveštavanja.**

Ključne reči – finansije, računovodstvo, organizacija, menadžment, ERP sistem, digitalizacija, optimizacija procesa.

I. UVOD: ERP SISTEMI

ERP sistemi predstavljaju sveobuhvatan softver dizajniran da integriše i upravlja svim poslovnim funkcijama organizacije [1], poput prodaje, proizvodnje, logistike, finansija, računovodstva i ljudskih resursa [2]. Ove funkcije su ranije bile pohranjene u autonomnim softverskim jedinicama [3], a implementacijom ERP sistema se nastoji da sve funkcije budu centralizovane u jedinstveni računarski sistem [1], odnosno jedinstvenu centralnu bazu podataka [4], kako bi se prikupljale informacije na jednom istom mestu i učinile dostupnim unutar cele organizacije [5]. Ova rešenja su od ključnog značaja za poslovanje u savremenom dobu [6].

ERP sistemi su standardizovana, integrisana softverska rešenja zasnovana na „najboljim praksama“ iz različitih industrija [7] koja pomažu u rešavanju problema fragmentacije organizacionih informacija, automatizujući i integrišući ključne poslovne procese, obezbeđujući informacije u realnom vremenu o tim procesima i omogućavajući međusektorsko deljenje zajedničkih podataka i praksi unutar preduzeća [8]. Jedan su od ključnih koraka digitalne transformacije, koja podrazumeva upotrebu digitalnih tehnologija u svrhu kreiranja novih ili modifikovanja postojećih poslovnih modela i procesa i predstavljaju podršku transformaciji organizacione strukture, resursa, kao i internih i eksternih odnosa [9].

Feran i Salim [10] su istakli da je implementacija ERP sistema usko povezana sa povećanjem nivoa produktivnosti, smanjenjem troškova i povećanjem efikasnosti [2], a Barna i Igna [11] navode još neke glavne prednosti implementacije ERP sistema:

- potpunost i tačnost poslovnih podataka, generisanih u realnom vremenu;
- niski operativni troškovi;
- unapređenje u deljenju podataka korisnicima finansijskih informacija;
- automatizacija aktivnosti;
- smanjenje rizika zahvaljujući implementiranim finansijskim kontrolama.

II. ERP SISTEMI I FINANSIJSKA FUNKCIJA U ORGANIZACIJI

Razvoj informacionih sistema, naročito u domenu finansija i računovodstva, prepoznat je kao ključni pokretač operativne efikasnosti, a empirijski dokazi ukazuju na to da ERP sistemi igraju ključnu ulogu u optimizaciji poslovnih procesa, doprinoseći tačnosti, pouzdanosti i relevantnosti finansijskih podataka [6]. ERP sistemi zamenjuju odvojene računarske sisteme u oblastima finansija, ljudskih resursa, proizvodnje i skladištenja jedinstvenim softverskim programom sa modulima koji u velikoj meri podsećaju na prethodne sisteme. Nekada odvojeni softveri ovim postaju integrisani, što omogućava skoro trenutno deljenje informacija koje nastaju unosom podataka u različitim funkcijama [1].

Implementacija ERP sistema unapređuje finansijsko-računovodstvene procese značajnim smanjenjem kašnjenja u izveštavanju, kao i izradom preciznijih i verodostojnijih finansijskih izveštaja. Pored toga, obim podataka sa kojima kompanije moraju da se nose, imajući u vidu različite procese

koji se realizuju u magacinskom poslovanju, prodaji, proizvodnji itd., postaje lakše upravljiv uz ERP sisteme, jer prestaje potreba za višestrukim unosom podataka [3]. Jednom unet podatak u određenoj funkciji, na određenom radnom mestu, ostaje u sistemu i postaje potencijalno vidljiv svim ulogama koje imaju odobreno pravo pristupa konkretnoj informaciji. ERP sistemi omogućavaju automatizaciju velikog broja aktivnosti [12], čime se smanjuje vreme potrebno za finansijske procedure, poput izrade finansijskih izveštaja, jer se oni generišu na automatizovan način. Time se zaposlenima na finansijsko-računovodstvenim poslovima obezbeđuje efikasnije korišćenje raspoloživog vremena, budući da sada provode manje vremena na zatvaranju stavki i finansijskih knjiga, a donosioci odluka imaju više vremena da detaljno analiziraju te izveštaje [3]. Korišćenje ERP rešenja vodi ka smanjenju manuelnog rada, obezbeđujući podršku u obradi svakodnevnih transakcija, podršku u internom donošenju odluka i ispunjavanju obaveza vezanih za administraciju [12].

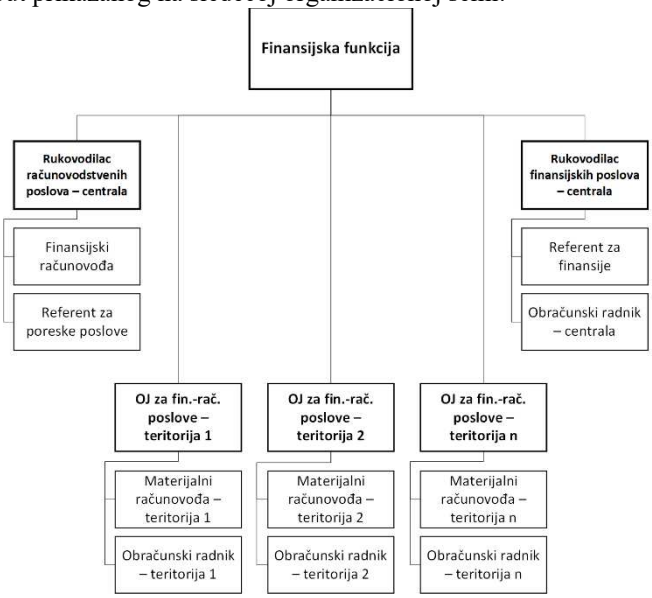
ERP sistemi većine dobavljača omogućavaju implementaciju pojedinačnih modula, bez potrebe za kupovinom celog paketa, a mnoge organizacije inicijalno implementiraju upravo modul za finansije, dok od ostalih u potpunosti odustaju ili ih uvode u kasnijoj fazi implementacije ERP sistema [1]. Neke od glavnih prednosti ERP rešenja koje se tiču finansijske funkcije su [12]:

- brza obrada finansijsko-računovodstvenih informacija;
- čuvanje informacija u jedinstvenoj bazi podataka i kontrolisan pristup podacima;
- bolja kontrola i praćenje informacija obrađenih pomoću ovih sistema;
- efikasnije upravljanje resursima (finansijskim, materijalnim i ljudskim);
- tačnost i potpunost informacija;
- smanjenje troškova;
- ušteda vremena;
- ograničavanje ljudskih grešaka.

III. Uticaj uvođenja ERP sistema na organizacionu strukturu u domenu finansijskih poslova

Iznesena zapažanja temelje se na analizi nekolicine studija slučaja, razvijenih u okviru konsultantske prakse autora u oblasti organizacionog restrukturiranja i unapređenja procesa. Konsultantski projekti su obuhvatili velike kompanije sa sedištem u Srbiji, koje svoje operativne aktivnosti realizuju na geografski dislociranim lokacijama.

Organizacije koje svoju delatnost obavljaju na više različitih lokacija i imaju dislocirane proizvodne i administrativne jedinice često organizuju i finansijsku funkciju decentralizovano, u skladu sa geografskom distribucijom svojih kapaciteta. Na ovim lokacijama tradicionalno postoje posebne organizacione celine sa radnim mestima na kojima se obavljaju pretežno poslovi materijalnog (pogonskog) računovodstva i obračuna zarada za zaposlene koji se fizički nalaze na toj lokaciji. Organizaciona struktura takvih kompanija obično ima oblik poput prikazanog na sledećoj organizacionoj šemi:



Slika 1. Ilustracija organizacione šeme finansijske funkcije sa teritorijalno rasprostranjenim organizacionim jedinicama.

U centrali kompanije se nalaze radna mesta čiji je zadatak objedinjavanje podataka i realizacija finansijsko-računovodstvenih operacija koje obuhvataju organizaciju u celini, dok se na svakoj od teritorija obavljaju zadaci karakteristični za tu teritoriju i prikupljaju pojedinačni podaci koji se prosleđuju centrali na objedinjavanje, obradu i analizu.

Zaposlenima na finansijsko-računovodstvenim poslovima u teritorijalno rasprostranjenim organizacionim jedinicama je obično organizaciono nadređen rukovodilac finansijsko-računovodstvenih poslova u centrali kompanije, iako ovako dislocirane organizacione jedinice otežavaju upravljanje iz centrale. Retke su situacije gde svaka finansijsko-računovodstvena organizaciona jedinica ima svog rukovodioca na toj teritoriji, iz razloga što tu postoji samo nekoliko tipova radnih mesta, sa po jednim ili jako malim brojem izvršilaca na njima, pa ne postoji potreba za poslovima rukovođenja i koordinacije. Sa druge strane, ovako dislocirane organizacione jedinice su nekada otežavale upravljanje rukovodiocima iz centrale, jer su oni mogli da vide samo produkte njihovog rada, ali ne i da organizaciju sam način rada, ukoliko nisu fizički prisutni na mestu rada.

Postojeća tehnologija i način na koji su organizovani finansijsko-računovodstveni procesi u određenim kompanijama uslovljavaju potrebu za zaposlenima koji obavljaju poslove materijalnog računovodstva, a koji se fizički nalaze u pogonu ili neposredno uz sam pogon. Osnovna uloga zaposlenih koji obavljaju poslove materijalnog računovodstva je tačna i pravovremena evidencija materijalnih tokova, predstavljajući tako osnovu za internu kontrolu i izveštavanje iz tog domena [13]. Najveći deo njihovih radnih zadataka predstavlja evidentiranje materijalnih tokova u skladištu i u toku proizvodnog procesa, ručnim knjiženjem papirne dokumentacije kroz lokalna softverska rešenja, dostupna na mestu rada. Ovakav način rada podrazumeva višestruku proveru dokumentacije, što

povećava rizik od dupliranih aktivnosti i/ili šifara, pojave grešaka i kašnjenja u izveštavanju.

Generalno, informacione i komunikacione tehnologije (IKT) transformišu ključne parametre organizacione komunikacije — frekvenciju, vremensku usklađenost i mesto komunikacije. Digitalizacijom, učestalost izveštavanja i sastanaka postaje manje resursno zahtevna, a razmena informacija značajno efikasnija u odnosu na komunikaciju licem u lice. IKT podržavaju i sinhronu i asinhronu komunikaciju, pri čemu omogućavaju da asinhrona razmena informacija dostigne gotovo isti nivo efikasnosti kao sinhrona, u situacijama kada zadatak ne zahteva simultano delovanje. Istovremeno, omogućena je komunikacija među dislociranim timovima, nezavisno od fizičke udaljenosti i vremenskih zona, što značajno doprinosi agilnosti organizacione koordinacije [14].

Tehnološki napredak izaziva promene u organizacionoj strukturi [15]. U tradicionalnim organizacionim strukturama, izvršioci finansijskih procesa su uglavnom organizaciono objedinjeni ispod hijerarhije finansijskog rukovodstva, u cilju ispunjenja zahteva za neprekidnošću i kontrolom procesa. U mnogim slučajevima, ERP sistemi ne mogu poboljšati produktivnost i efikasnost organizacije ukoliko ona ne prilagodi svoje poslovne procese pravilima ERP sistema [7]. Implementacija ERP sistema bi trebalo da dovede i do promene organizacionog dizajna, gde digitalizacija omogućava obezbeđivanje neprekidnosti i kontrole procesa i onda kada se posao obavlja od strane izvršilaca koji su izvan finansijske funkcije, na primer, u okviru funkcije nabavke ili logistike. Upotreba ERP rešenja omogućava ublažavanje dejstva Alenove krive [16], pre svega usled unapređenja u sferi komunikacije, odnosno delovanja na parametre frekvencije i vremenske usklađenosti [14].

U modelu poslovanja pre implementacije ERP sistema, funkcionalna podela između materijalnog i finansijskog knjigovodstva je bila jasno izražena – prvi su vodili količinske evidencije, dok su drugi upravljali finansijskim aspektima (računi, PDV, amortizacija, glavna knjiga). Koordinacija između ova dva segmenta je podrazumevala učestalu razmenu obimne papirne dokumentacije. Implementacijom ERP sistema, određene transakcije automatski generišu finansijske zapise u finansijskom modulu, čime se uspostavlja automatska i direktna veza između količinskih i finansijskih evidencija. Ovo omogućava ukidanje pozicije materijalnog računovođe ili njegovu integraciju sa drugim radnim mestima, kao što su magacioner-računopolagač ili finansijski knjigovođa. Deo operativnih zadataka, koje nije moguće automatizovati, preuzima magacioner uz dodatnu obuku i ovlašćenje za rad u ERP sistemu, koji olakšava realizaciju rutinskih poslovnih procesa [1]. Time magacioner postaje računopolagač sa proširenim nadležnostima i odgovornošću za tačnost unosa u sistem. Ovakvom reorganizacijom, magacioner-računopolagač postaje odgovoran za deo procesa materijalnog računovodstva, odnosno evidentiranje prijema robe na osnovu otpremnice i ažuriranje zaliha u realnom vremenu uz formiranje izveštaja o stanju zaliha. Nasuprot tome, poslovi koji ostaju u domenu centralnog računovodstva uključuju usklađivanje količinskog i finansijskog stanja, knjiženje vrednosti i kontiranje zaliha, knjiženje viška i manjka i pripremu internih obračuna i izveštaja. Takvom organizacijom procesa i integracijom je omogućeno zaposlenom na računovodstvenim poslovima da koristi podatke koje je uneo magacioner za sprovođenje kontrole, zatvaranje faktura i konsolidaciju finansijskih izveštaja [13].

Procesi usmeravaju tok informacija kroz organizacioni sistem [17]. Sama digitalizacija poslovnih procesa dovodi do toga da kod još jednog segmenta finansijskih poslova prestane da postoji potreba za zaposlenima koji obavljaju te poslove na teritorijalno dislociranim mestima, a to su poslovi obračuna zarada. Jedan od osnovnih procesa obračuna zarada, beleženje prisustva na poslu, se sve više izvršava automatski, kroz sistem za čekiranje sa identifikacionim karticama koji je direktno povezan na ERP sistem. Samim tim, prestaje da postoji potreba za ručnim beleženjem prisustva na poslu, manuelnim unosom i redovnom iscrpljujućom kontrolom podataka koji se unose u sistem. Pored toga, poslovi poput overe različitih potvrda za kredite i administrativne zabrane se takođe mogu obavljati dislocirano, ukoliko se internom uredbom tako propiše, gde bi se zahtevi od strane zaposlenih slali elektronski ili poštom, umesto ličnom dostavom obračunskim radnicima na terenu.

Opisane promene značajno utiču na organizacionu strukturu preduzeća usled transformacije poslovnih procesa finansijske funkcije, naročito u segmentima koji su se prethodno realizovali lokalno. Automatizacijom većine aktivnosti, kao i preraspodelom zadataka na druga radna mesta, stvaraju se uslovi za reorganizaciju i racionalizaciju procesa, koja se odnosi na ukidanje radnog mesta koje obavlja poslove isključivo materijalnog računovodstva, budući da se ključne funkcije ovog radnog mesta više ne obavljaju manuelno. Istovremeno, implementacijom sistema za automatsko evidentiranje prisustva zaposlenih i uvođenjem elektronske dostave dokumenata u vezi sa obračunom zarada, eliminiše se potreba za zaposlenima koji obavljaju poslove obračuna zarada na pojedinačnim lokacijama. Kao posledica ukidanja ovih radnih mesta, prestaje potreba za postojanjem određenih organizacionih jedinica, čime se otvara prostor za redefinisanje organizacione strukture kroz model centralizovanog obavljanja finansijsko-računovodstvenih poslova.

Dodatno, usled centralizacije i povećane tačnosti podataka kreiraju se podloge za naprednije analitike i izveštavanja. Ovo stvara potrebu za kvalifikovanijim radnicima sa znanjima iz oblasti analize podataka i upotrebe različitih *business intelligence* (BI) alata, što je dodatna izmena u strukturi radnih mesta, kroz otvaranje novih pozicija analitičara koje ranije nisu postojale ili izmenu opisa poslova odgovarajućih radnih mesta. Takve promene se mogu javiti u samoj finansijsko-računovodstvenoj funkciji, ili u organizacionoj celini zaduženoj za poslove finansijskog kontrolinga.

Takođe, javljaju se i pozicije nadležne za koordinaciju i administraciju ERP sistema u domenu finansijsko-računovodstvenih poslova, dok kod većine radnih mesta dolazi do izmena u opisima poslova i zahtevanim kompetencijama izvršilaca angažovanim na njima, kako bi se obezbedila adekvatna primena novih softverskih rešenja.

IV. ZAKLJUČAK

Automatizacija procesa kroz primenu softverskih rešenja zahteva i odgovarajuće promene u organizaciji rada [18]. Implementacija ERP sistema u organizacijama sa složenom strukturom i dislociranim organizacionim jedinicama transformiše njen organizacioni dizajn, a naročito način na koji je organizovana finansijska funkcija u kompaniji. Saznanja koja su predstavljena u ovom radu, a naročito kroz primere iz pomenutih studija slučaja, pokazuju da uvođenje ERP sistema zahteva određene promene u strukturi, a koje se uglavnom odnose na:

1. *Ukidanje određenih strogo operativnih i visokospecijalizovanih pozicija, čiji je posao isključivo pohranjivanje sistema*. Automatizacijom velikog broja operacija kroz ERP sistem i alokacijom pojedinih aktivnosti materijalnog računovodstva na druga radna mesta u procesu, određena radna mesta ostaju bez realnog radnog zadatka u toku radnog dana. Automatizacija operacija dodatno utiče na smanjenje specijalizacije izvršilaca finansijsko-računovodstvenih poslova, a samim tim i broja različitih radnih mesta u ovoj funkciji. Takođe, značajno se povećava i produktivnost [11] rada na finansijsko-računovodstvenim poslovima, jer je sada moguće obraditi i proknjižiti značajno veći broj dokumenata nego pre implementacije ERP sistema, što navodi da je, pri istom obimu posla, potreban značajno manji broj zaposlenih za obavljanje finansijsko-računovodstvenih aktivnosti.
2. *Uvođenje pojedinih analitičkih pozicija na finansijsko-računovodstvenim poslovima, za kojima se javlja potreba usled novih mogućnosti i podloga za analitiku koje pruža ERP*. U novoj strukturi se projektuju radna mesta sa širim obuhvatom posla, a zaposleni na ovim radnim mestima imaju minimalni broj ručnih transakcija koje izvršavaju, dok sve više preuzimaju kontrolnu i korektivnu ulogu u procesu, ali nad većim brojem transakcija [3], kao i poslove napredne analitike i izveštavanja. Dodatno, na osnovu mnogobrojnih predefinisanih izveštaja koje je moguće da ERP kreira, zaposleni na ovim radnim mestima su u mogućnosti da tumače podatke i definišu različita scenarija, za čiju izradu je nekada bilo potrebno neuporedivo više vremena.
3. *Uvođenje specijalizovanih pozicija, koje imaju ulogu koordinatora ERP sistema i čine sponu između dobavljača i korisnika sistema*. Njihova uloga jeste da na najbolji mogući način učestvuju u implementaciji i prilagođavanju konkretnih modula ERP sistema specifičnostima procesa u kompaniji, prenose korisničke zahteve dobavljačima rešenja i predstavljaju internog stručnjaka za sistem, kome mogu da se obrate korisnici ukoliko imaju neki problem ili zahtev.
4. *Promene u potrebnim kompetencijama izvršilaca na finansijsko-računovodstvenim poslovima*. Svi zaposleni koji će obavljati finansijsko-računovodstvene poslove se moraju obučiti za rad u ERP sistemu [6], jer se obrada bilo kog finansijskog dokumenta mora realizovati kroz isti. Međutim, imajući u vidu da se automatizacijom aktivnosti izgubila potreba za najjednostavnijim operativnim finansijsko-računovodstvenim poslovima, ali se javila potreba za drugim poslovima, koji podrazumevaju dublje poznavanje finansija, računovodstva i finansijsko-računovodstvenih aspekata poslovanja. To znači da se pored obuke za rad u ERP sistemu, zaposleni moraju konstantno nadograđivati i u pogledu domenskog znanja, kako bi ispratili i iskoristili sve mogućnosti koje implementacija ERP sistema pruža.
5. *Organizaciona i teritorijalna centralizacija poslova i radnih mesta*. Decentralizovani poslovi materijalnog računovodstva, obračuna zarada i sl. se ukidaju ili redefinišu, dok se aktivnosti koje je potrebno realizovati konsoliduju u centralnim timovima, čime se stvara osnov za efikasnije upravljanje resursima i kontrolu procesa.

Ovo istraživanje sprovedeno je na uzorku ograničenog broja kompanija iz Srbije, što može uticati na opštu primenu dobijenih zaključaka. Potrebno je sprovesti dodatna istraživanja koja bi obuhvatila međunarodna iskustva, kako bi se omogućilo poređenje rezultata u različitim ekonomskim, organizacionim, društvenim i kulturnim kontekstima. Takođe, preporučuje se detaljnije ispitivanje promena u velikim kompanijama, malim i srednjim preduzećima, kao i u različitim državama, kako u razvijenim, tako i u nerazvijenim ekonomijama, budući da se karakteristike radnih mesta i organizacionih struktura mogu značajno razlikovati u zavisnosti od veličine preduzeća, stepena ekonomskog razvoja i opšte informatičke pismenosti.

LITERATURA

[1] F. Tuli and S. Kaluvakuri, "Implementation of ERP Systems in Organizational Settings: Enhancing Operational Efficiency and Productivity", Asian Business Review, 2022, str. 89-96.

[2] O. Abdulfattah, „The social impacts of ERP implementation on employees and work environments in higher education institutions", International Journal of Advanced and Applied Sciences, 2020, str. 78-85.

[3] G. Spyridon, D. Charamis, and E. Tabouratzi, "Accounting Benefits of ERP Systems across the Different Manufacturing Industries of SMEs", Theoretical Economics Letters 8, 2018, str. 1232-1246.

[4] N. Dechow and J. Mouritsen, "Enterprise resource planning systems, management control and the quest for integration", Accounting, Organizations and Society 30, 2005, str. 691–733.

[5] B. Johansson and M. Bystrom "The role of organizational culture in ERP implementation - The case of replacing an old ERP in a retail organization", Procedia Computer Science 239, 2024, str.1911–1918.

[6] H. Tran and N. Nguyen, “Impact of ERP System Implementation on Accounting Information Quality in Vietnamese SMEs”, Journal of Accounting, Finance and Auditing Studies, 2024, str. 1-9.

[7] M. Jovanović, S. Komazec and L. Miloš, “Design Parameters Adjustment To Implementation Period: A Case Study From Serbia”, Zbornik radova 19. konferencije “Symorg 2024: Unlocking the Hidden Potentials of Organization Through Merging of Humans and Digitals”, University of Belgrade – Faculty of Organizational Sciences, Zlatibor, 2024, str. 373-377.

[8] L. Bailey, L. Seymour and J. Van Belle, “Impact of ERP implementation on the quality of work life of users: A sub-Saharan African study”, The African Journal of Information Systems, 2017, str. 192-212.

[9] I. Todorović, M. Jovanović, J. Krivokapić, D. Milković, V. Lučanin and J. Tanasković, “Digital Transition of the Maintenance Process: Case of Rail Transport Company”, Conference Proceedings from 43rd International Conference on Organizational Science Development, University of Maribor, Maribor, 2024, str. 987-998.

[10] C. Ferran and R. Salim, “Enterprise Resource Planning for Global Economies: Managerial Issues and Challenges”, New York: Information Science Reference, 2008.

[11] L. Barna, and R. Igna, “The influence of the implementation of ERP systems on the performance of an organization”, International Conference on Business Excellence, Sciendo, 2021, str. 268-279.

[12] B. Ionescu, L. Barna, “Digitalization in the Accounting and Auditing Profession through ERP Systems”, Audit Financiar, 2021, str. 769-778.

[13] L. Reddi, “Transforming Management Accounting: Analyzing The Impacts Of Integrated SAP Implementation”, International Research Journal of Modernization in Engineering Technology and Science, 2023, str. 1786-1793.

[14] O. Jaško, M. Čudanov, M. Jevtić, i J. Krivokapić, “Organizacioni dizajn – pristupi, metode i modeli”, Beograd: Fakultet organizacionih nauka, Univerzitet u Beogradu, 2017.

[15] I. Todorović, M. Čudanov and S. Komazec, “Interrelationships of Changes in Organizational Structure and Technology”, In Ferjan, M., Kljajić Borštnar, M., Marič, M., Pucihar, A., Bernik, M. (Ed.) Quality, Inovation, Future: Proceedings of the 31st International Conference on Organizational Science Development, Moderna organizacija, Kranj, Slovenia, 2012, str. 1264-1271.

[16] T. Allen, “Managing the Flow of Technology: Technology Transfer and the Dissemination of Technological Information Within the R&D Organization”, Cambridge: MIT Press, 1984.

[17] I. Todorović, S. Komazec i O. Jaško, “Model podataka za informacioni sistem koji optimizuje angažovanje kadrova na održavanju tehničkih sistema”, Zbornik radova 27. konferencije “YU INFO 2021”, Društvo za informacione sisteme i računarske mreže, Kopaonik, 2021.

[18] I. Todorović, S. Komazec and M. Marič, “Organizational Preconditions for Turning Maintenance Planning into Smart Automated Process”, Proceedings of the XVII International Symposium of Organizational Sciences: Business and Artificial Intelligence, Faculty of Organizational Sciences, Belgrade, Serbia, 2020, str. 437-444.

The Impact of ERP System Implementation on the Internal Organization of the Financial Function within a Company

Milan Jovanović, Ivan Todorović, Ondrej Jaško, Petar Stanimirović and Miha Marič

ABSTRACT

Finance represents one of the key functions in every organization. Regardless of the organizational structure model, finance emerges as an organizational unit providing one of the core support functions. During the implementation of an ERP solution, the financial module is almost always among the first to be purchased and deployed, and often the only one implemented in the initial phase. Accurate and rapid financial reporting is one of the main drivers for acquiring an ERP system. In addition to these benefits, the digitalization of financial processes leads to a range of other changes that impact all key elements of the organization. This paper will present how the introduction of an ERP solution affects organizational design within the domain of the financial function. The results are generated based on an analysis of changes in the organization of financial operations in several companies that have implemented ERP systems. These findings can be utilized by company managers and experts engaged in ERP implementation projects to gain a better understanding of the organizational changes that must accompany the process, as well as the potential additional indirect benefits that are reflected through organizational improvements, rather than solely through direct enhancements in financial controlling, analytics, and reporting.

CHATGPT KAO PODRŠKA INTELIGENTNIM RUDNICIMA

Vidosav Majstorović
Mašinski fakultet, Univerzitet u Beogradu
vidosav.majstorovic@gmail.com
[0000-0001-9534-846]

Vladimir Simeunović
Institut „Mihajlo Pupin“ Beograd
vladimir.simeunovic@pupin.rs

Dragan Stošić
Institut „Mihajlo Pupin“ Beograd
dragan.stosic@pupin.rs

Rastko Negoicić
AD EPS, Ogranak TE-KO Kostolac
rastko.negoicic@te-ko.rs

Filip Todorović
AD EPS, Ogranak TE-KO Kostolac
filip.todorovic@te-ko.rs

***Apstrakt* - Industrija 4.0 kao model u rudarstvu počinje sve više da se primenjuje, tako da danas imamo studije ali i prva pozitivna iskustva u ovoj oblasti. Trend razvoja u ovoj oblasti su pametni rudnici, a oni se zasnivaju na primeni alata i tehnika veštačke inteligencije (AI): modeli učenja (ML/DML), generativna AI i chatGPT. Ovaj rad analizira mogućnosti za razvoj i primenu modela chatGPT u pametnom rudarstvu, sa posebnim osvrtom na mogućnost primene u već ranije razvijenom i primenjenom modelu Industrije 4.0 za površinski kop „Drmno“, na kome ovaj tim radi već nekoliko godina.**

Ključne reči: Industrija 4.0, Pametno rudarstvo, Veštačka inteligincija, ChatGPT

I. UVODNE NAPOMENE

Budućnost rudarenja su pametni rudnici, ili drugačije, puna primena modela Industrije 4.0 u njima. U takvom sledu događaja, uloga veštačke inteligencije (AI) će biti velika i značajna, što znači da će ona primenjivati za [1]: (i) digitalnu rudarsku proizvodnju, koja će biti zasnovana na AI, gde će kreirati digitalne verzije fizičkih rudarskih polja i entiteta, vršiti njihova simulacija i online praćenje proizvodnje, (ii) planiranje u upravljanje proizvodnjom, AI će pomoći da se stvore uslovi, koji će zadovoljiti sve standarde zelene proizvodnje uz štednju resursa, (iii) optimizacija rudarskih operacija proizvodnje, AI može analizirati proizvodne podatke kako bi se optimizirali proizvodni i transportni procesi, smanjujući potrošnju energije, otpad i troškove, (iv) optimizacija lanca snabdevanja, se pomoću AI vrši tako da se optimiziraju zalihe u rudnuku (proizvodnje, potrošnje), logistika i rute isporuke, smanjujući troškove i skarćujući vreme isporuke, kao i upravljanje rizikom, (v) optimizacija održavanja kroz primenu modela prediktivnog održavanja a uz pomoć proširene stvarnosti (AR) i digitalnih blizanaca (DT), gde se koriste algoritmi mašinskog učenja za analizu podataka o performansama opreme, predviđajući kada je održavanje potrebno i proaktivno planirajući održavanje, smanjujući zastoje i povećavajući ukupnu efikasnost rudarske opreme, (vi) kontrola i obezbeđenje kvaliteta rude, AI se koristiti za analizu podataka o kvalitetu rude u realnom vremenu, otkrivanje nedostataka, uz omogućavanje korektivnih radnji za poboljšanje kvaliteta rude, (vii) analitika podataka zasnovana na AI će analizirati velike količine proizvodnih podataka (BDA), kako bi se identifikovali trendovi, obrasci ponašanja i korelacije, omogućavajući donošenje odluka na osnovu podataka, za sve poslovno-proizvodne aktivnosti u rudniku, i (viii) sajber bezbednost - sigurnosni sistemi podržani AI mogu otkriti prijetnje i odgovoriti na njih u realnom vremenu, štiteći se od sajber-napada, a u rudniku se to još više potencira primenom privatnog cloud SaaS modela.

Cilj rada je da ukaže na jednu veoma važnu oblast buduće primene AI u rudarstvu, kao što je ChatGTP, sa dva aspekta: mogućnosti primene i jednog razvojnog Projekta.

Ovo su neki pravci budućih oblasti primene AI u pametnom rudniku, a kako tehnologija AI (GenAI, ChatGTP) nastavlja da se razvija, možemo očekivati da će se u budućnosti pojaviti još inovativnije aplikacije, kao platforme AI za pametne rudnike. Ovaki modeli su već sada u eksperimentalnoj fazi razvoja (Goldspot Discoveries, Earth AI, Minerva Intelligence, DroneDeploy, Hikvision, Imago i druge).

Ovaj rad ima nekoliko celina: (i) uvod sa elementima pametnog ridarstva, ciljem i sadržajem rada, (ii) pregled literature sa mogućnostima primene ChatGTP, (iii) primer jednog razvojnog Projekta, i (iv) zaključci i buduća istraživanja.

II. PREGLED LITERATURE

ChatGPT je moćan jezički model AI, koji se može primeniti u različitim industrijama za automatizaciju zadataka, poboljšanje efikasnosti raličitih aktivnosti, kao i poboljšanje korisničkog iskustva. Tako na primer u pametnom rudniku, ovaj alat AI nam može biti od koristi za: (i) generiranjem izvještaja iz MES o radnom nalogu, pružanjem povratnih informacija u realnom vremenu i omogućavanjem daljinskog nadzora i kontrole proizvodnih procesa u rudniku, (ii) generisanjem izveštaja iz ERP sistema, pružanjem personalizovane korisničke podrške i na primer optimizacijom logistike lanca snabdevanja, (iii) pomoć QMS (IMS) i ESG modelima u analizi podataka iz procesa kontrole kvaliteta, ekologije, emisije CO2 i drugih pokazatelja, a radi identifikacije nedostataka i poremećaja ovih parametara, i (iv) pomoć digitalnim entitetima rudnika generisanjem digitalnih blizanaca ovih entiteta, omogućavajući praćenje i simulaciju u realnom vremenu, proizvodnje i održavanja i drugih

ruraskih aktivnosti. Jedan pregled moguće primene ovog modela AI je prikazan u tabeli 1 [2-5].

Tabela 1. Elementi Industrije 4.0 na površinskom kopu "Drmno"

Oblast u rudniku	Aktivnost	Operacije u rudniku podržane ChatGTP	Prednosti primene ChatGTP
Geološka istraživanja na rudarskom polju	Automatizacija analize podataka u rudarskoj industriji	Analiza geoloških podataka radi identifikacije nalazišta minerala.	Smanjenje vrema i troškova geološkim istraživanjima, kroz automatizaciju obrade i predloge zaključaka o rezultatima geoloških istraživanja.
	Automatska identifikacija minerala	Identifikacija specifičnih minerala koristeći obradu prirodnog jezika i mašinsko učenje.	
	Automatska geotehnička analiza	Analiza geotehničkih podataka i predviđanje stabilnosti tla, smanjujući rizik od urušavanja i drugih opasnosti.	
	Automatsko mapiranje i geodetska snimanja	Analiza kartografskih i geodetskih podataka radi identifikacije potencijalna ležišta minerala.	
	Analiza tržišta minerala u realnom vremenu	Analiza tržišnih podataka u realnom vremenu kako bi se identifikovale profitabilne mogućnosti rudarenja i donele pravilne odluke za investiranje.	
Rudarske operacije	Praćenje proizvodnje u realnom vremenu	Praćenje rudarskih operacija u realnom vremenu i optimizacija proizvodnje uz povećavanje efikasnosti.	Smanjenje troškova automatizacijom praćenja proizvodnje.
	Automatsko planiranje bušenja	Analiza geoloških podataka i planiranje operacija bušenja.	Smanjenje troškova automatizacijom planiranja bušenja.
	Inteligentno sortiranje rude	Analizu uzoraka minerala i njihova klasifikacija prema njihovom sastavu i procentu učešća.	Smanjenje troškova optimizacijom klasifikacije rude.
	Autonomne rudarske operacije	Kontrolu i praćenje autonomne rudarske opreme (autonomna vozila).	Smanjenje troškova automatizacijom rudarskih operacija.
	Inteligentna optimizacija opreme	Optimizaciju rudarske opreme analizom podataka o korištenju opreme, performansama i potrebama održavanja.	Smanjenje troškova optimizacijom korištenja opreme.
	Nadzor sigurnosti u realnom vremenu	Praćenje rudarskih operacija u realnom vremenu i identifikaciju potencijalnih opasnosti, poboljšavajući sigurnost radnika.	Smanjenje troškova zbog iznenadnih događaja.
	Automatske sigurnosne inspekcije	Praćenje rudarskih operacija i obavljanje automatskih sigurnosnih inspekcija.	Smanjenje troškova inspekcije.
	Inteligentno upravljanje flotom	Optimizacija upravljanja rudarskom flotom BDA analizom podataka o korištenju vozila, potrebama održavanja i potrošnji goriva.	Smanjenje troškova optimizacijom korištenja voznog parka uz smanjenje potrošnje goriva.
	Inteligentna rudarska logistika	Optimizacija rudarske logistike analizom podataka o rudarskim operacijama, transportnim rutama i dostupnosti resursa.	Smanjenje troškova optimizacijom logistike uz smanjenje troškova transporta.
Održavanje mašina i opreme	Prediktivno održavanje u rudarstvu	Analizu podataka senzora iz rudarske opreme i predviđanje kada je potrebno održavanje, smanjujući zastoje i troškove.	Smanjenje troškova održavanja i povećanje efektivnosti rudarske opreme.
	Predviđanje otkaza opreme	Analizu podataka senzora i predviđanje kada će oprema verovatno otkazati, omogućavajući proaktivno održavanje i smanjenje vremena zastoja.	Smanjenju troškova održavanja i gubitaka proizvodnje zbog iznenadnog zastoja.
Kontrola kvaliteta rude	Inteligentna kontrola kvaliteta	Analiza uzoraka minerala i vršenje kontrole kvaliteta, smanjujući rizik od slanja rude niskog kvaliteta u dalji proces.	Smanjenje troškova automatizacijom kontrole kvaliteta.
	Automatsko obezbeđenje kvaliteta	Praćenje rudarskih operacija i izvođenje automatskih provera parametara kvaliteta, obezbeđujući tako usklađenost sa industrijskim standardima i propisima.	Smanjenje troškova automatizacijom obezbeđenja kvaliteta.
Upravljanje nabavkom u rudniku	Inteligentna optimizacija lanca snabdevanja	Optimizacija lanca nabavke u rudniku analizom podataka o dostupnosti resursa, transportnim rutama i performansama dobavljača.	Smanjenje troškova optimizacijom lanca nabavke i smanjenjem troškova transporta.
	Automatsko upravljanje zalihama	Analiza podataka o zalihama i optimizacija nivoa zaliha.	Smanjenje troškova automatizacijom upravljanja zalihama.
	Inteligentna optimizacija nabavke	Optimizacija nabavke u rudniku, BDA analizom podataka o performansama entiteta dobijenih od dobavljača, dostupnosti resursa i tržišnim trendovima.	Smanjenje troškova optimizacijom nabavke.
Podrška rudarskim operacijama	Inteligentno upravljanje radnom snagom	Analizu podataka o veštinama zaposlenih, produktivnosti radnog mesta i potrebama za rasporedom radi optimizacije korišćenja radne snage i smanjenja troškova rada.	Smanjenje troškova optimizacijom korištenja radne snage i smanjenjem troškova rada.

	Automatsko praćenje usklađenosti	Praćenje rudarskih operacija i obezbeđenja usklađenosti sa zakonskim propisima i industrijskim standardima.	Smanjenje troškova kroz ispunjenje zahteva.
	Automatsko praćenje parametara životne sredine	Praćenje rudarskih operacija i obezbeđenje usklađenosti sa ekološkim propisima i industrijskim standardima.	Smanjenje troškova kroz primenu EMS, ESG i CBAM dokumenata.
	Automatsko upravljanje rizicima	Analiza podataka i identifikaciju potencijalnih rizika, omogućavajući proaktivno upravljanje rizikom i smanjenje verovatnoće neželjenih događaja.	Smanjenje troškova (ISO 31000).

Analiza prikazana u prethodnoj tabeli nam omogućava da definišemo sledeće zaključke: (i) chatGTP će biti jedan od najvažnijih alata u pametnim rudnicima za generisanje različitih vrsta izveštaja za sve aktivnosti u rudniku, na bazi pouzdanih i tačnih podataka, koji su dobijeni iz digitalnog modela rudnika, na platformi Industrije 4.0, i (ii) ovaj koncept (chatGTP) biće platforma za izgradnju koncepta pametnog rudinika druge generacije.

III. PRIMER ZA RUDARSTVO – NAŠE ISTRAŽIVANJE

U tabeli 2, je dat primer stanja višegodišnjeg Projekta Industrije 4.0 za površinski kop „Drmno“. Zajednički projektni tim, ove Kompanije, Instituta „Mihajilo Pupin“ i Mašinskog fakulteta su redovno obaveštavali naučno stručnu javnost u zemlji i inostranstvu, o napretku ovog Projekta [7-10]. U donjoj tabeli se po prvi put, u koloni sasvim desno govori o mogućnosti primene ChatGTP alata u ovom Projektu.

U desnoj koloni tabele 2 su dati elementi Industrije 4.0, priminjeni ili u razvoju za ovaj površinski kop.

Tabela 2. Stanje razvoja i primene elemenata Industrije 4.0 na površinskom kopu „Drmno“, sa mogućnosću primene ChatGTP.

*Aktivnost (**primenjeno – P, u razvoju – R**)*	*Funkcija / oblast primene*	*Karakteristika*	*Softverska podrška / Element Industrije 4.0*	***Primena ChatGTP***
Upravljanje porudžbinama (R)	Rezervni delovi, gorivo, mazivo, procesne sirovine, ostali materijal	Optimalni nivo zaliha	ERP, MES (WIP)	*Generisanje izveštaja o porudžbinama.*
Praćenje radnih mašina (BTO) na površinskom kopu (P)	Praćenje mašina (bager – transport – odlagač) - CPS	Virtuelni GPS	GPS (MMS)	*Pregledi položaja radnih mašina i poruka po različitim osnovama.*
Kontrolna tabla pregleda stanja na površinskom kopu (P)	Praćenje proizvodnje u realnom vremenu	Virtuelno polje	GPS, GIS, DT	*Pregledi stanja na površinskom kopu.*
Održavanje mašina (R)	Radne mašine (BTO), rudarska oprema i postrojenja	Senzori	SCADA, ERP, MES (WIP), IoT, DT	*Različiti izveštaji o održavanja.*
Prediktivna analitika velikih podataka (BDA) (R)	Na nivou površinskog kopa, potrpšnja goriva, maziva i rezervnih delova, optimizacija održavanja radnih mašina (BTO) i ostalih postrojenja. Ceo lanac vrednosti na površinskom kopu	Analiza velikih podataka	ERP, BDA, AI (ML)	*Analiza podataka po različitim osnovama.*
On line upravljanje kvalitetom (proizvodnje uglja, rezervnih delova – guma i transportnih traka / BTO) (R)	Postrojenje za preradu, rudarska proizvodnja, sistem guma i transportnih traka (BTO)	Veliki podaci, vertikalna integracija	ERP, MES, QMS (IMS), BDA	*Analiza kvaliteta rude i kvaliteta održavanja.*
Veća bezbednost zaštita prirode (ergonomski uslovi) na radu kroz proširenu automatizaciju na bazi Industriju 4.0 (R)	Svi pogonski i proizvodni sistemi na kopu (prerada, geologija i planiranje rudnika)	Svi sistemi i oprema na površinskom su integrisani (BTO)	ERP, MES, ISO 14001, OH&S (IMS), ESG	*Analiza SMS i ESG pokazatelja.*
Poboljšan timski rad u proizvodnom okruženju kroz transparentnost i dostupnost podataka (P)	Rudarska proizvodnja (BTO), postrojenje za preradu, geologija i planiranje rudnika. Svi pogonski i proizvodni sistemi	Vertikalna integracija	ERP, MES, GPS, IMS	*Izveštaji o ERP i MES parametrima.*
Poboljšano okruženje kroz optimizovano korišćenje resursa, posebno energije (P)	Rudarska proizvodnja (BTO), postrojenje za preradu, geologija i planiranje rudnika	Integracija senzora za praćenje stanja. Pametni čitači.	MES, EMS	*Izvešatji o proizvodnji, po različitim osnovama.*
Proširene inovativne mogućnosti kroz nove tehnološke mogućnosti u proizvodnji (R)	Rudarska proizvodnja (BTO), postrojenje za preradu, geologija i planiranje rudnika	Vertikalni, horizontalni i slojevi integracije od kraja do kraja	ERP, MES, IMS, MMS, GPS, GIS	*Izvešatji o lancima nabavke.*
Izveštavanje o zalihama (P)	Svi rezervni delovi, gorivo, mazivo, procesne sirovine, ostali materijal	Izveštavanje u realnom vremenu, BDA, vertikalna integracija	ERP, MES	*Izveštaji o rezervnim delovima.*

IV. ZAKLJUČNA RAZMATRANJA

Pošto je osnovni kontekst ovog rada primena novih modela AI u pametnim rudnicima, u zaključku možemo da ukažemo i na trendove budućeg razvoja AI, koji će imati i direktnog uticaja na njenu primenu u pametnim rudnicima. Prema analizama iz [1-6], trendovi razvoja AI, koji važe i za njihovu primenu u pametnim rudnicima su: (i) opšta AI (GAI), je model koji se odnosi na hipotetičke sisteme AI, koji poseduju AI rešenja nalik ljudskoj sposobni da rasuđuje, uči i primenjuje znanje u širokom spektru zadataka. U tom slučaju GAI bi mogao nadmašiti ljudsku inteligenciju u mnogim područjima, (ii) edge AI, se odnosi na upotrebu AI i mašinskog učenja u graničnim oblastima, gde se podaci generišu i obrađuju u realnom vremenu, bez potrebe za centraliziranom obradom, (iii) transfer učenja, je tehnika u kojoj AI modeli mogu učiti iz jednog zadatka ili domena i primeniti to znanje na drugi povezani zadatak ili domen, bez potrebe za ponovnim učenjem, i (iv) kvantna AI, se zansiva ma kvantnom računarstvu, koje ima potencijal da unapredi AI, obezbeđujući nove metode za mašinsko učenje, optimizaciju i simulaciju, što bi moglo dovesti do novih otkrića u oblastima kao što su pametno rudarstvo, medicina, finansije i klimatsko modeliranje.

Sve navedene činjenice govore o velikoj perspektivi razvoja i primene naprednih AI modela i u pametnom rudniku.

ZAHVALNICA

Istraživanje opisano u ovom radu je delimično finansirano od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije. (Br. ugovora 451-03-136/2025-03/ 200034)

LITERATURA

[1] Smart mining: How Technology Can Improve Operations, [citirano Januara 2025. god.]. https://www.celona.io/5g-lan/smart-ining.

[2] Kristine Castillo, ChatGPT in Mining Operations, 2024. https://nexacu.com/insights-blog/chatgpt-in-mining-operations/.

[3] Smart Mines: The 6 Benefits of Making Your Mine Digital, https://www.worldsensing.com /knowledge-center/smart-mines-benefits/

[4] Smart Mining: Transforming the Mining Industry through Technological Innovations, https://horiz onpowered.com/smart-mining-and-technology/

[5] Wang, Fei-Yue & Yang, Jing & Wang, Xingxia & Li, Juanjuan & Han, Qing-Long. (2023). Chat with ChatGPT on Industry 5.0: Learning and Decision-Making for Intelligent Industries. IEEE/CAA Journal of Automatica Sinica. 10. 831-834. 10.1109/JAS.2023.123552.

[6]]. Gill, Sukhpal Singh & Kaur, Rupinder. (2023). ChatGPT: Vision and challenges. Internet of Things and Cyber-Physical Systems, Volume 3, 2023, Pages 262-271. 10.1016/j.iotcps.2023. 05.004.

[7] Vidosav Majstorovic, Vladimir Simeunovic, Zarko Miskovic, Radivoje Mitrovic, Dragan Stosic, Sonja Dimitrijevic, Smart Manufacturing as a framework for Smart Mining, Procedia CIRP, Volume 104, 2021, Pages 188-193, 10.1016/j.procir.2021.11.032.

[8] Vidosav, Majstorovic & Simeunović, Vladimir & Mitrovic, Radivoje & Stosic, Dragan & Dimitrijevic, Sonja & Miskovic, Zarko. (2022). How to apply the ERP model for Smart Mining?. MATEC Web of Conferences. 368. 10.1051/matecconf/202236801015

[9] Majstorovic, V., Simeunovic, V., Mitrovic, R., Stosic, D., Dimitrijevic, S., Miskovic, Z. (2023). Development of Cloud ERP Model and Its Application in Smart Mining. In: Burduk, A., Batako, A., Machado, J., Wyczółkowski, R., Antosz, K., Gola, A. (eds) Advances in Production. ISPEM 2023. Lecture Notes in Networks and Systems, vol 790. Springer, Cham. 10.1007/978-3-031-45021-1_3

[10] Majstorović, V. D., Simeunović, V. R., Stošić, D. P., Dimitrijević, S. T., Mitrović, R. M., & Mišković, Ž. Z. (2024). Pametno rudarstvo. Tehnika, 79(3), 279-286. 10.5937/tehnika2403279M.

Chatgpt As A Support For Intelligent Mines

Vidosav Majstorović, Vladimir Simeunović, Dragan Stošić, Rastko Negoicić, Filip Todorović

ABSTRACT

Summary: Industry 4.0 as a model in mining is starting to be applied more and more, so today we have, as studies, already the first positive experiences in this area. The development trend in this area is smart mines, and they are based on the application of artificial intelligence (AI) tools and techniques: learning models (ML/DML), generative AI and chatGPT. This paper analyzes the possibilities for the development and application of the chatGPT model in smart mining, with special reference to the possibility of application in the previously developed and applied Industry 4.0 model for the "Kostolac" surface mine, on which this team has been working for several years.

.

Unapređenje procesa obračuna utrošene električne energije u EPS AD Beograd

Jadranka Ristić
Akcionarsko društvo "Elektroprivreda Srbije" Beograd, Srbija
jadranka.ristic@eps.rs
ORCID broj 0009-0006-9674-104X

Apstrakt **- U skladu sa globalnom digitalizacijom, „Elektroprivreda Srbije" nastoji da ide u korak sa istom. Već duži niz godina uspešno se modernizuje i unapređuje poslovni sistem „Elektroprivrede Srbije", kroz proces digitalizacije. Najznačajniji digitalni projekat „Elektroprivrede Srbije", u ovom trenutku, jeste implementacija jedinstvene baze podataka i unapređenje procesa obračuna utrošene električne energije za kupce na garantovanom snabdevanju, kojih je više od 3,6 miliona. U radu su prikazani izazovi sa kojima se „Elektroprivreda Srbije" susrela u periodu unapređenja procesa obračuna, ali i benefiti koje je dobila sa novim načinom obračuna. Pored toga, detaljno su prezentovani dalji planovi na unapređenju istog.**

Ključne reči – obračun električne energije, EPS

I. UVOD

Akcionarsko društvo "Elektroprivreda Srbije" Beograd (u daljem tekstu: EPS AD Beograd) već duži niz godina uspešno modernizuje i unapređuje poslovni i tehnički sistem, kroz proces digitalizacije. Rezultat toga je digitalizacija mnogih procesa, dok su neki procesi trenutno u fazi prelaska na digitalni sistem.

Jedan od trenutno najznačajnih procesa u EPS AD Beograd, koji je u završnoj fazi unapređenja je proces obračuna utrošene električne energije za kupce na garantovanom snabdevanju, kojih je više od 3,6 miliona.

Glavni razlog unapređenja ovog procesa bio je veliki broj baza kupaca električne energije na području cele Republike Srbije. Pristup svakoj od baza podataka bio je omogućen preko različitih softverskih aplikacija, a svaka baza je imala posebnu proceduru za izradu obračuna utrošene električne energije. Ovakav sistem nosio je visok rizik od grešaka i zahtevao značajan vremenski angažman za testiranje i verifikaciju računa.

Nakon detaljne analize postojećeg stanja, EPS AD Beograd doneo je odluku da implementira jedinstvenu bazu podataka i objedinjeni sistem obračuna za sve kupce električne energije. Za realizaciju ovog projekta zaključen je ugovor sa kompanijom SAP, koja poseduje veliko iskustvo i znanje u implementaciji sistema u energetskom sektoru.

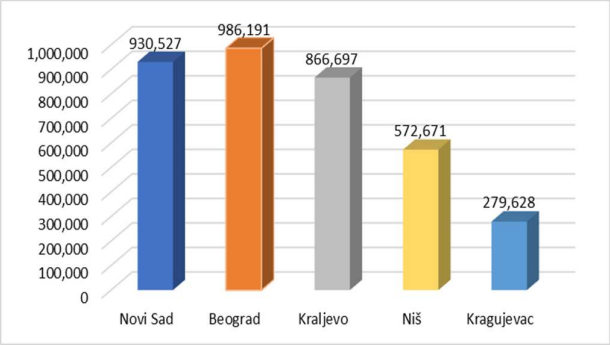
II. IZAZOVI U IMPLEMENTACIJI JEDINSTVENOG PROCESA OBRAČUNA UTROŠENE ELEKTRIČNE ENERGIJE

Svi kupci električne energije na garantovanom snabdevanju u Republici Srbiji, grupisani su u 5 regionalnih centara:

- Vojvodina (centar Novi Sad)
- Beograd (Centar Beograd)
- jugoistočna Srbija (Centar Niš)
- jugozapadna Srbija (centar Kraljevo) i
- centralna Srbija (centar Kragujevac).

Do sada su podaci o kupcima i pripadajućim mestima primopredaje iz svih regionalnih centara migrirani u jedinstvenu bazu, osim podataka kupaca sa područja Beograda. Planirano je da se do kraja maja 2025. godine završi i ovaj deo procesa.

Na slici 1. prikazan je broj mesta primopredaje po regionalnim centrima. Najveći broj je u Beogradu, a najmanji u centru Kragujevac.

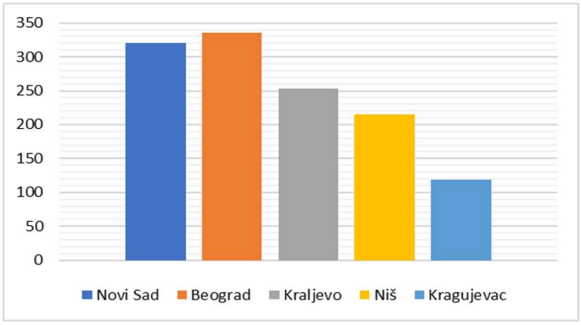


Slika 1. Pregled broja mesta primopredaje po regionalnim centrima

Migracije podataka vršene su postepeno, po regionalnim centrima, u odvojenim vremenskim intervalima. Ovi procesi bili su izuzetno zahtevni, jer je bilo neophodno da se prvo usklade svi šifarnici i razni tipovi podataka, a zatim izvrše migracije podataka.

Pre prve migracije podataka razvijen je proces obračuna električne energije, koji je rezultat sinergije timova EPS-a i SAP-a. Obračun električne energije u Republici Srbiji jedan je od najsloženijih. Pored toga, često (skoro svakog meseca) vrše se izmene u obračunu, poput dodavanja ili uklanjanja stavki na računu. Zbog toga je projektovanje i testiranje obračuna zahtevalo značajan vremenski period.

Dodatni izazov bio je obezbeđivanje obuke zaposlenih za rad u novom sistemu. U početku je bilo otpora prema promenama, što je uobičajena ljudska reakcija, ali postepenom edukacijom taj otpor je prevaziđen. Na slici 2. prikazan je broj korisnika koji rade na obračunu po regionalnim centrima.



Slika 2. Pregled broja korisnika angažovanih u procesu obračuna po regionalnim centrima

Najveći izazov bio je testiranje prvog ciklusa računa iz SAP sistema. Ovaj zadatak zahtevao je ogroman napor svih zaposlenih u EPS AD Beograd, ali je uspešno realizovan, a prvi računi iz SAP sistema podeljeni su kupcima iz centralne Srbije (centar Kragujevac).

III. FUNKCIONALNOSTI OBRAČUNA ELEKTRIČNE ENERGIJE

Proces obračuna utrošene električne energije je veoma složen i obavlja se na mesečnom nivou, rezultirajući sa preko 3,6 miliona računa za električnu energiju.

Pre početka, snabdevač mora da izvrši sledeće radnje, da bi sistem bio spreman za predstojeći obračun:

- unese parametre obračuna (cena električne energije, cene za pristup distributivnom sistemu, visina posebne naknade za podsticaj, visina naknade za unapređenje energetske efikasnosti, stopa akcize i PDV, iznos takse za JMS, prosečna potrošnja u domaćinstvima u Srbiji, kamatna stopa, rok plaćanja, rok za izradu računa i predaju pošti na dostavu kupcima
- učita spiskove: energetski ugroženih kupaca, korisnika javnih sredstava i mesta isporuke za obračun takse za javni medijski servis.
- proveri da li su evidentirane sve uplate kupaca.

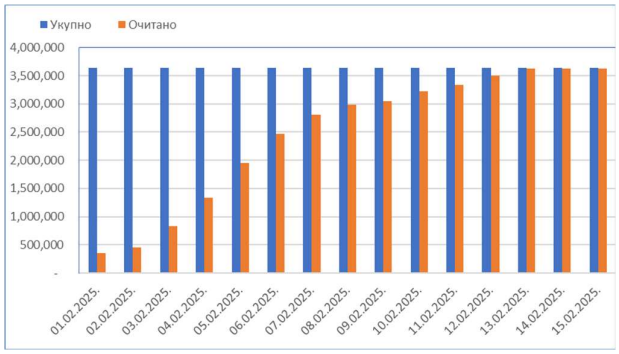
Jedinstvena baza podataka omogućila je da se svi navedeni spiskovi, kao i parametri unose jednom, što ranije nije bio slučaj.

Pre početka masovne izrade računa, u skladu sa procedurama kompanije, vrši se testiranje za sledeće karakterične tipove računa: novoprijavljeni kupac, neočitano brojilo, očitano brojilo sa 0 kWh, brojilo je očitano nakon x meseci, potrošnja do 350 kWh (zelena zona), potrošnja u sve tri zone, potrošnja na jednotarifnom i dvotarifnom brojilu, izvršena zamena brojila, kupac je utužen, kupac je platio dug, odobren popust 5% za redovno izmirenje duga, popust 50 din za elektronsku dostavu računa, popust 30 din za uplatu preko portala EPS Uvid u račun, umanjenje računa za energetski ugroženog kupca, obračunata kamata,... Pošto se svi računi formiraju u jedinstvenoj bazi podataka, primenom istih procedura, sada nije potrebno nekoliko puta testirati navedene tipove računa po mnogobrojnim regionalnim centrima i još većim brojem baza podataka.

Ukoliko se uoče određene greške prilikom testiranja, vrše se izmene i ponovo testira. Kada su testovi uspešno prošli, pristupa se masovnoj izradi računa.

Pored navedenih priprema za svaki obračun, neophodni ulazni podaci jesu merni podaci o očitavanju brojila, koje snabdevaču dostavlja operator distributivnog sistema počev od 1. do 15. u mesecu. Do sada, operator distributivnog sistema instalirao je 570.000 pametnih brojila od ukupno 3,8 miliona brojila. Ovaj broj se stalno povećava ugradnjom novih smart brojila i krajnji cilj jeste da sva brojila u Republici Srbiji budu pametna. Ovo je vrlo značajno za snabdevača, jer za ova brojila, snabdevač će 1. u mesecu dobiti očitavanja i to 100% tačna, dok je za klasična brojila potrebno da čitač izađe na teren, očita brojilo i da se podaci prenesu snabdevaču. U zavisnosti od načina očitavanja, postoji određena mogućnost greške, pogotovo ako čitač ručno upisuje očitana stanja.

Prenos mernih podataka između operatora distributivnog sistema i snabdevača, vrši se automatski market komunikacijom između dve baze podataka. Ranije su podaci dostavljani preko txt, csv, xls fajlova što je dosta usporavalo prenos i kontrolu podataka.



Slika 2. Pregled broja ukupno i očitanih mesta isporuke

Nakon preuzimanja mernih podataka, snabdevač u toku obračuna, vrši automatske provere ispravnosti podataka. Posebna pažnja posvećuje se:

• Validaciji računa sa iznosima iznad određenog limita

• Poređenju iznosa mrežarine sa podacima dostavljenim od strane operatora distributivnog sistema.

U slučaju da je određeni račun pao na nekoj od validacija, vrši se dodatna analiza. Ukoliko se analizom utvrdi da su preneti merni podaci logični tj. da je npr. utrošena električna energija u saglasnosti sa prethodnim mesecima, dozvoljava se dalja obrada. U suprotnom, obaveštava se operator distributivnog sistema da je potrebno izvršiti dodatnu proveru. I ova komunikacija sada se vrši preko market komunikacije, čime je ceo proces digitalizovan.

Masovna izrada računa omogućena je zahvaljujući višegodišnjoj optimizaciji sistema, što je dovelo do toga da se dnevno formira i do 700.000 računa.

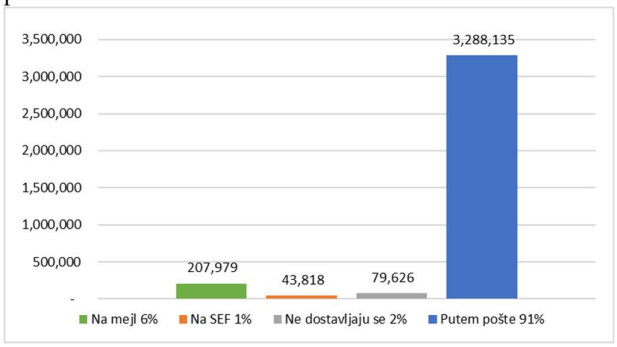
Na slici 3. prikazan je pregled ukupno očekivanih i dnevno formiranih faktura na garantovanom snabdevanju u periodu od 1. do 15. februara 2025. godine.

Slika 3. Pregled broja ukupno i dnevno formiranih faktura

Za sve formirane fakture u bazi podataka, EPS AD Beograd obezbeđuje uslove da se u najkraćem roku dostave kupcima na više načina:

- slanjem računa na mejl adresu kupca 6%
- slanjem računa na SEF (sistem elektronskih faktura) 1%
- preko protokola 2%.
- štampanjem i dostavom putem pošte 91%

Pored toga, na portalu EPS Uvid u račun, trenutno je preko 750.000 korisnika (21%) i svi oni mogu odmah pogledati i platiti račun preko portala. Oko 80.000 računa tj. 2% ne dostavlja se na ove načine, već u skladu sa dogovorenim protokolima.



Slika 4. Broj računa prema načinu dostave

Kao što se vidi na slici 4. svakog meseca, najveći broj faktura šalje se kupcima putem pošte, za čije sprovođenje je potrebno odraditi sledeće korake:

1. Formirati pdf fajl sa računima kupaca (sortiranje se vrši po adresnim podacima)
2. Štampanje i kovertiranje računa
3. Predaja računa pošti na dostavu
4. Dostava računa kupcima od strane pošte.

U ovom trenutku, najviše prostora za unapređenje procesa obračuna postoji u podprocesu formiranja pdf sa računima kupaca. Idealno bi bilo da EPS AD Beograd ima sva očitavanja 1. u mesecu, formira sve fakture u bazi podataka, sortira račune po adresnim podacima i na osnovu toga kreira pdfove sa sortiranim računima. To će biti moguće za desetak godina, kada sva brojila budu pametna.

Sada se radi na automatizaciji ovog procesa, ali na dnevnom nivou, kako se formiraju fakture. Predlog je sledeći: svakoga dana, posle ponoći kada je najmanje opterećenje sistema, automatski treba pustiti job, koji će prvo sortirati sve formirane račune u tom danu, i nakon toga formirati pdf-ove od po 5.000 računa. Do ujutru, job završava svoj zadatak, vrši se provera pdf fajlova i isti proseđuju na štampu i kovertiranje.

IV. Dodatne funkcionalnosti

U prethodnih nekoliko meseci, proces obračuna je dosta unapređen, što za rezultat ima posledicu da je već svakog 10. u mesecu formirano minimum 3 miliona faktura, a zaključno sa 15. u mesecu i preostalih 600.000 faktura. To daje mogućnost EPS AD Beograd, da u narednim mesecima skrati rok plaćanja računa (sa 28. na npr. 20. u mesecu) i na taj način ranije obezbedi veći priliv novca.

U EPS AD Beograd postoji kontakt centar, koji kupcima pruža odgovore na pitanja. Uvidom u jedinstvenu bazu podataka, operateri sada mogu efikasnije da odgovore kupcima i na taj način smanje potencijalni broj reklamacija.

Dalji planovi su da se formira posebna organizaciona celina u EPS AD Beograd, koja će se baviti izradom računa za električnu energiju. Postavljanje novih KPI, još više će doprineti unapređenju procesa obračuna.

V. Zaključak

Unapređenje procesa obračuna utrošene električne energije u EPS AD Beograd donelo je značajne benefite, uključujući smanjenje grešaka, efikasniju obradu podataka i bržu distribuciju računa. Implementacija jedinstvene baze podataka i objedinjeni sistem obračuna omogućili su centralizovanu kontrolu i bolju transparentnost procesa. Pored toga, svi podaci su u jednoj bazi, grupisani na isti način, što omogućava brže i jednostavnije izveštavanje.

Naredni koraci u unapređenju procesa uključuju automatizaciju formiranja PDF računa, što će dodatno skratiti vreme obrade i povećati tačnost. Daljim poboljšanjima EPS AD Beograd nastaviće da modernizuje svoje poslovanje i unapređuje korisničko iskustvo, što je ključno za održivu budućnost elektroenergetskog sektora u Srbiji.

VI. Literatura

[1] "Sajt EPS" [Online], dostupno na: www.eps.rs
[2] "Portal Uvid u račun" [Online], dostupno na: https://portal.eps.rs
[3] "Sajt EDS" [Online], dostupno na: www.elektrodistribucija.rs .

Improvement of the billing process of consumed electricity in EPS AD Belgrade

Jadranka Rsitić

ABSTRACT

In accordance with global digitization, "Elektroprivreda Srbije" strives to keep pace with it. For many years now, the business system of "Elektroprivreda Srbije" has been successfully modernized and improved through the digitalization process. The most significant digital project of "Electricity of Serbia", at the moment, is the implementation of a unique database and the improvement of the calculation process of consumed electricity for customers on guaranteed supply, of which there are more than 3.6 million. The paper presents the challenges that "Elektroprivreda Srbije" encountered during the period of improvement of the billing process, as well as the benefits it received with the new method of billing. In addition, further plans for its improvement were presented in detail.

YU #2: Sesija 2

Veštačka inteligencija

Neuromorfno računarstvo i tečne mašine stanja u sistemima za otkrivanje upada nad CIC-IDS2017 skupom podataka

Miloš Živadinović
Fakultet Organizacionih Nauka
Beograd, Republika Srbija
mzdv@protonmail.com
0000-0002-0342-340X

Dejan Simić
Fakultet Organizacionih Nauka
Beograd, Republika Srbija
dsimic@fon.bg.ac.rs
0000-0002-0744-5411

***Apstrakt* - Neuromorfno računarstvo predstavlja jednu od novijih paradigmi u pristupu rešavanja problema sajber bezbednosti primenom veštačke inteligencije. Ovaj rad opisuje primenu tečnih mašina stanja (LSM) sa Leaky Integrate-and-Fire (LIF) neuronima u sistemima za detekciju mrežnih napada, koristeći CIC-IDS2017 skup podataka. Izvršena je komparativna analiza LSM modela sa različitim veličinama rezervoara (50, 100, 500 i 1000 neurona) na devet klasa mrežnog saobraćaja primenom PyTorch i snnTorch biblioteka. Rezultati eksperimenta opisani su AUROC vrednostima iznad 0,99 za sve testirane konfiguracije, kao i brzu konvergenciju, dostižući AUROC vrednost od 0,998 nakon prve epohe obuke. Posebno se ističe varijanta LSM modela sa 100 neurona koja postiže tačnost od 0,94 i AUPRC vrednost od 0,97 posle deset epoha. Rezultati ukazuju na potencijal modela zasnovanih nad tečnim mašinama stanja u razvoju sistema za detekciju upada u realnom vremenu uz balans između brzine konvergencije i uspešnosti detekcije.**

Ključne reči – neuromorfno računarstvo, veštačka inteligencija, PyTorch, snnTorch, sajber bezbednost.

I. Uvod

Razvitak sajber pretnji u savremenim umreženim računarskim sistemima zahteva konstantan nadzor i unapređivanje sistema za otkrivanje mrežnih napada. Tradicionalni sistemi za otkrivanje upada (IDS) često se suočavaju sa izazovima u prepoznavanju novih i izmenjenih napada [1], kao i sa ograničenjima u pogledu brzine obrade mrežnog saobraćaja u stvarnom vremenu. Ovi izazovi su posebno izraženi u kontekstu uređaja Internet of Things (IoT) uređaja i edge computing računarskih sistema koji se najčešće izvršavaju u okruženjima sa ograničenim računarskim i energetskim resursima.

Neuromorfno računarstvo predstavlja pristup veštačke inteligenciji koji nastoji da oponaša građu i funkcionalnost bioloških nervnih sistema [2]. Njihova glavna karakteristika je procesiranje složenih obrazaca uz minimalnu potrošnju električne energije [3].

Tečne mašine stanja (LSM - Liquid State Machines) [4] predstavljaju podvrstu rekurentnih neuronskih mreža nastalih pod uticajem koncepata računarstva u rezervoaru (Reservoir Computing) [5]. LSM arhitektura sastoji se od tri sloja: (1) ulaznog sloja koji preslikava ulazne podatke na neurone u rezervoaru, (2) rezervoara koji sadrži rekurentno povezane neurone i pravila dinamike tečnosti, i (3) izlaznog sloja koji se obučava da tumači stanja rezervoara. Naročita prednost LSM modela je njihova sposobnost da delotvorno obrađuju vremenske serije i obrasce u podacima, što ih čini pogodnim za otkrivanje nepravilnosti u mrežnom saobraćaju.

U ovom radu istražujemo mogućnost primene LSM sa Leaky Integrate-and-Fire (LIF) neuronima u kontekstu otkrivanja mrežnih napada. LIF neuroni predstavljaju biološki verodostojan model koji oponaša ponašanje neurona kroz integraciju dolaznih signala do trenutka aktivacije neurona, nakon čega dolazi do okidanja i vraćanja neurona u početno stanje [6].

Glavni cilj našeg istraživanja je procena delotvornosti LSM arhitektura sa različitim veličinama rezervoara (50, 100, 500 i 1000 neurona) u klasifikaciji mrežnog saobraćaja na osnovu CIC-IDS2017 skupa podataka. CIC-IDS2017 [7] predstavlja skup podataka za obuku i testiranje sistema za otkrivanje mrežnih napada. Ovaj skup podataka je razvijen od strane Kanadskog instituta za sajber bezbednost (Canadian Institute for Cybersecurity) i sadrži realistične mrežne tokove koji uključuju benigne mrežne aktivnosti i različite kategorije mrežnih sajber napada. Skup podataka obuhvata petodnevni period mrežnog saobraćaja sa detaljno označenim instancama napada, uključujući DDoS napade, brute force napade, infiltracije i web napade [8]. Posebna vrednost ovog skupa podataka leži u tome što kombinuje realne mrežne uslove sa precizno kategorisanim podacima, što omogućava istraživačima da razvijaju i evaluiraju algoritme mašinskog učenja za detekciju napada u kontrolisanim, ali realističnim uslovima.

II. Prikaz pojmova

A. Neuromorfno računarstvo

Neuromorfno računarstvo predstavlja interdisciplinarnu oblast koja spaja saznanje iz računarskih nauka sa saznanjima neuronauke u cilju razvoja računarskih sistema koji oponašaju strukturu i funkcionalnost bioloških nervnih sistema [9]. Neuromorfni sistemi odlikuju se paralelnim procesiranjem, integrišu obradu i prikupljanje podataka, kao i poseduju sposobnost samoorganizacije i prilagođavanja u skladu sa promenom ulaza i izlaza.

Osnovna svojstva neuromorfnih sistema uključuju: (1) implementaciju biološki inspirisanih modela neurona kao što su Integrate-and-Fire neuroni koji akumuliraju ulazne signale i generišu impuls kada se dostigne prag aktivacije, (2) asinhronu komunikaciju između neurona kroz diskretne impulsne događaje umesto kontinuiranih signalnih vrednosti, (3) kolokaciju memorije i računske obrade u istim

komponentama što eliminiše potrebu za stalnim transferom podataka između memorijskih i procesorskih jedinica, i (4) događajno-vođeno procesiranje gde se računanje izvršava samo kada se javi impuls što značajno smanjuje energetsku potrošnju.

Jedna od ključnih implementacija koncepta neuromorfnih sistema su impulsne neuronske mreže (spiking neural networks) [10] koje oponašaju funkcionisanje bioloških neurona. Ovi modeli, za razliku od tradicionalnih veštačkih neuronskih mreža, prenose informacije putem pojedinačnih impulsa (spike) umesto kontinualnih vrednosti, što rezultira računskim sistemima koji su energetski efikasniji i pogodniji za obradu podataka iz fizičkog okruženja.

B. Tečne mašine stanja (LSM)

Tečne mašine stanja predstavljaju specifičan oblik računarstva u rezervoaru, paradigme inspirisane dinamikom tečnosti. Koncept LSM prvi put je predstavljen 2002. godine, [11] predlažući računski model zasnovan na obradi informacija kroz dinamički rezervoar neurona. LSM metaforički oponaša efekat koji se dešava pomeranjem tečnosti nakon unošenja stranog tela - početni poremećaj stvara složene obrasce talasa koji se prostiru kroz tečnost, a ovi obrasci sadrže informacije o karakteristikama objekta.

Arhitektura LSM sastoji se od tri osnovna dela:

1. **Ulazni sloj** koji transformiše ulazne podatke u impulse koji se prosleđuju rezervoaru
2. **Rezervoar** sastavljen od slučajno povezanih rekurentnih impulsnih neurona sa nasumičnim vrednostima težina (tzv. "tečno stanje")
3. **Izlazni sloj** koji se obučava da prepoznaje obrasce u stanju rezervoara i vrši klasifikaciju

Ključna prednost LSM je što se obučava samo izlazni sloj, dok veze unutar rezervoara ostaju nepromenljive nakon početne inicijalizacije u vidu postavljanja nasumičnih težina neurona i stepena povezanosti među njima. Ovo značajno pojednostavljuje proces obučavanja u poređenju sa drugim rekurentnim neuronskim mrežama, omogućavajući efikasno učenje vremenskih nizova i prepoznavanje složenih obrazaca u podacima.

C. Leaky Integrate-and-Fire (LIF) neuroni

LIF neuroni predstavljaju fizički inspirisan računski model koji apstrahuje ponašanje bioloških neurona. Ovaj model opisuje dinamiku membranskog potencijala neurona kroz proces integracije ulaznih impulsa, postupno opadanje potencijala za vreme integracije, i generisanje izlaznog impulsa (akcionog potencijala) kada integracija prekorači određeni prag (prag akcionog potencijala) [6]. Potencijal membrane predstavlja razliku u voltaži između unutrašnjosti nervne ćelije i njenog okruženja.

Matematički, LIF neuron možemo opisati diferencijalnom jednačinom:

$$\tau \frac{dV(t)}{dt} = -(V(t) - V_{rest}) + RI(t)$$

gde je V potencijal membrane, V_{rest} potencijal membrane u stanju mirovanja, τ vremenska konstanta membrane, R otpornost membrane, a $I(t)$ ulazna struja u trenutku t. Kada V dostigne prag V_{th}, neuron generiše impuls i potencijal se vraća na vrednost V_{reset}.

Primena LIF neurona u LSM arhitekturi omogućava efikasno kodiranje vremenskih obrazaca u diskretne impulse što je posebno korisno za otkrivanje anomalija u mrežnom saobraćaju. Sposobnost LIF neurona da integrišu ulazne signale tokom vremena, uz postepeno opadanje uticaja starijih signala, omogućava modelu da uoči vremenske obrasce karakteristične za različite vrste mrežnih napada.

D. Sistemi za otkrivanje upada

Sistemi za otkrivanje upada (IDS) predstavljaju ključnu komponentu savremene sajber bezbednosne infrastrukture, sa zadatkom da prepoznaju nedozvoljene aktivnosti, neovlašćene pristupe i zlonamerne pokušaje napada na računarske sisteme i mreže [12].

Tradicionalno, IDS sistemi se dele na dve glavne kategorije [13]:

1. **Sistemi zasnovani na potpisu** koji otkrivaju napade upoređivanjem mrežnog saobraćaja sa poznatim obrascima napada (potpisima)
2. **Sistemi zasnovani na anomalijama** koji grade model normalnog ponašanja i otkrivaju odstupanja od tog modela

Sa porastom složenosti i raznovrsnosti sajber napada, tradicionalni IDS sistemi suočavaju se sa značajnim izazovima. Sistemi zasnovani na potpisima ne mogu otkriti nove, prethodno neviđene napade, dok sistemi zasnovani na otkrivanju anomalija često generišu veliki broj lažnih uzbuna.

U poslednjih nekoliko godina, pristupi zasnovani na mašinskom učenju pokazali su obećavajuće rezultate u prevazilaženju ovih ograničenja [1]. Duboke neuronske mreže, posebno rekurentne arhitekture poput Long Short-Term Memory (LSTM) [14] i Gated Recurrent Unit (GRU) [15], kao i veliki jezički modeli (Large Language Model - LLM) [16] primenjene su za otkrivanje anomalija i malicioznog sadržaja u mrežnom saobraćaju sa visokom tačnošću [17]. Međutim, ovi modeli zahtevaju značajne računarske resurse za obuku i izvršavanje u stvarnom vremenu, kao i potrebu za čestim dodatnim obukama [18].

Neuromorfni pristup, posebno primena LSM sa LIF neuronima, obećava značajno smanjenje računske složenosti i energetske potrošnje, uz zadržavanje visoke tačnosti otkrivanja napada. CIC-IDS2017 skup podataka, sa svojom raznovrsnošću realističnih mrežnih napada, predstavlja proveren okvir za procenu delotvornosti ovog pristupa u kontekstu izazova sajber bezbednosti i otkrivanja upada.

III. TRENUTNA ISTRAŽIVANJA

Oblast primene neuromorfnih računarskih arhitektura i impulsnih neuronskih mreža u otkrivanju sajber napada doživljava značajan razvoj poslednjih decenija. U ovom delu predstavljamo pregled najznačajnijih doprinosa u ovoj oblasti, sa posebnim osvrtom na primenu tečnih mašina stanja i srodnih paradigmi računarstva u rezervoaru.

A. *Neuromorfno računarstvo u sajber bezbednosti*

Tradicionalni pristupi u otkrivanju mrežnih napada zasnovani na mašinskom učenju često se oslanjaju na složene modele koji zahtevaju značajne računarske resurse. Ovo predstavlja izazov u kontekstu sistema koji moraju raditi u stvarnom vremenu sa ograničenim resursima. Neuromorfne arhitekture pojavljuju se kao obećavajuća alternativa zbog svoje energetske efikasnosti i sposobnosti obrade temporalnih podataka.

Alom i Taha [19] su prvi primenili koncept neuromorfnog računarstva kroz impulsne neuronske mreže unutar sistema za otkrivanje upada. Karakteristično je postojanje spoljne duboke neuronske mreže čije su težine primenjene na impulsnu neuronsku mrežu.

Isto tako, Zahm et al. [20] izvršili su istraživanje u vezi primene neuromorfnog računarstva unutar sistema bezbednosti u stvarnom vremenu, gde je utvrđeno da su uporedivi sa tradicionalnim neuronskim mrežama.

Ciljevi i metodologija ovog rada mogu se smatrati unapređenjem jednog od prethodnih radova autora [21] gde je razvijen model koji koristi impulsne neuronske mreže sa LIF neuronima nad skupom podataka BoT-IoT koji predstavlja pregled mrežnog saobraćaja IoT uređaja [22], [23], [23], [24], [25], [26], [27].

B. *Tečne mašine stanja i računarstvo u rezervoaru*

Koncept računarstva u rezervoaru, a posebno tečnih mašina stanja, privukao je pažnju istraživača u oblasti sajber bezbednosti zbog svoje sposobnosti da efikasno procesira vremenske nizove i prepozna složene obrasce.

Tečne mašine stanja predstavio je Maas [4] 2011. godine kao model koji je primenljiviji za opisivanje računanja kroz biološki inspirisane sisteme u odnosu na Turingovu mašinu. Isto tako, predložen model je primenljiviji za sisteme koji su po prirodi rekurentni, poput neuroračunarstva.

Prvobitne ideje iza mašina tečnih stanja definisali su Maas et al. [11] kao okvir za razvoj neuroračunarstva zasnovanog na perturbacijama, bez postojanja stabilnih stanja unutar sistema.

C. *Leaky Integrate-and-Fire neuroni u impulsnim neuronskim mrežama*

LIF neuroni predstavljaju jedan od najčešće korišćenih modela u impulsnim neuronskim mrežama zbog svog dobrog balansa između biološke verodostojnosti i računske efikasnosti. Njihova primena u neuroračunarstvu bila je predmet nekoliko značajnih istraživanja.

Način funkcionisanja bioloških neurona kroz propagaciju akcionih potencijala otkrili su Hodgkin i Huxley [28] eksperimentima nad gigantskim lignjama zbog jednog od najvećeg promera neurona unutar tela među živim bićima. Otkriveni mehanizmi delovanja predstavljaju prvi uvid u način funkcionisanja neuromorfnih sistema.

Prvobitni prikaz Leaky Integrate-and-Fire neurona izvršen je 1907. godine [6] od strane Louis Lapicque-a, pre nego što je potvrđeno funkcionisanje bioloških neurona kroz elektrofiziologiju. Jedna od glavnih prednosti LIF neurona u savremenom neuroračunarstvu je u jednostavnosti softverskog i hardverskog prikaza (pošto se može opisati RC strujnim kolom) uz zadovoljavajuće računske performanse i određeni nivo biološke tačnosti.

Sveobuhvatni računarski prikaz rada bioloških neurona prikazao je Izhikevich [29] zajedno sa njihovim ograničenjima u računarstvu. Predložio je Izhikevich-ev model neurona koji omogućava veći stepen biološke tačnosti uz efikasnije performanse.

IV. METODOLOGIJA

A. *CIC-IDS2017 skup podataka*

CIC-IDS2017 kao skup podataka obuhvata oko 2,8 miliona uzoraka, podeljenih u 11 različitih kategorija (od čega je deset maliciozno, a jedna obuhvata benigni saobraćaj) sa 80 parametara. Za potrebe našeg eksperimenta, koristimo svih 80 različitih parametar i 11 različitih izlaznih kategorija, nad kojima je primenjen one-hot encoding radi efikasnije i efektivnije obuke modela. Primenjeno je min-max skaliranje radi uniformnosti numeričkih podataka, dok batching nije korišćen zbog mogućnosti hardverskih resursa nad kojima se izvršavao eksperiment. Rezultujući skup podataka podeljen je u odnosu 80:20 u korist podataka za obuku.

B. *Arhitektura tečne mašine stanja*

U našem istraživanju implementirali smo LSM arhitekturu sa LIF neuronima koristeći PyTorch [30] i snnTorch [10] biblioteke. Osnovna arhitektura našeg modela sastoji se od tri komponente:

1. **Ulazni sloj** koji vrši kodiranje parametara CID-IDS2017 skupa podataka u nizove impulsa (spike train). Koristili smo pristup kodiranja brzinom (rate coding) [31] gde je verovatnoća generisanja impulsa proporcionalna intenzitetu ulaznog parametra.
2. **Rezervoar** sastavljen od rekurentno povezanih LIF neurona sa beta vrednošću od 0,95 i stopom međupovezanosti neurona od 0,2. Veze između neurona u rezervoaru su nasumično inicijalizovane i ostaju nepromenjene tokom obuke.
3. **Izlazni sloj** koji vrši linearnu transformaciju stanja rezervoara u verovatnoće kategorija. Ovaj sloj se obučava koristeći Stochastic Gradient Descent (SGD) sa Adam optimizacijom i primenom cross-entropy loss kao funkcije gubitaka.

Opisani LSM model je obučen i testiran sa četiri različite veličine rezervoara: 50, 100, 500 i 1000 LIF neurona, dok su ostali parametri konstantni. Obuka se vrši kroz 10 epoha, gde svaka epoha sadrži 100 temporalnih koraka. Temporalni koraci simuliraju kontinualnost i dolazni tok mrežnog saobraćaja.

C. *Eksperimentalno okruženje*

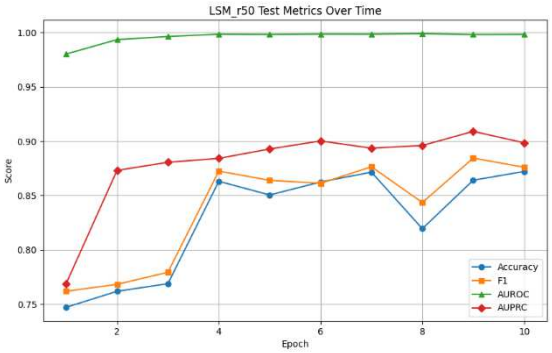
Obuka je odrađena unutar Oracle Cloud Infrastructure okruženja koje se nalazi u javnom cloud regionu eu-jovanovac-1 unutar Državnog Data Centra Republike Srbije. Korišćena je mašina BM.GPU.A100-v2.8 sa osam NVIDIA A100 grafičkih kartica i ukupno 640 GB VRAM memorije. Iako nije neophodna hardverska mašina ovakvih performansi

za potrebu eksperimenta, iskorišćeni su njeni potencijali radi brže iteracije kroz razvoj i postavku modela.

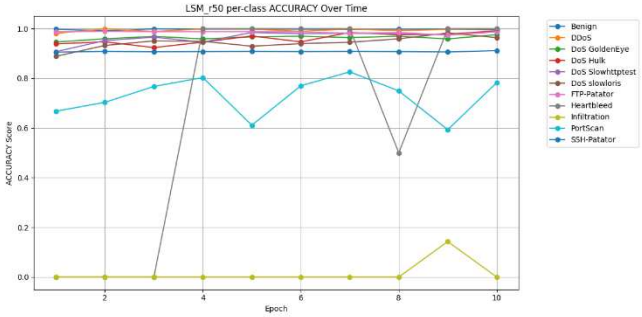
Mašina je korišćena kroz Oracle Data Science Service koji predstavlja okruženje za obradu podataka i razvoj rešenja veštačke inteligencije. Eksperiment je razvijen upotrebom PyTorch i snnTorch biblioteka uz primenu sklearn [32] biblioteke za potrebe numeričkih transformacija podataka.

V. EKSPERIMENTALNI REZULTATI

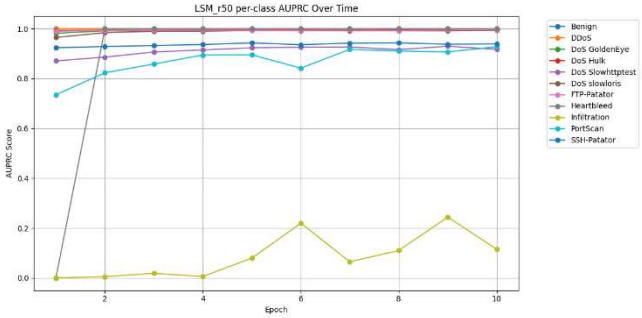
A. Rezervoar od 50 neurona



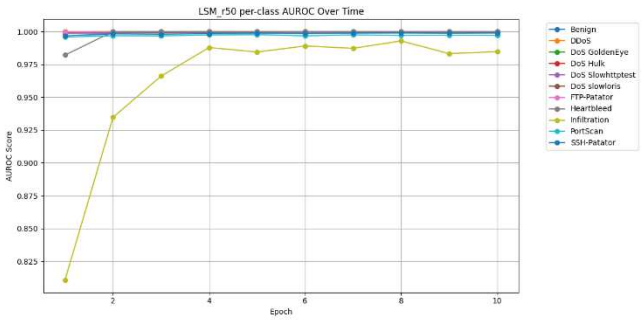
Slika 1. Kretanje eksperimentalnih metrika kroz obuku za rezervoar od 50 neurona



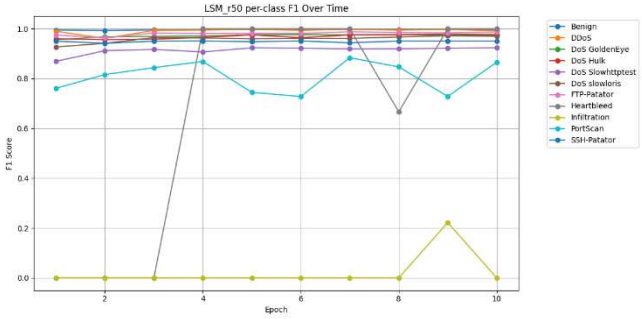
Slika 2. Kretanje tačnosti kroz obuku za rezervoar od 50 neurona



Slika 3. Kretanje AUPRC kroz obuku za rezervoar od 50 neurona



Slika 4. Kretanje AUROC kroz obuku za rezervoar od 50 neurona

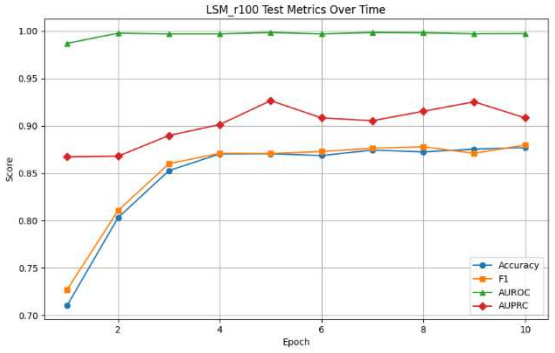


Slika 5. Kretanje F1 vrednosti kroz obuku za rezervoar od 50 neurona

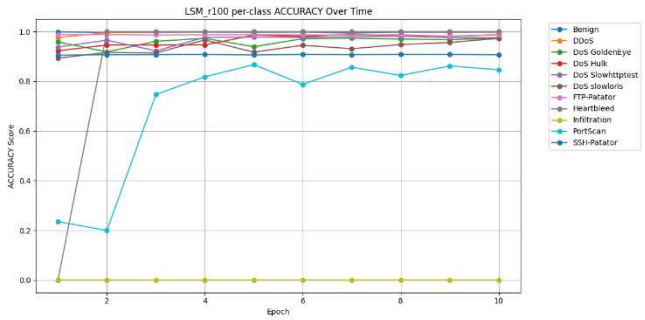
Analiza eksperimentalnih rezultata otkriva varijacije u performansama LSM modela pri klasifikaciji različitih kategorija mrežnih napada tokom procesa obuke. Model demonstrira visoke performanse za većinu analiziranih klasa napada, sa AUROC vrednostima koje dostižu opseg između 0.98 i 1.0 za kategorije Benign, DDoS varijante, Infiltration, PortScan i SSH-Patator napade. Konvergencija performansi se ostvaruje relativno brzo, sa stabilizacijom rezultata već nakon druge epohe obuke. Međutim, uočljive su značajne razlike u detekciji specifičnih tipova napada, posebno DoS GoldenEye kategorije, koja beleži početne AUROC vrednosti od 0.81 sa postupnim poboljšanjem do 0.985 u trećoj epohi.

Evaluacija preko F1 metrike otkriva dodatne probleme u postupku klasifikacije, gde DoS GoldenEye napadi pokazuju kritično niske performanse sa F1 skorom koji perzistira na nivou 0.0 kroz većinu epoha obuke. Heartbleed kategorija takođe manifestuje nestabilno ponašanje sa varijacijama između 0.65 i 0.98 tokom različitih epoha.

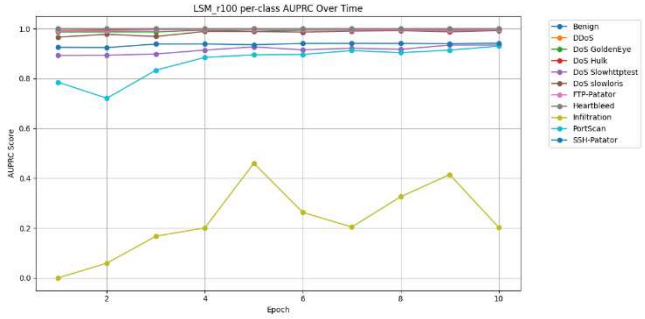
B. Rezervoar od 100 neurona



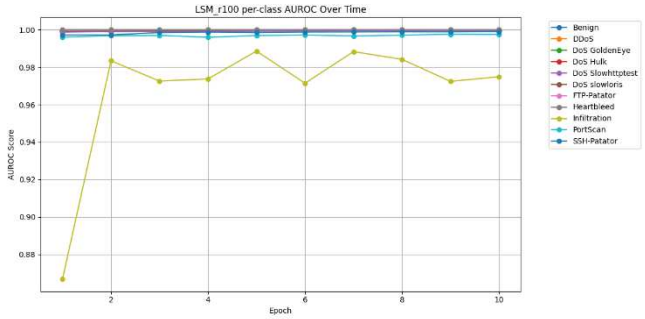
Slika 6. Kretanje eksperimentalnih metrika kroz obuku za rezervoar od 100 neurona



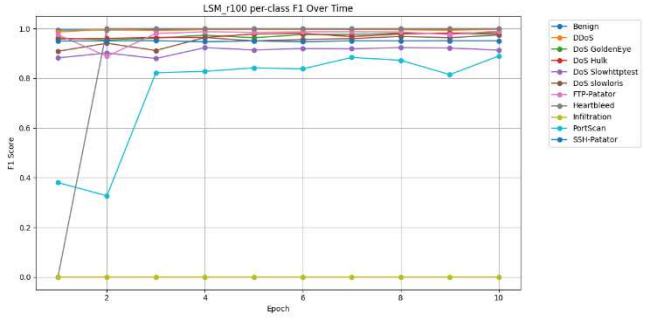
Slika 7. Kretanje tačnosti kroz obuku za rezervoar od 100 neurona



Slika 8. Kretanje AUPRC kroz obuku za rezervoar od 100 neurona



Slika 9. Kretanje AUROC kroz obuku za rezervoar od 100 neurona

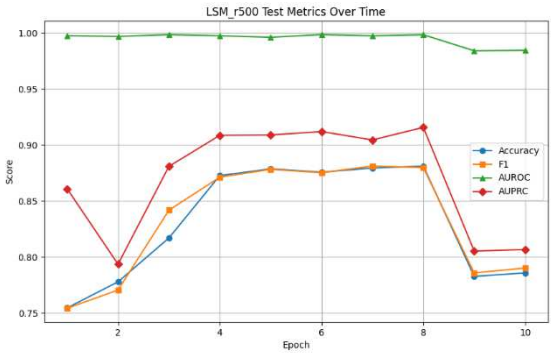


Slika 10. Kretanje F1 vrednosti kroz obuku za rezervoar od 100 neurona

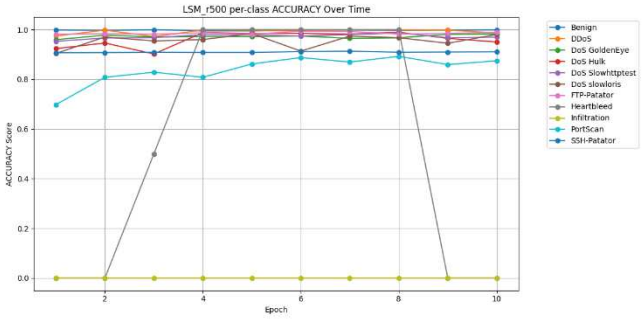
Rezultati eksperimentalnog istraživanja LSM modela sa 100 neurona demonstriraju značajno poboljšanje performansi u odnosu na konfiguraciju sa 50 neurona, posebno u kontekstu problematičnih kategorija napada identifikovanih u prethodnoj analizi. Model sa povećanom mrežnom kompleksnošću postiže stabilnije i konzistentnije rezultate kroz sve evaluacione metrike, sa uočljivim poboljšanjem u detekciji DoS GoldenEye napada koji su ranije predstavljali izazov za sistem.

AUROC analiza otkriva da je model sa 100 neurona uspešno prevazišao ograničenja uočena kod manje konfiguracije, postižući vrednosti između 0.97 i 1.0 za sve kategorije napada. Posebno je značajno poboljšanje u detekciji DoS GoldenEye kategorije, gde se AUROC vrednosti kreću između 0.87 i 0.99 sa stabilnom konvergencijom već nakon druge epohe. SSH-Patator napadi takođe pokazuju poboljšane performanse sa početnih 0.37 u prvoj epohi do stabilizacije na nivou između 0.81 i 0.89 tokom naknadnih iteracija. Konsolidovani rezultati potvrđuju unapređenje proširene arhitekture, sa AUROC vrednošću koja dostiže maksimum na vrednosti 0.99, dok se ostale metrike stabilizuju na visokom nivou između 0.87 i 0.92, što predstavlja značajno poboljšanje u odnosu na prethodnu konfiguraciju.

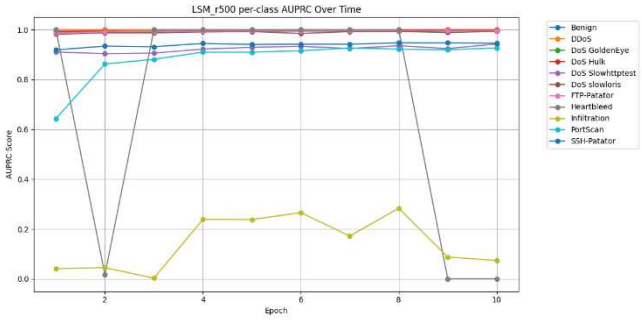
C. Rezervoar od 500 neurona



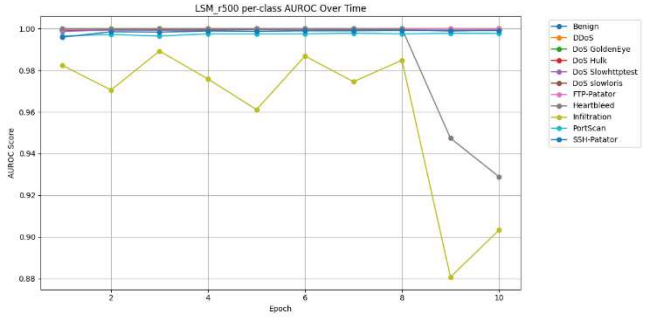
Slika 11. Kretanje eksperimentalnih metrika kroz obuku za rezervoar od 500 neurona



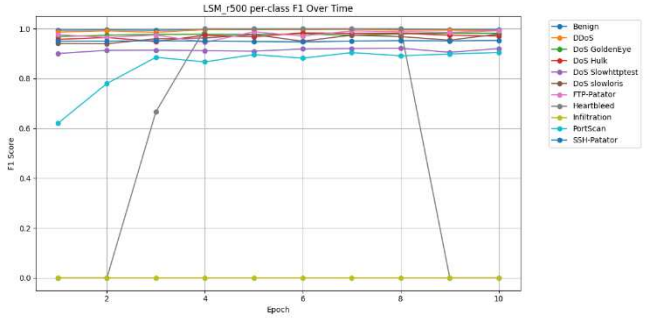
Slika 12. Kretanje tačnosti kroz obuku za rezervoar od 500 neurona



Slika 13. Kretanje AUPRC kroz obuku za rezervoar od 500 neurona



Slika 14. Kretanje AUROC kroz obuku za rezervoar od 500 neurona

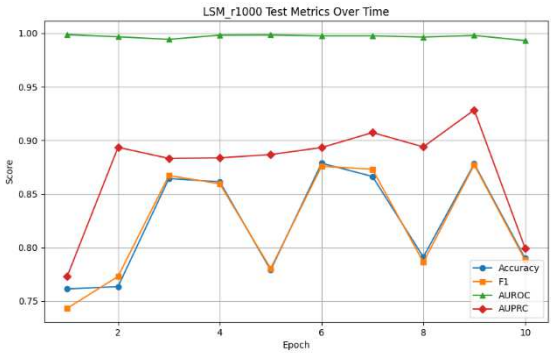


Slika 15. Kretanje F1 vrednosti kroz obuku za rezervoar od 500 neurona

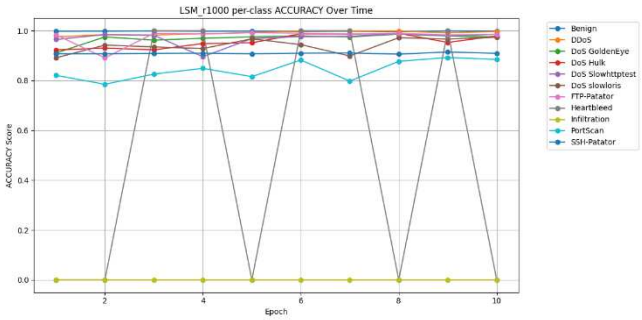
Rezultati eksperimentalnog istraživanja LSM modela sa 500 neurona pokazuju dalja poboljšanja u performansama sistema, sa unapređenim poboljšanjima u detekciji prethodno problematičnih kategorija napada. Model sa proširenom arhitekturom postiže izuzetno visoke performanse kroz sve evaluacione metrike, uz značajno poboljšanje stabilnosti tokom procesa obuke. Posebno je značajno napredovanje u detekciji SSH-Patator napada koji su u prethodnim konfiguracijama predstavljali izazov za sistem.

Analiza AUROC vrednosti otkriva da model sa 500 neurona održava konzistentno visoke performanse sa vrednostima između 0.96 i 1.0 za sve kategorije napada. SSH-Patator napadi pokazuju dramatično poboljšanje sa početnih vrednosti od 0.61 u prvoj epohi do stabilizacije na nivou između 0.83 i 0.89 tokom naknadnih iteracija. Heartbleed kategorija takođe demonstrira poboljšanu stabilnost kroz većinu epoha obuke, mada se uočava pad performansi u krajnjim iteracijama. Međutim, objedinjeni test rezultati otkrivaju trend degradacije performansi u poslednjim epohama, gde se AUROC vrednost smanjuje sa prethodno dostignute vrednosti od 0.99 na 0.98, dok ostale metrike beleže značajniji pad sa tačnošću i F1 skorom koji se smanjuju sa 0.87 na 0.79. Ovaj fenomen ukazuje na mogućnost preobuke modela (overfitting) ili potrebe za implementacijom tehnika regularizacije kako bi se održale konzistentne performanse kroz produženi period obuke.

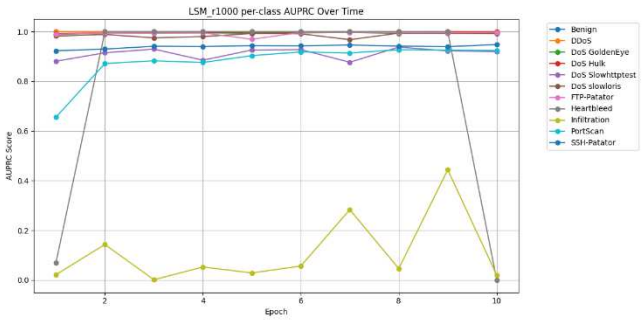
D. Rezervoar od 1000 neurona



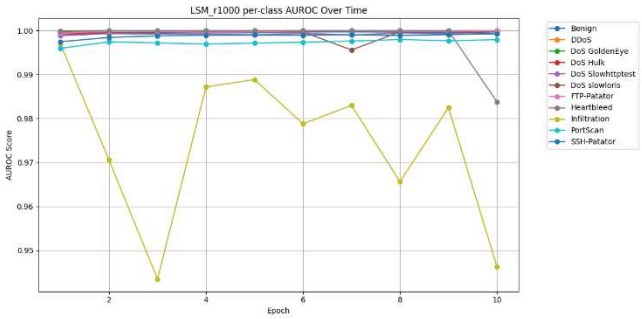
Slika 16. Kretanje eksperimentalnih metrika kroz obuku za rezervoar od 1000 neurona



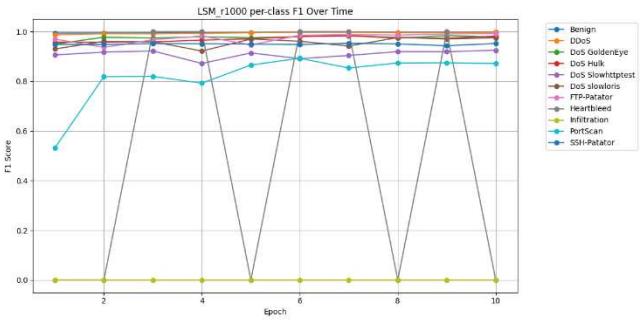
Slika 17. Kretanje tačnosti kroz obuku za rezervoar od 1000 neurona



Slika 18. Kretanje AUPRC kroz obuku za rezervoar od 1000 neurona



Slika 19. Kretanje AUROC kroz obuku za rezervoar od 1000 neurona



Slika 20. Kretanje F1 vrednosti kroz obuku za rezervoar od 1000 neurona

Rezultati eksperimentalnog istraživanja LSM modela sa 1000 neurona otkrivaju složene tendencije u ponašanju sistema koje ukazuju na značajne izazove skalabilnosti pri povećanju mrežne kompleksnosti. Analiza demonstrira da većina kategorija napada održava izuzetno visoke AUROC performanse sa vrednostima između 0.99 i 1.0, što predstavlja konstantnu karakteristiku kroz sve testirane arhitekture. Međutim, uočljive su značajne fluktuacije u performansama SSH-Patator kategorije koja pokazuje početne vrednosti od 0.54 sa konvergencijom na nivou između 0.81 i 0.89, slično prethodnim konfiguracijama.

Kritičan aspekt ovih rezultata manifestuje se kroz izraženu nestabilnost Heartbleed kategorije, koja demonstrira dramatične varijacije u performansama sa potpunim kolapsom detekcije u određenim epohama. Ovaj fenomen se posebno izražava u osmoj epohi gde F1 skor, tačnost i AUPRC za ovu kategoriju dostiže nulte vrednosti. Konsolidovani rezultati dodatno potvrđuju trend nestabilnosti, gde se uočava značajna varijabilnost u performansama kroz epohe obuke sa AUROC vrednostima koje se kreću između 0.99 i 0.985, dok ostale metrike pokazuju oscilatorne tendencije sa vrednostima između 0.74 i 0.92. Ovi nalazi sugerišu da povećanje broja neurona u rezervoaru ne garantuje linearno poboljšanje performansi.

VI. Diskusija

Rezultati ovog istraživanja pružaju značajan uvid u potencijal primene tečnih mašina stanja sa LIF neuronima u sistemima za otkrivanje mrežnih napada. Značajne performanse postignute na CIC-IDS2017 skupu podataka, uz brzu konvergenciju i efikasnost, pokazuju da neuromorfni pristupi mogu predstavljati obećavajuću alternativu tradicionalnim metodama u ovoj oblasti.

AUROC vrednosti preko 0,99 postignute za sve testirane konfiguracije LSM modela ukazuju na izuzetnu sposobnost diskriminacije između legitimnog saobraćaja i različitih tipova napada. Sposobnost LSM modela da efikasno procesiraju vremenske nizove omogućava im da uhvate suptilne vremenske obrasce karakteristične za mrežne napade. Za razliku od tradicionalnih pristupa koji često tretiraju svaki mrežni tok nezavisno, LSM kroz svoju rekurentnu strukturu može modelirati vremenske zavisnosti između uzastopnih paketa ili tokova, što je ključno za otkrivanje složenijih ili novih vrsta napada. Nasumična povezanost neurona u rezervoaru stvara bogato dinamičko okruženje koje može predstavljati složene nelinearne transformacije ulaznih podataka, omogućavajući LSM modelima dodatne mogućnosti za obučavanje nad složenijim ili novim vrstama napada.

Jedan od značajnijih aspekata naših rezultata je brzina konvergencije LSM modela. Postizanje AUROC vrednosti od 0,998 već nakon prve epohe obuke predstavlja značajnu prednost u odnosu na tradicionalne duboke neuronske mreže koje često zahtevaju desetine ili stotine epoha za postizanje zadovoljavajućih performansi. Brzina konvergencije može se objasniti osobinama LSM modela gde se veze unutar rezervoara ne menjaju tokom obuke, dok se samo obučava linearni izlazni sloj da interpretira stanja generisana od strane fiksnog rezervoara. Praktične implikacije ove karakteristike su značajne za sisteme za otkrivanje upada u realnom vremenu, jer sposobnost modela da se brzo adaptira na nove obrasce napada znači da IDS sistemi zasnovani na LSM mogu biti brže ažurirani.

Uprkos izuzetnim opštim performansama, naša analiza otkriva značajne izazove u detekciji određenih tipova napada, posebno Infiltration i Heartbleed klasa. Infiltration klasa pokazuje najlošije performanse sa F1 skorom koji ostaje blizu nule za manje konfiguracije i dostiže samo 0,41 za najveći rezervoar. Slabe performanse mogu se objasniti činjenicom da ovakvi napadi često oponašaju legitimne aktivnosti korisnika, praveći minimalne promene u mrežnom saobraćaju. Heartbleed klasa takođe predstavlja izazov, sa značajnim fluktuacijama u performansama tokom epoha. Nestabilnost performansi ukazuje na to da LSM rezervoar povremeno generiše stanja koja omogućavaju detekciju ovog napada, ali ta stanja nisu konzistentno dostupna. Za obe klase napada je karakteristična manja zastupljenost u skupu podataka u odnosu na ostale.

Analiza uticaja veličine rezervoara otkriva nekoliko važnih obrazaca. Značajno poboljšanje performansi između konfiguracije sa 50 i 100 neurona, gde se tačnost povećava sa 0,87 na 0,94, sugeriše da manji rezervoari mogu biti nedovoljni za adekvatno modeliranje složenosti mrežnog saobraćaja u CIC-IDS2017 skupu podataka. Međutim, marginalna poboljšanja postignuta daljim povećanjem rezervoara na 500 i 1000 neurona ukazuju na postojanje praga nakon kojeg dodatni neuroni ne doprinose značajno poboljšanju performansi (overfitting).

Posebno je zanimljivo poboljšanje performansi za problematične klase sa povećanjem veličine rezervoara, kao što je Heartbleed klasa koja pokazuje poboljšanje F1 skora sa 0,72 na 0,93, što sugeriše da veći rezervoari mogu bolje modelirati retke i složene obrasce, uprkos njihovoj manjoj zastupljenosti u skupu podataka.

Poređenje sa tradicionalnim algoritmima mašinskog učenja otkriva nekoliko ključnih prednosti LSM pristupa. Prva je energetska efikasnost - dok naši eksperimenti nisu direktno merili potrošnju energije, teorijski osnovi neuromorfnog računarstva sugerišu značajno smanjenje energetskih zahteva u odnosu na tradicionalne pristupe, što je posebno bitno za implementaciju na uređajima sa ograničenim resursima poput IoT ili edge computing uređaja. Još jedna prednost je otpornost na šum u podacima, jer LIF neuroni kroz svoj

mehanizam integracije i curenja prirodno filtriraju kratkotrajne perturbacije u ulaznim signalima.

Eksperiment sproveden nad jednom skupu podataka može ograničiti opštu prirodu rezultata, jer CIC-IDS2017, uprkos svojoj sveobuhvatnosti, predstavlja specifično mrežno okruženje koji se mogu razlikovati od onih u drugim kontekstima. Takođe, trenutni pristup koristi statičke parametre za LIF neurone kroz celokupne varijacije broja neurona u rezervoaru, dok adaptivni pristupi koji dinamički podešavaju parametre neurona na osnovu karakteristika ulaznih podataka mogu učiniti potencijalno poboljšanje performansi, posebno za problematične klase napada.

Kodiranje ulaznih podataka u impulse predstavlja izazov koji može značajno uticati na performanse, jer trenutno istraživanje koristi jednostavno kodiranje brzinom, dok sofisticiraniji pristupi kodiranja mogu poboljšati sposobnost modela da reprezentuje složene obrasce u podacima.

VII. Zaključak

Ovo istraživanje predstavlja sveobuhvatnu analizu primene tečnih mašina stanja sa Leaky Integrate-and-Fire neuronima u sistemima za otkrivanje mrežnih napada, koristeći CIC-IDS2017 skup podataka. Rezultati potvrđuju da neuromorfni pristupi mogu postići visoke performanse u klasifikaciji benignog i malicioznog mrežnog saobraćaja, sa AUROC vrednostima preko 0,99 za sve testirane konfiguracije.

Jedan od značajnijih rezultata istraživanja je prikaz brzine konvergencije LSM modela, koji dostižu AUROC vrednost od 0,998 već nakon prve epohe obuke. Ova karakteristika čini LSM pristup posebno pogodnim za implementaciju u sistemima za otkrivanje upada u realnom vremenu, gde je brza adaptacija na nove tipove napada ključna za efikasnu zaštitu. Konfiguracija sa 100 neurona u rezervoaru pokazala se kao najefikasnija od predloženih, postižući tačnost od 0,94 i AUPRC vrednost od 0,97, što predstavlja dobar kompromis između performansi klasifikacije i računskih zahteva.
Analiza uticaja veličine rezervoara otkriva da povećanje sa 50 na 100 neurona donosi značajno poboljšanje performansi, dok dalje povećanje na 500 i 1000 neurona rezultira samo marginalnim poboljšanjima. Ovo ukazuje na postojanje optimalnog praga složenosti nakon kojeg dodatni neuroni ne doprinose značajno poboljšanju performansi u odnosu na priloženi skup podataka i okruženje eksperimenta. Napadi koji pripadaju Infiltration i Heartbleed klasama, pokazuju slabije performanse u poređenju sa drugim tipovima napada, što ukazuje na potrebu za dalju optimizaciju pristupa u slučaju mrežnih napada koji nisu komparativno podjednako zastupljeni.

Poređenje sa tradicionalnim algoritmima mašinskog učenja pokazuje da LSM modeli postižu uporedive ili bolje performanse uz značajno kraće vreme obuke u odnosu na duboke neuronske mreže. Ove karakteristike, zajedno sa teoretskim osnovama neuromorfnog računarstva koje sugerišu nisku energetsku potrošnju omogućavaju predočeni LSM model atraktivnim za implementaciju na uređajima sa ograničenim računarskim resursima.

Literatura

[1] M. Lopez-Martin, B. Carro, and A. Sánchez-Esguevillas, "Application of deep reinforcement learning to intrusion detection for supervised problems," *Expert Syst. Appl.*, vol. 141, p. 112963, Mar. 2020, doi: 10.1016/j.eswa.2019.112963.

[2] C. Bartolozzi, G. Indiveri, and Elisa Donati, "Embodied neuromorphic intelligence," *Nat. Commun.*, vol. 13, no. 1, Feb. 2022, doi: 10.1038/s41467-022-28487-2.

[3] W. Maass and H. Markram, "On the computational power of circuits of spiking neurons," *J. Comput. Syst. Sci.*, vol. 69, no. 4, pp. 593–616, Dec. 2004, doi: 10.1016/j.jcss.2004.04.001.

[4] W. Maass, "Liquid State Machines: Motivation, Theory, and Applications," in *Computability in Context*, IMPERIAL COLLEGE PRESS, 2011, pp. 275–296. doi: 10.1142/9781848162778_0008.

[5] M. Lukoševičius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Comput. Sci. Rev.*, vol. 3, no. 3, pp. 127–149, Aug. 2009, doi: 10.1016/j.cosrev.2009.03.005.

[6] N. Brunel and M. C. W. van Rossum, "Quantitative investigations of electrical nerve excitation treated as polarization: Louis Lapicque 1907 · Translated by:," *Biol. Cybern.*, vol. 97, no. 5, pp. 341–349, Dec. 2007, doi: 10.1007/s00422-007-0189-6.

[7] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization:," in *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, Funchal, Madeira, Portugal: SCITEPRESS - Science and Technology Publications, 2018, pp. 108–116. doi: 10.5220/0006639801080116.

[8] Zafar Iqbal Khan, Mohammad Mazhar Afzal, and Khurram Naim Shamsi, "A Comprehensive Study on CIC-IDS2017 Dataset for Intrusion Detection Systems," *Int. Res. J. Adv. Eng. Hub IRJAEH*, vol. 2, no. 02, pp. 254–260, Feb. 2024, doi: 10.47392/IRJAEH.2024.0041.

[9] K. Roy, A. Jaiswal, and P. Panda, "Towards spike-based machine intelligence with neuromorphic computing," *Nature*, vol. 575, no. 7784, pp. 607–617, Nov. 2019, doi: 10.1038/s41586-019-1677-2.

[10] J. K. Eshraghian *et al.*, "Training Spiking Neural Networks Using Lessons From Deep Learning," Aug. 13, 2023, *arXiv*: arXiv:2109.12894. Accessed: Sep. 03, 2023. [Online]. Available: http://arxiv.org/abs/2109.12894

[11] W. Maass, T. Natschläger, and H. Markram, "Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations," *Neural Comput.*, vol. 14, no. 11, pp. 2531–2560, Nov. 2002, doi: 10.1162/089976602760407955.

[12] S. E. Smaha, "Haystack: an intrusion detection system," pp. 37–44, Dec. 1988, doi: 10.1109/acsac.1988.113412.

[13] K. A. Scarfone and P. M. Mell, "Guide to Intrusion Detection and Prevention Systems (IDPS)," Gaithersburg, MD: National Institute of Standards and Technology, 2007, p. NIST SP 800-94. doi: 10.6028/NIST.SP.800-94.

[14] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks," Sep. 12, 2019, *arXiv*: arXiv:1909.09586. Accessed: Feb. 24, 2023. [Online]. Available: http://arxiv.org/abs/1909.09586

[15] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," Dec. 11, 2014, *arXiv*: arXiv:1412.3555. doi: 10.48550/arXiv.1412.3555.

[16] A. Vaswani *et al.*, "Attention is All you Need," presented at the Neural Information Processing Systems, Jun. 2017. Accessed: Dec. 10, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Attention-is-All-you-Need-Vaswani-Shazeer/204e3073870fae3d05bcbc2f6a8e263d9b72e776

[17] C. Xu, J. Shen, X. Du, and F. Zhang, "An Intrusion Detection System Using a Deep Neural Network With Gated Recurrent Units," *IEEE Access*, vol. 6, pp. 48697–48707, 2018, doi: 10.1109/ACCESS.2018.2867564.

[18] M. Chen *et al.*, "Evaluating Large Language Models Trained on Code," Jul. 14, 2021, *arXiv*: arXiv:2107.03374. Accessed: Sep. 02, 2023. [Online]. Available: http://arxiv.org/abs/2107.03374

[19] Z. Alom and T. M. Taha, "Network intrusion detection for cyber security on neuromorphic computing system," *IEEE Int. Jt. Conf. Neural Netw.*, pp. 3830–3837, May 2017, doi: 10.1109/ijcnn.2017.7966339.

[20] W. Zahm *et al.*, "Cyber-Neuro RT: Real-time Neuromorphic Cybersecurity".

[21] M. Živadinović and D. Simić, "Resource efficient Internet-of-Things intrusion detection with spiking neural networks," presented at the

19th Conference on Computer Science and Intelligence Systems, Nov. 2024, pp. 73–78. doi: 10.15439/2024F8800.

[22] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the Development of Realistic Botnet Dataset in the Internet of Things for Network Forensic Analytics: Bot-IoT Dataset," Nov. 01, 2018, *arXiv*: arXiv:1811.00701. doi: 10.48550/arXiv.1811.00701.

[23] N. Koroniotis, N. Moustafa, E. Sitnikova, and J. Slay, "Towards Developing Network Forensic Mechanism for Botnet Activities in the IoT Based on Machine Learning Techniques," in *Mobile Networks and Management*, J. Hu, I. Khalil, Z. Tari, and S. Wen, Eds., in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Cham: Springer International Publishing, 2018, pp. 30–44. doi: 10.1007/978-3-319-90775-8_3.

[24] N. Koroniotis, N. Moustafa, and E. Sitnikova, "A new network forensic framework based on deep learning for Internet of Things networks: A particle deep framework," *Future Gener. Comput. Syst.*, vol. 110, pp. 91–106, Sep. 2020, doi: 10.1016/j.future.2020.03.042.

[25] N. Koroniotis and N. Moustafa, "Enhancing network forensics with particle swarm and deep learning: The particle deep framework," May 02, 2020, *arXiv*: arXiv:2005.00722. doi: 10.48550/arXiv.2005.00722.

[26] N. Koroniotis, N. Moustafa, F. Schiliro, P. Gauravaram, and H. Janicke, "A Holistic Review of Cybersecurity and Reliability Perspectives in Smart Airports," *IEEE Access*, vol. 8, pp. 209802–209834, 2020, doi: 10.1109/ACCESS.2020.3036728.

[27] N. Koroniotis, "Designing an effective network forensic framework for the investigation of botnets in the Internet of Things," Thesis, UNSW Sydney, 2020. doi: 10.26190/unsworks/21942.

[28] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *J. Physiol.*, vol. 117, no. 4, pp. 500–544, Aug. 1952, doi: 10.1113/jphysiol.1952.sp004764.

[29] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Trans. Neural Netw.*, vol. 14, no. 6, pp. 1569–1572, Nov. 2003, doi: 10.1109/TNN.2003.820440.

[30] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," Dec. 03, 2019, *arXiv*: arXiv:1912.01703. doi: 10.48550/arXiv.1912.01703.

[31] W. Guo, M. E. Fouda, A. M. Eltawil, and K. N. Salama, "Neural Coding in Spiking Neural Networks: A Comparative Study for Robust Neuromorphic Systems," *Front. Neurosci.*, vol. 15, 2021, Accessed: May 30, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.638474

[32] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," Jun. 05, 2018, *arXiv*: arXiv:1201.0490. doi: 10.48550/arXiv.1201.0490.

Neuromorphic computing and liquid state machines in intrusion detection systems over the CIC-IDS2017 dataset

Miloš Živadinović, Dejan Simić

ABSTRACT

Neuromorphic computing represents an innovative approach to solving complex problems in cyber security. This paper investigates the application of liquid state machines (LSM) with Leaky Integrate-and-Fire (LIF) neurons in network attack detection systems, using the CIC-IDS2017 dataset. A comparative analysis of LSM architectures using the PyTorch and snnTorch libraries with different reservoir sizes (50, 100, 500 and 1000 neurons) on nine classes of network traffic was conducted. The analysis shows results with AUROC values above 0.99 for all tested configurations, as well as fast convergence, reaching an AUROC value of 0.998 after the first training epoch. Of particular note is the configuration with 100 neurons that achieves an accuracy of 0.94 and an AUPRC value of 0.97. The results indicate the potential of the LSM architecture in the development of real-time intrusion detection systems, demonstrating the balance between convergence speed and detection success in basic configurations.

Digitalna transformacija u šumarstvu – primena TinyML tehnologije i veštačke inteligencije na ivici

Dejan L. Pavlović, dipl.inž.el.
JP "Nacionalni park Đerdap"
Donji Milanovac, Srbija
dejanpav@yahoo.com
ORCID broj: 0009-0000-5522-8908

***Apstrakt* - Razvojem veštačke inteligencije, minijaturizacijom hardverskih komponenti, primenom senzorskih i drugih sistema postiže se razvoj novih digitalnih tehnologija. Integracijom ovakvih naprednih tehnologija dolazimo do koncepta pametnog šumarstva čiji je cilj unapređenje aktivnosti u svim oblastima šumarstva. Sa druge strane, digitalna transformacija ne obuhvata samo primenu naprednih tehnologija već i promenu celokupnog načina funkcionisanja i upravljanja radi postizanja zadatih ciljeva. Upravo primenom veštačke inteligencije na ivici i TinyML tehnologije, moguće je unaprediti rad mnogih oblasti šumarstva - poslove planiranja, organizovanja i upravljanja šumskim resursima u cilju očuvanja biodiverziteta, održivog korišćenja i očuvanja zdravlja šuma.**

Ključne reči – Digitalna transformacija, pametno šumarstvo, TinyML, veštačka inteligencija na ivici, računarstvo na ivici.

I. UVOD

Digitalna transformacija u pametnom šumarstvu koja integriše napredne tehnologije poput veštačke inteligencije i senzorskih sistema u cilju upravljanja i očuvanja šuma, ključna je za unapređenje upravljanja šumama uz minimizovanje uticaja na životnu sredinu tj. zaštitu biodiverziteta [1].

Na osnovu [2], predviđa se da će do kraja 2025. godine biti oko dvadeset milijardi IoT (eng. Internet of Things) uređaja povezanih na internet. Ovi uređaji generišu veliku količinu podataka u sekundi, imaju ograničenu računarsku snagu, male memorijske kapacitete i nedostaje im samodovoljna inteligencija da lokalno obrađuju neobrađene podatke i donose nezavisne odluke. Tradicionalno korišćenje IoT uređaja podrazumevalo je prebacivanje svih podataka u centralizovani računarski sistem – oblak (eng. Cloud Computing) radi dalje obrade. Nedostatak ovakvog pristupa dovodi do zasićenja propusnog opsega i velikog kašnjenja. Takođe, bezbednost privatnih podataka se dovodi u pitanje jer se podaci moraju čuvati u oblaku na neodređeno vreme [3].

Za rešavanje ovih problema predlažu se dva efikasna rešenja [3]. Jedno od rešenja podrazumeva da se podaci obrađuju što bliže njihovom izvoru dok se samo bitni podaci prenose udaljenim serverima u oblaku radi dalje obrade. Ovakvo rešenje poznato je kao računarstvo na ivici (eng. Edge Computing). Ono podrazumeva korišćenje manjih servera na lokacijama blizu korisnika ili pak mikroservera na bazi ARM procesora – npr. mobilnih uređaja za primenu specifičnih aplikacija. Takođe za primenu računarstva na ivici mogu biti korišćene i kamere sa integrisanim procesorima kao i razni senzorski sistemi. Drugo rešenje je pokretanje ML (eng. Machine Learning) algoritma mašinskog učenja na krajnjim uređajima. Primena ovakvog rešenja omogućava donošenje odluka u realnom vremenu kroz analizu podataka. Takođe, ovakvo rešenje sprečava zagušenje propusnog opsega i omogućava zaštitu podataka korisnika. Jedno od takvih rešenja je primena TinyML (eng. Tiny Machine Learning) tehnologije.

Ostatak rada organizovan je na sledeći način – nakon uvodnog dela predstavljena je arhitektura računarstva na ivici kao i prednosti korišćenja u odnosu na tradicionalno računarstvo u oblaku. Zatim sledi poglavlje koje obrađuje praktičnu primenu veštačke inteligencije na ivici u šumarstvu, a nakon toga poglavlja o koncepciji i primeni TinyML tehnologije. U narednom poglavlju su obrađene uporedne karakterisitke i primene veštačke inteligencije na ivici i TinyML tehnologije. Zatim sledi poglavlje u kome je dat pregled najčešće korišćenih senzora za realizaciju koncepcije pametnog šumarstva. Na kraju sledi zaključak u kome je objašnjena suština digitalne transformacije u šumarstvu i pregled korišćene literature.

II. RAČUNARSTVO NA IVICI

Na osnovu [3, 4] pod računarstvom na ivici se može smatrati bilo koji računarski ili mrežni resurs između izvora podataka i oblaka. To mogu biti pametni telefoni, mrežni prolazi, mikrocentri podataka i sl. Računarstvo na ivici smanjuje količinu podataka između čvorova, omogućava lokalnu analizu podataka i donošenje odluka u realnom vremenu. Takođe, omogućava smanjeno oslanjanje na resurse u oblaku kao i poboljšanu autonomiju rada.

Na osnovu [4], arhitektura računarstva na ivici sastoji se iz tri sloja – *krajnjeg ili terminalnog sloja*, *ivičnog sloja* i *sloja oblaka*.

Krajnji (terminalni) sloj – sastoji se od svih tipova uređaja povezanih na ivicu mreže uključujući mobilne terminale i IoT uređaje (senzore, pametne telefone, kamere i sl.). Ovaj sloj pomoću senzora vrši digitalizaciju fizičkih veličina okoline, a zatim prosleđuje ivičnom sloju na dalju obradu i čuvanje. Takođe, vrši prijem obrađenih podataka od ostalih slojeva prema potrebi krajnjih korisnika.

Ivični sloj – nalazi se na ivici mreže i sastoji se od ivičnih čvorova široko raspoređenih između terminalnih uređaja i oblaka. Obično uključuje bazne stanice, rutere, pristupne tačke, mrežne prolaze i sl. Ivični sloj podržava pristup terminalnih uređaja nadole i izračunava i skladišti podatke koje prenose terminalni uređaji. Obzirom da je ivični sloj blizu korisnika, prenos podataka do ivičnog sloja je pogodniji za analizu podataka u realnom vremenu i inteligentnu obradu, što je bezbednije i efikasnije od računarstva u oblaku.

Sloj oblaka – sastoji se od niza servera i uređaja za skladištenje visokih performansi i ima veliku ulogu u oblastima koje zahtevaju analizu velikih količina podataka.

U poređenju sa tradicionalnim računarstvom u oblaku, računarstvo na ivici ima mnoge jedinstvene prednosti, koje podrazumevaju [5] - *malu (nisku latenciju)*, *uštedu energije*, *kontekstnu svesnost* kao i *privatnost i bezbednost*.

Mala (niska) latencija - Pošto su ivični uređaji postavljeni bliže krajnjim uređajima koji su obično i izvor podataka, kašnjenje prenosa može biti u velikoj meri smanjeno u poređenju sa korišćenjem računarstva u oblaku. Na primer, kašnjenje prenosa se obično kreće u desetinama ili stotinama milisekundi između krajnjeg korisnika i servera u oblaku, dok je kašnjenje korišćenjem računarstva na ivici reda veličine od nekoliko milisekundi pa čak i na nivou mikrosekundi.

Ušteda energije – Računarstvo na ivici omogućava milijardama IoT uređaja da računske zadatke koji troše najviše energije prebace na ivične servere što ne samo da smanjuje potrošnju energije već i poboljšava efikasnost obrade.

Kontekstna svesnost – Kontekstno-svesno računarstvo ima važnu ulogu u IoT i ivičnim računarskim aplikacijama pošto se dobro modelovanje i rezonovanje prikupljenih podataka može u velikoj meri osloniti na kontekst podataka. Zbog svoje blizine, ivični serveri mogu prikupiti više informacija o kontekstu i na taj način podržati obradu podataka.

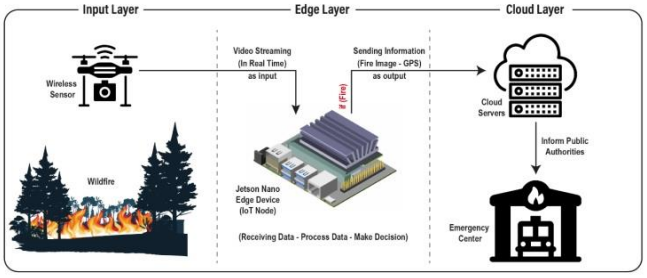
Privatnost i bezbednost – U poređenju sa računarstvom u oblaku, računarstvo na ivici je efikasnije u zaštiti privatnosti podataka i bezbednosti aplikacija.

III. PRIMENA VEŠTAČKE INTELIGENCIJE NA IVICI U ŠUMARSTVU

Jedna od praktičnih primena veštačke inteligencije na ivici u šumarstvu objašnjena je radom [6]. Predstavljen je višeslojni sistem rane detekcije šumskih požara korišćenjem YOLOv5 (eng. You Only Look Once version 5) [7] modela dubokog učenja za primenu u ivičnom okruženju. Sistem se sastoji iz sledećih slojeva:

1. *Ulaznog sloja* - sastavljenog od IoT uređaja opremljenih kamerama. IoT uređaji nalaze se na bespilotnim letelicama - dronovima (UAV eng. Unmanned Aerial Vehicle),
2. *Ivičnog sloja* – sa NVIDIA Jetson Nano Developer razvojnim alatom za primenu u ivičnom okruženju,
3. *Sloja oblaka* – sačinjenog od servera u oblaku.

Slikom 1 [6] predstavljen je princip rada sistema.



Slika 1. Princip rada sistema za otkrivanje požara

Naporima autora rada [6], pod nazivom WILDFIRE-I napravljen je skup podataka sačinjen od slika šumskih požara. Proces stvaranja WILDFIRE-I skupa podataka uključivao je četiri glavna koraka – prikupljanje slika, tehnike predobrade slika radi uklanjanja šuma i ispravljanja grešaka, označavanje slika i razdvajanje podataka. Modifikovani model dubokog učenja (zasnovan na YOLOv5 mreži) usvojen je za obuku, testiranje i validaciju modifikovanog modela. Obrada strimovanog video materijala vrši se u ivičnom sloju uključujući detekciju višestrukih objekata na svakom pojedinačnom frejmu u okviru zadatka binarne klasifikacije. Konačno, slike sa objektima vatre šalju se na server u oblaku (Google drive) koji predstavlja sloj donošenja odluka.

Hardverski zahtevi predstavljenog sistema – Lenovo laptop računar (CPU: Ryzen 5, GPU: RTX 3050 4 GB, RAM: 16 GB). Za potrebe rada ivičnog sloja NVIDIA Jetson Nano Developer razvojni alat (GPU 128-core Maxwell, CPU Quad-core ARM A57 1.43 GHz, memorija 4 GB 64-bit LPDDR4 25.6 GB/s), Logitech USB kamera: 5MP, 720 HD, 5V 3A napajanje Type-C / Micro USB, SanDisk 64GB Ultra Micro SD HC Class 10 memorijska kartica i USB WiFi adapter.

Softverski zahtevi predstavljenog sistema – Operativni sistem Windows 10 za laptop, Ubuntu 18.04 za Jetson Nano i Python programski jezik. Dodatni softverski zahtevi su: PyCharm Community, Anaconda 3, PyTorch 1.9 sa Torchvision 0.11.0, Cuda 10.2, najnovija verzija biblioteka za Python, Image Annotation Lab i FastStone Photo Resizer (za obradu slika).

IV. TINYML TEHNOLOGIJA

TinyML tehnologija omogućava primenu modela mašinskog učenja na uređajima sa ultra-niskom potrošnjom, omogućavajući obradu podataka u realnom vremenu i donošenje odluka bez oslanjanja na računarstvo u oblaku [8]. Međutim, integracijom računarstva u oblaku, računarstva na ivici i TinyML koncepta stvara se moćan okvir za primenu IoT aplikacija. Računarstvo u oblaku obezbeđuje skalabilnu moć skladištenja i obrade podataka, računarstvo na ivici približava obradu podataka ka izvoru smanjujući kašnjenje i poboljšavajući donošenje odluka u realnom vremenu, dok TinyML svojim efikasnim modelima mašinskog učenja olakšava analizu čineći IoT uređaje osetljivijim i pametnijim [8].

Paralelno razvoju mašinskog učenja i veštačke inteligencije došlo je i do razvoja IoT uređaja. Blizina izvoru fizičkih informacija putem senzora i nedovoljno iskorišćene mogućnosti obrade podataka čine IoT uređaje veoma pogodnim za primenu lakih algoritama zaključivanja. Ovaj trend doveo je do razvoja TinyML tehnologije [9]. Dva ključna izazova u primeni neuronskih mreža na mikrokontrolerima su mala veličina memorije i kratko trajanje baterije. Pokretanje složenih modela mašinskog učenja kao što su neuronske mreže na visoko ograničenom ugrađenom hardveru zahteva pažljivo projektovanje harvera i softvera. Veličina modela mašinskog učenja mora biti dovoljno mala da se uklopi u ograničenja uređaja. Tipični mikrokontroleri imaju izuzetno ograničenu integrisanu memoriju (SRAM) koja se kreće u opsegu od 192-512 KB i fleš memoriju u opsegu od 256 KB – 2 MB. Ceo model neuronske mreže sa svojim težinama, neuronskim vezama i pratećim kodom mora da stane u malu fleš memoriju.

Tabelom 1 prikazane su vrednosti tehničkih karakteristika pojedinih MCU platformi [9].

Tabela 1. Reprezentativni uređaji koji podržavaju TensorFlow Lite za mikrokontrolere

MCU platforma	Procesor	Frekvencija	SRAM	Fleš memorija
Arduino Nano 33 BLE Sense	ARM Cortex M4	64 MHz	256 KB	1 MB
ESP 32	Tensilica Xtensa LX6	160 MHz	512 KB	2 MB
Sparkfun Edge Appolo 3 Blue	ARM Cortex M4F	48 MHz	384 KB	1 MB
ST Nucleo Boards	ARM Cortex M7	216 MHz	320 KB	1 MB
Adafruit EdgeBadge	ATSAMD51	120 MHz	192 KB	512 KB

Da bi se modeli mašinskog učenja mogli uklopili u hardverska ograničenja samih uređaja, neophodno je koristiti odgovarajuće softverske okvire ili platforme koje pružaju osnovne alate, biblioteke i funkcionalnosti potrebne za razvoj i implementaciju modela veštačke inteligencije i algoritama. Neke od popularnih platformi su [3,8,9]: TensorFlow Lite, uTensor, Edge Impulse, NanoEdge AI Studio, PyTorch Mobile, MediaPipe, Web of Science, OpenVINO, Embedded Learning Library (ELL), ARM-NN i dr.

Za prevazilaženje izazova ograničene memorije i niske brzine obrade ugrađenih uređaja koriste se pristupi *redukcije modela* i *laganih okvira* [9].

1. Pristup redukcije modela – podrazumeva smanjenje veličine neuronskog modela kako bi se prilagodio mikrokontroleru. Postiže se *kompresijom modela* (smanjivanjem broja slojeva neuronske mreže), *obrezivanjem modela* (postavljanjem težina s niskim vrednostima na nulu) ili *kvantizacijom parametara*.

Kompresija modela kombinuje metode primenjene iz različitih oblasti – mašinskog učenja, obrade signala, računarske arhitekture i optimizacije. Metoda obrezivanja podrazumeva uklanjanje nepotrebnih ili manje bitnih parametara, čime se postiže veća efikasnost modela, dok proces kvantizacije parametara podrazumeva proces kvantizacije mreže gde se veličina mreže smanjuje smanjivanjem broja bita potrebnih za predstavljanje svake težine.

2. Pristup laganih okvira – podrazumeva primenu alata otvorenog koda za dizajniranje i implementaciju algoritama mašinskog učenja na uređajima sa ograničenim resursima.

V. PRIMENE TINYML TEHNOLOGIJE U ŠUMARSTVU

TinyML koristi lagane algoritme optimizovane za ivične uređaje sa ograničenim resursima. Najčešći algoritmi uključuju varijante neuronskih mreža kao što su CNN (eng. Convolutional Neural Networks), rekurentnih LSTM (eng. Long Short-Term Memory) mreža kao i stabla odlučivanja (eng. Decision Tree). Ključne karakteristike ovih modela uključuju kompresiju modela, obrezivanje i kvantizaciju parametara. Aplikacije zasnovane na viziji koriste TinyML za zadatke klasifikacije slika, prepoznavanje objekata i detekcije pokreta [8].

Tabelom 2 [3] prikazane su različite hardverske arhitekture i softverski okviri sa prednostima i nedostatacima korišćenih algoritama mašinskog učenja, primenjenih u *pametnom transportu* i *pametnoj poljoprivredi*. Ovakva ili delimično modifikovana arhitektura sa istim ili modifikovanim algoritmima mašinskog učenja, takođe bi mogla biti primenjena u pojedinim oblastima *pametnog šumarstva*.

Tabela 2. Prednosti i nedostaci primenjenih algoritama na različitim arhitekturama

Primena	ML model	Hardver / Softverski okvir	Prednosti	Nedostaci
Pametni transport	Random Forest	Raspberry Pi 3B + Smartphone	Smanjeno kašnjenje	Niži nivo privatnosti
Pametna poljoprivreda	LSTM	TensorFlow i Keras sa Raspberry Pi 4 Model B	Smanjeno opterećenje	Manja tačnost
	ResNet - 50 (CNN)	TensorFlow, Keras i OpenCV sa NVIDIA Jetson Nano i Logitech WebCam	Visoka tačnost i detekcija u realnom vremenu	Ne uzima u obzir raznovrsnost podataka
	CNN-SVM	TensorRT sa NVIDIA Jetson TX1	Visoka tačnost i brzo donošenje odluka	Skalabilnost nije rešena
	LSTM i GRU	TensorFlow Lite i Pytorch sa Sensors i Raspberry Pi 3 B+	Poboljšano donošenje odluka i poboljšana održivost	Visoka kompleksnost modela i bezbedonosni rizici

VI. KOMPARACIJA VEŠTAČKE INTELIGENCIJE NA IVICI I TINYML TEHNOLOGIJE

Na osnovu [10], tabelom 3 prikazane su uporedne karakteristike ključnih razlika, performansi i efikasnosti primene veštačke inteligencije na ivici i TinyML tehnologije.

Tabela 3. Uporedne karakteristike primene veštačke inteligencije na ivici i TinyML tehnologije

Definicija i obim	
Veštačka inteligencija na ivici	**TinyML tehnologija**
Odnosi se na primenu algoritama veštačke inteligencije na ivici mreže bliže izvoru.	Fokusira se na pokretanje algoritama mašinskog učenja na uređajima sa ekstremno ograničenim resursima kao što su mikrokontroleri.
Performanse i korišćenje resursa	
Obično zahteva više računarskih resursa i može da iskoristi moćniji hardver za obavljanje složenih zadataka.	Dizajnirana da radi na uređajima sa minimalnim resursima često koristeći kvantizovane modele optimizovane za nisku potrošnju energije i smanjenu upotrebu memorije.
Slučajevi upotrebe	
Najčešće se koristi u aplikacijama koji zahtevaju trenutnu obradu	Koriste se u aplikacijama kao što su prenosivi monitori zdravlja,

podataka i nisko kašnjenje kao što su pametna vozila, pametne kamere i industrijska automatizacija.	pametni kućni uređaji i senzori životne sredine.
Povezivanje i rukovanje podacima	
Može da radi povremenim povezivanjem sa oblakom, omogućavajući da se podaci obrađuju lokalno, a da i dalje ima mogućnost slanja objedinjenih podataka nazad u oblak radi dalje analize.	Često radi u potpunom off-line režimu obrađujući podatke lokalno bez potrebe za interakcijom sa oblakom. Od posebnog je značaja za primenu u udaljenim oblastima gde je veza nepouzdana ili nepostojeća.

VII. PAMETNO ŠUMARSTVO

Digitalna transformacija u pametnom šumarstvu predstavlja značajan pomak ka korišćenju naprednih tehnologija za održivo upravljanje i očuvanje šumskih resursa. Ono što je značajno u vezi ove transformacije je generisanje i efikasno korišćenje podataka, što zahteva različite vrste senzora i drugih ulaznih uređaja. Na primer, senzori temperature su ključni za detekciju šumskih požara, monitoring divljih životinja ili procenu zdravlja drveća. Ili npr. senzori vlage doprinose razumevanju sadržaja vode u šumskom zemljištu [11].

Tabelom 4 [11] prikazane su prednosti i nedostaci senzora najčešće primenjivanih u pametnom šumarstvu.

Tabela 4. Primena senzora u pametnom šumarstvu

Tip senzora	Primena	Prednosti	Nedostaci
GNSS i WiFi	Lokalna	Određivanje tačne pozicije.	Problemi u pokrivanju terena zbog nekonzistentnosti signala.
Temperaturni	Merenje temperature	Pružanje uvida u fluktuacije temperature. Može se integrisati sa drugim ulazima.	Bez integracije sa drugim podacima, ima veoma usku primenu u šumarstvu.
Senzor vlage	Merenje sadržaja vode u vazduhu ili zemljištu	Generisanje klimatskih podataka. Mogućnost lakog povezivanja sa temperaturnim ulazima. Primene variraju od merenja vlažnosti ispod krošnji stabala do određivanja prohodnosti šumskih puteva.	Monitoring u realnom vremenu moguć je samo korišćenjem pouzdane mreže. Podešavanje može biti komplikovano.
Senzor pH vrednosti zemljišta	Određivanje rastvorljivosti raznih nutrijenata u zemljištu	Razumevanje sposobnosti apsorpcije nutrijenata u odnosu na zemljište.	Terenska merenja sa ugrađenim senzorima ne mogu postići istu preciznost kao merenja u laboratorijskim uslovima.
RGB kamera	Generisanje slike okoline	Praćenje divljih životinja, razumevanje stanja šumskih puteva i pomoć u navigaciji.	Obrada podataka slike može biti izazov. Generisanje velike količine podataka
			korišćenjem video materijala ili fotografija visoke rezolucije.
Termalna kamera	Merenje toplotnih profila okoline	Razumevanje temperaturnih gradijenata. SAR detekcija šumskih požara.	Moguće generisanje velike količine podataka. Obrada slika može biti izazov.
LiDAR senzor	Obezbeđivanje detaljnih topografskih podataka	Pomoć pri proceni biomase, analizi strukture šuma i pomoć u navigaciji.	Skupa tehnologija. Problem sa zaprljanošću sočiva za vreme kiše ili usled drugih atmosferskih zagađenja.
Multispektralna kamera	Istovremeno merenje više spektara u cilju jednostavnog povezivanja podataka	Jednostavno podešavanje radi prikupljanja širokog spektra međusobno direktno povezanih podataka.	Istovremeno korišćenje više talasnih dužina može dovesti do lošijih performansi svakog detektora u poređenju sa specijalizovanim detektorima.

VIII. ZAKLJUČAK

Koncept pametnog šumarstva obuhvata upotrebu najsavremenijih tehnologija u upravljanju šumama. Te tehnologije podrazumevaju primenu IoT uređaja, senzorskih sistema, upotrebu algoritama mašinskog učenja, veštačke inteligencije i savremenih mrežnih koncepata. U zavisnosti od vrste zadatih ciljeva vrši se i izbor odgovarajućih koncepata, odnosno izbor hardverske arhitekture, softverskih okvira/alata i primenjenih algoritama. Dakle, primenom naprednih tehnologija vrši se unapređenje konkretnih aktivnosti u šumarstvu kao što je monitoring, analiza šumskih ekosistema, precizno upravljanje resursima, automatizacija raznih procesa i sl. Sa druge strane, digitalna transformacija u šumarstvu nije samo usvajanje tehnologija, već i integracija ovih tehnologija u kohezivni sistem koji unapređuje održivo upravljanje šumama. Ona uključuje digitalizaciju svih procesa, integraciju svih sektora šumarstva, promenu strategija i organizacionih struktura. Jednom rečju, digitalna transformacija predstavlja promenu svih aspekata upravljanja šumama.

LITERATURA

[1] Andreas Holzinger, Anna Saranti, Alessa Angerschmid, Carl Orge Retzlaff, Andreas Gronauer, Vladimir Pejakovic, Francisco Medel-Jimenez, Theresa Krexner, Christoph Gollob, Karl Stampfer „Digital Transformation in Smart Farm and Forest Operations Needs Human-Centered AI: Challenges and Future Directions“, Sensors 2022, 22, 3043. https://doi.org/10.3390/s22083043

[2] https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/ [pristupljeno dana 03.01.2025.]

[3] Oumayma Jouini, Kaouthar Sethom, Abdallah Namoun, Nasser Aljohani, Meshari Huwaytim Alanazi, Mohammad N. Alanazi „A Survey of Machine Learning in Edge Computing: Techniques, Frameworks, Applications, Issues, and Research Directions“, Technologies 2024, 12, 81. https://doi.org/10.3390/technologies12060081

[4] Keyan Cao, Yefan Liu, Gongjie Meng, Qimeng Sun „An Overview on Edge Computing Research“, IEEEAccess, Volume 8, 2020. DOI: 10.1109/ACCESS.2020.2991734

[5] Fangxin Wang, Miao Zhang, Xiangxiang Wang, Xiaoqiang Ma, Jiangchuan Liu „Deep Learning for Edge Computing Applications: A State-of-the-art Survey“, IEEE Access, Volume 8, 2020. DOI: 10.1109/ACCESS.2020.2982411

[6] Ahmed Saleem Mahdi, Sawsen Abdulhadi Mahmood „An Edge Computing Environment for Early Wildfire Detection“, Annals of Emerging Technologies in Computing (AETiC), Vol. 6, No. 3, 2022. DOI: 10.33166/AETiC.2022.03.005

[7] https://iq.opengenus.org/yolov5/ [pristupljeno dana 09.01.2025.]

[8] Abdussalam Elhanashi, Pierpaolo Dini, Sergio Saponara, Qinghe Zheng „Advancements in TinyML: Applications, Limitations, and Impact on IoT Devices“, Electronics 2024, 13, 3562. https://doi.org/10.3390/electronics13173562

[9] Stanislava Soro „TinyML for Ubiquitous Edge AI“, MTR200519, MITRE Technical Report, September 2020. https://doi.org/10.48550/arXiv.2102.01255

[10] https://www.restack.io/p/edge-ai-vs-tinyml-answer-cat-ai [pristupljeno dana 13.01.2025.]

[11] Florian Ehrlich-Sommer, Ferdinand Hoenigsberger, Christoph Gollob, Arne Nothdurft, Karl Stampfer, Andreas Holzinger „Sensors for Digital Transformation in Smart Forestry“, Sensors 2024, 24, 798. https://doi.org/10.3390/s24030798

Digital Transformation in Forestry - Application of TinyML Technology and Edge Artificial Intelligence

Dejan L. Pavlović

ABSTRACT

The development of artificial intelligence, miniaturization of hardware components and the application of sensor and other systems lead to the development of new digital technologies. By integrating these advanced technologies, we arrive at the concept of smart forestry, which aims to improve activities in all areas of forestry. On the other hand, digital transformation encompasses not only the application of advanced technologies but also the change of the entire way of functioning and management to achieve the set goals. Through the application of edge computing and TinyML technology, it is possible to improve operations in many areas of forestry - planning, organizing, and managing forest resources with the goal of preserving biodiversity, sustainable exploitation, and maintaining forest health.

Sistemi za podršku odlučivanju u eri veštačke inteligencije

Boris Delibašić
Univerzitet u Beogradu – Fakultet organizacionih nauka
boris.delibasic@fon.bg.ac.rs
0000-0002-6153-5119

Sandro Radovanović
Univerzitet u Beogradu – Fakultet organizacionih nauka
sandro.radovanovic@fon.bg.ac.rs
0000-0002-9975-8844

Milija Suknović
Univerzitet u Beogradu – Fakultet organizacionih nauka
sandro.radovanovic@fon.bg.ac.rs
0009-0003-0449-5258

***Apstrakt* - Razvoj veštačke inteligencije i metode mašinskog učenja značajno su promenile način na koji se donose odluke u savremenim organizacijama. Iako se često naglašava potreba za automatizacijom i prediktivna tačnost procesa odlučivanja, razumljivost, transparentnost i podrška ljudskom odlučivanju sve više dobijaju na značaju. U ovom radu razmatra se uloga sistema za podršku odlučivanju (DSS) u savremenoj eri veštačke inteligencije, sa posebnim naglaskom na hijerarhijske, kvalitativne modele zasnovane na ekspertskom znanju. Prikazuje se kako modeli DSS mogu da zadovolje ključne kriterijume kvalitetne podrške odlučivanju i da obezbede dodatnu vrednost u odnosu na modele mašinskog učenja. Kroz ilustrativni primer predikcije odliva korisnika diskutuje se komplementarnost DSS i AI pristupa, kao i njihove praktične implikacije u realnim organizacionim okruženjima.**

Ključne reči – sistemi za podršku odlučivanju, mašinsko učenje, DEX model, hijerarhijski modeli odlučivanja, odlazak korisnika.

I. Uvod

Sistemi za podršku odlučivanju (Decision Support Systems – DSS) predstavljaju jednu od najstarijih oblasti informacionih sistema, razvijenu sa ciljem da pomogne donosiocima odluka u rešavanju polustrukturiranih i nestrukturiranih problema odlučivanja [2]. Klasična DSS arhitektura obuhvata sistem znanja, sistem za obradu problema i sistem za komunikaciju sa korisnikom. Za razliku od automatizovanih sistema, DSS su od samog početka bili usmereni ka podršci ljudskom odlučivanju, a ne njegovoj zameni.

Tokom poslednje decenije, razvoj velikih skupova podataka i naprednih algoritama mašinskog učenja doveo je do prevage pristupa zasnovanima nad podacima i veštačkoj inteligenciji nad ekspertskim modelovanjem procesa odlučivanja. Modeli mašinskog učenja ostvaruju visoke performanse u prediktivnim zadacima i omogućavaju značajan stepen automatizacije poslovnih procesa. Međutim, ovakav pristup često zanemaruje ključna pitanja odgovornosti, razumljivosti i usklađenosti modela sa stvarnim ciljevima organizacije.

U savremenom kontekstu, gde se sve više govori o odgovornoj i objašnjivoj veštačkoj inteligenciji, javlja se potreba za ponovnim razmatranjem uloge sistema za podršku odlučivanju. Ovaj rad polazi od pretpostavke da DSS i AI tehnologije ne treba da budu posmatrani kao konkurentske tehnologije, već kao komplementarni pristupi koji zajedno mogu obezbediti kvalitetnije odlučivanje.

Rezultati ovog istraživanja prvi put su prikazani na ovoj konferenciji, a nastavak rada je objavljen u radu [4].

II. Sistemi za podršku odlučivanju i kriterijumi kvaliteta

Jedan od osnovnih načina za procenu kvaliteta DSS modela jeste analiza njihovih karakterističnih svojstava. U literaturi se često ističe skup od pet ključnih kriterijuma, poznatih kao 5C kriterijumi [3]: tačnost, potpunost, konzistentnost, razumljivost i pogodnost za korišćenje.

Tačnost podrazumeva da model ne treba samo da daje tačne rezultate, već i da bude usklađen sa problemom odlučivanja. Kod sistema AI često se sreće problem neusklađenosti (eng. Misalignment). Potpunost se odnosi na sposobnost modela da funkcioniše za sve moguće kombinacije ulaznih vrednosti, što je ograničenje određenih AI modela koji rade isključivo nad sličnim podacima nad kojim su učeni. Konzistentnost zahteva da model DSS ne proizvodi kontradiktorne odluke za slične situacije, što se može dešavati kod određenih modela AI. Razumljivost omogućava donosiocima odluka da shvate logiku modela i da mu veruju, dok pogodnost za korišćenje obuhvata praktične aspekte primene, uključujući analizu osetljivosti i protivčinjenično zaključivanje, što većina DSS ima u sebi, a kod AI modela tek postaje standard.

Iako modeli mašinskog učenja često postižu visoke prediktivne tačnosti, samo delimično zadovoljavaju ostale 5C kriterijume. DSS modeli su, sa druge strane, projektovani tako da inherentno poštuju svih pet kriterijuma, čime postaju posebno pogodni za kompleksne i odgovorne odluke.

III. Hijerarhijski DSS modeli i metod DEX

DEX (Decision EXpert) metod predstavlja kvalitativni hijerarhijski pristup višekriterijumskom odlučivanju koji se svrstava u klasu hijerarhijskih modela DSS [3]. U ovom radu će ovaj biti prikazan ovaj sistem DSS i njegove praktične pogodgnosti. DEX umesto numeričkih vrednosti, koristi skup uređenih kvalitativnih skala (ordinalni podaci), dok se odluke donose na osnovu eksplicitno definisanih pravila tipa „AKO–ONDA“ koje povezuje vrednosti skala različitih atributa u nove, izvedene atribute, ili ciljni atribut, omogućavajući da se donose odluka o vrednosti ciljnog atributa na osnovu korisnički-definisanih pravila i hijerarhijske strukture. Ovakav pristup omogućava modelovanje odluka na način koji je blizak ljudskom rezonovanju.

Hijerarhijska struktura DEX modela sastoji se od elementarnih atributa, koji predstavljaju direktno posmatrane karakteristike, i agregiranih atributa, koji opisuju složenije koncepte dobijene kombinovanjem nižih nivoa. Pravila agregacije jasno definišu kako se vrednosti atributa

propagiraju kroz hijerarhiju do konačne odluke.

Jedna od najvažnijih prednosti DEX modela jeste mogućnost analize „ŠTA-AKO“ scenarija, što predstavlja svojevrsni pandan protivčinjenične analize, sa važnim ograničenjem da se ovde radi o strukturalnoj protivčinjeničnoj analizi zasnovanoj na modelu koji je definisao ekspert donosilac odluke ili grupa eksperata donosilaca odluke. Tokom primene modela DSS, donosioci odluka mogu sistematski menjati skalirane vrednosti pojedinačnih atributa i posmatrati kako te promene utiču na krajnji ishod, što omogućava identifikaciju ključnih faktora i potencijalnih intervencija.

IV. PRIMER: PREDIKCIJA ODLIVA KORISNIKA

Predikcija odliva korisnika predstavlja tipičan problem u kome se često primenjuju modeli mašinskog učenja [1]. Cilj je identifikovati korisnike sa visokim rizikom od napuštanja sistema kako bi se blagovremeno preduzele mere zadržavanja korisnika.

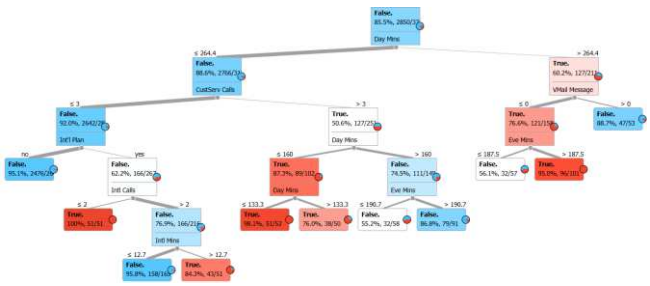
Modeli AI u ovom kontekstu postižu visoke vrednosti standardnih metrika performansi, ali često ne pružaju dovoljno informacija o razlozima koji dovode do odliva. DSS modeli zasnovani na hijerarhijskom DEX pristupu omogućavaju eksplicitno modelovanje odnosa između ponašanja korisnika, karakteristika usluge i konačne odluke.

Osobinom šta-ako analize modeli DSS omogućavaju da se za svakog korisnika sistema zna da li je sklon ka prekidanju korišćenja usluge ili ne, ali isto tako i koji parametri utiču na takvu njegovu/njenu odluku i šta minimalno dovodi do promene takve odluke. Iako se ove osobine mogu dograditi i korišćenjem algoritama mašinskog učenja, posebno korišćenjem protivčinjenične analize, ovakvo rezonovanje je već sastavni deo većine modela DSS.

Posebno značajna prednost modela DSS ogleda se u mogućnosti da odnos pogrešne klasifikacije propuštanja stvarnog odliva i lažne uzbune, direktno ugrađuje u modele DSS, bez potrebe da se eksplicitno navodi veličina tih troškova. Kod modela AI je potrebno eksplicitno navesti ove veličine da bi sistem bio svestan da pragove odlučivanja treba da prilagodi zahtevima donosilaca odluke.

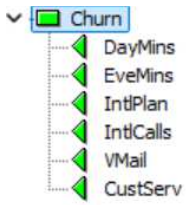
U nastavku prikazujemo eksperiment sa tri modela za sprečavanje odliva korisnika i to:

1. Model zasnovan na modelima mašinskog učenja, korišćenjem standardnih parametara modela.
2. DEX model razvijen za isti problem od strane eksperta za modelovanje DEX sistema.
3. DEX model naučen iz podataka, kao što je prikazano u radu [6].



Slika 1. Model stabla odlučivanja napravljen u softveru Orange.

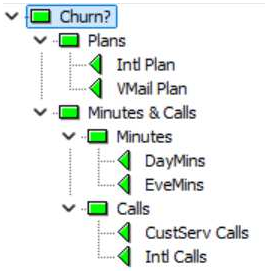
Model stabla odlučivanja (Slika 1) postiže tačnost od 90,7%, odziv 56,7%, i preciznost 73,3%. Smanjen odziv u odnosu na preciznost govori da bi greška neprepoznavanja odlaska klijenta mogla da se smanji na račun povećavanja lažnih pretpostavki o odlasku klijenta. Jedan od načina da se to postigne jeste da se definišu troškovi pogrešne klasifikacije, i da se shodno tim troškovima prag odlučivanja pomeri, ne bi li uzeo u obzir te zahteve.



Slika 2. DEX model o odlasku klijenta. Izabrano je šest najbitnijih atributa i na osnovu kombinacije vrednosti ordinalnih vrednosti definisana su pravila o odlasku klijenta (Npr. jedno pravilo bi moglo da bude da mali broj dnevnih minuta, umereni broj noćnih minuta, korisćenje međunarodnih poziva, umereni broj međunarodnih poziva, nekorišćenje govorne pošte i mali broj poziva ka korisničkom servisu uslovljava da klijent ne napušta kompaniju)

DEX model je postigao tačnost od 74,71%, odziv od 62,94%, preciznost od 31,4%. Vidi se da su prediktivne moći DEX modela niže u odnosu na model stabla odlučivanja, ipak je model uspeo da ima bolji odziv, tj. da bolje prepoznaje klijente koji zaista planiraju da napuste kompaniju.

Na slici 3. može da se vidi DEX model koji je naučen podataka korišćenjem DIDEX algoritma [4].



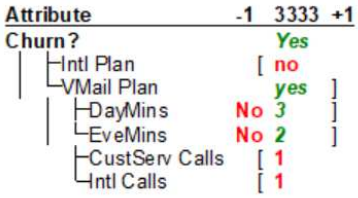
Slika 3. DEX model napravljen korišćenjem algoritma DIDEX.

DEX model izgrađen metodom DIDEX je postigao tačnost od 64,66%, odziv od 79,09%, preciznost od 26,18%. Ovaj model ima najnižu prediktivnu moć u smislu tačnosti modela, ali pokazuje najviše stope odziva. Ipak, u slučaju modela 2 i 3 cena velikog odziva je plaćena velikim brojem lažnih uzbuna.

Postavlja se pitanje koji je model najbolji za posmatrani problem. Uzimajući pretpostavku [7] da je trošak neprepoznavanja odlazećeg korisnika 5 do 7 puta veći od troška lažne uzbune, dobija se da je najbolje koristiti DEX model modelovan od strane eksperta, potom DEX model koji je generisao DIDEX algoritam, a na kraju stablo odlučivanja. Naravno, ove zaključke treba uzeti sa oprezom, sa obzirom da model mašinskog učenja nije učen precizno i da bi uključivanjem svih mogućnosti rafinisanja modela, model svakako postao još svrsishodniji.

Dodatna prednost koje DEX modeli imaju u odnosu na modele mašinskog učenja, jeste mogućnost strukturirane protivčinjenične analize zasnovane na pretpostavkama

modela, tzv. Šta-ako analize.



Slika 4.Protivčinjenična analiza DEX modela u softveru DEXi.

Sa Slike 4 se vidi protivčinjenična analiza za korisnika 3333. Može da se vidi da je odluka korisnika da napusti kompaniju **Da**. Ipak, može da se vidi šta je najmanje potrebno uraditi da bi korisnik promenio svoje mišljenje. To je tzv. korak -1, tj. najmanji korak koji je potrebno preduzeti da bi korisnik promenio odluku. U ovom slučaju se vidi da bi se odluka promenila na **Ne**, ukoliko bi korisnik smanjio broj dnevnih minuta sa visokog nivoa (3) na srednji nivo, ili ako bi visok nivo večernjih minuta (2) smanjio na niži nivo. U oba slučaja, bi u realnom svetu, pomogla verovatno neka bolja tarifa za dnevne ili večernje minute za korisnika kako bi promenio/promenila svoju odluku.

V. Praktične implikacije i diskusija

Primena DSS modela u organizacijama ima važne praktične implikacije. Transparentni modeli olakšavaju komunikaciju između analitičara, menadžera i operativnih timova, jer omogućavaju zajedničko razumevanje razloga iza donetih odluka. Ovakav pristup povećava poverenje u sistem i smanjuje otpor prema njegovoj primeni.

DSS modeli su takođe pogodni za edukativne i simulacione svrhe, jer omogućavaju korisnicima da eksperimentišu sa različitim scenarijima i razviju dublje razumevanje problema. U kombinaciji sa AI tehnikama, oni mogu predstavljati osnovu za razvoj hibridnih sistema koji objedinjuju snagu podataka i strukturu znanja.

VI. Zaključak

U ovom radu ukazano je na trajnu relevantnost sistema za podršku odlučivanju u eri veštačke inteligencije. Hijerarhijski DSS modeli, poput onih zasnovanih na DEX metodu, omogućavaju transparentno, razumljivo i analitičko donošenje odluka. Umesto da budu posmatrani kao zastarela tehnologija, DSS predstavljaju neophodan komplement savremenim AI sistemima, naročito u kontekstima gde su odgovornost, objašnjivost i mogućnost intervencije ključni zahtevi. Buduća istraživanja treba da budu usmerena ka razvoju integrisanih DSS–AI rešenja koja kombinuju automatizaciju sa ljudskom kontrolom odlučivanja. Jedan primer takvog istraživanja prikazan je u radu [5].

Zahvalnica

Ovo istraživanje je delimično podržalo Ministarstvo nauke, tehnološkog razvoja i inovacija Republike Srbije kroz institucionalno finansiranje (broj projekta 200151). Rad doprinosi, između ostalog, 9. cilju samoodrživog razvoja Ujedinjenih Nacija (Industrija, inovacije i infrastruktura).

Literatura

[1] Amin, A., Khan, C., Ali, I., & Anwar, S. (2014). Customer churn prediction in telecommunication industry: With and without counter-example. In Nature-Inspired Computation and Machine Learning: 13th Mexican International Conference on Artificial Intelligence, MICAI 2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings, Part II 13 (pp. 206-218). Springer International Publishing.

[2] Applegate, L. M., Holsapple, C. W., Kalakota, R., Radermacher, F. J., & Whinston, A. B. (1996). Electronic commerce: building blocks of new business opportunity. *Journal of organizational computing and electronic commerce*, *6*(1), 1-10.

[3] Bohanec, M. (2021). From data and models to decision support systems: Lessons and advice for the future. In *EURO Working Group on DSS: A tour of the DSS developments over the last 30 years* (pp. 191-211). Cham: Springer International Publishing.

[4] Delibašić B, Radovanović S, Bohanec M, Suknović M (2025) A comparison between DSS and ML models for churn prediction, 11th International Conference on Decision Support System Technology, ICDSST 2025, Belgrade, Serbia, May 26–29.

[5] Delibašić, B., Radovanović, S., & Vukanović, S. (2023). A Decision Support System for Internal Migration Policy-Making. https://ipsitransactions.org/journals/papers/tir/2023jul/p7.pdf

[6] Radovanović, S., Bohanec, M., & Delibašić, B. (2023). Extracting decision models for ski injury prediction from data. *International Transactions in Operational Research*, *30*(6), 3429-3454.

[7] Taskin, N. (2023). Customer Churn Prediction Model In Telecommunication Sector Using Machinelearning Technique. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1774181&dswid=-3518

Decision Support Systems in the AI Era

Boris Delibašić, Sandro Radovanović, Milija Suknović

ABSTRACT

The development of artificial intelligence and machine learning methods has significantly changed the way decisions are made in modern organizations. Although the need for automation and predictive accuracy of decision-making processes is often emphasized, understandability, transparency and support for human decision-making are increasingly gaining importance. This paper discusses the role of decision support systems (DSS) in the modern era of artificial intelligence, with a special emphasis on hierarchical, qualitative models based on expert knowledge. It shows how DSS models can satisfy key criteria for quality decision support and provide additional value compared to machine learning models. Through an illustrative example of customer churn prediction, the complementarity of DSS and AI approaches is discussed, as well as their practical implications in real organizational environments.

Primjena mašinskog učenja za opisivanje slika pomoću teksta

Aleksije Mićić
Bravo Systems d.o.o. Banja Luka
Banja Luka, Republika Srpska, BiH
aleksije.micic@oddbytes.com
0009-0004-3650-4697

prof. dr Zoran Đurić
Elektrotehnički fakultet, Univerzitet u Banjoj Luci
Banja Luka, Republika Srpska, BiH
zoran.djuric@etf.unibl.org
0000-0003-4824-2765

***Apstrakt* - Moderni informacioni sistemi za e-trgovinu integrišu sisteme za preporuku proizvoda kako bi upotpunili korisničko iskustvo. U ovom radu su analizirane primjene opisivanja slika pomoću tehnika mašinskog učenja, s ciljem generisanja boljih preporuka za e-trgovine koje se bave prodajom odjevnih predmeta. Posebna pažnja je posvećena problemima sa tradicionalnim pristupom za pretragu proizvoda, koja je obično bazirana na analizi sličnosti slika. U ovom radu evaluirani su opisi generisani pomoću osam različitih modela, pri čemu su se najbolje pokazali opisi generisani sa velikim jezičkim modelima. Evaluirano je i predloženo rješenje koje kombinuje tradicionalni pristup sa analizom sličnosti opisa na prethodno opisanim problemima, gdje se pokazala opravdanost korištenja novog pristupa.**

Ključne riječi – Sistemi za preporuku, Opisivanje slika pomoću teksta, Pretraga unutar prodavnice, Pretraga od potrošača do prodavnice, Veliki jezički modeli.

I. Uvod

Kupovina proizvoda putem informacionih sistema za e-trgovinu je u stalnom porastu. Ovakav vid kupovine omogućava generisanje preporuka za proizvode na osnovu korisničkih akcija, što može pozitivno uticati na zaradu prodavača [1], [2]. S druge strane, korisnička očekivanja, vezana za kvalitet preporuka, su sa godinama samo porasla. Za pronalazak sličnih i povezanih proizvoda, u daljnjem tekstu pretraga prodavnice, koristi se analiza sličnosti slika [3], [4]. Kao ključ pretrage uzima se slika na kojoj su prikazani odjevni predmeti od interesa, a kao rezultat pretrage dobijaju se dostupni slični i povezani odjevni predmeti iz kataloga proizvoda e-trgovine. Ovaj pristup veoma brzo generiše preporuke koje su, vizuelno slične predmetima koji su dio ključa pretrage. Takođe, memorijski zahtjevi za čuvanje vektorske reprezentacije slika svih proizvoda su relativno male. Ipak, proces prikupljanja i labelisanja podataka koji će se koristiti tokom obučavanja modela je mukotrpan i iziskuje mnogo vremena, ali i zahtjeva obučene anotatore, zbog čega može biti skup. Takođe, nije dovoljno samo obučiti model koji vrši analizu sličnosti slika odjevnih predmeta, jer dati model kao ulaz podrazumijevano dobija pojedinačne predmete. Zbog toga je potrebno i obučiti model koji detektuje pojedinačne predmete, tako što određuje njihov granični okvir (eng. *bounding box*) [3], [4], [5].

U zavisnosti od složenosti ključa pretrage, može doći do detekcije lažno pozitivnih predmeta, npr. kada je ključ majica koja prikazuje ljude koji na sebi imaju nove odjevne predmete za koje se onda takođe traže slični proizvodi. Takođe, ako je ključ pretrage složeniji, odnosno ako je na slici osoba koja nosi višeslojnu garderobu, tada je često nemoguće identifikovati sve odjevne predmete. Kada je detekcija predmeta uspješna, problem ponekad predstavlja i to što model pronalazi vizuelno sličnu odjeću, bez razumijevanja konteksta. Ako je npr. ključ pretrage košulja koju nosi neka osoba, onda će košulje iz kataloga koje su na slici savijene imati manju sličnost, čak i kada je riječ o identičnom predmetu. Ovaj problem je još više izražen, kada je kao ključ pretrage data majica sa nekim likom, jer će se kao rezultat dobiti majice koje prikazuju druge, vizuelno slične likove, ne nužno povezane sa datim likom. Ovdje bi bilo adekvatnije prikazati majice sa drugim likovima povezanim sa datim likom, ili čak majice sa natpisima vezanim za istu franšizu.

Za prevazilaženje ovih mana predloženi su pristupi koji kombinuju analizu sličnosti slika sa opisima slika generisanim pomoću tehnika mašinskog učenja. U drugom poglavlju je analiziran tradicionalni pristup za pretragu prodavnice baziran na detekciji objekata i analizi sličnosti slika, kao i problemi sa ovakvim pristupom. U trećem poglavlju je dat opisa rješenja koje se temelji na kombinovanju starog pristupa sa generisanjem opisa. Objašnjen je postupak gene-risanja opisa, te su navedene vektorske reprezentacija teksta, koje su korištene u eksperimentalnom dijelu rada. U četvrtom poglavlju su za sve predložene modele dati rezultati evalua-cije na metrikama za ocjenu kvaliteta generisanih opisa. Nakon toga, analizirani su rezultati primjene generisanja opisa za probleme navedene u drugom poglavlju. Peto poglavlje sadrži zaključak.

II. Opis problema

Prvi korak kod tradicionalnog pristupa za pretragu prodavnice je detekcija graničnih okvira na ulaznoj slici I, gdje se koristi model za detekcije objekata, poput YOLO, da identifikuje skup graničnih okvira $\mathrm{B} = \{\mathrm{b}_1, \mathrm{b}_2, \ldots, \mathrm{b}_k\}$ pri čemu svaki okvir b_i definiše koordinate (x, y, w, h) [6], [7]. Za svaki identifikovani okvir b_i računa se vektorska reprezentacija v_i primjenom konvolucione neuronske mreže (eng. *convolutional neural network* - CNN), rezultujući skupom vektora $V = \{v_1, v_2, \ldots, v_k\}$. Naredni korak je pretraga kataloga proizvoda $C = \{c_1, c_2, \ldots, c_m\}$, gdje svaki proizvod c_j ima svoju vektorsku reprezentaciju. Za svaki vektor v_i izračunava se sličnost sa proizvodima c_j korišćenjem metričke funkcije, kao što je kosinusna sličnost, čime se identifikuje skup N najsličnijih proizvoda $S_i = \{s_{i1}, s_{i2}, \ldots, s_{iN}\}$. Konačno, korisniku se prikazuje skup preporuka $S = \{S_1, S_2, \ldots, S_k\}$.

A. Obučavanje modela za detekciju objekata

Neka je dat skup podataka $D = \{(x_i, y_i)\}_{i=1}^{N}$, gdje je $x_i \in \mathbb{Z}^{H \times W \times K}$ slika, a y_i pripadajuće oznake za tu sliku. Svaka

oznaka y_i sastoji se od skupa graničnih okvira i odgovarajućih klasa za tu sliku. Ako se na slici x_i nalazi n_i objekata garderobe, tada je:

$$y_i = \{(b_{i1}, c_{i1}), (b_{i2}, c_{i2}), \dots, (b_{in_i}, c_{in_i})\}, \quad (1)$$

gdje je b_{ij} skup koordinata koje definišu granični okvir objekta j na slici i, a c_{ij} je klasa tog objekta. Cilj je obučiti model detekcije objekata f_θ, koji mapira ulaznu sliku x_i u skup predviđenih graničnih okvira i klasa:

$$\hat{y}_i = \{(\widehat{b_{ik}}, \widehat{c_{ik}})\}_{k=1}^{\widehat{n_i}}, \quad (2)$$

gdje je $\widehat{n_i}$ broj predviđenih objekata na slici x_i, $\widehat{b_{ik}}$ su predviđeni granični okviri, a $\widehat{c_{ik}}$ su predviđene klase. Model f_θ se trenira da minimizuje ukupnu funkciju gubitka $L(\theta)$, koja kvantifikuje razliku između predviđanja modela i stvarnih oznaka. Funkcija gubitka sastoji se iz dva dijela: gubitka klasifikacije i gubitka regresije i data je izrazom:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\left(L_{\text{cls}}(c_{ij}, \widehat{c_{ij}}) + \lambda L_{\text{reg}}(b_{ij}, \widehat{b_{ij}})\right), \quad (3)$$

gdje je n_i broj objekata na slici x_i, λ je hiperparametar koji balansira uticaj između gubitka klasifikacije i regresije, c_{ij} i $\widehat{c_{ij}}$ su stvarna i predviđena klase, a b_{ij} i $\widehat{b_{ij}}$ su stvarni i predviđeni granični okviri.

B. Obučavanje modela za analizu sličnosti slika

Nakon detekcije objekata i izdvajanja pojedinačnih predmeta garderobe sa slika, cilj je obučiti model koji može da mapira ove predmete u vektorski prostor gdje su više slični predmeti bliži jedni drugima, a manje slični predmeti udaljeniji jedni od drugih. Neka je dat model g_ϕ, koji mapira ulaznu sliku predmeta x_{ij} (određenu graničnim okvirom b_{ij}) u vektorsku reprezentaciju $z_{ij} \in \mathbb{R}^d$, odnosno $z_{ij} = g_\phi(x_{ij})$.

Obučavanje modela g_ϕ zasniva se na korišćenju funkcije gubitka tipa *triplet loss* zajedno sa strategijom *hard mining* [4]. Formiraju se trojke (a, p, n), gdje su:

- a sidro (eng. *anchor*), slika nekog predmeta,
- p pozitivan uzorak, slika istog predmeta,
- n negativni uzorak, slika različitog predmeta.

Nakon što se svaka od tih slika proslijedi kroz model g_ϕ, dobiju se vektorske reprezentacije $z_a = g_\phi(a)$, $z_p = g_\phi(p)$ i $z_n = g_\phi(n)$. *Triplet loss* funkcija definiše se kao:

$$L_{\text{triplet}}(\phi) = \sum_{(a,p,n)}\left[|z_a - z_p|_2^2 - |z_a - z_n|_2^2 + \alpha\right]_+, \quad (4)$$

gdje je $\alpha > 0$ margina koja definiše minimalnu razliku između pozitivnih i negativnih parova, a $[x]_+ = \max(0, x)$ se koristi u funkciji kako bi se osiguralo da samo pozitivni tj. problematični slučajevi doprinesu gubitku, dok se negativni zanemaruju.

Hard mining strategija podrazumijeva izbor najtežih pozitivnih i negativnih uzoraka tokom obučavanja. Za svako sidro a, bira se pozitivni primjer p koji je najudaljeniji od a u trenutnom vektorskom prostoru, što znači da ga model teže razlikuje od negativnih. Za svako sidro a, bira se negativni primjer n koji je najbliži a, tj. onaj koji je za model najteže razlikovati od pozitivnih. Iterativnim ažuriranjem parametara ϕ, model uči vektorske reprezentacije koje bolje odražavaju semantičku sličnost predmeta. Na ovaj način, model je sposoban kodirati predmete tako da su slični predmeti blizu u vektorskom prostoru, dok su različiti predmeti udaljeni.

Za potrebe rada dotrenirana su tri modela za detekciju objekata, bazirani na Detectron2 Mask-RCNN, YOLOv8 i YOLOv11, nad *DeepFashion2* [1] – *DF2* skupu podataka. Pored toga, obučen je jedan model za potrebe analize sličnosti slika baziran na ResNet50 CNN nad istim skupom podataka.

C. Mane tradicionalnog pristupa za pretragu proizvoda

Razlikuju se dvije vrste nedostataka kod tradicionalnog pristupa za pretragu proizvoda.

1) Nedostatci modela za detekciju objekata

Pošto je prvi korak u sistemu pretrage prodavnice, detekcija odjevnih predmeta i njihovih graničnih okvira, greške u ovoj fazi mogu nepovratno poremetiti čitav proces. Najčešće greške su [8]:

- Pogrešno određen granični okvir, kada se može desiti da se ne pronađu dovoljno slični predmeti.
- Pogrešna klasifikacija graničnog okvira, kada model dodijeli pogrešnu klasu nekom graničnom okviru. Moguće je i da je okvir pravilno određen, ali da se zbog pogrešne klase ne pronađu dovoljno slični predmeti.

Ove dvije greške se mogu uglavnom prevazići analizom pogrešnih uzoraka i proširivanjem trening skupa. Ipak postoje i greške koje se ne mogu prevazići na ovakav način:

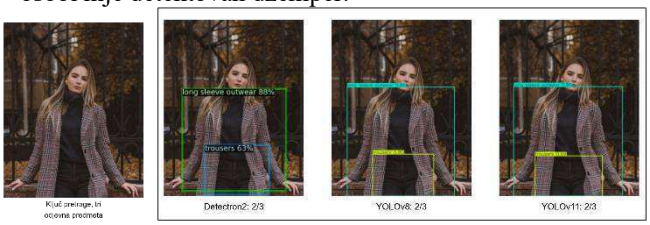
- Detekcija lažno pozitivnih uzoraka. Ako je ključ pretrage majica na kojoj su naslikani ljudi, model će detektovati granične okvire za garderobu koju ti naslikani ljudi nose, što rezultuje velikim brojem lažno pozitivnih uzoraka (slika 1).

Slika 1. *Primjer detekcije lažno pozitivnih uzoraka, kada je na majici prikazan neki broj ljudi*

- Detekcija slojevite odjeće predstavlja veliki problem za modele za detekciju. Kako su određeni slojevi odjeće samo djelimično vidljivi, model ih "ne vidi", obično se u takvim situacijama samo detektuje posljednji sloj garderobe koji je i najvećim dijelom vidljiv. Ovo je problem u situacijama kada potrošač želi u potpunosti da rekreira stil garderobe iz

[1] DF2 sadrži slike odjevnih predmeta s detaljnim oznakama, uključujući granične okvire, maske segmentacije, atribute odjeće i informacije o kategorijama.

ključa pretrage. Primjer problema sa slojevitom garderobom je prikazan na slici 2, gdje od vidljive garderobe ni na jednoj osobi nije detektovan džemper.



Slika 2. *Primjer problema sa detekcijom sve odjeće kod slojevite garderobe*

2) Nedostatci modela za analizu sličnosti slika

Ako su pravilno detektovani granični okviri sljedeći korak je da se za njih kreiraju vektorske reprezentacije i da se pronađu najsličniji odjevni predmeti iz kataloga proizvoda. Ipak i u ovom koraku postoje određeni problemi:

- Nemogućnost razumijevanja konteksta iz ključa pretrage. Pošto je model obučen da pronalazi vizuelno slične slike, on nije u stanju da pravilno odredi sličnost odjevnih predmeta koji možda nisu toliko vizuelno slični, ali pripadaju zajedničkom kontekstu. Npr. ako je ključ pretrage majica sa likom iz neke serije, a katalog proizvoda ima samo majice sa natpisima iz te serije, tada katalog vjerovatno neće odabrati te majice među najsličnijima, a važi i obrnuto. Takođe, ako je ključ pretrage majica sa nekim likom specifičnog izgleda iz određene franšize, sistem će kao najsličnije da vrati majice sa likovima sličnog izgleda, a koji su možda iz potpuno drugih franšiza, a koje uopšte ne interesuju potrošača. Na slici 3 je ilustrovan dati problem, gdje majica sa likom čarobnjaka i majica sa engleskom riječi za čarobnjaka nisu uopšte vizuelno slične pa se ne bi pojavili u rezultatima pretrage jedna za drugu.

Slika 3. *Primjer problema sa nerazumijevanjem konteksta kod pretrage majica u prodavnici*

- Nedovoljna invarijantnost na različite načine prikaza nekih odjevnih predmeta. Kako bi se postigla invarijantnost u pretrazi prodavnice, potrebno je da se u trening skupu podataka nađe dovoljan broj uzoraka koji su prikazani pod različitim orijentacijama i u različitim okolnostima. Ipak model ima tendenciju da preferira vizuelno slične odjevne predmete i potpunu invarijantnost je nemoguće postići. Ako je ključ pretrage košulja prikazana na nekoj osobi, a u katalogu proizvoda prodavača se nalazi ista ta košulja, ali na slici na kojoj je savijena, tada se ona vjerovatno neće naći među najsličnijim košuljama koje su vraćene kao rezultat pretrage, ova situacija je ilustrovana na slici 4.

Slika 4. *Primjer problema sa nerazumijevanjem konteksta i nedovoljne invarijantnosti kod pretrage košulja u prodavnici*

III. PRIJEDLOG RJEŠENJA

Za rješavanje prethodno navedenih problema moguće je primijeniti opisivanje slika pomoću tehnika mašinskog učenja, pomoću kojeg se može simulirati razumijevanje konteksta. Slično kao i kod sistema tradicionalnog sistema prvi korak uključuje detekciju graničnih okvira na ulaznoj slici I, kako bi se identifikovao skup $B = \{b_1, b_2, \dots, b_k\}$. Paralelno s tim, koristi se model za generisanje opisa slike koji analizira ulaznu sliku i generiše skup tekstualnih opisa $O = \{o_1, o_2, \dots, o_l\}$, od kojih svako o_i opisuje jedan od detektovanih odjevnih predmeta. Sljedeći korak uključuje kombinovanje i ispravljanje rezultata, gdje se rješavaju prethodno opisane mane. Problem generisanja lažno pozitivnih uzoraka rješava se tako što se, u slučaju kada model detektuje više okvira $|B| > 1$, a generiše samo jedan opis $|O| = 1$, zadržava samo najveći okvir $b_{\max}$, definisan površinom $w \cdot h$. Problem slojevite odjeće rješava se tako što se, u slučaju da model detektuje manje okvira $|B|$ nego što ima generisanih opisa $|O| > |B|$, opisi i okviri uparuju prema kategorijama (npr. "jakna", "majica"), dok se za opise koji ostanu neupareni generišu samo tekstualni vektori. Za svaki par (b_i, o_i), vektor slike v_i^s se računa pomoću CNN, dok se vektor opisa v_i^t generiše korišćenjem nekog NLP modela. Rezultujući vektor v_i se dobija kao linearna kombinacija: $v_i = w_1 v_i^s + w_2 v_i^t$, gdje su w_1 i w_2 hiperparametri koji se određuju eksperimentalno. U slučaju da vektori v_i^s i v_i^t nisu iste dimenzije, vektor veće dimenzije se redukuje na dimenziju manjeg vektora. Nakon što su svi parovi (b_i, o_i) obrađeni, rezultujući vektori se upoređuju sa vektorskim reprezentacijama proizvoda iz kataloga $C = \{c_1, c_2, \dots, c_m\}$. Slični proizvodi se pronalaze na isti način kao i ranije. Problem razumijevanja konteksta se rješava tako što generisani opisi sadrže informacije koje su od značaja za neki odjevni predmet, opisujući ujedno i neke kontekstualne informacije vezane za taj predmet.

A. Generisanje opisa slika

Opisivanje slika pomoću teksta (eng. *image captioning* - IC) je zadatak generisanja deskriptivnih tekstualnih opisa za slike koristeći ML tehnike. Ovaj zadatak spada u multimodalne probleme jer integriše vizuelni i jezički domen. Cilj ovog postupka je razviti model koji, datu sliku $I \in \mathcal{I}$, mapira na tekstualnu sekvencu $S = \{w_1, w_2, \dots, w_T\}$, gdje je svaka riječ w_t element predefinisanog vokabulara V. Formalno, cilj je modelovati uslovnu vjerovatnoću $P(S|I)$,

odnosno pronaći sekvencu riječi S koja maksimizuje ovu vjerovatnoću. Stoga, cilj je optimizacija sljedećeg izraza:

$$S^* = \arg\max_S P(S|I) \quad (5)$$

gdje S^* predstavlja optimalnu sekvencu riječi koja najbolje opisuje sliku I. Da bi se riješio ovaj zadatak, IC sistem se tipično sastoji od dvije glavne komponente: enkodera i dekodera. Enkoder je komponenta koja obrađuje ulaznu sliku I i ekstrahuje njene značajne vizuelne karakteristike $F \in \mathbb{R}^d$, gdje d predstavlja dimenzionalnost vektora karakteristika. Tipično, enkoder je implementiran korišćenjem vizuelnih transformatora, koji su prethodno trenirani na velikim skupovima podataka za zadatke klasifikacije slika. Formalno:

$$f_{\text{enc}}: \mathcal{I} \to \mathbb{R}^d, \quad F = f_{\text{enc}}(I), \quad (6)$$

gdje F predstavlja vektorsku reprezentaciju slike I. Dekoder, s druge strane, ima zadatak da generiše tekstualni opis $S = \{w_1, w_2, \dots, w_T\}$ na osnovu vektorske reprezentacije F. Dekoder je obično implementiran korišćenjem transformatora. Dekoder modeluje uslovnu vjerovatnoću $P(S|I)$ kao produkt uslovnih vjerovatnoća svake riječi u sekvenci:

$$P(S|I) = \prod_{t=1}^{T} P(w_t|w_1, w_2, \dots, w_{t-1}, F), \quad (7)$$

gdje w_t predstavlja t-tu riječ u sekvenci. Da bi dekoder pravilno obradio izlaz enkodera, često je neophodno uskladiti dimenzije vektorskih reprezentacija enkodera sa očekivanim ulaznim dimenzijama dekodera. Zbog toga se u praksi, posebno kod multimodalnih LLM-ova, primjenjuje linearna projekcija na izlaz enkodera. Na ovaj način se sprječava problem da dekoder prima vektor neodgovarajuće dimenzije, što bi onemogućilo normalnu propagaciju signala tokom treninga ili generisanja opisa. Dekoder se formalno može predstaviti kao funkcija:

$$f_{\text{xe}^{\text{c}}}: (\mathbb{R}^{\text{d}}, w^1, \dots, w_{\text{t}}{}^{-1}) \to P(w_{\text{t}}|w^1, \dots, w_{\text{t}}{}^{-1}, F). \quad (9)$$

Kombinovanjem enkodera i dekodera, cjelokupan model može se posmatrati kao funkcija:

$$f_{\text{IC}}: \mathcal{I} \to \mathcal{S}, \quad S = f_{\text{IC}}(I), \quad (10)$$

gdje $\mathcal{S}$ predstavlja prostor svih mogućih tekstualnih sekvenci.

Za potrebe rada analizirani su razni IC modeli, ali su se najbolje pokazali multimodalni LLM jer su u stanju da identifikuju i detaljno opišu više objekata nezavisno jedan od drugih, dok se ostali modeli mogu koristiti za jednostavne ključeve pretrage gdje je na slici prikazan jedan odjevni predmet.

B. *Analizirani modeli za generisanje opisa slika*

Modeli za IC koji su prethodili multimodalnim LLM nisu bili direktno namijenjeni za detaljno opisivanje svih predmeta koji su prikazani na slici. Umjesto toga cilj je bio da se uopšteno opiše slika i sve što se dešava na slici. Što većinu ovakvih modele čini nepraktičnim za rješavanje problema poput pretrage prodavnice.

Tabela 1. Analizirani modeli za generisanje opisa slika

Model	Primjena u radu	Opis
GenerativeImage2Text	Razumijevanje konteksta	Dotreniran za opisivanje majica i košulja
Florence-2		
BLIP-2		
gpt-4o-mini-2024-07-18	Svi navedeni problemi	Pretreniran komercijalni model
gpt-4o-2024-08-06		
claude-3-5-sonnet-		
claude-3-haiku-20240307		
llama3.2-vision-11b		Pretreniran javno

Ipak određeni modeli poput Florence-2 su u stanju da generišu relativno detaljne opise, koji obuhvataju većinu predmeta na slici, ali ni ti modeli ne generišu struktuiran opis, gdje se jasno izdvajaju pojedinačni predmeti [9]. Zbog toga su u radu stariji modeli isključivo analizirani, zajedno sa multimodalnim LLM, za problem razumijevanja konteksta gdje je kao ključ pretrage data majica, odnosno gdje je majica glavni predmet za koji se opis generiše. Primjenom multimodalnih LLM moguće je opisati sve predmete na slici, jer se kao odgovor može generisati struktuirani izlaz, npr. u JSON formatu, gdje je moguće na lak način izdvojiti opise za pojedinačne odjevne predmete. U tabeli 1. su dati svi analizirani modeli.

C. *Analizirane vektorske reprezentacije teksta*

Pomoću obrade prirodnog jezika (eng. *natural language processing* - NLP) i ML tehnika se rješavaju mnogi zadaci kao što je analiza sličnosti teksta, koju je potrebno sprovesti da bi se izvršila pretraga kataloga proizvoda na osnovu generisanih opisa proizvoda.

Tabela 2. Analizirani modeli za analizu sličnosti teksta

Model	Tip
Word2Vec	Ugrađivanje riječi
FastText	
GloVe	
ELMo	Kontekstualni jezički model
BERT	
text-embedding-3-small	

Tekst je vremenski nezavisan sadržaj sa linearnom strukturom, a u računarima je kodovan u binarnom obliku. Kako bi se tekstualni sadržaj tumačio pomoću NLP metoda, neophodno je kreirati odgovarajuću vektorsku reprezentaciju. Što je reprezentacija bogatija informacijama, to se efikasnije može primijeniti u različitim NLP zadacima. Reprezentacije variraju od jednostavnih, koje su bazirane na frekvenciji riječi, do onih složenijih, baziranih na ugrađivanju riječi ili na kontekstualnim jezičkim modelima koji trenutno predstavljaju najnapredniji pristup. U tabeli 2 su navedene analizirane vektorske reprezentacije teksta. Pošto su vektori predloženih tekstualnih reprezentacija veće dimenzije nego vektori korištene vizuelne reprezentacije, vršena je redukcija dimenzionalnosti, prije računanja linearne kombinacije.

IV. EKSPERIMENTALNI DIO

Za potrebe obučavanja modela za detekciju objekata i modela za analizu sličnosti slika korišten je DF2 skup podataka.

A. *Metrike za valuaciju kvaliteta generisanih opisa slika*

Za evaluaciju generisanih opisa korištene su metrike: BLEU, METEOR, ROUGE-L, CIDEr, SPICE, SPIDER, najčešće korištene metrike u radovima koji se bave problemom generisanja opisa pomoću tehnika mašinskog

učenja. Za potrebe dotreniravanja transformatorskih modela iskorišten je skup podataka sa opisima odjevnih predmeta formiran iz *Eureka-Attr*[2] kataloga proizvoda. Za potrebe evaluacije generisanih opisa, detekcije slojevite odjeće, kao i za validiranje kvaliteta preporuka korištena su tri druga skupa podataka (disjunktni u odnosu na prethodne), a svi oni predstavljaju podskup *Eureka-PC*[3] kataloga proizvoda. U tabeli 3. su dati rezultati evaluacije različitih modela korištenjem opisanih metrika, na skupu od 10k slika od čega je 5k iz kategorije *t-shirts* i 5k iz kategorije *shirts*, nasumično odabranih iz *Eureka*-PC skupa. Najbolje rezultate ostvaruju OpenAI ChatGPT modeli, ali su dobri rezultati ostvareni i od strane Antrophic Claude modela i Llama modela.

Tabela 3. Rezultati evaluacije analiziranih modela na Eureka-PC skupu podataka

Model	BLEU	METEOR	ROUGE-L	CIDEr	SPICE	SPIDEr
GIT	18,45	50,78	54,67	1,876	0,278	1,077
Florence2	20,12	53,34	57,23	2,101	0,301	1,201
BLIP-2	23,01	56,89	60,34	2,678	0,345	1,511
gpt-4o-mini	34,57	78,84	79,19	4,997	0,578	2,788
gpt-4o	36,91	81,23	81,34	5,123	0,593	2,858
claude-3-5-s	33,12	76,45	77,56	4,678	0,541	2,609
claude-3-h	31,45	73,67	74,89	4,321	0,512	2,416
llama3.2-v-11b	29,78	70,89	72,34	4,012	0,478	2,245

B. Analiza rezultata za predloženo rješenje

Rezultati evaluacije tri predložena modela za detekciju graničnih okvira i objekata, kao i pet predloženih LLM modela su dati u tabeli 4. Korišten je podskup proizvoda iz *Eureka-PC* kataloga, gdje je analizirano po pet hiljada proizvoda iz kategorije *t-shirts* i pet hiljada iz kategorije *shirts*, na svakoj slici je prikazan tačno jedan odjevni predmet, ali neki od odjevnih predmeta imaju na sebi naslikane ljudske likove.

Tabela 4. Rezultati evaluacije predloženih modela na problemu detekcije lažno pozitivnih uzoraka

Model	Prosječan broj		Tačnost	
	t-shirts	*shirts*	*t-shirts*	*shirts*
gpt-4o	1,003	1,002	99,60%	99,98%
gpt-4o-mini	1,005	1,0076	99,58%	99,46%
claude-3.5-s	1,003	1,0004	99,60%	99,60%
claude-3-h	1,007	1,0083	99,30%	99,30%
llama3.2-v-11b	1,007	1,0013	99,30%	99,30%
YOLOv8	1,275	1,1712	77,90%	95,60%
YOLOv11	1,272	1,164	78,34%	95,60%
Mask R-CNN	1,28	1,1698	77,50%	77,50%

Najbolji rezultati su ostvareni korištenjem GPT-4o modela, ali i svi ostali predloženi modeli su ostvarili visoku tačnost na datom testnom skupu.

Rezultati evaluacije predloženih modela na problemu detekcije slojevite odjeće su dati u tabeli 5. Korišten je podskup iz *Eureka-PC* kataloga proizvoda, od hiljadu proizvoda. Slike proizvoda prikazuju ljude koji nose slojevitu garderobu, gdje je ručno određen tačan broj predmeta na svakoj slici iz trinaest kategorija koje su obrađene u DF2 skupu podataka, dato ograničenje je uvedeno zbog poređenja sa modelima za detekciju koji su obučeni na DF2 skupu i samo mogu da detektuju navedene kategorije odjeće, dok u praksi LLM modeli mogu da detektuju i predmete iz drugih kategorija. Pravilna detekcija je slučaj kada neki model detektuje odgovarajuće odjevne predmete, ali i pravilne kategorije za date predmete.

Tabela 5. Rezultati evaluacije predloženih modela na problemu detekcije slojevite odjeće

Model	Tačnost
gpt-4o	70,80%
gpt-4o-mini	71,00%
claude-3.5-s	71,00%
claude-3-h	68,70%
llama3.2-v-11b	67,10%
YOLOv8	35,40%
YOLOv11	36,40%
Mask R-CNN	41,40%

Najbolji rezultati su ostvareni pomoću GPT-4o modela, dok modeli za detekciju graničnih okvira znatno zaostaju za LLM modelima.

Za potrebe analize razumijevanja konteksta, kreiran je skup od 100 ključeva pretrage, pri čemu polovinu čine majice s printom, odnosno sa dodatnim kontekstom, a drugu polovinu obične majice bez printa. Skup je pažljivo izbalansiran kako bi rezultati analize bili validni. Naime, da je analiza vršena isključivo na ključevima pretrage s print majicama, dobiveni rezultati bi bili preprilagođeni specifičnom problemu. S druge strane, za eksperiment je korišten katalog proizvoda od 10.000 majica, formiran na osnovu *Eureka-PC* kataloga proizvoda, gdje je za svaki ključ pretrage određeno 20 najsličnijih majica. Poređeni su rezultati pretrage proizvoda u dva scenarija:

1. Korišćenje samo vizuelne reprezentacije slike.
2. Korišćenje linearne kombinacije vizuelne reprezentacije v_1 i tekstualne reprezentacije generisanog opisa v_2.

Eksperiment je uključivao poređenje dva pristupa: YOLOv11+ResNet50 kao najboljeg predstavnika starog pristupa i ChatGPT-4o modela kao najboljeg predstavnika LLM modela. Za svaki ključ pretrage pronađeno je dvadeset najsličnijih predmeta pomoću oba pristupa, a rezultati su upoređeni kako bi se utvrdilo koji model vraća više relevantnih predmeta. Kako bi se postigla optimalna linearna kombinacija vektora v_1 i v_2, težine w_1 i w_2 su određene tako da zadovoljavaju uslov $w_1 + w_2 = 1$. Vrijednosti težina su ispitivane unutar intervala $[0, 1]$ s korakom $\text{EPS} = 0{,}01$. Na taj način, za svaku kombinaciju težina testirani su rezultati kako bi se pronašao optimalan omjer. Eksperiment je pokazao da se bolji rezultati postižu korištenjem linearne kombinacije obje vektorske reprezentacije (v_1 i v_2) u odnosu na korištenje samo jedne od njih. U tabeli 6. su prikazani rezultati poređenja novog pristupa sa starim za različite modele vektorske reprezentacije opisa.

Tabela 6. Rezultati evaluacije za problem razumijevanje konteksta nad majicama

Model	% Bolji stari	% Isti	% Bolji novi	Optimalan omjer $w1 : w2$
t-e-3-s	15	10	75	14 : 86
Word2Vec	12	26	62	52 : 48
FastText	9	27	64	41 : 59
GloVe	11	26	63	38 : 62
ELMo	8	27	65	43 : 57
BERT	14	16	70	34 : 66

Najbolje performanse postignute su s modelom *text-embedding-3-small*, pri čemu je optimalan omjer težina bio

[2] U pitanju je skup podataka koji sadrži 120.000 slika kategorisanih kao majice i 12.000 slika iz kategorisanih kao košulje.

[3] U pitanju je skup podataka koji sadrži po 100.000 kategorisanih kao majice i košulje i 10.000 slika na kojima je prikazana slojevita odjeća.

$w_1 : w_2 = 14 : 86$.

Konačno analizirana je primjena na problemu razumijevanja konteksta i nedovoljne invarijantnosti kad su ključevi pretrage košulje. Za potrebe analize, kreiran je skup od 50 ključeva pretrage, od čega polovinu čine savijene košulje, a drugu polovinu košulje na ljudima. Cilj je bio ispitati performanse različitih pristupa u pronalaženju sličnih predmeta unutar kataloga koji sadrži 10.000 predmeta, također formiranog na osnovu *Eureka-PC* kataloga proizvoda. Važno je napomenuti da ovaj katalog nije isti kao u prethodnom primjeru za majice. Za svaki ključ pretrage identifikovano je deset najsličnijih predmeta. Analiza je rađena isto kao i u prethodnom slučaju.

Testirani su sljedeći modeli: YOLOv11+ResNet50 i ChatGPT-4o. Jedan od izazova u analizi je bio taj što detekcija graničnih okvira kod starog pristupa, nije uvijek dobro funkcionisala na savijenim košuljama. Uzrok ovog problema vjerovatno leži u nedostatku primjera sa savijenim košuljama u trening skupu podataka, zbog čega detekcija nije uvijek moguća, ili se loše odredi granični okvir. Iz tog razloga je, prilikom analize u slučaju da nije došlo do detekcije graničnog okvira, korištena cijela ulazna slika bez dodatnog isjecanja. Na ovaj način osigurano je da se ne donose zaključci isključivo na osnovu problema sa detekcijom graničnih okvira, već da se analizira i širi kontekst performansi modela. Kao i u prethodnom primjeru, korišćenje linearne kombinacije vizuelne reprezentacije v_1 i reprezentacije generisanog teksta v_2 dalo je bolje rezultate u poređenju sa korišćenjem samo vizuelne reprezentacije. U tabeli 7. su prikazani rezultati poređenja novog pristupa sa starim za različite modele vektorske reprezentacije opisa.

Tabela 7. Rezultati evaluacije za problem razumijevanje konteksta i nedovoljne invarijantnosti nad košuljama

Model	% **Bolji stari**	% **Isti**	% **Bolji novi**	**Optimalan omjer $w1 : w2$**
t-e-3-s	20	10	70	73 : 27
Word2Vec	16	26	58	86 : 14
FastText	12	27	61	82 : 18
GloVe	14	26	60	83 : 17
ELMo	14	24	62	84 : 16
BERT	15	22	63	79 : 21

Najbolji rezultati postignuti su s modelom *text-embedding-3-small*, pri čemu je optimalan omjer bio $w_1 : w_2 = 73 : 27$. Međutim, analiza je pokazala zanimljiv obrazac:

- Kod savijenih košulja, kombinacija vizuelne i tekstualne reprezentacije značajno je nadmašila pristupe bazirane samo na vizuelnim modelima. Ovo je očekivano jer tekstualni opisi omogućavaju bolji kontekstualni uvid u karakteristike savijenih košulja.
- Kod košulja na ljudima, pristup (YOLO+ResNet50) pokazao se boljim u određenim slučajevima. Razlog za to je što je teško riječima precizno opisati šaru na košulji, dok vizuelni modeli poput ResNet50 bolje prepoznaju te suptilne vizuelne detalje. Zbog toga je omjer takav da mnogo više ide u korist vizuelne reprezentacije.

V. ZAKLJUČAK

U ovom radu opisan je prijedlog rješenja za primjenu opisa generisanih pomoću tehnika mašinskog učenja za problem pretrage proizvoda. Ovim radom pokazano je da je za uspješniju pretragu bolje kombinovati model za analizu sličnosti slika zajedno sa modelom za opisivanje slika, odnosno linearno kombinovanje vektorske reprezentacije dobijene na osnovu vizuelnih karakteristika sa vektorskom reprezentacijom generisanog opisa. Posebna pažnja je posvećena problemima sa klasičnim pristupom za pretragu proizvoda koji se oslanja samo na analizu sličnost slika, gdje problemi potiču ili od modela za detekciju koji se nalazi na početku takvog sistema, ili od modela za analizu sličnosti koji se nalazi na kraju tog sistema.

LITERATURA

[1] D.-C. Pahonțu and E.-Ș. Enache, "An Overview of AI-driven Recommendation Systems: Enhancing Personalization & User Experience (Qualitative Study)," STAR, vol. 3, br. 2, novembar 2024.

[2] "THE ROLE OF AI IN ENHANCING PERSONALIZATION IN ECOMMERCE: A STUDY ON CUSTOMER ENGAGEMENT AND SATISFACTION", ASPAREV, vol. 17, br. 2, str. 160–177, novembar 2024.

[3] Y. Ge, R. Zhang, L. Wu, X. Wang, X. Tang, and P. Luo, „DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images", 23. januar 2019., arXiv: arXiv:1901.07973.

[4] D. Morelli, M. Cornia, and R. Cucchiara, "FashionSearch++: Improving consumer-to-shop clothes retrieval with hard negatives," CEUR Workshop Proceedings, vol. 2947, CEUR-WS, 2021.

[5] P. Alirezazadeh, F. Dornaika i A. Moujahid, „Deep Learning with Discriminative Margin Loss for Cross-Domain Consumer-to-Shop Clothes Retrieval", Sensors, vol. 22, br. 7, str. 2660, mart 2022, doi: 10.3390/s22072660.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, „You Only Look Once: Unified, Real-Time Object Detection", 09. maj 2016., arXiv: arXiv:1506.02640. doi: 10.48550/arXiv.1506.02640.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, „Mask R-CNN", 24. januar 2018., arXiv: arXiv:1703.06870.

[8] D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing Error in Object Detectors," Computer Vision – ECCV 2012, vol. 7574, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, str. 340–353, doi: 10.1007/978-3-642-33712-3_25.

[9] B. Xiao et al., "Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks," arXiv preprint, arXiv:2311.06242, novembar 2023, doi: 10.48550/arXiv.2311.06242.

Using Machine Learning for Image Captioning

Aleksije Mićić and Prof. Zoran Đurić, PhD

ABSTRACT

Modern information systems for e-commerce integrate product recommendation systems to enhance the user experience. This paper analyzes the application of image captioning using machine learning to generate better recommendations for e-commerce platforms that sell clothing. Special attention is given to issues with traditional product retrieval methods based on image similarity analysis. Image captions generated by eight different models were evaluated, with large language models performing the best. Following this, the proposed solution, which combines the traditional retrieval approach with similarity analysis of captions, was evaluated on the previously described problems, demonstrating the validity of using the new approach. Finally, potential directions for further research are provided.

Predviđanje uspešnosti memorizacije slike pomoću modela mašinskog učenja

Valerijan Matvejev
Elektrotehnički fakultet Univerziteta u Beogradu
Beograd, Srbija
mv245055p@student.etf.rs
ORCID: 0009-0000-7125-2078

Dražen Drašković
Elektrotehnički fakultet Univerziteta u Beogradu
Beograd, Srbija
drazen.draskovic@etf.bg.ac.rs
ORCID: 0000-0003-2564-4526

***Apstrakt* - Vizuelna pažnja, odnosno automatska selekcija najvažnijih informacija unutar vizuelnog stimulusa, tema je unutar računarske vizije koja zahvata veliko interesovanje naučnika. U ovom istraživanju, fokus je na proučavanju komplikovanog odnosa između implicitne, skrivene (engl. *Covert*) i eksplicitne, otvorene (engl. *Overt*) pažnje i ljudske memorije pomoću eksperimenta memorabilnosti na statičnim slikama. 30 slika za eksperiment je odabrano iz *FIGRIM* [1] skupa slika. Sakupljeni podaci gledanja slika (engl. *Eye Tracking Data*) korišćeni su najpre za izračunavanje fiksacionih i vizuelnih mapa upečatljivosti, a potom i *IOVC* metrike (engl. *Inter Observer Visual Congruency*), koje su zajedno sa procentom uspešne memorizacije slike (memorabilnog skora) bili ključni parametri istraživanja. Značajna korelacija između IOVC skora i skora memorabilnosti ukazala je na mogućnost automatizacije predikcije uspešnosti gledanja slike samo na osnovu njene fiksacione mape. Inspirisano ovim uvidom, najpre je treniran model mašinskog učenja na sakupljenom skupu podataka gledanja, koji je na osnovu fiksacione mape trebalo da zaključi o kojoj od 30 datih slika je reč. Cilj je bio da se istrenira klasifikator koji uspešno može da razlikuje slike samo na osnovu podataka o njihovom gledanju. *MLP* arhitektura se pokazala najpreciznijom za ovaj problem višestruke klasifikacije, i taj model je dalje korišćen za nalaženje praga zaključivanja (engl. *Inference Thresholding*) uspešnosti memorizacije za svaku od 30 slika. Istrenirani model je imao tačnost blizu 90% u određivanju da li proizvoljna fiksaciona mapa vodi uspešnoj memorizaciji date slike.**

Ključne reči – Vizuelna pažnja, Memorabilnost, Klasifikacija, Predikcija

I. UVOD

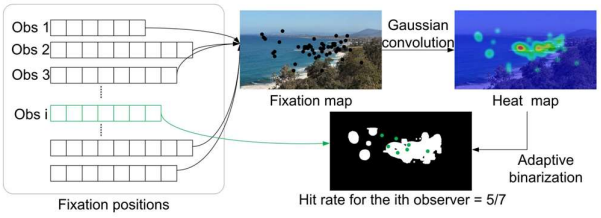
Unutar diskursa kognitivnih nauka, dobro je poznata podela ljudske vizuelne pažnje na implicitnu (engl. *Covert*), i eksplicitnu (engl. *Overt*) pažnju. Implicitna pažnja se odnosi na ljudsku mentalnu fokusiranost na svoje vidno polje, i pokrenuta je najčešće dubokim psihološkim efektima. S druge strane, eksplicitna pažnja podrazumeva samo fizičku usmerenost zenica oka [2]. Interes ovog istraživanja bio je na povezivanju ove dve vrste pažnje najpre teorijski pomoću znanja kognitivne psihologije, a potom i pomoću eksperimenta memorabilnosti statičnih slika, na osnovu kojeg je razvijen model mašinskog učenja za predikciju (oba tipa) ljudske vizuelne pažnje. Uticaj ovakvog jednog istraživanja može biti ogroman, kako na detekciju poremećaja pažnje i pamćenja, praćenju uspešnosti studentskog učenja, tako i na generalno opšte predviđanje ljudske psihološke spremnosti. Budućim istraživanjima na ovu i slične interdisciplinarne teme vezane za ljudsku pažnju ostaje da nakon kombinovanja domenskih znanja iz više oblasti sa mašinskim učenjem, uspešno i implementiraju svoja rešenja u realnom svetu zarad opšteljudske koristi.

II. GLAVNI KONCEPTI

Vizuelna pažnja se nalazi na raskrsnici nekoliko nauka: psihologije, neurofiziologije, komputacionih neuronauka i računarstva. Istraživačima u oblasti računarstva najčešće je cilj da naprave komputacione modele i algoritme koji predviđaju lokacije unutar vizuelnog stimulusa na koje ljudi najviše obraćaju pažnju. Samo neke od modernih primena vizuelne pažnje obuhvataju: autonomnu vožnju, upečatljivost veb-sajtova, interakciju čovek-računar pomoću pogleda, medicinske dijagnoze, ljudske emocije [3].

Osnovu svih algoritama vizuelne pažnje čine fiksacione mape i vizuelne mape upečatljivosti (engl. *Visual Saliency Maps*). Za generisanje ovih mapa, koriste se uređaji za praćenje oka (engl. *Eye Trackers*) tokom posmatranja stimulusa. Važna metrika memorabilnosti vizuelnog sadržaja izražena je pomoću stope pogotka (engl. *Hit Rate*), izražene kao odnos broja uspešnih prepoznavanja i ukupnog broja prikazivanja stimulusa.

Inicijalna ideja za povezivanje memorabilnosti, izražene preko stope pogotka, i fiksacionih mapa, dobijena je izračunavanjem visoke korelacije između stope pogotka slike i *IOVC* metrike, što je otvorilo mogućnost predviđanja memorabilnosti slike na osnovu fiksacionih mapa (detaljnije u poglavlju V ovog rada). *IOVC* metrika reprezentuje sličnost gledanja stimulusa, tj. koliko se jedno posmatranje slike razlikuje od svih ostalih posmatranja, i izračunava se kao na Slika 1. Shema izračunavanja IOVC metrike .



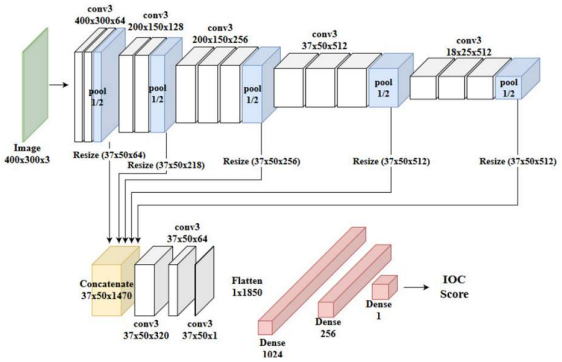
Slika 1. Shema izračunavanja IOVC metrike [4]

Cilj ovog istraživanja nije bio samo predviđanje eksplicitne ljudske pažnje, već duboke, implicitne pažnje i čak ljudske memorije. Za dokazivanje ove hipoteze, korišćena su znanja kognitivnih nauka Majkla Posnera [5], za povezivanje eksplicitne i implicitne pažnje, kao i Holingvorta [6], za povezivanje ljudske memorije i implicitne pažnje. Zaključeno je da postoji teorijska veza između sva tri fenomena, te se predviđanje implicitne pažnje i ljudske memorije može uraditi implicitno na osnovu predikcije eksplicitne pažnje. Rečju, razvijeni model uspešno predviđa celokupnu fokusiranost korisnika, ne samo lokacije gledanja.

III. Povezana istraživanja

Temeljni članak ovog istraživanja jeste [1], u kome je sakupljen *FIGRIM* skup podataka, korišćen upravo i u centralnom eksperimentu. Jedan od zaključaka ovog članka jeste da je memorabilnost slika instrinzična, ali i ekstrinzična osobina slike. U poglavlju 7 navedenog članka, opisan je trening *RUSBoost* klasifikatora za predikciju uspešnosti memorizacije slike na osnovu podataka praćenja oka, koji je kao krajnju tačnost imao 66% balansirane tačnosti, što je prilično bolje od nasumičnosti (50%).

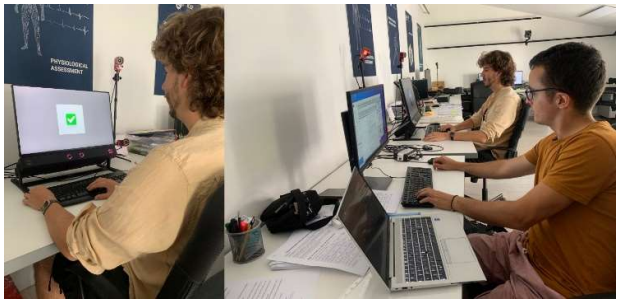
Zatim su istraženi članci koji obrađuju temu automatske predikcije memorabilnosti slike i njenog *IOVC* skora na osnovu različitih karakteristika same slike. [7] je prvi u nizu članaka, u kome su najpre korelisane jednostavne osobine slike (broj objekata, površina), a zatim i semantičke, komplikovanije osobine, sa memorabilnošću, ali je visoka korelacija izostala ($\rho = 0.45$). U radu [8], istraživači su otišli i korak dalje, jer su u priču ubacili i vizuelnu pažnju, te su rezultati bili za nijansu bolji ($\rho = 0.48$). Rad [4] je članak u kojem je prvi put i definisan pojam *IOVC* metrike. Novija istraživanja [9] i [10] su izračunala najviše korelacione koeficijente između predviđenog *IOVC* skora slike, dobijenog pomoću dubokih fičera slike, i njenog stvarnog *IOVC* skora. Specifično, u radu [10] autori su dobili $\rho = 0.611$, što predstavlja srednju ka jakoj korelaciji, primenom pretrenirane *VGG19* [11] duboke mreže za ekstrakciju fičera na skupove podataka praćenja oka. Potom je izračunat i predviđeni *IOVC* skor, regresijom, i upoređen je sa stvarnim skorom slike. Shematski prikaz ovog modela dat je na Slika 2. Arhitektura korišćena za ekstrakciju fičera i IOVC regresiju



Slika 2. Arhitektura korišćena za ekstrakciju fičera i IOVC regresiju [10]

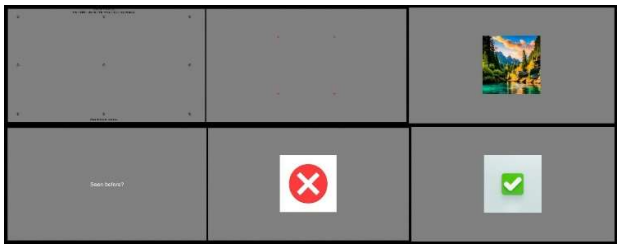
IV. Metodologija

Za svrhe ovog istraživanja, organizovan je eksperiment testiranja memorije učesnika, u prostorijama laboratorije Inovacionog centra Elektrotehničkog fakulteta u Beogradu, u kojem je svakom od 48 učesnika prikazano 120 različitih slika u jednoj rundi. Svaki učesnik je radio dve ili tri runde, a svaka od slika je prikazana dve sekunde na ekranu, praćena sa tekstualnim pitanjem da li je učesnik već video datu sliku. Nakon očekivanog odgovora preko tastature, eksperiment prikazuje na ekranu zeleni ili crveni indikator tačnog, odnosno netačnog odgovora respektivno. Pre prikazivanja sledeće slike, četiri crvena markera su prikazana na ekranu, kako bi se pažnja učesnika održala samo na regionu slike. Fizički izgled eksperimenta, sa učesnikom i voditeljem eksperimenta dat je na Slika 3. Fizička postavka eksperimenta memorabilnosti.



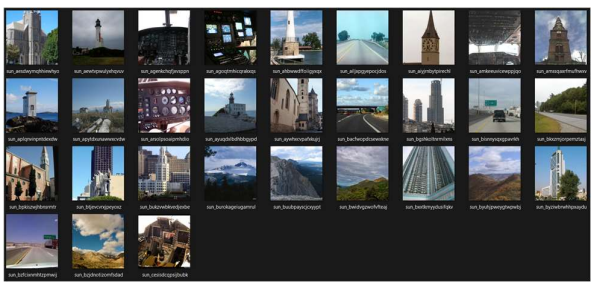
Slika 3. Fizička postavka eksperimenta memorabilnosti

Test memorije dizajniran je u softveru *Psychopy*, psihološkom softveru sa otvorenim pristupom baziranom na programskom jeziku *Python*. Korišćen je *Tobii* uređaj za praćenje oka, a izgled različitih ekrana testa vidljiv je na Slika 4. Ekrani testa (sleva nadesno, odozgo nadole) – kalibracija, fiksacioni markeri, stimulus (slika), tekstualni upit, netačan odgovor, tačan odgovor.



Slika 4. Ekrani testa (sleva nadesno, odozgo nadole) – kalibracija, fiksacioni markeri, stimulus (slika), tekstualni upit, netačan odgovor, tačan odgovor

Slike su prikazivane u rezoluciji 700x700 piksela, baš kao u originalnom članku [1], a za potrebe ovog eksperimenta izdvojeno je 30 slika iz *FIGRIM* skupa podataka i to na sledeći način. Najpre je izdvojeno 6 najlošije rangiranih kategorija slika (po memorabilnom skoru) iz [1], i to su kategorije kokpit, put, planina, svetionik, kula, soliter. Iz svake od ovih kategorija uzeto je po 5 najlošije rangiranih slika kao osnova za eksperiment. Pored ovih 30 slika, dodato je i proizvoljnih 100 slika takođe iz *FIGRIM* skupa podataka, koje su ili služile samo da popune ostatak nedostajućih slika u svakom testu ili kao lažne provere da li učesnik idalje prati test. Uz njih, dodato je i 90 slika „pojačavača" sa Interneta iz tri kategorije: poznate ličnosti, umetnička dela, popularni brendovi. Motivacija za odabir 6 najlošije rangiranih kategorija slika leži u ideji da bi postavljanje slika „pojačavača" ispred nisko memorabilnih slika trebalo da poveća njihov memorabilni skor, što je i dokazano. Na Slika 5. 30 najlošije rangiranih slika iz 6 najlošije rangiranih kategorija FIGRIM skupa podataka. prikazano je 30 odabranih ciljnih slika.

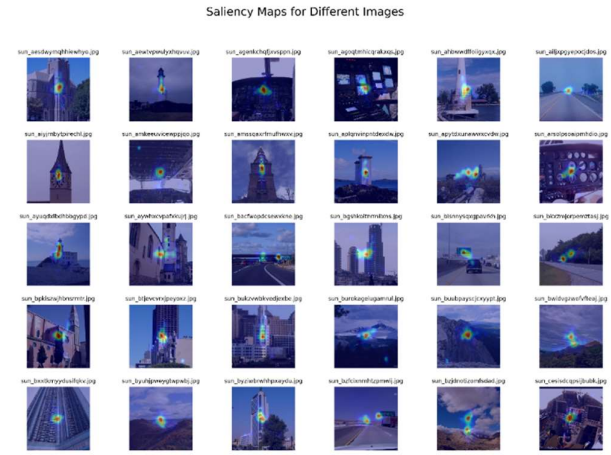


Slika 5. 30 najlošije rangiranih slika iz 6 najlošije rangiranih kategorija FIGRIM skupa podataka

Redosled prikazivanja slika izračunat je na sledeći način. U prvih 30 slika u okviru jednog testa prikazano je 10 proizvoljno izabranih ciljnih slika, i upravo pri kraju testa, odnosno u poslednjih 30 slika u testu, prikazana su i njihova ponavljanja u nasumičnom redosledu. Ispred svakog prikazivanja ciljne slike, dodata je proizvoljna slika pojačavač, iz jedne od tri kategorije proizvoljno odabrane za dati test. Između ovih slika, prikazivane su razne slike, koje su služile da učesnicima održe pažnju. Jedan eksperiment je trajao oko 9, odnosno 18 minuta ukoliko je učesnik bio voljan da uradi dva testa odjednom.

V. REZULTATI I TRENIRANJE MODELA

Najpre su izračunate fiksacione, a zatim i vizuelne mape upečatljivosti za svaku od 30 ciljnih slika. Dobijene mape upečatljivosti prikazane su na Slika 6. Vizuelne mape upečatljivosti za 30 ciljnih slika

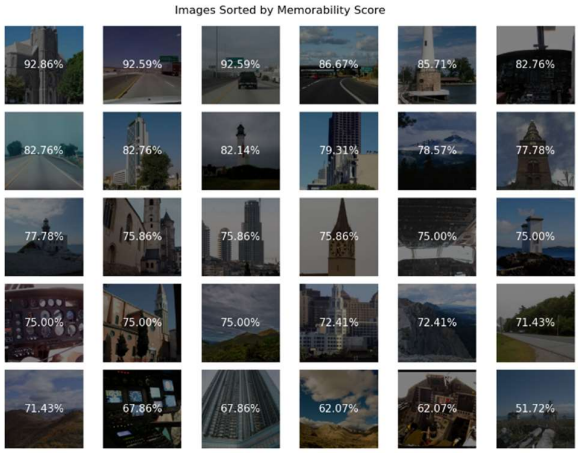


Slika 6. Vizuelne mape upečatljivosti za 30 ciljnih slika

Zatim su izračunati *IOVC* skorovi i skor memorabilnosti (Slika 7. Srednja vrednost i standardna devijacija 4 tipa IOVC skora (prvog i drugog gledanja slike, samo uspešnog prvog, samo uspešnog drugog i samo uspešnog prvog gledanja sa AUC metrikom) i Slika 8. Memorabilni skorovi ciljnih slika u opadajućem redosledu, respektivno).

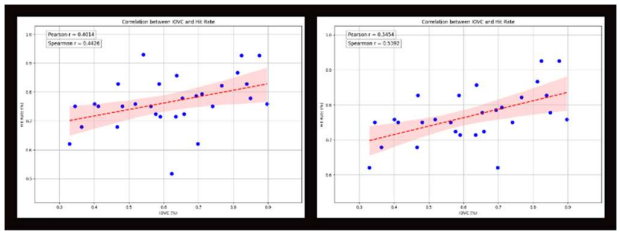
TYPE OF IOVC SCORE	MEAN	STD
GENERAL IOVC	0.8375	0.0935
SUCCESSFUL FIRST VIEW IOVC	0.6186	0.1621
SUCCESSFUL SECOND VIEW IOVC	0.8249	0.1293
SUCCESSFUL FIRST VIEW IOVC USING AUC	0.6983	0.1028

Slika 7. Srednja vrednost i standardna devijacija 4 tipa IOVC skora (prvog i drugog gledanja slike, samo uspešnog prvog, samo uspešnog drugog i samo uspešnog prvog gledanja sa AUC metrikom)



Slika 8. Memorabilni skorovi ciljnih slika u opadajućem redosledu

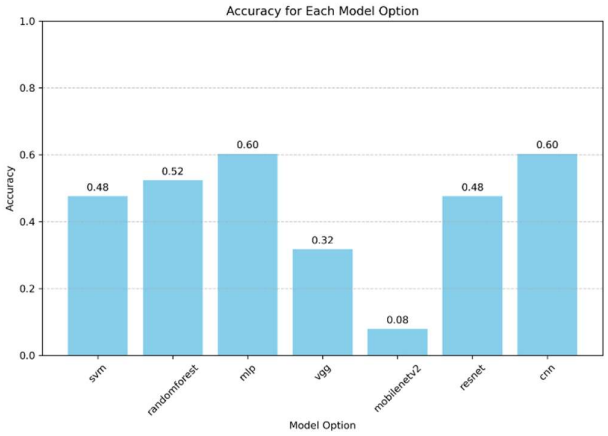
Korelacioni koeficijent između *IOVC* i skorova memorabilnosti je bio vrlo značajan (Slika 9. Korelacioni koeficijenti (Pirson i Spirman) između IOVC i skora memorabilnosti slika), te je treniranje modela mašinskog učenja izgledalo kao logičan sledeći korak.



Slika 9. Korelacioni koeficijenti (Pirson i Spirman) između IOVC i skora memorabilnosti slika

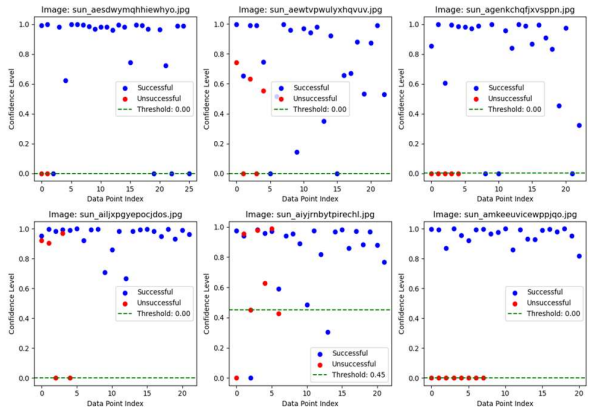
Reper koji je trebalo prestići jeste spomenuti *RUSBoost* klasifikator od 66% balansirane tačnosti [1]. Iako je sakupljena količina od 840 podataka mala, klasične tehnike augmentacije skupa podataka nisu bile primenjive, zbog senzitivnosti podataka na prostorne manipulacije. Stoga, treniranje je prosto započeto na ovolikom skupu podataka. Ideja je bila da se najpre istrenira klasifikator koji bi na osnovu proizvoljne fiksacione mape na ulazu, znao sa velikom preciznošću da kaže kojoj od 30 ciljnih slika to gledanje pripada. Preciznije, ulaz u model bila je fiksaciona mapa, a izlaz je labela 0 do 29, koja označava pripadnost ciljnoj slici, te je problem multiklasni.

Pošto je uočljiva centralna pristrasnost (engl. *Central Bias*) kod podataka gledanja slika (Slika 6. Vizuelne mape upečatljivosti za 30 ciljnih slika), primenjena je diskretizacija podataka na velike regione, kao i Gausov filter. Veličina regiona, Gausovog filtera, jednog beča, kao i test skupa su varijable koje su fino podešene tako da maksimiziraju tačnost klasifikatora. Na Slika 10. Rezultat treniranja modela različitih arhitektura dat je rezultat treniranja modela na 7 različitih arhitektura (*SVM, RandomForest, MLP, VGG, MobileNetV2, ResNet, CNN*), gde se vidi da *MLP* i *CNN* model imaju najveću tačnost, premda je *MLP* arhitektura ipak za nijansu tačnija.



Slika 10. Rezultat treniranja modela različitih arhitektura

Finalni korak bio je nazvan *Inference Thresholding*, gde je svaki podatak iz skupa reprocesuiran najboljim modelom (*MLP*), kako bi model izbacio svoj nivo pouzdanosti u pripadanje podatka datoj klasi. Zatim je izračunata najmanja vrednost nivoa pouzdanosti iznad koje treba verovati modelu, kako bi procenat ispravno razdvojenih podataka (uspešnih i neuspešnih gledanja slike) bio maksimalan. Pod pojmom ispravno gledanje podrazumeva se ono gledanje slike koje je dovelo uspešnom prepoznavanju slike kasnije u eksperimentu. Rezultat ovog koraka prikazan je na Slika 11. Rezultat Inference Thresholding procesa na 6 od 30 slika iz skupa. Prosečna tačnost svih 30 klasa bila je 89.35%, što je odličan rezultat u poređenju sa svim prethodnim reperima.



Slika 11. Rezultat Inference Thresholding procesa na 6 od 30 slika iz skupa

VI. ZAKLJUČAK

U ovom istraživanju, prezentovan je teorijski kognitivni okvir, a zatim i eksperimentalno potvrđena studija mogućnosti predviđanja ljudske pažnje. Istrenirani klasifikator sa preciznošću od 90% može sa velikom sigurnošću može da predvidi da li će određeno gledanje slike voditi uspešno kasnijem zapamćivanju iste, odnosno da li je ljudska fokusiranost prilikom gledanja bila dovoljno velika. Moguće primene ovakvog rešenja su ogromne, od obrazovnog sistema, marketinga, medicine, psihologije, sporta. Ono što je definitivno jeste da je dodatno istraživanje na ovu temu preko potrebno, pogotovu u mnogo većem obimu i sa većim resursima. Popularnost istraživanja kao što je ovo dokazuje neophodnost naučne interdisciplinarnosti.

ZAHVALNICA

Istraživanje sprovedeno uz podršku Fonda za nauku Republike Srbije, broj projekta 11113 - „*Software for Text Offences Prevention in Serbian: AI-driven Hate Speech Detection*" - STOP. Istraživanje je delimično sprovedeno i u prostorijama Palate nauke – Zadužbine Miodraga Kostića, u Beogradu, u okviru Centra za primenu veštačke inteligencije.

LITERATURA

[1] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva, "Intrinsic and extrinsic effects on image memorability," *Vision Research*, vol. 116, pp. 165–178, Nov. 2015, doi: 10.1016/j.visres.2015.03.005.

[2] Y. Rai and P. L. Callet, "Chapter 3 - Visual attention, visual salience, and perceived interest in multimedia applications," in *Academic Press Library in Signal Processing, Volume 6*, R. Chellappa and S. Theodoridis, Eds., Academic Press, 2018, pp. 113–161. doi: 10.1016/B978-0-12-811889-4.00003-8.

[3] A. Bruckert, L. Lévêque, M. Perreira Da Silva, and P. Le Callet, "A Dataset of Gaze and Mouse Patterns in the Context of Facial Expression Recognition," in *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, Nantes France: ACM, Jun. 2023, pp. 157–164. doi: 10.1145/3573381.3596153.

[4] O. Le Meur, T. Baccino, and A. Roumy, "Prediction of the inter-observer visual congruency (IOVC) and application to image ranking," in *Proceedings of the 19th ACM international conference on Multimedia*, Scottsdale Arizona USA: ACM, Nov. 2011, pp. 373–382. doi: 10.1145/2072298.2072347.

[5] M. Posner, "Orienting of Attention," *The Quarterly journal of experimental psychology*, vol. 32, pp. 3–25, Mar. 1980, doi: 10.1080/00335558008248231.

[6] A. Hollingworth, C. Williams, and J. Henderson, "To See and Remember: Visually Specific Information is Retained in Memory from Previously Attended Objects in Natural Scenes," *Psychonomic Bulletin & Review*, vol. 8, Sep. 2000, doi: 10.3758/BF03196215.

[7] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?," in *CVPR 2011*, Colorado Springs, CO, USA: IEEE, Jun. 2011, pp. 145–152. doi: 10.1109/CVPR.2011.5995721.

[8] M. Mancas and O. Le Meur, "Memorability of natural scenes: The role of attention," in *2013 IEEE International Conference on Image Processing*, Melbourne, Australia: IEEE, Sep. 2013, pp. 196–200. doi: 10.1109/ICIP.2013.6738041.

[9] S. Rahman and N. D. B. Bruce, "Factors underlying inter-observer agreement in gaze patterns: predictive modelling and analysis," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, Charleston South Carolina: ACM, Mar. 2016, pp. 155–162. doi: 10.1145/2857491.2857495.

[10] A. Bruckert, Y. H. Lam, M. Christie, and O. L. Meur, "Deep Learning For Inter-Observer Congruency Prediction," in *2019 IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan: IEEE, Sep. 2019, pp. 3766–3770. doi: 10.1109/ICIP.2019.8803596.

[11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 10, 2015, *arXiv*: arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556.

Evolucija veštačke inteligencije, izazovi i novi trendovi u zdravstvu

Rajko Terzić
Gradski zavod za javno zdravlje
Beograd, Srbija
rajko.tezic@gmail.com
ORCID 0000-0003-0772-8291

Milosav Majstorović
Visoka škola strukovnih studija za informacione tehnologije
Beograd, Srbija
majstorovic1962@gmail.com
ORCID 0000-0003-0787-7625

Vladan Pantović
Fakultet za projektni i inovacioni menadžment
Beograd, Srbija
vladan@patovic.rs
ORCID 0009-0001-6319-8822

Dušan Terzić
Republički fond za zdravstveno osiguranje
Beograd, Srbija
dusan.terzic88@gmail.com
ORCID 0000-0002-5882-5027

***Apstrakt* - Veštačka inteligencija (VI) postala je ključni alat u modernoj medicini, omogućavajući analizu kompleksnih medicinskih podataka, unapređenje dijagnostike i personalizaciju terapija. Ovaj rad pruža pregled istorijskog razvoja VI u zdravstvu, počevši od ranih ekspertnih sistema do savremenih algoritama mašinskog učenja i dubokog učenja. Analizirane su trenutne primene, uključujući prepoznavanje obrazaca na medicinskim slikama, razvoj lekova i telemedicinske usluge, kao i izazovi poput etike, privatnosti podataka i nedostatka transparentnosti u VI sistemima. Perspektive razvoja ukazuju na dalju integraciju VI u dijagnostiku, robotiku i globalnu zdravstvenu dostupnost, uz naglasak na potrebu adekvatne regulative i edukacije zdravstvenih radnika. Rad ističe potencijal VI tehnologija da transformišu zdravstvo, dok se suočavaju s izazovima u njihovoj implementaciji.**

Ključne reči – veštačka inteligencija, zdravstvo, trendovi u zdravstvu

I. UVOD

Veštačka inteligencija (VI) odnosi se na sposobnost mašina, posebno računarskih sistema, da obavljaju zadatke koji zahtevaju ljudsku inteligenciju. To uključuje razumevanje jezika, prepoznavanje obrazaca, rešavanje problema, donošenje odluka i učenje iz iskustva. VI sistemi koriste tehnologije kao što su mašinsko učenje, duboko učenje i prirodni jezik za obradu velikih količina podataka i automatsko unapređivanje performansi.

Prema jednoj od najčešće korišćenih definicija, VI se može podeliti na [1, 2]:

- Uska VI (*Weak AI*): Specijalizovana za obavljanje određenih zadataka (npr. prepoznavanje slika).
- Opšta VI (*Strong AI*): Ima potencijal da razmišlja i rešava probleme slično ljudima (trenutno teorijska ideja).
- Super inteligencija: Hipotetički oblik VI koji bi nadmašio ljudsku inteligenciju u svim aspektima.

Veštačka inteligencija igra ključnu ulogu u različitim aspektima modernog društva, transformišući industrije, poboljšavajući efikasnost i omogućavajući nova otkrića.

Neke od ključnih oblasti gde se VI koristi prikazane su u tabeli 1.

Tabela 1. Ključne oblasti primene VI

Oblast	**Opis**
Zdravstvo	Dijagnostika zasnovana na analizi medicinskih snimaka. Personalizovane terapije. Praćenje pacijenata pomoću senzora i VI algoritama.
Obrazovanje	Personalizovani programi učenja za studente. Virtuelni asistenti i analize obrazovnih trendova.
Industrija i ekonomija	Automatizacija proizvodnje. Analitika tržišta i predviđanje trendova. Optimizacija lanaca snabdevanja.
Transport	Autonomna vozila. Optimizacija javnog prevoza i logistike.
Privatnost i bezbednost	Prepoznavanje lica i nadzor. Borba protiv sajber kriminala.
Ekologija	Održivo upravljanje resursima. Predviđanje klimatskih promena i rešavanje ekoloških problema.

II. KRATAK ISTORIJSKI PREGLED RAZVOJA VI U ZDRAVSTVU

Razvoj veštačke inteligencije u zdravstvu pratio je evoluciju računarskih tehnologija i algoritama od sredine 20. veka do danas. U nastavku su dati ključni momenti i tehnologije koje su definisale njen napredak [3, 4].

Početak VI je vezan za 1950-te i 1960-te godine. VI je formalizovan kao polje istraživanja 1956. godine na konferenciji u *Dartmout*-u [5], što je postavilo temelje za primenu u različitim disciplinama, uključujući zdravstvo.

Prvi koncepti u zdravstvu bili su zasnovani na "ekspertskim sistemima" koji su koristili pravila i logiku za donošenje dijagnostičkih odluka. Ovaj period obuhvata 1970-te i 1980-te godine, i karakterišu ga sledeći ekspertni sistemi:

- MYCIN [6]: Jedan od prvih ekspertskih sistema razvijen za dijagnostiku infektivnih bolesti i preporuke za antibiotsku terapiju.
- CADUCEUS i INTERNIST [7]: Ovi sistemi su bili fokusirani na dijagnozu složenih internih bolesti, ali su se suočavali sa izazovima u prikupljanju podataka i interakciji s korisnicima.

Sledeći period je vezan za 1990-te i karakteriše ga uspon mašinskog učenja. Uvođenje mašinskog učenja omogućilo je sistemima VI da analiziraju velike količine medicinskih podataka. Sistemi su počeli da se integrišu u kliničke tokove rada, pružajući preporuke lekarima za dijagnozu i terapiju. Tako, ovi sistemi su davali svojevrsnu podršku odlučivanju.

Napredak u analizi podataka i pojava „Big Data“ vezani su za 2000-te godine. Elektronski zdravstveni zapisi (*EHR - Electronic Health Record*) omogućili su prikupljanje digitalnih podataka i razvoj modela VI za predikciju rizika od bolesti. Ovaj period karakteriše i razvoj softverskih platformi specijalizovanih za alate VI primenjivane za analizu medicinskih podataka i istraživanja od strane velikih kompanija kao što su *Google* i *IBM Watson*.

Moderna era u ovoj oblasti počinje od 2010-ih i traje do danas. Algoritmi dubokog učenja unapredili su prepoznavanje slika, što je unapredilo dijagnostiku u radiologiji, dermatologiji i patologiji. Korišćenje VI za analizu genetskih podataka, omogućilo je prilagođenu terapije za pacijente. Ovaj koncept je poznat kao *personalizovana medicina*. *Chatbot*-ovi i sistemi za daljinsko praćenje pacijenata (telemedicina i virtuelni asistenti) postaju ključni u pružanju medicinskih usluga na daljinu. Ovo se pokazalo kao posebno korisno tokom pandemije COVID-19.

III.TRENDOVI I PRIMENE VI U MEDICINI

Trendovi i primene VI u medicini uključuju brojne inovacije i pristupe koji poboljšavaju dijagnostiku, lečenje i upravljanje zdravstvenim podacima [8, 9, 10, 11]. U nastavku je dato nekoliko ključnih oblasti primene.

A. Dijagnostika i prepoznavanje obrazaca

VI igra važnu ulogu kod postavljanja medicinskih dijagnoza – algoritmi analiziraju podatke u realnom vremenu, nakon čega se sprovodi brzo uspostavljanje dijagnoze.

Unapređeni algoritami za analizu medicinskih slika i podataka omogućavaju ranije otkrivanje bolesti i precizniju predikciju ishoda. Automatska analiza medicinskih slika uz pomoć AI, posebno dubokog učenja (deep learning) ubrzava i olakšava analizu rendgenskih snimaka, CT skenova, MRI slika, ultrazvuka i drugih medicinskih slika. Algoritmi mogu prepoznavati bolesti poput raka, kardiovaskularnih bolesti i neuroloških poremećaja abnormalnosti poput tumora, fraktura, dijabetesa, i drugih patoloških stanja. Često se koristi u onkologiji (prepoznavanje raka), neurologiji (prepoznavanje moždanih bolesti), oftalmologiji (detekcija bolesti oka), i dermatologiji (prepoznavanje kožnih oboljenja).Sistemi poput *Google DeepMind* i *IBM Watson* omogućavaju brzu i tačnu dijagnostiku pomoću dubokog učenja.

B. Personalizovana medicina

Personalizovana medicina predstavlja pristup medicinskim tretmanima prilagođenih individualnim karakteristikama svakog pacijenta. Pristup na osnovu specifičnih karakteristika pacijenta, kao što su genomika, biometrijski podaci, stil života i istorija bolesti pored ostalog, omogućuju osnovna i klinička istraživanja vezana za ljudsku genetiku i molekularne osnove raznih bolesti. Poseban doprinos daju i metode i tehnike VI. Ovo je posebno važno u oblasti onkologije (izbor odgovarajuće terapije za pacijente sa rakom na osnovu njihovog genoma).

Tako, algoritmi mašinskog učenja analiziraju genetske informacije kako bi identifikovali individualne faktore rizika i optimizovali terapije. Primer za to su terapije zasnovane na sekvenciranju genoma kod karcinoma gde precizniji izbor lekova i doza su efikasnije i imaju minimalne nuspojave.

Slika 1. Novi način lečenja koje VI nudi

C. Virtuelni asistenti i telemedicina

Chatbot-ovi pomažu pacijentima u zakazivanju pregleda, odgovaranju na osnovna medicinska pitanja i podsećanju na uzimanje lekova. Takođe, VI može da prati vitalne parametre pacijenata u realnom vremenu putem senzora i nosivih uređaja.

Telemedicina je postala važna tokom pandemije koronavirusa, jer su pacijenti mogli da se konsultuju s lekarima preko video-poziva. U kombinaciji sa uređajima za praćenje stanja pacijenata na daljinu, lekari su mogli da ih posavetuju, a da sebe ne dovedu u opasnost od zaražavanja.

D. Farmaceutika i razvoj lekova

Veštačka inteligencija igra značajnu ulogu u razvoju novih lekova, a njene primene su višestruke:

VI analizira velike baze podataka i identifikuje nove molekule koje bi mogle delovati kao lekovi. Algoritmi mašinskog učenja predviđaju kako će se različiti molekuli ponašati i analizira njihovu moguću efikasnost i sigurnost.

U razvoju lekova, VI se koristi za optimizaciju hemijskih procesa. To može uključivati izračunavanje najboljih reakcija za sintetske puteve ili predviđanje osobina supstanci.

Veštačka inteligencija omogućava simulacije interakcija između lekova i bioloških sistema, čime se smanjuje potreba za eksperimentalnim ispitivanjima, što ubrzava proces.

VI analizira podatke iz kliničkih ispitivanja i otkriva obrasce koje bi ljudi mogli prevideti, pomažući u razumevanju efikasnosti i sigurnosti lekova.

Ove primene doprinose bržem i ekonomičnijem razvoju lekova, a mnoge farmaceutske kompanije koriste ove tehnologije u svojim istraživačkim procesima.

E. Nosivi uređaji

Nosivi uređaji su značajno napredovali u poslednjih deset godina i postali su važna komponenta e-zdravstvene nege u realnom vremenu. Našli su brojne primene u zdravstvu, u

rasponu od fizioloških bolesti, kao što su kardiovaskularne bolesti, hipertezzija i poremećaji mišića do neurokognitivnih poremećaja, kao što su Alchajmerova i Parkinsonova bolest, kao i druge psihološke bolesti. U tu svrhu se koriste različiti tipovi nosivih uređaja.

Tako, pametni satovi, fitnes trakeri, čak pametni komadi odeće, prate vitalne funkcije, otkucaje srca, krvi pritisak, nivo šećera u krvi, a informacije odmah šalju lekarima. Na taj način se doktorima omogućava da reaguju na vreme.

F. 5G tehnologija

Razvoj 5G mreža je transformisao zdravstvo, budući da se informacije brže razmenjuju, a konekcije su pouzdanije. To je omogućilo i brži razvoj aplikacija koje rade u realnom vremenu, čak izvođenje hirurških operacija na daljinu uz pomoć robota. Takođe, promovisanje i implementacije 5G pametne zdravstvene zaštite može smanjiti nedoslednosti u alokaciji medicinskih resursa i generalno ubrzati medicinski napredak.

G. Robotska automatizacija procesa

Robotska automatizacija procesa je tehnologija koja koristi softverske robote (botove) kako bi automatizovala rutinske i ponavljajuće zadatke unutar poslovnih procesa, pa utiče i na administrativne poslove u zdravstvenim ustanovama. Tako zdravstvene radnike oslobađa zadataka koji se ponavljaju i ostavlja im više prostora za brigu o pacijentima. Moguće oblasti primene, pored ostalih, su upravljanje terminima, zakazivanje pacijenata, upravljanje zahtevima i automatizaciju medicinskih zahteva, obradu faktura, upravljanje inventarom zdravstvene zaštite, automatizaciju u kontakt centru, upravljanje ciklusom lečenja, itd.

H. Proširena i virtuelna stvarnost

Proširena stvarnost (*AR - Augmented Reality*) i virtualna stvarnost (*VR - Virtual Reality*) tehnologije donose revoluciju u obuci medicinskih radnika, ali i u postavljanju dijagnoza i utvrđivanju metoda lečenja. Simulacije omogućavaju treninge u okruženju koje liči na situacije iz stvarnog života, pa tako, na primer, mogu da se vežbaju rizične procedure bez opasnosti da se neko povredi. VI ima potencijal da poboljša efikasnost AR/VR sistema i tako dovede do boljih ishoda za pacijente i pruženu zdravstvenu negu.

IV. IZAZOVI U PRIMENI VI U MEDICINI

Svaka era u istoriji medicine imala je svoje izazove [8, 9, 10, 11, 12]. Savremena medicina se takođe suočava sa mnogim izazovima, kao što su ekonomska ograničenja, rastuća populacija i produženi životni vek, da spomenemo samo neke.

Postoji alarmantan porast broja kancera [13], kardiovaskularnih [14] i neurodegenerativniih bolesti [15], od kojih bi se mnoge mogle kontrolisati i adekvatno lečiti ako se dijagnostikuju na vreme. S druge strane, napredak u mnogim oblastima medicine rezultirao je visokim nivoom specijalizacije [16]. Dijagnostikovanje retkih bolesti zahteva godine učenja. Štaviše, neke bolesti su tek nedavno otkrivene, a dokumentovano je da su agenti VI dijagnostikivali bolesti koje samo nekoliko specijalista može da pronađe [17, 18]. Iako je ovaj napredak generalno pozitivan, on ukazuje na potrebu za skupom opremom i potrebu za analizom složenih podataka.

Savremena medicina mora voditi računa o etici i privatnosti podatka [19] jer su medicinski podaci osetljivi, a VI sistemi često zahtevaju pristup velikim količinama informacija. Osiguranje privatnosti pacijenata i zaštita podataka ostaje veliki izazov.

Tu se često javlja nedostatak transparentnosti [19, 20] (problem „crna kutija“). Duboko učenje često stvara modele čije odluke nije lako objasniti, što može izazvati nepoverenje kod lekara i pacijenata.

Nedostatak univerzalnih standarda i zakona za upotrebu VI u medicini otežava uvođenje ovih tehnologija u kliničku praksu pa su regulatorni okviri [20] vrlo važni.

Postoji tehnički i organizacioni jaz između VI sistema i postojećih bolničkih informacionih sistema (*HIS – Hospital Information System*). Zato se mora raditi na stvaranju intergracije [20].

Mnogi lekari i medicinski tehničari nemaju dovoljno znanja o veštačkoj inteligenciji, što otežava njenu implementaciju. Iz toga proističe da je *edukacija zdravstvenih radnika* [21] neminovnost.

V. PERSPEKTIVE VI U MEDICINI

Veštačka inteligencija, sa svojom sposobnošću obrade podataka i donošenja odluka, igra ključnu ulogu u unapređenju efikasnosti različitih sektora društva [8, 9,10].

Sa sveprisutnom integracijom VI dolazi i transformacija načina rada. Rutinske zadatke preuzimaju implementirani algoritmi, čime se oslobađa vreme za ljudski rad koji zahteva kreativnost, analitičke veštine i emocionalnu inteligenciju. Nova radna mesta se stvaraju u oblastima razvoja i održavanja VI, prateći brz rast tehnologije i potrebu za stručnjacima u ovoj oblasti.

Veštačka inteligencija u medicini nudi širok spektar perspektiva koje pokrivaju različite aspekte od dijagnostike, tretmana, upravljanja zdravstvenim podacima, pa do etičkih i društvenih pitanja.

Ranije pomenuti trendovi i primene VI u medicini kao što su dijagnostika i prepoznavanje obrazaca, personalizovana medicina, farmaceutika i razvoj lekova će nastaviti da se razvijaju i u budućnosti.

Pored opisanih trendova u poglavlju III u otvorenoj literaturi prepoznaju se i druge perspektive VI u medicini [22]:

A. Robotika i automatizacija u hirurgiji

Kombinacija VI i medicinskih robota može unaprediti hirurške intervencije i rehabilitaciju.

Robotizovane hirurške procedure i roboti razvijeni uz pomoć VI mogu asistirati hirurzima u izvođenju preciznih operacija. Na primer, roboti kao što su da Vinci sistem omogućavaju minimalno invazivne hirurške procedure sa većom preciznošću i manjim rizikom od komplikacija.

Simulicija hirurške operacije i automatsko praćenje uz pomoć VI može se koristiti za trening lekara, kao i da pruži asistenciju u operacijama u realnom vremenu.

B. *Telemedicina i dijagnostika na daljinu*

Telemedicina podržana VI tehnologijama može omogućiti pružanje medicinskih usluga u ruralnim i udaljenim područjima.

Virtuelni asistenti i chat-botovi razvijeni uz pomoć VI se koriste za pružanje osnovnih medicinskih saveta, odgovarati na pitanja pacijenata, pa čak i pomoći u dijagnostici na osnovu simptoma.

VI se koristi za omogućavanje dijagnostike i konsultacija na daljinu, što je naročito korisno u udaljenim područjima ili tokom pandemijskih situacija. Algoritmi mogu analizirati medicinske slike ili druge podatke koje šalju pacijenti i pružiti savete ili preporuke lekarima.

C. *Praćenje i upravljanje hroničnim bolestima*

Analizom podataka o pacijentima u realnom vremenu (poput podataka sa pametnih uređaja) se pomaže u upravljanju bolestima i pružaju preporuke za praćenje stanja hronične bolesti. Pomoć u upravljanju bolestima poput dijabetesa i hipertenzije se sprovodi tako što uređaji koji prate nivo šećera u krvi mogu koristiti AI za automatsko prilagođavanje terapije.

Ugrađene tehnologije VI omogućavaju kontinuirano praćenje vitalnih funkcija pacijenata, prepoznavanje promena koje mogu ukazivati na pogoršanje stanja, i pravovremeno upozoravati medicinsko osoblje.

D. *Upravljanje zdravstvenim podacima*

VI se koristi za analizu ogromnih količina podataka (*Big Data*) u zdravstvenim sistemima, što omogućava bolje upravljanje zdravstvenim resursima, praćenje bolesti na globalnom nivou i efikasnije donošenje odluka. Na primer, analize velikih podataka mogu pomoći u predviđanju izbijanja epidemija ili identifikovanju novih zdravstvenih rizika.

VI pomaže u organizaciji i analizi podataka iz elektronskih zdravstvenih kartona (*EHR*), automatizuje unos podataka, prepoznaje obrasce i pomaže lekarima da bolje razumeju pacijentovu istoriju lečenja i pruža odgovarajuću negu.

E. *Poboljšanje zdravlja i prevencija bolesti*

VI se koristi za analiziranje podataka i predviđanje potencijalnih rizika za bolesti pre nego što se pojave, omogućavajući implementaciju preventivnih mera na vreme. Na primer, analiza obrazaca ponašanja i životnih navika može pomoći u prevenciji bolesti kao što su srčana oboljenja, dijabetes ili mentalne bolesti.

VI se koristi za pružanje personalizovanih saveta u vezi sa zdravim načinom života, ishranom, fizičkom aktivnošću i mentalnim zdravljem (Zdravstvena edukacija i promocija).

F. *VI u kliničkom istraživanju*

VI se koristi u istraživanjima lekova kako bi se ubrzala otkrića novih tretmana, identifikovali biomarkeri, analizirali podaci iz kliničkih ispitivanja i optimizovali procesi razvoja lekova.

VI pomaže u analizi podataka iz kliničkih studija kako bi se identifikovali potencijalni kandidati za lekove, a takođe može pomoći u prepoznavanju novih terapijskih pristupa i rizičnih faktora.

G. *Etika i regulacija VI u medicini*

Razvoj transparentnijih i odgovornijih sistema VI omogućiće bolje prihvatanje tehnologije među zdravstvenim radnicima i pacijentima.

Jedan od ključnih izazova u primeni VI u medicini je zaštita privatnosti i sigurnost podataka, posebno s obzirom na osjetljivost medicinskih podataka. Pitanje kako obezbediti sigurnost podataka i usklađenost sa zakonima poput *GDPR*-a i *HIPAA* je veoma važno.

Postavlja se pitanje ko je odgovoran ako VI sistem pogrešno dijagnostikuje ili pogrešno donese odluku u vezi sa tretmanom pacijenta. Takođe, postoji zabrinutost u vezi sa "crnim kutijama" VI modela, jer nije uvek jasno kako algoritmi donose svoje odluke.

Postoji zabrinutost da VI modeli mogu odražavati ili čak pojačavati postojeće pristrasnosti u zdravstvenoj industriji. Na primer, neki algoritmi su pokazali tendenciju da favorizuju određene demografske grupe, što može dovesti do nepravednog tretmana pacijenata.

Veštačka inteligencija ima ogromnu potencijalnu primenu u medicini i, iako postoje izazovi u vezi sa etikom, sigurnošću i implementacijom, njena uloga u poboljšanju zdravstvene zaštite i efikasnosti je neosporna. Razvijaju se i nova istraživanja koja omogućavaju sve širu i sveobuhvatniju upotrebu VI u medicinskoj praksi.

VI. ZAKLJUČAK

Veštačka inteligencija u zdravstvu prešla je dug put od jednostavnih ekspertnih sistema do današnjih sofisticiranih algoritama koji transformišu dijagnostiku, terapiju i brigu o pacijentima. Svaka faza razvoja odražava tehnološki napredak i sve veću sposobnost da se nosi s kompleksnošću medicinskih podataka. Budući napredak obećava dodatne inovacije koje će unaprediti kvalitet i dostupnost zdravstvenih usluga.

VI nudi prednosti pojedincima, kompanijama i medicinskom sektoru. Postoje neke poteškoće, kao što je integracija podataka, zaštita privatnosti pacijenata, rešavanje pravnih pitanja i održavanje bezbednosti pacijenata.

Primenjena VI može da obavlja različite funkcije, uključujući dijagnozu, terapiju, razmenu informacija, zaštitu, konsultacije, praćenje, prikupljanje podataka, pa čak i operacije na daljinu. Ovaj rad pruža uvid u trenutno stanje istraživanja veštačke inteligencije, kao i njegovu primenu u zdravstvenoj industriji u stvarnom svetu.

Upotreba VI je donela revolucionarne promene u medicini, omogućavajući naprednu analizu podataka, preciznu dijagnostiku i personalizovane terapije. Njena primena u zdravstvu evoluirala je od ranih ekspertnih sistema do savremenih metoda dubokog učenja i integracije sa "*big data*" tehnologijama. Uprkos značajnim benefitima, izazovi poput zaštite privatnosti, nedostatka transparentnosti i regulatornih ograničenja ostaju prepreka za širu implementaciju.

Perspektive razvoja VI u medicini obećavaju još veću dostupnost zdravstvenih usluga, raniju detekciju bolesti i efikasnije terapijske pristupe, posebno kroz integraciju s robotikom i telemedicinskim tehnologijama. Međutim, kako bi VI u potpunosti ispunila svoj potencijal, neophodna je

saradnja tehnoloških i medicinskih stručnjaka, uz jasno definisane etičke i pravne okvire. Ovaj rad ističe ne samo trenutni značaj VI za javno zdravlje već i njenu ključnu ulogu u budućim inovacijama koje će oblikovati modernu medicinu.

Takođe je važno naglasiti da je edukacija zdravstvenih radnika ključna za uspešnu integraciju ovih tehnologija u svakodnevnu praksu. Pravilna obuka će omogućiti medicinskim stručnjacima da efikasno koriste VI kao alat koji im pomaže u donošenju informisanih odluka, analizi podataka i poboljšanju brige o pacijentima.

VI nije razvijena da bi zamenila lekare, već da im olakša rad, smanji opterećenje i poveća produktivnost, omogućavajući im da se više fokusiraju na neposredne potrebe pacijenata i pružanje visoko kvalitetne nege. Kroz saradnju između tehnologije i ljudske stručnosti, možemo očekivati značajne napretke u zdravstvenim sistemima i poboljšanje ishoda lečenja.

Na osnovu zaključaka ovog rada, upotreba VI u medicinskom sektoru je i dalje prilično ograničena, uprkos činjenici da VI ima širok spektar potencijalnih primena i prednosti. Stoga je moguće uraditi više istraživanja o aspektima koji utiču na strategije usvajanja VI u zdravstvenoj industriji. U daljim istraživanjima, tema o tome kako se tehnički, organizacioni, etički, podaci, politika, politički i pravni izazovi mogu efikasno smanjiti, trebalo bi da bude primarni fokus.

LITERATURA

[1] S. Russell and P. Norvig, "Artificial Intelligence: A Modern Approach", ISBN-13: 978-0-13-604259-4, 1995. Prentice Hall.

[2] N. Bostrom, „Superintelligence : paths, dangers, strategies“, Oxford University Press, 2015.

[3] Istorija veštačke inteligencije, Računarski fakultet, [Online] dostupno na: https://raf.edu.rs/citaliste/istorija-vestacke-inteligencije/

[4] R. Shouval et al. „Machine learning and artificial intelligence in haematology“. British Journal of Haematology, 2021, 192, 239–250

[5] Artificial Intelligence Coined at Dartmouth, 1956, [Online] dostupno na : https://home.dartmouth.edu/about/artificial-intelligence-ai-coined-dartmouth

[6] Mycin, Wikipedija, [Online] dostupno na: https://en.wikipedia.org/wiki/Mycin

[7] I. Mašić, Z. Riđanović, H. Pandža, „Medical expert systems“, February 1995Medical Archives 49(3-4):107-12, SourcePubMed

[8] S. Tovar-Arriaga, G. Israel Pérez-Sot, K. Anhel Camarillo-Gómez, M. Aviles and J. Rodríguez-Reséndiz, „Perspectives, Challenges, and the Future of Biomedical Technology and Artificial Intelligence“, Facultad de Ingeniería, Universidad Autónoma de Querétaro, Mexico, Technologies 2024, 12(11), 212; [Online] dostupno na : https://doi.org/10.3390/technologies12110212, Accepted: 16 October 2024 / Published: 24 October 2024 .

[9] D. Milovanović, R. Terzić, „Responsible application of Artificial Intelligence in healthcare: Ethical principles of practice and new regulations“. Proceedings of the ICT Conference YU INFO 2023, Beograd: Informaciono društvo Srbije, 46-49, [In Serbian].

[10] D. Milovanović, R. Terzić, Č.Vučetić, „Trends of application Artificial Intelligence in healthcare: Challenges and regulation issues“, Proceedings of the YU INFO 2024, 95-100, [In Serbian].

[11] F. Kitsios, G. Scholar,M. Kamariotou, A. I. Syngelakis and M. A. Talias, „Recent Advances of Artificial Intelligence in Healthcare“: A Systematic Literature Review [Online] dostupno na : https://doi.org/10.3390/app13137479, 25 June 2023.

[12] S. Tovar-Arriaga, G. Israel Pérez-Soto, K. Anhel Camarillo-Gómez, M. Aviles, and J. Rodríguez-Reséndiz, “Perspectives, Challenges, and the Future of Biomedical Technology and Artificial Intelligence” by Facultad de Ingeniería, Mexico, 2024, 12(11), [Online] dostupno na: https://doi.org/10.3390/technologies12110212, 12 October 2024

[13] R. Ortiz-Feregrino, S. Tovar-Arriaga, J.C. Pedraza-Ortega, J. Rodríguez-Reséndiz, “Segmentation of retinal blood vessels using focal attention convolution blocks in a UNET”. Technologies 2023, 11, 97.

[14] E.R.P. de Leon. Sanchez, J.D.Mendiola-Santibáñez, O.A. Dominguez-Ramirez, A.M. Herrera-Navarro; A. Vazquez-Cervantes, H. Jimenez-Hernandez, H. Senties-Madrid, “Fuzzy logic system for classifying multiple sclerosis patients as high, medium, or low responders to interferon-beta”. Technologies 2023, 11, 109.

[15] A.V. Cerón, E.L. Domínguez, S.D. Isidro, M.A.M Nieto, J. De La Calleja, S.E.P. Hernández, “Level of technological maturity of telemonitoring systems focused on patients with chronic kidney disease undergoing peritoneal dialysis treatment”: A systematic literature review. Technologies 2023, 11, 129.

[16] P. Moltó-Balado, S. Reverté-Villarroya, V. Alonso-Barberán, C. Monclús-Arasa, M.T. Balado-Albiol, J. Clua-Queralt, J.-L. Clua-Espuny, “Machine learning approaches to predict Major Adverse Cardiovascular Events in atrial fibrillation”. Technologies 2024, 12, 13.

[17] T. Chandel, V. Miranda, A. Lowe, T.C. Lee, “Blood pressure measurement device accuracy evaluation: Statistical considerations with an implementation in R”. Technologies 2024, 12, 44.

[18] M.A. Hasan, F. Haque, S.R. Sabuj, H. Sarker, M.O.F. Goni, F. Rahman, M.M. Rashid, “An end-to-end lightweight multi-scale CNN for the classification of lung and colon cancer with XAI integration”. Technologies 2024, 12, 56.

[19] “Etičke smernice za razvoj primenu i upotrebu pouzdane i odgovorne VI” – usvojeni tekst, 26-27 oktobar 2023. Master Class, Fakultet organizacionih nauka.

[20] S. Pasricha, „AI Ethics in Smart Healthcare“, Colorado State University, [Online] dostupno na :https://arxiv.org/pdf/2211.06346

[21] D. Jha A. Rauniyar, A. Srivastava, D. Haileselassie Hagos, N. Kumar Tomar, V. Sharma, E. Keles, Z. Zhang, U. Demir, A. Topcu, A. Yazidi, J. Erik Håakegård, U. Bagci, „Ensuring Trustworthy Medical Artificial Intelligence through Ethical and Philosophical Principles“, [Online] dostupno na : https://doi.org/10.48550/arXiv.2304.11530

[22] G. Parker, Ph.D., C. D. Parker, Artificial Intelligence in Healthcare: Future Benefits and Challenges, Health Affairs.

Artificial intelligence evolution, challenges and new trends in health

Rajko Terzić, Milosav Majstorović, Vladan Pantović, Dušan Terzić

ABSTRACT - Artificial intelligence (AI) has become a key tool in modern medicine, enabling the analysis of complex medical data, the improvement of diagnostics and the personalization of therapies. This paper provides an overview of the historical development of AI in healthcare, starting from early expert systems to modern machine learning and deep learning algorithms. Current applications are analyzed, including pattern recognition in medical images, drug development and telemedicine services, as well as challenges such as ethics, data privacy and lack of transparency in AI systems. Development perspectives point to the further integration of AI into diagnostics, robotics and global health accessibility, with an emphasis on the need for adequate regulation and education of health workers. The paper highlights the potential of AI technologies to transform health, while facing challenges in their implementation.

*K*eywords - artificial intelligence, healthcare, trends in healthcare

ENERGETSKA EFIKASNOST PRISTUPNE MOBILNE MREŽE TELEKOM SRBIJA A.D.
ENERGY EFFICIENCY OF TELEKOM SRBIJA A.D. RADIO ACCESS NETWORK

Danijela Aleksić, Snežana Elčić
Telekom Srbija a.d.

Sadržaj – *Kapacitet mobilne mreže dizajnirane za vršni dnevni saobraćaj neminovno dovodi do neiskorišćenja resursa u satima nižeg saobraćaja. Optimizacija korišćenja raspoloživih mrežnih resursa može poboljšati energetsku efikasnost u pristupnoj radio mreži, koja je prepoznata kao energetski najzahtevnija. Dugoročni ciljevi održivosti i energetske efikasnosti oblikuju načine na koje se mobilne mreže modernizuju i planiraju. Ovaj rad daje generalne okvire projekta energetske efikasnosti sprovedenog u kompaniji Telekom Srbija.*

Abstract - *The capacity of a mobile network designed to handle peak daily traffic inevitably leads to excess capacity during off-peak hours. Optimizing the use of available network resources can improve energy efficiency in the radio access network part, which is recognized as the most energy-intensive. Long-term goals of sustainability and energy efficiency shape the ways in which mobile networks are modernized and planned. This paper presents the general framework of the energy efficiency project conducted in Telekom Srbija.*

1. UVOD

Evolucija telekomunikacija uslovljava sve širi spektar finansijskih statistika i mera razvoja telco operatera. Smanjenje operativnih troškova (OPEX) je stalna težnja, a dugoročni klimatski akcioni ciljevi su realnost koja dodatno usmerava OPEX telco operatera. Kako je kvalitet korisničkih performansi imperativ koji ne podrazumeva degradaciju korisničkog iskustva, ostvarivanje zadatih agendi iziskuje sve zahtevnije dimenzionisanje OPEX-a. Energetska efikasnost je jedna od akcija koja je prepoznata kao potencijal za ostvarivanje više zadatih ciljeva. Tranzicija na čistije energije i jačanje energetske stabilnosti, kao i scenario neto nulte emisije, posebno aktuelizuju energetsku efikasnost. U oblasti mobilnih mreža dodatne izazove energetskoj efikasnosti nameće potreba za implementacijom tehnologija novih generacija, uz zadržavanje postojećih tehnologija starijih generacija. Viši stepen automatizacije i modernizacije ne uslovljava nužno i poboljšanje energetske efikasnosti. Periodična nadogradnja hardvera i softvera donosi inkrementalne dobitke u energetskoj efikasnosti. Za šire sagledavanje aktuelne teme energetske efikasnosti u oblasti mobilnih mreža potreban je fokus na više aspekta i domena uticaja, uz stalno rastući uticaj veštačke inteligencije (AI) .

Mobilni operateri se u sprovođenju modela energetske efikasnosti uveliko oslanjaju na preporuke i statistiku sprovedenih uporednih analiza performansi mobilnih mreža sa stanovišta *Mobile Energy Efficiency* (MEE). Učesnicima *MEE Benchmark*-a je garantovana poverljivost, dok su eksterna poređenja sprovedena anonimno.

U GSMA izveštaju [1] je istaknuto 6 bitnih prednosti *MEE Benchmark*-a:

- Detaljna analiza relativnih performansi mobilnih mreža za veliki skup podataka uz proračun potencijalnog dobitka u energetskim troškovima i redukciji karbonskih emisija.
- Jedinstveni "normalizacioni" pristup koji omogućava *like-to-like* poređenje upotrebom multivarijabilnih regresionih metoda.
- Godišnje praćenje poboljšanja i kvantifikacija inicijativa za smanjenje troškova.
- Uvid u poboljšanje energetske efikasnosti, uključujući pristup studijama za mreže sa najboljim učinkom.
- Demonstracija pozitivnih akcija i redukcije emisija zainteresovanim stranama.
- Mogućnost učešća u optimizacionim procesima u okviru projekata energetske efikasnosti mobilnih mreža.

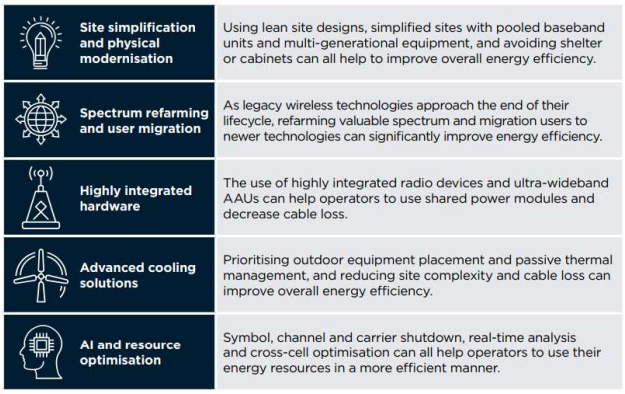
Kao KPIs (*Key Performance Indicators*) korišćene su vrednosti energetske efikasnosti po:

- mobilnoj konekciji,
- jedinici mobilnog saobraćaja,
- mobilnoj ćeliji,
- jedinici mobilnog prihoda.

Podaci koje dostavljaju mobilni operateri su:

- Potrošnja električne energije u mobilnoj mreži i upotreba dizela.
- Broj fizičkih lokacija i tehnologija po sajtu.
- Procentualnu pokrivenost teritorije (geografske i populacione statistike).
- Broj mobilnih korisnika.
- Prihod mobilnog operatera.
- Minuti mobilnog govornog saobraćaja i bajtovi saobraćaja mobilnih podataka.

Izazovi energetske efikasnosti u telekomunikacijama se mogu različito postaviti, od jednostavnijih pa do ekstremnih. *GSMA Intelligence* u izveštaju za energetsku efikasnost mobilnih mreža [2] daje smernice planiranja, modernizacije i optimizacije resursa energetski efikasnih mobilnih mreža (Slika 1.).

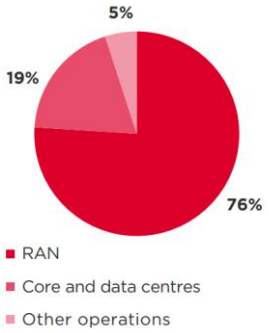


Slika 1. Glavne oblasti u kojima se može ostvariti značajno poboljšanje energetske efikasnosti

Isti izveštaj [2] daje i kategorizaciju potrošnje energije mobilnih operatera, kao i procentualni udeo svake od kategorija u ukupnoj potrošnji (Slika 2).

Prepoznate su tri glavne kategorije i to:

- RAN (*Radio Accesss Network*) – Pored baznih stanica BS (*Base Station*) ovoj kategoriji je pridodata i prateća pristupna infrastruktura.
- Core & Data Centre – Uključuje sve komutacione i aplikacione mrežne elemente sa pripadajućom transportnom mrežom; IT i intranet infrastrukturu kao i Data Centre koji su u vlasništvu samih operatera (izuzeti su Data Centri kod kojih se iznajmljuju resursi kao što su AWS, Google, Microsoft …).
- Other operations – U ovoj kategoriji su sagledani objekti operatera (kancelarije, poslovnice, magacini i ostali poslovni i logistički objekti).
-

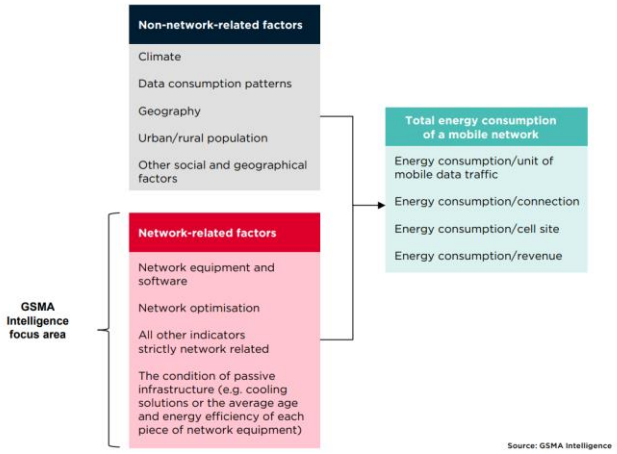


Slika 2. Procentualni udeo svake kategorije u ukupnoj potrošnji energije mobilnih operatera

Sa Slike 2. se može primetiti da je RAN energetski najzahtevnija kategorija, koju dominantno treba sagledati sa stanovišta energetske efikasnosti.

2. MOTIVACIJA

U kompaniji Telekom Srbija a.d. su aktivne mrežne tehnologije 2G, 3G i 4G. Za sagledavanje ukupne energetske efikasnosti svih aktivnih tehnologija, potrebno je posebno analizirati dominantan uticaj brojnih faktora podeljenih u dve osnovne grupe (Slika 3) [2].



Slika 3. Faktori koji utiču na energetsku efikasnost mobilnih mreža

Kako bi se osigurala konkurentnost mobilnog operatera na tržištu, infrastrukturna i tehnološka ulaganja u sve aktivne tehnologije su neophodna i u uslovima različitih geografskih i socijalnih karakteristika. Svi faktori iz prve grupe, na koje se ne može direktno uticati, imaju direktan i neposredan uticaj na ukupnu potrošnju energije. Druga grupa faktora, prepoznata kao oblast od interesa za poboljšanja energetske efikasnosti mobilnih mreža, je predmet skupa optimizacionih aktivnosti mobilnih operatera.

U kompaniji Telekom Srbija a.d. se kao jedan od najznačajnijih strateških ciljeva za tekuću 2025. godinu postavlja uvođenje 5G tehnologije. Izveštaj [2] naglašava da je 5G energetski efikasnija tehnologija na duži vremenski rok, ali da ova energetska efikasnost ne mora nužno odmah biti uočljiva. Pun potencijal energetske efikasnosti 5G mreže je očekivan tek kada se postigne odgovarajuća penetracija u broju korisnika, kao i u gustini 5G mobilne mreže. Upravo uvođenje 5G tehnologije je dodatni motiv za analizu i sprovođenje mera za poboljšanje energetske efikasnosti.

U izveštaju [3] su date smernice za planiranje i optimizaciju mobilnih mreža kod uvođenja 5G tehnologije. Značajnije promene već aktivnih mobilnih mreža uključuju ukidanje tehnologija starijih generacija [3]. Naprednije metode optimizacije mobilnih mreža bazirane su na korišćenju veštačke inteligencije i mašinskog učenja AI/ML i ove metode su fokusirane na 4G i 5G tehnologije.

3. ENERGETSKI EFIKASNIJI RAN

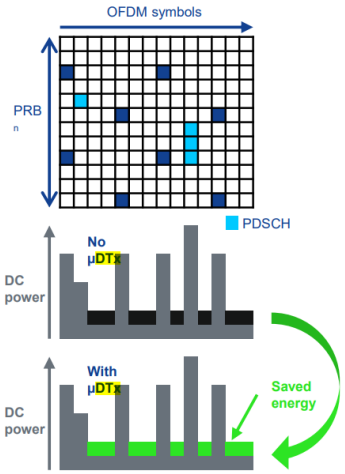
Kako je u mobilnim mrežama najveći utrošak energije u RAN kategoriji, energetski efikasniji RAN se može prepoznati kao prioritet. Prethodna uvođenja novih tehnologija od 2G do 4G, iziskivala su pažljivo planiranje i modernizaciju radio opreme kako bi se poboljšala osnovna metrika energetske efikasnosti. Refarming spektra, proširenje frekventnog opsega za 4G, kao i uvođenje SRAN (*Single RAN*) sa RRU (*Remote Radio Unit*) koji se mogu koristiti za 4G i neku od starijih

tehnologija, značajno su doprineli na planu energetski efikasnijeg RAN-a.

U kompaniji Telekom Srbija a.d. je u RAN delu mobilne mreže prisutna oprema različitih vendora. Za opremu svakog od vendora su određeni geografski poligoni ili klasteri u kojima je sprovedeno testiranje funkcionalnosti koje doprinose poboljšanju energetske efikasnosti. Implementacija i aktivacija samih funkcionalnosti (*feature*-a) je sprovođena u izdvojenim vremenskim prozorima kako bi se ispratili pojedinačni učinci *feature*-a kroz mrežne KPI-eve od interesa.

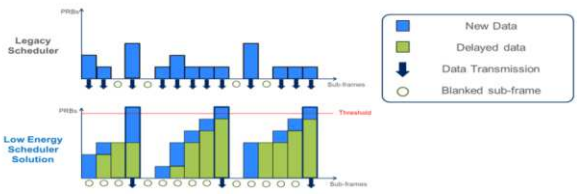
Neke od najznačajnijih funkcionalnosti su:

- *µDTX feature* (*micro sleep*) omogućava uštede energije isključivanjem pojačavača (PA) tokom *idle* transmisionih perioda. *µDTX* kontroliše PA zavisno od stanja RF simbola. Tokom trajanja tzv. *empty* simbola, PA se isključuju, dok se pri pojavi transmisije podatka, PA ponovo uključuju.



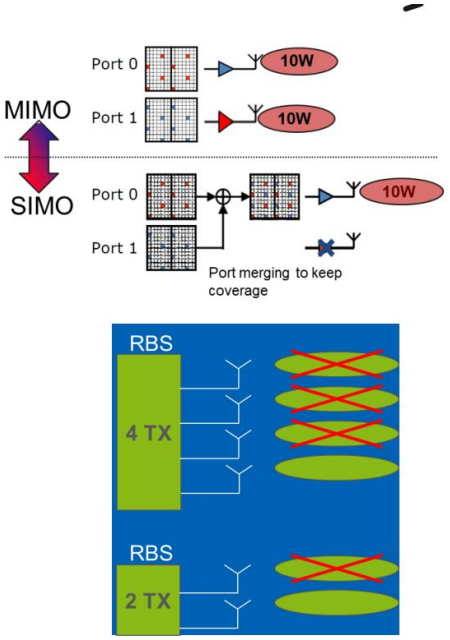
Slika 4. *µDTX feature* (*micro sleep*)

- *Low Energy Scheduler Solution (LESS)* odlaže *DL* (*downlink*) transmisije koje su u okviru definisanih margina bez uticaja na korisničko iskustvo. Aktivacija ove funkcionalnosti je moguća po *QCI* (*QoS Class Indentifier*) i naročito je primenljiva za saobraćaj koji nije posebno osetljiv na kašnjenja u vremenu.



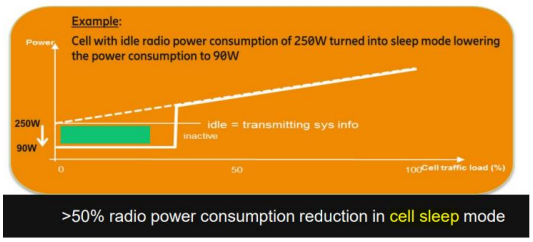
Slika 5. *Low Energy Scheduler Solution*

- *MIMO Sleep Mode* isključuje *Tx* grane tokom slabog i srednjeg saobraćaja uz dinamičku promenu moda 4*Tx*, 2*Tx* i SIMO.



Slika 6. *MIMO Sleep Mode*

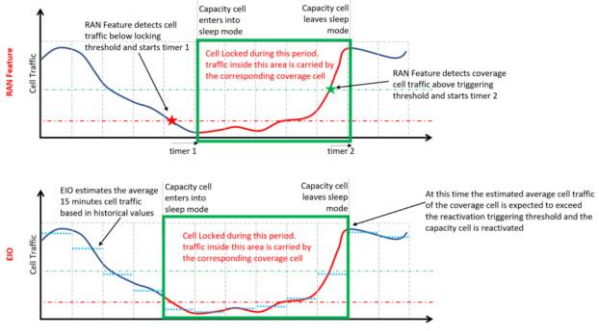
- *Cell Sleep Mode* gde se ćelije za obezbeđivanje kapaciteta automatski aktiviraju i deaktiviraju zavisno od trenutnog saobraćaja. Takođe se automatski konfigurišu i kontinualno ažuriraju odnosi između ćelija za obezbeđivanje kapaciteta i ćelija za obezbeđivanje pokrivenosti.



Slika 7. *Cell Sleep Mode*

- *AI Powered MIMO Sleep Mode* gde *embedded AI* automatski otkriva i uči saobraćajne obrasce ćelija tokom vremena i primenjuje optimalnu MIMO konfiguraciju kako bi se redukovala potrošnja energije. Sa Slike 8. se može primetiti da se primenom ove funkcionalnosti dodatno povećava period tokom koga su aktivne samo ćelije za obezbeđivanje pokrivenosti.

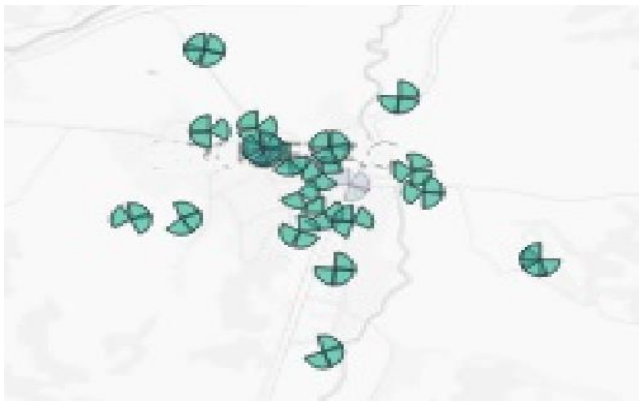
-



Slika 8. *RAN features vs. AI powered Cell/Site EM*

4. REZULTATI TESTIRANJA

U Telekom Srbija je sprovedeno testiranje seta funkcionalnosti za poboljšanje energetske efikasnosti na urbanim poligonima za opremu svih prisutnih RAN vendora. Poligoni su uključivali gradove sa okolinom sa projektovanih 50.000-100.000 stanovnika. Na Slici 9. je prikazan jedan od testnih poligona sa pripadajućim baznim stanicama.



Slika 9. Testni poligon

Testiranja su potvrdila da je zavisno od tipa i broja radio modula na sajtu, dominantan utrošak električne energije (do 90%), predstavljala upravo potrošnja radio modula.

Kako je u mobilnoj mreži postignuta visoka penetracija LTE korisnika, kao i kvalitetno pokrivanje svih teritorija LTE signalom, najveća pažnja je posvećena testiranju funkcionalnosti koje utiču na energetsku efikasnost LTE mreže. Najbolji rezultati postignuti su aktivacijom funkcionalnosti *μDTX*, a u zavisnosti od RAN vendora, ostvarena ušteda energije se kretala 12-14% na dnevnom nivou. Osetne uštede su postignute i korišćenjem *MIMO Sleep Mode* funkcionalnosti i to od 3% u doba većeg saobraćaja do 12% u doba manjeg saobraćaja.

Napomenimo da sprovedena testiranja nisu bila ekstremna, jer je postignut primarni cilj očuvanja korisničkog iskustva i ukupnih mrežnih performansi. Takođe je sagledana i ekonomska efikasnost aktivacije seta funkcionalnosti za energetsku efikasnost. I u ekonomskom smislu aktivacija ovih funkcionalnosti se pokazala kao opravdana, jer su ulaganja za aktivaciju licenci testiranih funkcionalnosti oko 5 puta niža od potencijalnih ušteda.

LITERATURA

[1] GSMA, Mobile Energy Efficiency Explained

[2] GSMA Intelligence, Going green: measuring the energy efficiency of mobile networks

[3] GSMA Intelligence, 5G energy efficiencies Green is the new black

YU #3: Sesija 3

Mašinsko učenje i veliki jezički modeli

Snimanje bilingvalne baze AI-SPEAK za multimodalno prepoznavanje govora

Tijana Nosek
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
tijana.nosek@uns.ac.rs
0000-0002-3707-0286

Siniša Suzić
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
sinisa.suzic@uns.ac.rs
0000-0002-0511-6729

Vuk Stanojev
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
vukst@uns.ac.rs
0000-0003-2517-3728

Nikša Jakovljević
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
jakovnik@uns.ac.rs
0000-0002-7283-3939

Lidija Krstanović
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
lidijakrstanovic@uns.ac.rs
0000-0001-7958-5846

Milan Sečujski
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
secujski@uns.ac.rs
0000-0002-3426-3277

***Apstrakt* – U ovom radu opisan je postupak snimanja, kao i krajnji sadržaj bilingvalne audio-vizuelne baze na srpskom i engleskom jeziku. Iako za engleski jezik postoji veći broj audio-vizuelnih baza različte veličine, ovo je prvi primer jedne takve baze na srpkom jeziku. Pored bilingvalnosti, još jedna prednost predstavljene baze ogleda se i u činjenici da su video snimci dostupni iz više različitih uglova-anfas, kao i s leve i desne strane u odnosu na glavnu kameru. Detaljno je opisan proces razvoja baze, kao i naknadna obrada i svi problem koji su se javljali.**

Ključne reči – multimodalno, audio-video, baza, govorna komunikacija.

I. Uvod

U svakodnevnom govoru, ljudi ostvaruju komunikaciju na multimodalan način, ne oslanjajući se samo na zvuk koji stiže do njih, već i na čitanje sa usana govornika, kao i na izraze njegovog lica [1]. Zbog toga se razvijaju audio-vizuelne baze podataka, koje pored govora sadrže i informacije o govornikovom licu ili samo usnama. Ovakve baze podataka omogućavaju istraživanja u različitim oblastima obrade govora, kao što su multimodalno prepoznavanje govora, multimodalna sinteza govora, konverzija video-snimaka u tekst, rekonstrukcija govora na osnovu video-snimaka i rekonstrukcija pokreta usana i ostatka lica na osnovu govora.

Iako su istraživanja o multimodalnom prepoznavanju govora sve češća, ovakvih baza podataka je trenutno veoma malo. Većina ovakvih baza podataka namenjena je isključivo za engleski jezik, uz nekoliko izuzetaka kao što su kineski i nemački jezik [2]. Obrada govora je u velikoj meri specifična za svaki jezik i da bi se postigao odgovarajući kvalitet, neophodno je razvijati kvalitetne resurse za taj jezik.

U ovom radu, biće predstavljene dostupne multimodalne govorne baze podataka, kao i detaljan način snimanja i obrade jedne takve baze, AI-SPEAK korpusa, koja pored snimaka govora 25 govornika na srpskom i engleskom jeziku, sadrži i video snimke delova njihovih lica.

II. Javno dostupne multimodalne baze

Istraživanja u oblasti prepoznavanja govora uz dodatnu informaciju u vidu video snimaka pokreta usta dovela su do formiranja različitih audio-vizuelnih korpusa snimanih u kontrolisanim uslovima. Među takvim bazama su AVLetters [3], CUAVE [4], OuluVS [5] i GRID [6], međutim, radi se o skupovima podataka sa malim brojem govornika i ograničenim rečenikom. AVLetter i CUAVE sadrže izgovore engleske abecede i 10 cifara, dok GRID i OuluVS sadrže kratke rečenice, ograničenog rečenika, iste za svakog govornika.

OuluVS2 [7] je složeniji korpus, govornici izgovaraju duže rečenice sa proširenim rečnikom i snimani su iz više uglova. AVICAR korpus [8] je po obimu sličan OuluVS2, ali se razlikuje po tome što su govornici snimani tokom vožnje automobila, iz više uglova. Da bi se premostio jaz između engleskog i ostalih jezika, razvijeni su RUSAVIC [9] korpus na ruskom jeziku i CI-AVSR [10] na kineskom jeziku, pandani AVICAR korpusu.

Ograničenje baza snimanih u kontrolisanim uslovima jeste što imaju mali broj sati govora, kao i ograničen skup reči, što nije dovoljno za potpunu obuku današnjih modela baziranih na dubokom učenju.

Da bi se razvili veći korpusi sa slobodnim rečnikom koji bolje predstavlja govor iz stvarnog sveta, razvijeni su korpusi koji uzimaju govor sa interneta sa više od 100 sati govora i velikim brojem govornika. LRW [11], LRS2-BBC [12] i LRS3-TED [13] su obimni korpusi na engleskom jeziku za audio-vizuelno prepoznavanje govora i razlikuju se pre svega po izvoru podataka i veličini. LRW-1000 [14] je obiman korpus za kineski jezik dobijen obrađivanjem TV emisija na kineskom jeziku, GLips [15] za nemački, dobijen obrađivanjem snimaka iz parlamenta. Jedna od mana ovih baza jeste što u većini slučajeva govornici nisu snimani iz više uglova.

Da bi se prevazišli nedostaci korpusa snimljenih u kontrolisanim uslovima i korpusa preuzetih sa interneta, razvijen je OLKAVS [16] korpus na korejskom jeziku sa 1107 govornika i 1150 sati govora. Govornici su snimani iz 9 različitih uglova.

III. Priprema AI-SPEAK baze

U fazi pripreme AI-SPEAK korpusa definisano je da će ovaj korpus sadržati audio snimke govora na srpskom i engleskom jeziku od 25 odraslih govornika oba pola, kao i video snimke pokreta njihovih usana iz tri različita ugla.

Planirana količina govora po govorniku je 10 minuta. Na kraju procesa snimanja snimljeno je 35 govornika, ali finalna verzija baze, nakon procesa obrade, sadrži 30 govornika.

A. *Korpus*

Pripremljen je fiksan broj fonetski uravnoteženih rečenica na oba jezika, uključujući skup rečenica identičan za sve govornike i skup rečenica koje su jedinstvene za svakog govornika. Deo korpusa zajednički za sve govornike sadrži izgovore slova srpske azbuke odnosno engleske abecede, izgovore 10 cifara, 7 dana u nedelji i 13 komandnih reči („napred, nazad, levo, desno, gore, dole, potvrdi, odustani, obriši, pošalji, dalje, početak, kraj“, odnosno „forward, back, left, right, up, down, confirm, cancel, delete, send, next, home, end“), kao i 25 rečenica po jeziku. Deo korpusa jedinstven po govorniku sadrži 50 rečenica za svaki od dva jezika, i posebno je kontrolisan u pogledu fonetske pokrivenosti (oko 350 reči na srpskom i oko 400 na engleskom jeziku).

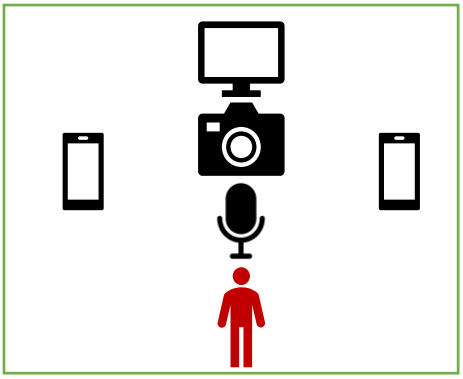
Za svakog govornika pripremljena je PPT prezentacija na kojoj svaki slajd sadrži po jednu rečenicu, a pri prelasku na sledeći slajd čuje se zvučni signal koji će se u kasnijim koracima obrade koristiti za sinhronizaciju (sinusoida frekvencije 1 kHz, trajanja 0.5 s). Govornici su instruisani da rečenicu prvo pročitaju u sebi kako ne bi pravili greške pri izgovoru te da rečenicu pročitaju normalnom brzinom, u neutralnom tonu i sa neutralnim izrazom lica, bez bilo kakvog naglašavanja u pogledu rečenične intonacije.

B. *Snimanje*

Snimanje je izvedeno u IAC Mini anehoičnoj komori Univerziteta u Novom Sadu. U pitanju je komora oblika kocke stranice oko 1.5m, koja je iznutra obložena modularnim akustičkim panelima koji upijaju zvuk. Po specifikaciji ova akustička izolacija smanjuje buku u odnosu na spoljašnjost komore za 50dB. Ovakve komore često se nazivaju i gluvim sobama. Ovako male gluve sobe obično su namenjene za razne eksperimente i ispitivanja u oblasti buke ili vibracija [17]. U ovom istraživanju korišćena je za snimanje govorne baze jer, pored toga što minimizuje pozadinsku buku, obezbeđena je i ponovljivost postavke opreme za snimanje i osvetljenja u uslovima kada je snimanje zbog većeg broja govornika obavljeno u dužem vremenskom periodu.

Za snimanje audio signala korišćen je Rode Podmic mikrofon. Za glavni video snimak govornikovog lica spreda korišćena je Sony VLOG ZV-1 kamera. Uz to, pomoćni audio i video snimci prikupljeni su pomoću mobilnih telefona Samsung A33 i Samsung S10, smeštenih dijagonalno ispred govornika s leve i desne strane, respektivno (Slika 1). Za osvetljenje je korišćen reflektor ugrađen u komoru, ali je prekriven odgovarajućim materijalom kako bi se formiralo difuzno svetlo.

Govornici su sedeli u anehogenoj komori leđima naslonjeni na zadnji unutrašnji zid komore, a ispred njih bio je postavljen stalak sa laptopom na kom je prikazana prezentacija sa rečenicama koje se snimaju, stalci sa kamerama i stalak za mikrofon. Celokupno snimanje



Slika 1. Šema pozicije opreme tokom snimanja u anehogenoj komori u odnosu na govornika.

po govorniku trajalo je u proseku 30-45 minuta. Govornici su instruisani da se što manje pomeraju tokom izgovora rečenice, kao i da se trude da ne menjaju položaj čitavog tela tokom celokupne sesije snimanja. U slučaju grešaka prilikom izgovora, govornici su instruisani da ponove ceo segment. Celokupna postavka u komori može se videti na slici 2. Tokom procesa snimanja, govornici su bili zatvoreni u komori, a tok snimanja praćen je van komore, i snimanje je zaustavljano ukoliko bi govornik prijavio neki problem.

Slika 2. Postavka opreme tokom snimanja u anehogenoj komori.

C. *Obrada*

a) *Obrada audio snimaka*

Kako je naknadno utvrđeno, zbog blizine električnog transformatora u susednoj prostoriji, u snimcima postoji interferencija u vidu stalno prisutnog šuma na niskim frekvencijama, pa je prvi korak obrade snimaka bila redukcija ovog šuma. Ona je urađena primenom programa *Audacity*, odnosno, odgovarajuće opcije *Noise Reduction*. U prvoj fazi bira se deo gde nema govora, odnosno gde postoji samo signal koji sadrži pomenuti šum, kako bi se izračunao spektar snage šuma. U drugoj fazi, korišćenjem infromacija o pomenutom spektru, vrši se uklanjanje šuma iz audio signala. U pojedinačnim frekvencijskim opsezima

procenjuju se snage govornog signala i šuma, a zatim se u pojedine opsege unosi odgovarajuće slabljenje u zavisnosti od toga u kojoj meri u njima dominira šum. Nakon toga vrši se vremensko ujednačavanje da bi se dobile spore promene pojačanja za svaku frekvenciju, a prati ga ujednačavanje frekvencija kako bi se postiglo da se nijedna frekvencija ne potiskuje ili pojačava izolovano.

Nakon redukcije šuma, usledio je proces anotacije koji uključuje manuelno označavanje segmenta u Audacity softveru, sa tri moguća ishoda: neupotrebljiv segment, segment sa greškama ali uglavnom ispravnim izgovorom, i segment sa ispravnim izgovorom. Segmenti su razdvojeni markerima, koje su anotatori postavljali u bilo koji trenutak tokom trajanja sinhronizacionog signala.

Tokom pregleda podataka uočene su moguće greške, kao što su: nepravilno pročitane reči, oštećeni snimci, dodati negovorni elementi kao što su zamuckivanje, tzv. lažni počeci (govornik počeo da izgovara određenu celinu, prekinuo i počeo od početka), te izgovori preklopljeni sa sinhronizacionim signalom. U zavisnosti od vrste greške, segment može biti odbačen, anotacija ispravljena, ili govornik isključen iz baze ako kod njega ima previše grešaka.

Snimci se naknadno automatski obrađuju prema podacima dobijenim manuelnom anotacijom. Prvo se vrši fino podešavanje pozicije markera na početak odnosno kraj segmenta, i to traženjem korelacije referentnog sinhronizacionog signala sa verzijom iz audio snimka. Verzija koja sadrži sinhronizacioni signal u snimku dobijena je isecanjem signala u intervalu ±1s od vremenske odrednice markera postavljenog od strane anotatora. Ovo fino podešavanje granica segmenata sprečava da se sinhronizacioni signal nađe u isečenim segmentima. Prema oznakama koje je anotator postavio, segment se ili u potpunosti zadržava, ili se odseca deo sa pogrešnim izgovorom te zadržava samo onaj sa konačnim ispravnim izgovorom, ili se u potpunosti odbacuje (Slika 3). Za delimično automatsku fonetsku i prozodijsku anotaciju (akcenti, pauze, naglasci) koristi se samo audio-snimak sa glavnog mikrofona, a dobijene transkripcije se automatski propagiraju na komplementarne snimke.

b) Obrada video snimaka

Video snimci sa svih kamera sinhronizovani su pomoću zvučnih signala snimljenih kamerama. Takođe, svi video snimci su pregledani manuelno, pri čemu su primećeni problemi sa prekomernim pomeranjem govornika tokom izgovora rečenica, kao i tehnički problemi u vidu nedostatka nekih snimaka sa pomoćnih kamera. U slučaju ozbiljnih problema, snimak se kompletno isključuje iz baze, ali se u slučaju manjih grešaka beleži napomena.

Manuelno je na video snimcima utvrđena vremenska odrednica pozicije poslednjeg sinhronizacionog signala, te su pozicije ostalih sinhronizacionih signala utvrđene odgovarajućim vremenskim pomerajem u odnosu na taj, a prema utvrđenim vremenskim odrednicama na glavnom audio-snimku. Za sinhronizaciju je odabran poslednji sinhronizacioni signal jer je on sigurno poslednji na svim kamerama i glavnom audio-signalu, dok prvi nije nužno isti jer su se neki od govornika na početku snimanja samo upoznavali sa radom prezentacije za prikaz rečenica.

Problem je takođe postojao u slučajevima kada je tokom snimanja sesije dolazilo do prekida, te ponovnog uključivanja neke od kamera iz razloga kao što su automatska zaštita od pregrevanja, što je moralo biti ručno pregledano i na odgovarajući način sinhronizovano.

Konačno, korišćenjem *MediaPipe* alata postavljene su maske preko svih video snimaka, kako bi se, radi zaštite privatnosti govornika, u AI-SPEAK korpusu našao samo donji deo lica govornika, gde se nalaze usne.

D. Najčešće greške

Oštećeni delovi audio i video snimaka su izbačeni, ali je ovoga bilo izuzetno malo (ukupno 2 rečenice). Govornici koji su se previše pomerali tokom snimanja i to tokom izgovora same rečenice (klimanje glavom, promene položaja sedenja) izbrisani su iz baze, i takvih slučajeva bilo je 3. Najveći problem bio je šta uraditi sa snimcima rečenica u kojima je pogrešna jedna reč, a govornik, suprotno uputstvima, nije ponovio celu rečenicu nego je ponovio samo tu jednu reč ili je nastavio dalje i bez pokušaja ispravke. U slučaju audio-snimka, bilo bi moguće (iako ne idealno) da se problematični deo rečenice izbaci, ali takav diskontinuitet u video snimcima apsolutno ne bi bio prihvatljiv. Kod 2 govornika ovakve greške bile su vrlo česte, te su oni izbačeni iz baze. Međutim, kod govornika kod kojih su ovakve greške bile sporadične, odgovarajući snimak bio bi izbačen u slučaju da je rečenica u delu korpusa zajedničkom za sve govornike, a u slučajevima kada se radilo o rečenici koja je namenjena samo jednom govorniku, a kada je to bilo moguće, transkript je bio modifikovan prema onome što je govornik zaista izgovorio. Utvrđeno je da postoji 2% nedostajućih vrednosti i u srpskom i u engleskom delu baze sa rečenicama zajedničkim za sve govornike (uključujući i jednu rečenicu koja je izbačena zbog oštećenog audio-snimka). U srpskom delu baze sa personalizovanim rečenicama postoji 0,3% nedostajućih vrednosti (uključujući jednu rečenicu sa oštećenim audio-snimkom), osnosno 1% rečenica sa izmenjenim transkriptom. U engleskom delu baze sa personalizovanim rečenicama ima po 1% nedostajućih rečenica i rečenica sa izmenjenim transkriptom.

IV. Rezultat

Finalni korpus se sastoji od 160 foldera po govorniku, gde svaki folder sadrži audio-snimak i tri video-snimka iste rečenice, uz transkripciju i prateće metapodatke (putanja do foldera, tekst rečenice na koju se folder odnosi, napomene

Slika 3. Korigovanje granica segmenata i prikaz mogućih scenarija (sleva nadesno): segment za brisanje, potpuno ispravan segment i segment koji sadrži i više puta pogrešno izgovorenu rečenicu, a zatim ispravno izgovorenu rečenicu.

ukoliko postoje neka odstupanja od očekivanog). Iako je snimljeno 35 govornika, u bazi je zadržano 30 govornika, 15 muških i 15 ženskih, od čega kod 3 govornika nedostaju snimci sa jedne od pomoćnih kamera. Okvirno postoji 10 minuta govora za svakog govornika za svaki od dva jezika, dakle ukupno oko 20 minuta po govorniku. Za svaki jezik je snimljeno po govorniku 80 snimaka (75 rečenica plus alfabet, cifre, dani u nedelji, pravci i komande), ali u proseku postoji po 1-2% nedostajućih vrednosti po grupi rečenica.

V. ZAKLJUČAK

AI-SPEAK baza će u okviru AI-SPEAK projekta biti javno otvorena naučnoj zajednici za istraživanje. Koliko su autori upućeni, ovo će biti prva javno dostupna bilingvalna multimodalna baza za istraživanje govora. Takođe, veoma značajan, a jedinstven deo ovakve baze čini dostupnost video-snimka usana govornika iz tri različita ugla što otvara razne mogućnosti za istraživanje. Ovaj korpus predstavlja vredan resurs za dalja istraživanja u oblasti govornih tehonologija i multimodalne analize govora.

ZAHVALNICA

Istraživanje sprovedeno uz podršku Fonda za nauku Republike Srbije, #7749, AI-SPEAK.

LITERATURA

[1] Sumby, William H., and Irwin Pollack. "Visual contribution to speech intelligibility in noise." *The journal of the acoustical society of america* 26.2 (1954): 212-215.

[2] Nemani, Praneeth, G. Sai Krishna, and Supriya Kundrapu. "Automated Speaker Independent Visual Speech Recognition: A Comprehensive Survey." *arXiv preprint arXiv:2306.08314* (2023).

[3] Matthews, Iain, et al. "Extraction of visual features for lipreading." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.2 (2002): 198-213.

[4] Patterson, Eric K., et al. "CUAVE: A new audio-visual database for multimodal human-computer interface research." *2002 IEEE International conference on acoustics, speech, and signal processing*. Vol. 2. IEEE, 2002.

[5] Zhao, Guoying, Mark Barnard, and Matti Pietikainen. "Lipreading with local spatiotemporal descriptors." *IEEE Transactions on Multimedia* 11.7 (2009): 1254-1265.

[6] Cooke, Martin, et al. "An audio-visual corpus for speech perception and automatic speech recognition." *The Journal of the Acoustical Society of America* 120.5 (2006): 2421-2424.

[7] Anina, Iryna, et al. "Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis." *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. Vol. 1. IEEE, 2015.

[8] Lee, Bowon, et al. "AVICAR: audio-visual speech corpus in a car environment." *Interspeech*. 2004.

[9] Ivanko, Denis, et al. "RUSAVIC Corpus: Russian audio-visual speech in cars." *Proceedings of the thirteenth language resources and evaluation conference*. 2022.

[10] Dai, Wenliang, et al. "Ci-avsr: A cantonese audio-visual speech dataset for in-car command recognition." *arXiv preprint arXiv:2201.03804* (2022).

[11] Chung, Joon Son, and Andrew Zisserman. "Lip reading in the wild." *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*. Springer International Publishing, 2017.

[12] Afouras, Triantafyllos, et al. "Deep audio-visual speech recognition." *IEEE transactions on pattern analysis and machine intelligence* 44.12 (2018): 8717-8727.

[13] Afouras, Triantafyllos, Joon Son Chung, and Andrew Zisserman. "LRS3-TED: a large-scale dataset for visual speech recognition." *arXiv preprint arXiv:1809.00496* (2018).

[14] Yang, Shuang, et al. "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild." *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*. IEEE, 2019.

[15] Schwiebert, Gerald, et al. "A multimodal German dataset for automatic lip reading systems and transfer learning." *arXiv preprint arXiv:2202.13403* (2022).

[16] Park, Jeongkyun, et al. "OLKAVS: an open large-scale Korean audio-visual speech dataset." *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

[17] https://www.iac-nordic.dk/downloads/lyddoede-rum/IAC%20Mini%20Anechoic%20Chambers.pdf.

Recording bilingual database AI-SPEAK for multimodal speech recognition

Tijana Nosek, Siniša Suzić, Vuk Stanojev, Nikša Jakovljević, Lidija Krstanović i Milan Sečujski

ABSTRACT

This paper describes the recording procedure as well as the final content of a bilingual audio-visual database in Serbian and English. Although there are numerous audio-visual databases of varying sizes for the English language, this is the first example of such a database in Serbian. In addition to its bilingual nature, another advantage of the presented database lies in the fact that the video recordings are available from multiple angles — frontal, as well as from the left and right sides relative to the main camera. The development process of the database is described in detail, along with the post-processing and all the issues that arose during the process.

Analiza indeksa ljudskog razvoja i njegovih komponenti korišćenjem metoda mašinskog učenja

Dragana Radojičić
Ekonomski fakultet, Univerzitet u Beogradu
Beograd, Republika Srbija
dragana.radojicic@ekof.bg.ac.rs
0000-0001-7850-2623

Mladen Stamenković
Ekonomski fakultet, Univerzitet u Beogradu
Beograd, Republika Srbija
mladen.stamenkovic@ekof.bg.ac.rs
0000-0002-3838-878X

***Apstrakt* - Indeks ljudskog razvoja je pokazatelj koji služi za rangiranje zemalja prema nivou njihovog ljudskog razvoja i predstavlja meru napretka zemlje u pogledu kvaliteta životnog standarda njenih stanovnika. U okviru ovog istrazivanja koristimo bazu podataka koja pruža informacije o indeksu ljudskog razvoja za 195 zemalja, kao i podatke o očekivanom životnom veku, predviđenim godinama školovanja i bruto nacionalnom dohotku, koji će biti ključni za dalja istraživanja. Ideja ovog rada je koristeći različite tehnike mašinskog učenja analiziramo komponente indeksa ljudskog razvoja, kao i socio-ekonomskih faktora koji utiču na razvoj zemalja. Rezultati grupisanja metodom K-srednjih vrednosti ukazuju da zemlje sa višim indeksom ljudskog razvoja i bruto domaćim proizvodu po glavi stanovnika pripadaju klasterima koji se razlikuju u poređenju sa onima sa nižim vrednostima, naglašavajući značajne socio-ekonomske razlike između dobijenih klastera. Dalje, posmatrane podatke analiziramo koristeći algoritam slučajnih šuma kako bi ispitali uticaj posmatranih komponenti na indeks ljudskog razvoja.**

Ključne reči – Mašinsko učenje, Indeks ljudskog razvoja, Analiza glavnih komponenti, Klasterizacija metodom K-srednjih vrednosti, Algoritam slučajnih šuma.

I. UVOD

Indeks ljudskog razvoja, skraćeno IHR, (engl. Human Development Index (HDI)) predstavlja meru koja se koristi za procenu razvoja država, uzimajući u obzir tri ključna aspekta, naime: zdravlje stanovništva, obrazovanje, i poslednja komponenta je bruto nacionalni dohodak po glavi stanovnika (izražen u američkim dolarima). IHR služi kao sredstvo za poređenje socioekonomskog stanja različitih zemalja, i može pružiti korisne informacije za analizu politika posmatranih zemalja i unapređenje kvaliteta života. IHR se određuje kao geometrijska sredina normalizovanih indeksa tri ključne komponente. Zemlje se na osnovu vrednosti IHR-a obično svrstavaju u četiri kategorije:

- Vrlo visok ljudski razvoj (IHR viši ili jednak od 0,800)
- Visok ljudski razvoj (IHR od 0,700 do 0,799)
- Srednji ljudski razvoj (IHR između 0,550 i 0,699)
- Nizak ljudski razvoj (IHR ispod 0,550).

IHR je važna mera za praćenje napretka zemalja u pogledu ljudskog razvoja, ali se često koristi u kombinaciji s drugim indikatorima kako bi se dobila sveobuhvatna slika o stanju u određenoj zemlji. Program Ujedinjenih naroda za razvoj (UNDP) je 1990. godine u godišnjem izveštaju o ljudskom razvoju (Human Development Report) prvi put predstavio IHR. Kasnije, 2010.godine UNDP uvodi IHR prilagođen nejednakostima (engl. Inequality-adjusted Human Development Index (IHDI)), koji uzima u obzir nejednakosti u području zdravlja, obrazovanja i dohotka unutar pojedinih zemalja. IHR je prihvaćen kao značajnija mera razvoja, a njegova važnost i relevantnost su prepoznatljive. Kako bi se IHR proširio i obuhvatio više aspekata ljudskog razvoja, UNDP je uveo dodatne indekse, kao što su Indeks humanog razvoja prilagođen za nejednakost polova, Indeks humanog razvoja prilagođen za nejednakost raspodele, itd. UNDP i dalje radi na modernizaciji IHR-a kako bi odražavao savremene izazove, poput: klimatskih promena, digitalne revolucije, ekoloških razmatranja i itd.

Mnogi istraživači su pokazali interesovanje za proučavanje Indeksa ljudskog razvoja zbog njegovog značaja kao sveobuhvatne mere ljudskog blagostanja i razvoja jedne zemlje. Njihove studije često se usmeravaju na analizu faktora koji utiču na IHR, njegove promene kroz vremena i uticaj na oblikovanje politike i održivi razvoj. U radu [1] autori analiziraju IHR, razmatrajući njegove komponente, strukturu i kritike, i predlažu alternativne indekse za merenje ljudskog razvoja bazirane na poboljšanju komponenti indeksa. Studija u radu [2] koristi kvantitativni pristup sa deskriptivnom analizom i prostornom regresijom kako bi se ispitali faktori koji utiču na indeks ljudskog razvoja u Indonezii pre i tokom pandemije COVID-19. U literaturi [3] istražuje se razvoj Indeksa ljudskog razvoja, kao mera društveno-ekonomskog napretka, i prvi put se uvodi Indeks političke slobode. Rad [4] ističe da rast BDP-a značajno poboljšava IHR doprinoseći poboljšanju blagostanja ljudi, na osnovu analiza na podacima iz Indonezije.

Ograničenja i nedostaci izveštaja o humanom razvoju analizirani su u radu [5], s posebnim osvrtom na odstupanja izveštaja od njegove prvobitno predviđene uloge, i na to da IHR ne uspeva verno da odrazi stvarnost koju bi trebalo da meri. Brojni naučni radovi ispituju potencijalna poboljšanja indeksa, pri čemu je ova tema prvi put pokrenuta neposredno nakon izveštaja iz 1990. godine u nekoliko akademskih članaka ([6], [7], [8], [9], [10], itd.). Kako je mašinsko učenje počelo da se primenjuje u različitim domenima, istraživači su takođe iskoristili njegov potencijal da istraže indeks ljudskog razvoja. Klasifikacija IHR-a u literaturi često uključuje subjektivno prosuđivanje i podložna je kritici, autori rada [11] koriste klasterizaciju metodom K-srednjih vrednosti i algoritme K-medoida vođene podacima da grupišu IHR u tri klastera, minimizirajući subjektivnost u procesu klasifikacije. U radu [12] autori koristeći klasterizaciju metodom K-srednjih vrednosti dolaze do tri različite grupe regencija u Centralnoj Javi: oblasti visog IHR, oblasti srednjeg IHR i oblasti niskog IHR. Koristeći dva modela slučajnih šuma,

jedan za regresiju i jedan za klasifikaciju, studija [13] analizira indeks ljudskog razvoja u cilju razumevanja dinamike globalnog razvoja. Autori rada [14] su implementirali različite algoritme mašinskog učenja za predviđanje indeksa ljudskog razvoja i testirali su klasterizaciju metodom K-srednjih vrednosti i hijerahijska klaster analizu da grupiše vrednosti IHR indikatora iz 186 zemalja u četiri oznake. Studija u radu [15] grupiše odabrane HDI indikatore u 6 klastera u svim okruzima u Istočnoj Nusa Tengari, kako bi bile identifikovane oblasti za sprovođenje odgovarajućih politika.

U ovom radu ideja je da korišćenjem algoritma klasterizacije metodom K-srednjih vrednosti identifikujemo klastere zemalja prema njihovim sličnostima u razvojnim pokazateljima i izdvojimo grupe zemalja koje imaju sličan nivo razvoja.

II. DESKRIPTIVNA ANALIZA

A. Baza podataka

Za potrebe ovog istraživanja korišćena je baza podataka „Human Development Index and Components 2021“, preuzeta sa sajta https://www.kaggle.com/ (datum pristupa: 01.12.2024.godine). Baza sadrži podatke o indeksu ljudskog razvoja, kao i podatke o njegovim komponenetama: očekivanom životnom veku, prosečnom broju godina školovanja i bruto nacionalnom dohotku, za 195 zemalja u 2021. godini. Posmatrana baza sadrži podatke o indikatorima za procenu i poređenje nivoa razvoja i kvaliteta života u različitim zemljama sveta.

B. Analize baze i deskriptivne statistike

Pre početka same analize podataka, proveravamo da li baza sadrži nedefinisane ili nedostajuće podatke. Da bismo sprečili negativan uticaj na rezultate, uklanjamo sve redove u kojima se pojavljuje barem jedna nedostajuća numerička vrednost. Kako bismo dobili jasniju sliku o posmatranim podacima, pregledajmo deskriptivne statistike numeričkih karakteristika prisutnih u posmatranoj bazi podataka koje su tabelarno prikazane na Slici 1.

	Human Development Index (HDI)	Life expectancy at birth	Expected years of schooling	Mean years of schooling	Gross national income (GNI) per capita
count	191.0000	191.0000	191.0000	191.0000	191.0000
mean	0.7206	71.3147	13.5304	8.9838	20249.0942
std	0.1507	7.6465	2.9260	3.1732	21825.2641
min	0.3850	52.5000	5.5000	2.1000	732.0000
25%	0.5995	65.7500	11.6000	6.2500	4593.0000
50%	0.7390	71.7000	13.4000	9.3000	12306.0000
75%	0.8350	76.7000	15.6000	11.5000	30079.5000
max	0.9620	85.5000	21.1000	14.1000	146830.0000

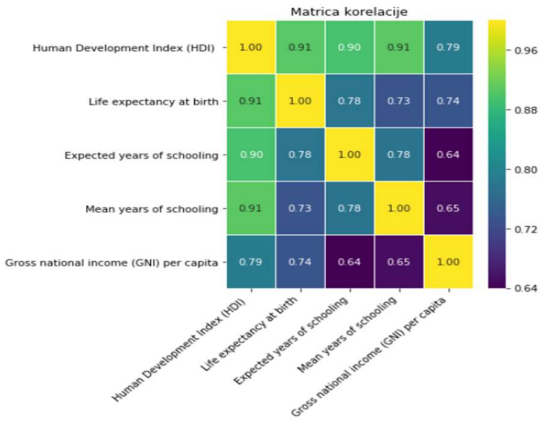
Slika 1. Prikaz deskriptivnih statistika numeričkih atributa

Konkretno, izdvajamo i podatke za Srbiju, što je prikazano na Slici 2.

Country	HUMAN DEVELOPMENT	Human Development Index (HDI)	Life expectancy at birth	Expected years of schooling	Mean years of schooling	Gross national income (GNI) per capita
Serbia	VERY HIGH	0.802	74.2	14.4	11.4	19123.0

Slika 2. Podaci o indikatorima o IHR za Srbiju

Kako bi se prikazala međusobna veza između posmatranih karakteristika i eventualna redundancija, na Slici 3 prikazane su međusobne individualne korelacije varijabli.



Slika 3. Međusobne individualne korelacije varijabli

Možemo primetiti jaku i veoma jaku korelaciju između varijabli, što je zapravo sasvim očekivan rezultat, s obzirom da posmatrani indikatora mere slične aspekte.

III. KLASTERIZACIJA METODOM K-SREDNJIH VREDNOSTI

Algoritam klasterizacije metodom K-srednjih vrednosti (eng. K-means) omogućava analizu razvojnih pokazatelja i prepoznavanje zemalja koje se svrstavaju u slične razvojne grupe. Klasterizacija metodom K-srednjih vrednosti je algoritam nenadgledanog učenja koji omogućava grupisanje opservacija u K klastera prema njihovim sličnostima.

Klasterizacija metodom K-srednjih vrednosti se može predstaviti u sledećim fazama:

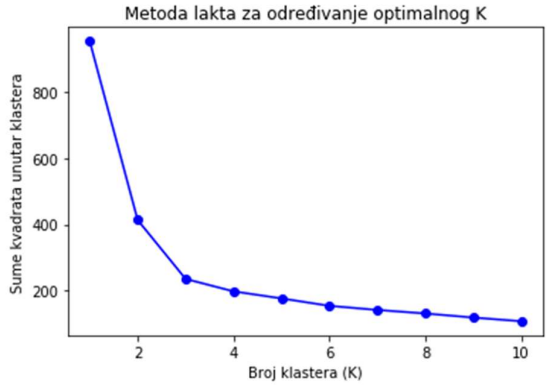
- *Odabir broja klastera* (može se odrediti na osnovu prethodnog znanja, metodom lakta, analizom podataka, itd.),
- *Inicijalizacija K početnih centroida* (središnjih tačaka) za klastere (obično se nasumično odredi K tačaka),
- *Dodeljivanje opservacija klasterima* (svaka opservacija iz posmatrane baze dodeljuje se najbližem klasteru najčešće korišćenjem Euklidske distance),
- *Ažuriranje centara klastera* (centroida svakog klastera se ponovo izračunava kao prosek svih opservacija koje su u tom klasteru),
- *Ponovljavanje koraka 3 i 4* (Ovi koraci se ponavljaju dok se pozicije centroida više ne menjaju ili dok se ne dostigne maksimalni broj iteracija).

Ovaj algoritam spada među najstarije algoritme za klasterovanje – njegova osnovna ideja datira iz 1956. godine (videti [16]), iako je formalno dobio naziv u radu [17].

Prednosti algoritma klasterizacije metodom K-srednjih vrednosti su jednostavnost za implementaciju i razumevanje, efikasnost, brzina, itd. Međutim, algoritam ima i nedostatke, kao što su: potrebno je unapred odrediti broj klastere, osetljivost na outliere, zavisnostost od odabira inicijalnih tačaka, itd. U radu [18] autori predlažu poboljšani algoritam

klasterizacije metodom K-srednjih vrednosti, koji kombinuje algoritam najveće minimalne udaljenosti i tradicionalni algoritam klasterizacije metodom K-srednjih vrednosti, koji prevazilazi nedostatke tradicionalnog algoritma klasterizacije metodom K-srednjih vrednosti za inicijalizaciju početnih centroidnih tačaka. Rad [19] prikazuje sveobuhvatnu sliku i analizu algoritma klasterizacije metodom K-srednjih vrednosti, ispitujući njegove prednosti i nedostatke.

Postupak klasterovanja metodom K-srednjih vrednosti počinjemo selekcijom numeričkih kolona, i standardizacijom njihovih vrednosti, zatim biramo broj klastera metodom lakta. Posmatranjem grafikona prikazanog na Slici 3 optimalan broj klastera K determinišemo koristeći metod lakta. Na grafikonu su prikazane promene suma kvadrata odstupanja tačaka od centroida u klasterima, i dalje tražeći prevojnu tačku određujemo broj klastera. Primećujemo da je uočljivo prelamanje linije oko vrednosti 3 na *x* osi, i nakon toga algoritam se pokreće za K=3, pa se podaci raspoređuju u tri klastera.



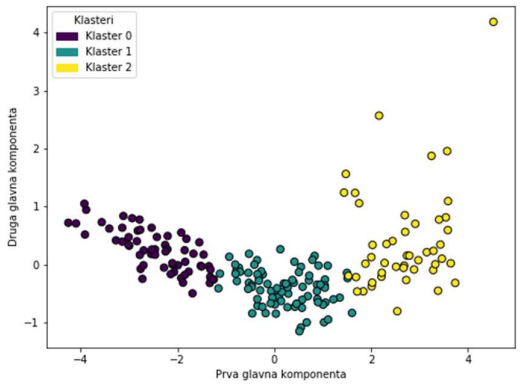
Slika 3. Determinisanje optimalnog broja klastera

Kako bismo utvrdili specifičnosti i razlike između klastera izračunavamo prosečne vrednosti obeležja po klasterima, a rezultati su prikazani na Slici 4. Treba imati na umu da indeksiranje serija i nizova u programskom jeziku Python počinje od broja 0, pa je na Slici 4 prvi klaster je indeksiran '0', itd. Možemo zaključiti da treći klaster obuhvata zemlje sa najvišim vrednostima svih posmatranih karakteristika, dok se u prvom klasteru nalaze zemlje sa najnižim vrednostima posmatranih parametara.

Cluster	Human Development Index (HDI)	Life expectancy at birth	Expected years of schooling	Mean years of schooling	Gross national income (GNI) per capita
0	0.535250	62.866667	10.363333	5.220000	3576.800000
1	0.748798	72.116667	13.992857	9.915476	14480.404762
2	0.906723	80.665957	16.746809	12.123404	51842.872340

Slika 4. Prosečne vrednosti karakteristika u dobijenim klasterima

S obzirom da su naši podaci visokodimenzionalni, kako bismo redukovali dimenzionalnost i vizuelno prikazali klastere, a da sačuvamo što više varijanse iz originalnog skupa podataka, primenjujemo analizu glavnih kompomnenti. Analiza glavnih komponenti omogućava projekciju višedimenzionalnih podataka uz pomoć najvažnijih glavnih komponenti. Prikaz raspodela zemalja po klasterima u odnosu na prve dve glavne komponente je prezentovan na Slici 5.



Slika 5. Vizualizacija klasterovanja metodom K-srednjih vrednosti pomoću glavnih komponenti

IV. SLUČAJNE ŠUME

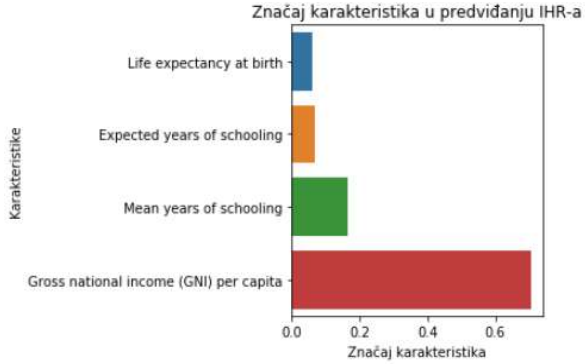
Slučajne šume (engl. random forest) je algoritam mašinskog učenja zasnovan na ansambl metodi, što zapravo znači da kombinuje više modela kako bi poboljšao tačnost i pouzdanost pri rešavanju kompleksnih zadataka. Ovo je algoritam proste agregacije i baziran je na treniranju više stabala odlučivanja (tzv. ansambli), koristeći nasumično izabrane podskupove stabala, a ponekad i nasumične podskupove posmatranih karakteristika.

Algoritam slučajne šume se može predstaviti u sledećim fazama:

- *Generisanje više stabala odlučivanja:* kreira veliki broj stabala odlučivanja, zatim se svako stablo obučava na različitim podskupovima podataka.
- *Slučajni izbor karakteristika:* slučajnim izborom selektuje se podskup karakteristika iz skupa svih prisutnih karakteristika.
- *Treniranje stabala odlučivanja:* svako stablo se trenira na drugačijem podskupu podataka i karakteristika.
- *Agregacija rezultata:* u slučaju regresionih zadataka rezultati pojedinačnih stabala se uprosečavaju, a u slučaju klasifikacionih zadataka odlučuje na osnovu većinskog glasanja.

Ključna ideja je da se za treniranje svakog stabla koriste različiti uzorci podataka i atributa, što omogućava smanjenje varijanse i poboljšanje tačnosti. Slučajne šume su korisne i imaju široku primenu zbog prednosti koje ima ovaj algoritam kao sto su: velika otpornost na nedovoljno prilagođavanje (engl. underfitting) i preterano prilagođavanje (engl. overfitting), otpornost na promene u podacima, pruža dobre rezultate i kada neki podaci nedostaju, može da oceni važnost svake posmatrane karakteristike u bazi podataka, itd. S druge strane ovaj algoritam ima i nedostatke kao što su: proces treniranja je nekada računarski zahtevan, gubitak interpretabilnosti, zahteva puno memorije, itd. S obzirom da se algoritam slučajnih šuma može koristiti za evaluaciju važnosti karakteristika, koristimo algoritam da analiziramo uticaj posmatranih komponenata iz naše baze na indeks humanog razvoja. Najpre definišemo novi kategorički atribut „HDI Rank“ i dodajemo ga u posmatranu bazu za svaku prisutnu opservaciju, tako što u zavisnosti od intervala kome vrednost IHR-a pripada (pomenutih u prvom delu) preslikavamo u vrednost ’4’ ukoliko je vrednost iz kategorije ’Vrlo visok ljudski razvoj’, u ’3’ ukoliko je vrednost iz

kategorije ’Visok ljudski razvoj’, u ’2’ ukoliko je vrednost iz kategorije ’ Srednji ljudski razvoj’, u ’1’ ukoliko je vrednost iz kategorije ’ Nizak ljudski razvoj’. Možemo uočiti da bruto nacionalni dohodak po glavi stanovnika (engl. Gross national income per capita) ima najznačajniji uticaj na "IHR Rank", pa samim tim i na vrednost IHR-a, što je i logično, imajući u vidu samu prirodu njegove definicije. Prikaz uticaja posmatranih atributa na varijablu 'HDI Rank' je prikazan na Slici 6. Posmatrani model zasnovan na algoritmu slučajnih šuma pokazuje da bruto nacionalni dohodak po glavi stanovnika ima izrazito značajan uticaj na IHR, što se poklapa i sa metodologijom izračunavanja IHR-a. Važnost karakteristika u modelu slučajne šume pokazuje koliki uticaj svaka pojedinačna karakteristika ima na donošenje odluka unutar modela.



Slika 6. Uticaj karakteristika na IHR

V. ZAKLJUČAK

Budući da indeks ljudskog razvoja odražava socioekonomski status zemlje, analiza njegovih komponenti i uticaja na sam indeks je ključna za identifikaciju oblasti koje zahtevaju poboljšanje kvaliteta života. Primenom metode K-srednjih vrednosti identifikove su grupe zemalja sličnim profilima razvoja koje se odlikuju sličnim razvojnim karakteristikama. Analiza je pokazala da zemlje sa višim vrednostima IHR-a i bruto nacionalnog dohodka po glavi stanovnika formiraju posebne klastere u odnosu na zemlje sa nižim vrednostima, upravo to ukazuje na socio-ekonomske razlike između posmatranih zemalja. Dalje, primenom algoritma slučajne šume ocenjujemo značaj karakteristika i zaključujemo da bruto nacionalni dohodak po glavi stanovnika ima najsnažniji uticaj u ocenjivanju kategorije IHR-a. Dalje, primenom algoritma slučajne šume ocenjujemo značaj karakteristika i zaključujemo da bruto nacionalni dohodak po glavi stanovnika ima najsnažniji uticaj u ocenjivanju kategorije IHR-a. S obzirom da tehnike mašinskog učenja mogu ukazati i izdvojiti obrasce između IHR-a i socio-ekonomskih indikatora, daljim razvojom i primenom tih modela na podacima o IHR-a i njegovim komponentama mogu se izvesti relacije što može dati smernice za unapredjenja.

LITERATURA

[1] F. Noorbakhsh. “The human development index: some technical issues and alternative indices,” Journal of International Development: The Journal of the Development Studies Association, vol. 10.5, str. 589-605, 1998.

[2] S. Astari. “Spatial Analysis of The Human Development Index in Indonesia Before and During The Covid-19 Pandemic,” IOP Conference Series: Earth and Environmental Science. IOP Publishing, str. 012002, 2024.

[3] M. Ul Haq, “Reflections on human development,” Oxford university Press, 1995.I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 2003, str. 271–350.

[4] M. B. Setiawan, and A. Hakim, “Indeks pembangunan manusia Indonesia,” Jurnal Economia, vol. 9(1), str. 18-26, 2015.

[5] S. Anand, and A. Sen, “Human Development Index: Methodology and Measurement,” 1994.

[6] D. P. Doessel, and R. Gounder, “International comparisons of the standards of living and the human development index,” Discussion Papers in Economics, vol. 72, str. 1212-1217, 1991.

[7] M. Hopkins, “Human development revisited: A new UNDP report,” World Development, vol. 19(10), str. 1469-1473, 1991.

[8] N. C. Lind, “Some thoughts on the human development index,” Social Indicators Research, vol. 27, str. 89-101, 1992.

[9] G. Pyatt, “Poverty: a wasted decade,” European economic review, vol. 35(2-3), str. 358-365, 1991.

[10] H. Wang, J.H. Feil, and X. Yu, “Let the data speak about the cut-off values for multidimensional index: Classification of human development index with machine learning,” Socio-Economic Planning Sciences, vol. 87, str. 101523, 2023.

[11] R.T. Vulandari, S. Siswanti, A.K. Kusumawijaya, and K. Sandradewi, “Classification of human development index using k-means,” Indonesian Journal of Applied Statistics, vol. 2(1), str. 1-9, 2019.

[12] J. Gsim, and M.Z. Es-sadek, “Machine Learning Projections for Human Development Index Anticipation,” 2024. https://doi.org/10.21203/rs.3.rs-4376154/v1D.

[13] F.B. Khan, and A. Noor, “Prediction and Classification of Human Development Index Using Machine Learning Techniques,” In 2021 5th International Conference on Electrical Information and Communication Technology (EICT), IEEE, str. 1-6, decembar 2021.

[14] J.E. Simarmata, D. Chrisinta, and M. Purnomo, “Implementation of K-Means Clustering to Human Development Indicators in East Nusa Tenggara,” Journal of Research in Mathematics Trends and Technology, vol. 6(2), str. 46-56, 2024.

[15] Y. Li, and H. Wu, “A clustering method based on K-means algorithm,” Physics Procedia, vol. 25, str. 1104-1109, 2012.

[16] H. Steinhaus, “Sur la division des corps matériels en parties,” Bull. Acad. Polon. Sci. (in French), vol. 4 (12), str. 801–804, 1957.

[17] J. B. MacQueen, “Some Methods for classification and Analysis of Multivariate Observations,” Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press, MR 0214227, Zbl 0214.46201, str. 281–297, 1967.

[18] J. Cui, J. Liu, and Z. Liao, “Research on K-means clustering algorithm and its implementation,” In Conference of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013), Atlantis Press, str. 1804-1806, mart 2013.

[19] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,” Information Sciences, vol. 622, str. 178-210, 2023.

Analysis of human development index and its components using machine learning methods

Dragana Radojičić, Mladen Stamenković

ABSTRACT

The Human Development Index is an indicator used to rank countries according to the level of their human development and is a measure of the country's progress in terms of the quality of living standards of its inhabitants. In this research, we use a database that provides information on the human development index for 195 countries, as well as data on life expectancy, projected years of schooling and gross national income, which will be key to further research. The idea of this paper is to use different machine learning techniques to analyze the components of the human development index, as well as socio-economic factors that influence the development of countries. The results of clustering using the K-means method indicate that countries with higher human development index and gross domestic product per capita belong to clusters that differ compared to those with lower values, highlighting significant socio-economic differences between the obtained clusters. Furthermore, we analyze the observed data using the random forest algorithm in order to examine the influence of the observed components on the human development index.

Integracija velikih jezičkih modela u obrazovanju: Percepcije studenata kroz praktične primene

Zdravko Ivanković
Visoka škola strukovnih studija: Sirmium
ivankovic.zdravko@gmail.com
0009-0003-4044-6445

Predrag Pecev
Visoka škola strukovnih studija: Sirmium
predrag.pecev@gmail.com
0009-0009-3784-442X

Saša Sudar
Visoka škola strukovnih studija: Sirmium
sasa.sudar@gmail.com
0009-0002-2601-2993

Jasmina Damnjanović
Visoka škola strukovnih studija: Sirmium
vs.jasmina.damnjanovic@gmail.com
0009-0003-5492-0849

Milana Ivanković
Visoka škola strukovnih studija: Sirmium
ivankovic.milana@gmail.com
0009-0002-5071-0935

Apstrakt **- Ova studija istražuje percepcije studenata o integraciji velikih jezičkih modela (Large Language Models - LLM) i sistema sa generacijom uz podršku pretrage (RAG - Retrieval Augmented Generation) u obrazovnom kontekstu. Predstavljena su dva studijska slučaja koja demonstriraju praktičnu primenu ovih tehnologija. Prvi slučaj koristi RAG metodologiju, oslanjajući se na model all-mpnet-base-v2 u kombinaciji sa LLaMA 2, kako bi omogućio interakciju sa udžbenikom o bazama podataka u formatu četa. On pruža korisnicima mogućnost pristupa informacijama putem prirodnih upita na jeziku korisnika. Drugi slučaj ističe primenu naprednog LLM-a, konkretno ChatGPT-4, u prevođenju prirodnih jezičkih upita u SQL izjave za direktnu interakciju sa bazom podataka. Povratne informacije od studenata prikupljene su kroz strukturisani upitnik, osmišljen za procenu njihovih iskustava i uvida u ove inovativne tehnološke pristupe.**

Ključne reči – LLM, RAG, Generativna AI, ChatGPT.

I. Uvod

Veliki jezički modeli (LLM) predstavljaju revolucionaran napredak u oblasti informacionih tehnologija, transformišući način na koji komuniciramo sa mašinama i pristupamo informacijama. LLM-ovi koriste arhitekturu dubokog učenja (Deep Learning), posebno transformator modele, koji primenjuju mehanizme za hvatanje kontekstualnih odnosa unutar jezika. Ova arhitektura omogućava LLM-ovima da obrađuju i generišu tekst sa izuzetnom koherentnošću i tečnošću, čineći ih pogodnim za širok spektar primena, od četbotova do generisanja sadržaja.

Obuka LLM-ova obuhvata ne nadgledano učenje, gde modeli uče obrasce i strukture jezika bez eksplicitnih oznaka. Ovo je dopunjeno finim podešavanjem za specifične zadatke, omogućavajući im prilagođavanje različitim domenima.

Iz perspektive informacionih tehnologija, LLM-ovi imaju značajne implikacije za razne primene, uključujući korisničku podršku, kreiranje sadržaja i analizu podataka. Oni mogu automatizovati zadatke koji su tradicionalno zahtevali razumevanje jezika od strane ljudi, pojednostavljujući tokove rada i povećavajući produktivnost. LLM-ovi se integrišu u četbotove, virtuelne asistente i sisteme za preporuke, pružajući korisnicima intuitivne interakcije i personalizovana iskustva.

Primena LLM-ova zahteva razmatranje infrastrukture, poput mogućnosti cloud servera i skalabilnih arhitektura. Organizacije često koriste moćne GPU-ove i distribuirane sisteme kako bi efikasno obučavale i koristile ove modele. Pored toga, integracija LLM-ova u postojeće sisteme zahteva pažnju na API-je, privatnost podataka i usklađenost sa regulatornim standardima.

II. Pregled oblasti istraživanja

Veliki jezički modeli (LLM) privukli su značajnu pažnju tokom proteklih godina zbog svojih izuzetnih sposobnosti u obradi prirodnog jezika (Natural Language Processing - NLP). Ovo poglavlje daje pregled ključnih razvoja u oblasti LLM-ova, ističući evoluciju arhitektura, metodologija obuke i primena.

Evolucija LLM-ova može se pratiti od uvođenja Transformer arhitekture, koju je u svom radu predstavio Vaswani [1]. Ova arhitektura zamenila je rekurentne neuronske mreže (RNN) mehanizmom samopažnje, značajno poboljšavajući efikasnost i efektivnost zadataka modeliranja jezika. Prva implementacija Transformera otvorila je put za različite adaptacije, uključujući BERT (Bidirectional Encoder Representations from Transformers) [2] i GPT (Generative Pre-trained Transformer) [3][4]. BERT, koji su predstavili Devlin i saradnici [2], koristio je modelovanje maskiranog jezika za postizanje vrhunskih rezultata na nekoliko NLP standardnih testova.

Dalji napredak postignut je uvođenjem GPT-2 i GPT-3, koji su pokazali potencijal skaliranja Transformer modela. Radford i saradnici [5] predstavili su GPT-2, demonstrirajući njegovu sposobnost generisanja koherentnog i kontekstualno relevantnog teksta. Na osnovu ovoga, Brown i saradnici [6] su predstavili GPT-3, model sa 175 milijardi parametara koji je postigao izvanredne performanse na različitim zadacima, ilustrujući prednosti skaliranja.

Metodologije obuke za LLM-ove takođe su značajno evoluirale. Paradigma predobuke i finog podešavanja postala je standardni pristup. Predobuka podrazumeva ne nadgledano učenje iz ogromnih količina tekstualnih podataka, dok fino podešavanje prilagođava model specifičnim zadacima korišćenjem označenih skupova podataka. Ovaj pristup je uspešno primenjen u BERT-u, koji je bio predobučen na BooksCorpus i engleskoj Vikipediji [2].

Koncept transfernog učenja u NLP-u dobio je na značaju, omogućavajući modelima obučenim za jedan zadatak da

dobro izvršavaju druge. Liu i saradnici [7] diskutovali su o različitim tehnikama transfernog učenja koje su uspešno primenjene na LLM-ove, naglašavajući važnost finog podešavanja specifičnog za zadatak. Dalje, inovativne tehnike kao što su učenje sa nekoliko primera (few-shot learning) i bez primera (zero-shot learning) istražene su, posebno sa pojavom GPT-3, koji je pokazao impresivne performanse sa minimalnim primerima specifičnim za zadatak [8].

LLM-ovi su našli primenu u širokom spektru oblasti, uključujući generisanje teksta, analizu sentimenta, mašinski prevod i odgovaranje na pitanja. U obrazovnom okruženju, LLM-ovi se sve više koriste za personalizovano učenje i podučavanje. Na primer, sistemi pokretani LLM-ovima mogu pružiti prilagođene povratne informacije i pomoć studentima, poboljšavajući njihovo iskustvo učenja.

U domenu zdravstvene zaštite, LLM-ovi su korišćeni za analizu kliničkog teksta i podataka o pacijentima. Pored toga, etička razmatranja u vezi sa primenom LLM-ova naširoko su diskutovana, posebno u vezi sa pristrasnošću i pravičnošću izlaza modela [9].

Uprkos impresivnim sposobnostima LLM-ova, i dalje postoje brojni izazovi. Troškovi obrade povezani sa obukom i primenom ovih modela su značajni, izazivajući zabrinutost za ekološku održivost. Nedavna studija [10] istakla je ugljenični otisak povezan sa obukom modela velikih razmera. Dodatno, problemi vezani za interpretaciju modela i rizik od generisanja štetnog sadržaja zahtevaju dalja istraživanja. Trenutni napori usmereni su na razvoj robusnih metrika evaluacije i tehnika kako bi se osigurala odgovorna upotreba LLM-ova u stvarnim aplikacijama [11].

Generacija uz podršku pretrage (RAG) predstavlja novi pristup u obradi prirodnog jezika, integrišući mehanizme pretrage sa generativnim modelima kako bi se poboljšao kvalitet i relevantnost generisanih odgovora. Ovo poglavlje razmatra ključne doprinose u ovoj oblasti, ilustrujući evoluciju RAG-a i njegov uticaj na različite primene. RAG kombinuje snage pretrage informacija i generisanja teksta, omogućavajući modelima da dinamički pristupe eksternim bazama znanja tokom procesa generacije.

Okvir RAG popularizovali su Piktus i saradnici [12], koji su pokazali da RAG modeli mogu iskoristiti velike skupove podataka kako bi proizveli preciznije i kontekstualno relevantne rezultate u poređenju sa tradicionalnim generativnim modelima.

Razvijeni su različiti pristupi za efektivnu implementaciju RAG-a. Na primer, rad [13] predstavio je dvostepeni okvir gde se relevantni dokumenti pretražuju iz korpusa, koji se zatim koriste za uslovljavanje generisanja odgovora. Ovaj pristup značajno poboljšava performanse na zadacima otvorenog odgovaranja na pitanja.

Poslednja decenija donela je izvanredan napredak u razvoju i primeni LLM-ova, sa značajnim implikacijama za različite domene. Buduća istraživanja će se verovatno fokusirati na povećanje efikasnosti modela, rešavanje etičkih pitanja i proširenje primenljivosti LLM-ova na nove izazove.

Takođe, uprkos obećavajućim rezultatima koje je postigao RAG, i dalje postoje brojni izazovi. Jedno od ključnih pitanja je efikasnost procesa pretrage, koji može postati usko grlo, posebno u aplikacijama u realnom vremenu. Buduća istraživanja treba da se fokusiraju na optimizaciju algoritama pretrage i istraživanje alternativnih tehnika indeksiranja kako bi se poboljšale performanse.

Integracija mehanizama pretrage sa generativnim modelima u RAG-u otvorila je nove mogućnosti za unapređenje aplikacija u obradi prirodnog jezika. Kako istraživanje napreduje, rešavanje postojećih izazova biće od suštinskog značaja za ostvarivanje punog potencijala RAG modela u praktičnim scenarijima.

III. Primena RAG i LLM modela

Modeli velikih jezičkih modela, poput GPT-a, mogu obavljati zadatke kao što su prevođenje, pisanje eseja, odgovaranje na pitanja i programiranje, prilagođavajući se različitim domenima kroz fino podešavanje na specifične podatke.

RAG modeli kombinuju sposobnosti generativnih modela (poput LLM-a) sa mehanizmima pretrage. Ovi modeli prvo pretražuju relevantne informacije iz eksternih izvora, a zatim koriste te podatke za generisanje odgovora ili teksta. Ova tehnika poboljšava tačnost i kvalitet informacija u odgovoru jer se generacija oslanja na ažurirane, preuzete informacije, čime se prevazilazi ograničenje tradicionalnih modela koji funkcionišu isključivo na unapred obučenim podacima.

Na osnovu akronima, RAG modeli imaju tri glavne komponente:

- Retrieval (Pretraga): Ova komponenta pretražuje velike baze podataka ili spoljne izvore, kao što su dokumenti ili veb stranice, kako bi pronašla relevantne informacije na osnovu zadatog upita, pružajući modelu pristup svežim i aktuelnim podacima.
- Augmented (Obogaćivanje): Informacije dobijene pretragom integrišu se sa prethodno obučenim znanjem modela, obogaćujući i proširujući kapacitet generativnog modela za preciznije i tačnije odgovore.
- Generation (Generisanje): Na osnovu informacija dobijenih pretragom i ulaznog upita, model generiše prirodan, koherentan tekst koji kombinuje prethodno znanje i aktuelne podatke.

RAG modeli mogu se koristiti za:

- Četbot za korisničku podršku: Korišćenjem postojeće dokumentacije korisničke podrške kao resursa, kada korisnik postavi pitanje, sistem može preuzeti relevantne delove dokumentacije i generisati odgovore pomoću LLM-a. Na primer, kompanija Klarna, koja se bavi finansijama, koristi ovakav sistem kako bi uštedela 40 miliona dolara godišnje na troškovima korisničke podrške.
- Analizu e-pošte: Na primer, osiguravajuća kompanija može imati dugačke nizove e-pošte između klijenata i agenata. Umesto da ručno pretražuje svaku pojedinačnu poruku, sistem može preuzeti relevantne odlomke i generisati strukturisane izlaze pomoću LLM-a.
- Čet o unutrašnjoj dokumentaciji kompanije: U velikim kompanijama ponekad je teško doći do odgovora. RAG sistem može indeksirati informacije o kompaniji i omogućiti LLM-u da odgovori na bilo koje postavljeno

pitanje. Prednost RAG-a je što pruža reference na resurse za dalje istraživanje ukoliko odgovor LLM-a nije dovoljan.

- Odgovori na pitanja iz udžbenika: Na primer, ukoliko se pripremate za ispite i stalno listate veliki udžbenik tražeći odgovore na svoja pitanja, RAG može pomoći tako što pruža odgovore i reference za dalje učenje.

U ovom radu predstavljene su dve vrste pretrage podataka. Prva je RAG metodologija gde se kroz četbot pretražuje knjiga, dok druga koristi OpenAI, sada u vlasništvu Microsofta, kako bi se pretraživala baza podataka prevodeći prirodni jezik u SQL upite.

A. Upiti nad knjigom

Arhitektura RAG modela kreirana je za postavljanje upita knjizi o bazama podataka. Knjigu su napisali Ramez Elmasri i Shamkant B. Navathe pod naslovom "Fundamentals of Database Systems", 7. izdanje, objavljena od strane PEARSON-a 2015. godine. Ova knjiga je odabrana jer ima 1273 stranice i obuhvata raznovrsne teme o bazama podataka sa primerima. Primenom kreirane arhitekture, postavljana su pitanja na koja je algoritam generisao odgovore koristeći sadržaj knjige.

U procesu obrade i preuzimanja podataka korišćeni su sledeći koraci arhitekture RAG modela:

Obrada dokumenata

- Učitavanje PDF dokumenta
- Formatiranje teksta, podela na manje delove (po 10 rečenica)
- Pretvaranje delova teksta u numeričke prikaze (vektore)

Pretraga i generisanje odgovora

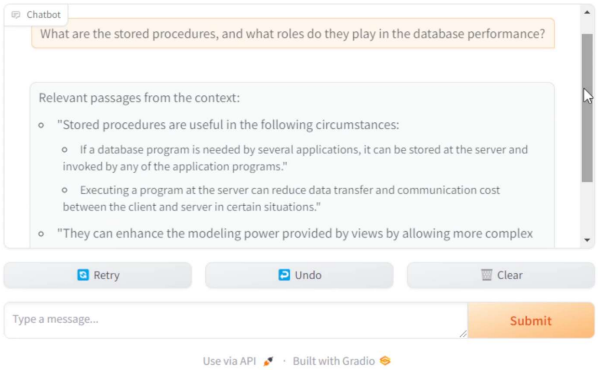
- Kreiranje sistema koji pretražuje vektore kako bi pronašao odgovarajuće odlomke teksta na osnovu upita
- Generisanje poruka koje uključuju pronađeni tekst
- Generisanje odgovora

Nakon uvoza prethodno pomenute knjige u PDF formatu i podele teksta na delove, korišćen je model all-mpnet-base-v2. All-mpnet-base-v2 je model veštačke inteligencije koji se koristi za razumevanje i poređenje rečenica. Njegov osnovni zadatak je da "prevede" rečenice u matematičke vektore, što pomaže u otkrivanju koliko su dve rečenice slične po značenju. Ovaj model je efikasan, precizan i koristi se za zadatke poput pretraživanja informacija, analize teksta, pa čak i detekcije plagijata.

Zatim, kako bi se izlaz all-mpnet-base-v2 modela preveo u prirodan govor (na engleskom jeziku), korišćen je model LLaMA 2. LLaMA 2 je model koji je razvila kompanija Meta (ranije poznata kao Facebook). Ovaj model je postao popularan jer je otvorenog koda, što znači da je dostupan svima za korišćenje i prilagođavanje. Jedna od najvećih prednosti modela LLaMA 2 je njegova fleksibilnost. Korišćen je za mnoge zadatke obrade jezika, uključujući:

- Prevođenje
- Sažimanje dugih tekstova
- Generisanje odgovora ili kreativnog pisanja

Na slici 1 prikazana je razvijena aplikacija koja sadrži četbot koji korisnik može da upita na prirodnom jeziku i dobije odgovor takođe na prirodnom jeziku. Baza podataka ovog četbota formirana je iz prethodno pomenute PDF knjige o bazama podataka.



Slika 1. Generisanje odgovora na upit pomoću RAG modela

B. Generisanje SQL upita nad bazom

Veliki jezički modeli (LLM) olakšavaju pisanje SQL upita omogućavajući korisnicima da postavljaju pitanja na prirodnom jeziku. To znači da nije potrebno da budete SQL ekspert kako biste dobili odgovore iz baze podataka – dovoljno je da objasnite šta želite, a LLM će sam generisati tačan upit. Ovaj proces je brži i jednostavniji jer automatski kreira složene upite bez potrebe za ručnim kodiranjem. Još jedna prednost je što LLM-ovi mogu raditi sa različitim tipovima baza podataka i prilagođavati upite njihovim pravilima. Takođe, smanjuje se potreba za obukom zaposlenih, jer više nije neophodno da znaju SQL, što štedi vreme i smanjuje broj grešaka.

Za implementaciju prethodno pomenutog LLM-a koji komunicira sa bazom podataka korišćen je upit: „Prikaži mi klijente za moje artikle, ali tako da vidim koliko klijenata imam u kojoj državi“. Kod je izvršen nad odabranom bazom podataka primenom OpenAI modela i dao je sledeće rezultate.

```
[('Argentina', 1), ('Australia', 1), ('Austria', 1), (
- Argentina: 1 customer
- Australia: 1 customer
- Austria: 1 customer
- Belgium: 1 customer
- Brazil: 5 customers
- Canada: 8 customers
- Chile: 1 customer
- Czech Republic: 2 customers
- Denmark: 1 customer
- Finland: 1 customer
- France: 5 customers
- Germany: 4 customers
- Hungary: 1 customer
- India: 2 customers
- Ireland: 1 customer
- Italy: 1 customer
- Netherlands: 1 customer
- Norway: 1 customer
- Poland: 1 customer
- Portugal: 2 customers
...
        Prompt Tokens: 1482
        Completion Tokens: 213
```

Slika 2. Generisanje odgovora na upit pomoću LLM modela

IV. Percepcija studenata kroz analizu upitnika

U anketiranju je učestvovalo 18 ispitanika, tj. studenata prve godine studija, studijskog programa Softversko inženjerstvo na Visokoj strukovnoj školi Sirmium u Sremskoj Mitrovici. Pre popunjavanja ankete, studenti su dali svoju pisanu saglasnost da prihvataju da učestvuju u navedenom istraživanju. Anketa je bila anonimnog karaktera. Ukupan broj pitanja sadržanih u anketi je 20. Svi ispitanici (18) su svoje odgovore dali na ukupno 19 pitanja, dok 1 ispitanik nije dao odgovor na jedno od ponuđenih pitanja u anketi. Anketom su ispitivani stavovi o primeni aplikacije za unapređenje kompetencija studenata u okviru kreiranog rešenja, „Jezički modeli (LLM) kao alat za unapređenje kompetencija studenata Visokih strukovnih škola“.

Polazeći od predmeta, cilja i zadataka, a na osnovu analize rezultata istraživanja, proizilaze sledeća zaključna razmatranja:

U okviru kategorije Opšta pitanja o tehnologiji, izdvajaju se ključni rezultati koji pokazuju da su:

- Studenti umereno upoznati sa konceptima tehnologije LLM i RAG (50%);

Kada se sagleda kategorija Korišćenje LLM i RAG tehnologija za tumačenje literature može se zaključiti da analizirani rezultate navode na sledeći zaključak:

- Studenti smatraju da LLM tehnologije mogu pomoći u boljem razumevanju kompleksne literature (66,7%);
- Približno isti broj studenata navodi da povremeno (27,8%), retko (27,8%) ili nikada (33,3%) ne koristi LLM tehnologije za dobijanje sažetaka i objašnjenja literature;
- Glavnim prednostima korišćenja LLM tehnologija u cilju dobijanja odgovora na postavljena pitanja, studenti smatraju brzinu (27,8%) i obuhvatnost informacija (27,8%);
- Interesantno je da studenti vrednuju tačnost podataka koji se generišu primenom LLM tehnologija kao približno istim sa ručnom pretragom (77,8%);

U analiziranim rezultatima kategorije Kreiranje upita nad bazama podataka uz pomoć LLM i RAG tehnologija može se zaključiti da:

- Da im LLM tehnologije ponekad olakšavaju formulisanje upita za pretragu baza podataka (72,9%);
- Studenti su zadovoljni mogućnošću kombinovanja informacija iz više izvora pomoću RAG tehnologija (66,7%);

Kada se sagleda kategorija Stavovi o tehnologijama u učenju, primetno je da se posebno izdvajaju sledeći značajni rezultati:

- Interesantno je da studenti smatraju da je najbolja kombinacija LLM i RAG tehnologija i tradicionalnog pristupa pretrazi informacija i procesu učenja (77,8%);
- Odgovor dobijen je u vezi sa stavovima studenata o korisnim stranama LLM i RAG tehnologija u odnosu na izučavanu nastavnu temu je da u specifičnim i naučnim disciplinama nisu korisne (33,3%) i da ne funkcionišu dobro u većini slučajeva (33,3%) ili da nemaju dovoljno iskustva (27,8%) sa njihovom primenom. Ovakvi rezultati navode na zaključak da studenti najpre treba da se upoznaju sa ključnim pojmovima izučavane teme kako bi pretraga bila olakšana.
- Približno isti broj studenata smatra da primena navedenih tehnologija može postati (38,9%) standardni deo akademskog učenja ili da nisu sigurni (44,4%) u to, a u prilog tome ide i činjenica da smatraju da ih je najbolje kombinovati sa tradicionalnim pristupom učenju.

Na osnovu pregleda značajnih rezultata dobijenih istraživanjem na ovom uzorku, može se izvesti jedan opšti zaključak: da studenti smatraju da LLM i RAG tehnologije mogu pomoći boljem razumevanju kompleksne literature, da su zadovoljni pretragom informacija iz različitih izvora, ali da ipak smatraju da je najbolji način za njihov proces učenja kombinacija primene navedenih tehnologija i tradicionalne metode.

V. Zaključak

U zaključku, integracija velikih jezičkih modela (LLM) i sistema sa generacijom uz podršku pretrage (RAG) predstavlja značajan napredak u oblasti obrade prirodnog jezika i pretraživanja baza podataka. Kroz istraživanje dva primera – jednog koji koristi RAG pristup za interakciju sa knjigom o bazama podataka u formatu četbota i drugog koji koristi ChatGPT-4 za konvertovanje prirodnog jezika u SQL – ilustrovali smo svestranost i efikasnost ovih tehnologija. Nalazi naglašavaju potencijal LLM-ova i RAG-a da ne samo unaprede korisničku interakciju sa složenim sistemima podataka, već i da demokratizuju pristup informacijama. Kako se ovi modeli nastavljaju razvijati, buduća istraživanja treba da se fokusiraju na optimizaciju njihove performanse, rešavanje izazova vezanih za tačnost i razumevanje konteksta, kao i na istraživanje novih primena u različitim domenima.

Zahvalnica

Autori se zahvaljuju za finansijsku podršku Pokrajinskom sekretarijatu za visoko obrazovanje i naučnoistraživačku delatnost Vojvodine, Republika Srbija (Projekat / Grant br. 000833972 2024 09418 004 000 000 001/1).

Literatura

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. (2017). Attention is All You Need. 1st Conference on Neural Information Processing Systems (NIPS 2017).

[2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[3] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. OpenAI.

[4] Goyal, A., Friesen, A. L., Weber, T., Banino, A., Ke, N. R., Puigdomenech Badia, A., Blundell, C. (2022). Retrieval Augmented Reinforcement Learning. 39th International Conference on Machine Learning, (p. PMLR 162).

[5] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI.

[6] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Amodei, D. (2020). Language Models are Few-Shot Learners. OpenAI.

[7] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Comput.Surv, 55(9), Article 195.

[8] Izacard, G., & Grave, E. (2021). Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. 16th Conference of the European Chapter of the Association for Computational Linguistics, (pp. 874–880).

[9] Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Proceedings of Machine Learning , (pp. 1–11).

[10] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. 57th Annual Meeting of the Association for Computational Linguistics, (pp. 3645–3650).

[11] Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., Huang, M. (2024). SafetyBench: Evaluating the Safety of Large Language Models. 62nd Annual Meeting of the Association for Computational Linguistics, (pp. 15537–15553)

[12] Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 34th Conference on Neural Information Processing Systems (NeurIPS 2020.

[13] Karpukhin, V., Oguz, B., Miny, S., Lewis, P., Wu, L., Edunov, S., Yih, W.-t. (2020). Dense Passage Retrieval for Open-Domain Question Answering. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), (pp. 6769–6781).

Integration of Large Language Models in Education: Student Perceptions Through Practical Applications

Zdravko Ivanković, Predrag Pecev, Saša Sudar, Jasmina Damnjanović, Milana Ivanković

ABSTRACT

This study investigates student perceptions of integrating Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) systems within educational contexts. Two case studies are presented to demonstrate the practical implementation of these technologies. The first case employs a RAG methodology, leveraging the all-mpnet-base-v2 model in conjunction with LLaMA 2 to facilitate a chatbot-based interaction with a database textbook, enabling users to access information through natural conversational queries. The second case highlights the application of an advanced LLM, specifically ChatGPT-4, in translating natural language inputs into SQL statements for direct database querying. Student feedback was collected through a structured questionnaire designed to assess their experiences and insights regarding these innovative technological approaches.

Uticaj augmentacija trening podataka na performanse doobučenog Whisper modela

Siniša Suzić
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
sinisa.suzic@uns.ac.rs
0000-0002-0511-6729

Tijana Nosek
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
tijana.nosek@uns.ac.rs
0000-0002-3707-0286

Nikola Simić
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
nikolasimic@uns.ac.rs
0000-0002-0748-4672

Darko Pekar
Alfanum Ltd,Novi Sad
darko.pekar@alfanum.co.rs

Vlado Delić
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
vlado.delic@uns.ac.rs
0000-0002-4558-9918

***Apstrakt* - Whisper je jedan od najpopularnijih višejezičnih modela za automatsko prepoznavanje i transkripciju govora u poslednje vreme. Njegova popularnost zasnovana je na otvorenosti samog modela i podršci prepoznavanju velikog broja jezika. Međutim, pošto su količine materijala za obuku ovog modela za različite jezike bile veoma nesrazmerne i performanse prepoznavanja se za različite jezike dosta razlikuju. U ovom radu je prikazan pozitivan uticaj doobuke medium varijante Whisper modela na vrednosti verovatnoće greške na nivou reči, i to za slučaj srpskog jezika, koji predstavlja tipičan primer slabo zastupljenog jezika u obuci inicijalnog modela. Posebno je ispitan i uticaj dodavanja novih podataka za obuku korišćenjem sledećih tipova augmentacija – kompresija, dodavanje šuma, reverberacija. Pokazano je da dodavanje augmentovanih podataka u slučaju prepoznavanja javno dostupnih baza dovodi do smanjenja verovatnoće greške. Sa druge strane, u slučaju testiranja doobučenog modela na realnim primerima, uticaj augmentacije podataka za obuku na performanse modela nije konzistentan, što naglašava potrebu za dodatnim testiranjima korišćenjem i nekih drugih postupaka augmentacije.**

Ključne reči – ASR, doobuka, augmentacija.

I. Uvod

Automatsko prepoznavanje govora (ASR, engl. Automatic Speech Recognition) predstavlja jednu od ključnih tehnologija u domenu interakcije između čoveka i mašine. Osnovna funkcija ASR sistema jeste konverzija govornog signala u tekstualni zapis, čime se omogućava efikasnija komunikacija sa softverskim i hardverskim sistemima. Zahvaljujući brzom razvoju u oblasti veštačke inteligencije i obrade prirodnog jezika, ovi sistemi nalaze široku primenu u različitim kontekstima — od digitalnih asistenata i korisničke podrške, do automatske transkripcije sastanaka, medicinske dokumentacije i medijskog sadržaja [1].

Poseban značaj ASR sistema ogleda se u sve većim zakonskim i društvenim zahtevima za inkluzivnošću, kao što je obavezno titlovanje televizijskog programa u cilju pristupačnosti za osobe sa oštećenim sluhom [2]. Takođe, organizacije sve češće koriste transkripciju govora radi arhiviranja i pretrage sadržaja, što je znatno efikasnije kada su podaci dostupni u tekstualnom obliku.

Pored toga, razvoj interneta stvari (IoT) dodatno naglašava potrebu za prirodnim načinima interakcije sa uređajima, pri čemu se govor nameće kao najintuitivniji modalitet [3]. U tom kontekstu, ASR tehnologija igra ključnu ulogu u omogućavanju ovakve komunikacije.

Tradicionalni ASR sistemi sastoje se od 3 komponente: akustičkog modela, leksikona i gramatike/jezičkog modela. Akustički model uspostavlja vezu između akustičkih obeležja i fonema ili neke druge jedinice za prepoznavanje, npr. trifona. Leksikoni omogućavaju da se jedinice za prepoznavanje povezuju u reči, dok gramatike, odnosno modeli jezika omogućavaju povezivanje reči u veće celine, tj. cele rečenice. Ovakva struktura ASR sistema činila ih je jezički zavisnim i bilo je neophodno razvijanje posebnih sistema za svaki jezik.

Kao i u mnogim drugim oblastima, rast popularnosti dubokih neuronskih mreža doveo je i do promena u pristupu prepoznavanju govora. Akustički modeli bazirani pre svega na skrivenim Markovljevim modelima [4], kao i jezički modeli bazirani na N-gramima [5], bivaju zamenjeni modelima na bazi neuronskih mreža [6, 7]. Međutim, možda najveću promenu u ASR sistemima u poslednje vreme predstavlja uvođenje tzv. end-to-end modela (E2E) [8]. U poređenju sa standardnim modelima E2E modeli integrišu sve različite komponente ASR sistema u jednu celinu i omogućavaju zajedničku obuku i predikciju [8]. Među E2E modelima u poslednje vreme veliku popularnost stekao je model pod nazivom Whisper [9], zahvaljujući činjenici da podržava veliki broj različitih jezika, kao i zahvaljujući dostupnosti za neograničeno korišćenje i doobuke. Iako Whisper podržava veliki broj jezika, performanse modela se značajno razlikuju za različite jezike, što je uglavnom u korelaciji sa količinom podataka dostupnih za obuku za određeni jezik. Takođe, performanse modela su obično merene na bazama na kojima je model i obučavan, iako ne nužno da istim rečenicama, jezički sadržaj je često ograničen temom, a govor sniman u ujednačenim akustičkim uslovima. Zbog toga se peformanse sistema u praktičnim primenama često razlikuju od onih koje su navedene u literaturi. U ovom radu biće prikazani rezultati adaptacije Whisper modela na srpski jezik, kao i uticaj primene odgovarajućih postupaka augmentacije na rezultate prepoznavanja.

Ostatak rada je organizovan na sledeći način. U Poglavlju II opisan je korišćeni ASR model. U poglavlju III dat je pregled korišćenih augmentacija, dok je u poglavlju IV dat pregled korišćenih baza i predstavljeni su dobijeni rezultati. Potom slede zaključak i pravci daljeg istraživanja.

II. WHISPER MODEL

Whisper model predstavlja E2E ASR sistem obučen na 680.000 sati višejezičnog audio materijala, što omogućava prepoznavanje, odnosno, prepoznavanje govora na 97 jezika. Očekivano, najveća količina materijala za obuku je na engleskom jeziku (preko 400.000 sati). Za srpski jezik korišćeno je svega 28 sati javno dostupnog audio materijala, uz dodatnih 91 sat na hrvatskom jeziku. Posledica neravnomerne distribucije podataka za obuku ogleda se u činjenici da model ne ostvaruje jednake performanse za sve jezike.

Pored prepoznavanja, odnosno, transkripcije govora ovaj model podržava i prevođenje, pri čemu je podržana samo opcija prevođenja na engleski. Odnosno, model za govor na srpskom jeziku može direktno da generiše rezultujući tekst na engleskom. Obrnuti pravac u ovom trenutku nije moguć, tj. nije moguće za govor na engleskom jeziku dobiti transkripciju na srpskom. Iako model nudi mogućnost generisanja vremenskih odrednica za segmente, ova opcija nije dovoljno pouzdana pa se u literaturi javilo nekoliko nadogradnji Whisper modela koje nude i automatsko poravnanje [10].

Obuka se sprovodi na parovima audio-tekst, pri čemu trajanje pojedinačnih audio snimaka ne prelazi 30 sekundi, uz učestanost odabiranja od 16 kHz. Kraći snimci se dopunjavaju tišinom. Tokom predikcije, model takođe obrađuje isključivo segmente trajanja do 30 sekundi, što može predstavljati izazov pri obradi dužih audio-zapisa i dovesti do grešaka na prelazima segmenata.

Whisper model zasnovan je na transformer arhitekturi u obliku koder-dekoder strukture. Ulaz kodera predstavljaju logmel-spektrogrami dobijeni iz audio-snimka. Izlaz kodera prosleđuje se dekoderu, koji generiše niz tokena. Ovi tokeni, koji mogu predstavljati foneme ili cele reči, preuzeti su iz ChatGPT-a [11]. Model stoga ne sadrži specifične jezičke modele za sve podržane jezike, što sa sobom nosi i prednosti i nedostatke. U slučaju srpskog jezika, jedan token može predstavljati jedan ili više fonema (uglavnom do tri), što može rezultirati generisanjem reči koje nisu deo srpskog jezika.

Whisper model je dostupan u više varijanata, koje se razlikuju po veličini arhitekture i broju parametara. Postoje modeli od najmanjeg („tiny“) sa 39 miliona parametara, do najvećeg („large“) sa 1550 miliona parametara. Veći modeli su robustniji i obučeni na većoj količini podataka, ali zahtevaju značajno snažniji hardver ne samo za obuku već i za primenu (engl. *inference)*. U ovom radu korišćen je srednji („medium“) model sa 769 miliona parametara. Za rad sa ovim modelom dovoljna je grafička kartica sa 10 GB video memorije.

III. METODE AUGMENTACIJE

Sistemi za automatsko prepoznavanje govora u slučaju jezika sa ograničenim resursima, poput srpskog jezika, suočavaju se sa nedostatkom podataka za obuku (audio snimaka praćenih odgovarajućim transkriptima), raznovrsnošću akustičkog okruženja i kanala, kao i varijabilnošću govora. Kako bi se unapredili postojeći modeli za ovakve jezike i razvili novi, neophodno je snimiti nove, pažljivo osmišljene baze podataka, što predstavlja proces zahtevan u pogledu potrebnih resursa. Povećanje raznovrsnosti postojećih skupova podataka metodama augmentacije može značajno uticati na performanse dodatno obučenog ASR modela, jer i relativno mali skupovi postaju reprezentativniji za realne uslove.

U tom kontekstu, kompresija audio-zapisa predstavlja efikasan način za unošenje dodatne varijabilnosti u postojeće podatke. Ideja je da se audio-zapis dekoduje i ponovo koduje u različitim formatima i sa različitim kvalitetima kompresije, čime se simuliraju realni uslovi u kojima korisnici koriste različite uređaje, aplikacije i mrežne protokole koji često komprimuju govor radi uštede protoka i prostora [12]. U slučaju finog podešavanja modela za prepoznavanje govora obučenog na čistim audio-podacima postoji velika verovatnoća da neće davati dobre rezultate na stvarnim korisničkim snimcima koji su prošli kroz različite oblike kompresije. Uvođenjem audio-fajlova koji su prethodno komprimovani u različitim formatima i pri različitim bitskim brzinama, model se izlaže raznovrsnijim akustičkim uslovima, čime se povećava njegova robustnost, odnosno, otpornost na degradaciju zvuka. Proces augmentacije obično podrazumeva konverziju audio-fajlova iz nekomprimovanog WAV formata u formate poput MP3 i AAC. Jedan od danas najpopularnijih alata za ovu vrstu manipulacije je FFMPEG [13], koji omogućava jednostavno kodovanje i dekodovanje audio-fajlova. Iako je moguće proizvesti više verzija istog zvučnog zapisa sa različitim bitskim brzinama, u ovom radu nasumično je birana po jedna vrednost bitske brzine iz unapred definisanog skupa za svaki audio-fajl koji se augmentujeme, kako ne bi bilo ponavljanja u skupu podataka za obuku. Korišćen je MP3 format sa sledećim bitskim brzinama: 20 kbps, 24 kbps, 32 kbps, 36kbps, 38 kbps i 64 kbps.

Pored kompresije, još jedan važan aspekt akustičke varijacije koji treba uzeti u obzir jeste reverberacija. Dok kompresija simulira degradaciju signala kroz digitalne kanale, reverberacija modeluje uticaj fizičkog prostora na kvalitet govora. Pojava vremenskog i spektralnog zamućenja snimljenog audio-signala usled višestrukih refleksija od različitih površina u prostoriji u kojoj se vrši snimanje naziva se reverberacija [14]. U mnogim realnim okruženjima poput kancelarija, učionica i holova, reverberantni govor je česta pojava. ASR sistemi obučeni isključivo na čistom govoru ne postižu visoke performanse u takvim uslovima pa se pribegava odgovarajućoj augmentaciji skupa podataka za obuku [15]. Augmentacija pomaže bolju generalizaciju u različitim akustičkim okruženjima i poboljšava otpornost u slučaju udaljenog izvora zvuka i u slučaju prisustva šuma. Kako je prikupljanje podataka koji sadrže reverberaciju kompleksan i vremenski zahtevan proces, korišćenje simuliranih impulsnih odziva prostorija predstavlja čest pristup u literaturi [15-16]. Ovi odzivi se konvoluiraju sa čistim audio-signalom da bi se simuliralo ponašanje zvuka u fizičkom prostoru. Za potrebe simulacije koristili smo skup predefinisanih vrednosti parametara koji se odnose na slabljenje i veličinu odgovarajuće prostorije (‘*damping factor*’ i ‘*room size*’).

Još jedna standardna metoda augmentacije podataka u cilju poboljšanja ASR sistema jeste dodavanje šuma [17], odnosno simulacija situacije u kojoj je govor snimljen u prisustvu različite pozadinske buke. Baza uzoraka šumova, odnosno, pozadinske buke dobijena je od preduzeća AlfaNum [18], a čini je oko 70.000 različitih šumova trajanja u proseku 6s, prikupljanih na različite načine (samostalno snimanje i preuzimanje sa interneta). U bazi se nalaze različite vrste šumova poput vetra, saobraćajne buke, buke sa gradilišta, do žamora u kafićima, pozadinskog zvuka sa televizije, i dr. Šumovi su dodati na pojedinačne audio-snimke nasumičnim

odabirom sa konstantnim SNR od 15dB tokom celog trajanja audio-snimka.

U ovom radu primenjene su tehnike augmentacije koje se odnose na kompresiju signala, reverberaciju i dodavanje šuma, i to simultano nad svim podacima iz baze, tako što su vrednosti relevantnih parametara za svaku od primenjenih tehnika nasumično odabrane iz unapred definisanih skupova.

IV. EKSPERIMENTI

A. *Baza za obuku*

U procesu doobuke modela korišćena je baza od 1500 sati transkribovanog audio-materijala na srpskom i hrvatskom jeziku, nastala manuelnom transkripcijom audio-materijala koji potiče iz audio-knjiga i radio-televizijskih emisija, koju je za potrebe ovog istraživanja takođe obezbedilo preduzeće AlfaNum. Prosečna dužina rečenice je 5 sekundi, a svi fajlovi imaju frekvenciju odabiranja 16 kHz.

Korišćenjem postupaka augmentacije opisanih u prethodnim odeljcima kreirano je još dodatnih 3000 sati audio materijala. Od ove količine materijala 1500 sati je dobijeno dodavanjem šuma, a preostalih 1500 kombinacijom kompresije i reveberacije.

B. *Test podaci*

Za potrebe testiranja korišćene su dve javno dostupne baze na srpskom jeziku – *Common Voice* [19] i *Fleurs* [20]. Pored ovih baza korišćen je i skup podataka koji potiče iz realnih primera iz prakse, a koji uključuju sastanke, emisije i telefonske razgovore. Konkretno, 25 snimaka, po 5 iz grupe sastanci i emisije, trajanja od 3-30minuta, i 15 telefonskih razgovora trajanja po 10-ak sekundi.

C. *Rezultati*

Za testiranje performansi modela korišćena je greška prepoznavanja na nivou reči (WER, engl. *Word Error Rate*). Ova mera računa se na osnovu sledeće formule:

$$WER = \frac{S+I+D}{W}. \quad (1)$$

U jednačini (1) W je ukupan broj reči koji je izgovoren u test rečenici. Sa S je označen broj reči koje su pogrešno prepoznate (npr. izgovoreno „mama“, a prepoznato „tama“), sa I broj dodatih reči (npr. izgovoreno „ja sam voleo“, a prepoznato „ja sam je voleo“), dok je sa D označen broj reči koje postoje u test-rečenici, ali ne i u prepoznatoj (npr. izgovoreno „ja sam je voleo“, a prepoznato „ja sam voleo“).

Vrednosti za WER dobijene na javnim bazama prikazane su u tabeli 1, dok su za test-rečenice koje potiču iz primera iz prakse prikazane u tabeli 2. Iz tabele 1 jasno se vidi u kojoj meri doobuka modela za specifičan jezik utiče na krajnje performanse. Za *Common Voice* bazu WER je sa 85.6% kod polaznog modela pao na svega 8.06% kod doobučenog, dok je kod *Fleurs* baze taj uticaj bio manji, ali i dalje veoma značajan – WER je kod doobučenog modela pao sa polaznih 44.9% na 10.61%. Kod obe baze uključivanje augmentovanog materijala je dovelo do dodatnog, ali ne tako drastičnog smanjenja WER. U tumačenju dodatnog poboljšanja vrednosti WER pod uticajem augmentovanih podataka u obzir treba uzeti i samu prirodu test-baza. I *Common Voice* i *Fleurs* sastoje se od veoma kratkih rečenica koje se često sastoje od 3 ili 4 reči. Kod takvih rečenica greška i u jednom slovu podiže vrednosti WER na 25% ili 33% (u konkretnim primerima, a uzimajući jednačinu 1 u obzir, W ima vrednost 3 ili 4, S je jednako 1, a preostale vrednosti su 0). Tako je u slučaju *Common Voice* baze ukupan broj sasvim tačno prepoznatih rečenica dobijenih korišćenjem modela doobučenog nad originalnom bazom 1282, dok je taj broj u slučaju modela dobijenog doobukom nad augmentovanom bazom 1305. Drugim rečima, ukupan broj sasvim tačno prepoznatih rečenica porastao je za 1.7%. Slično je kod *Fleurs* baze, gde je ukupan broj tačno prepoznatih rečenica zahvaljujući augmentaciji podataka za obuku porastao sa 226 na 240, tj. za oko 5.8%.

Rezultati prikazani za test-rečenice iz realnih situacija (tabela 2) ipak pokazuju manju konzistentnost u odnosu na rezultate dobijene nad javnim bazama. Naime, samo u slučaju telefonskih razgovora augmentacija podataka dovela je do poboljšanja rezultata, dok se u slučaju snimaka sastanaka i televizijskih emisija dobijaju nešto lošiji rezultati. Ovo se može objasniti činjenicom da degradacija signala korišćenjem MP3 kodovanja niskog bitskog protoka odgovara degradacijama koje nastaju prenosom signala telefonskim kanalom. Problemi koji su uočeni kod snimaka sastanaka i televizijskih emisija nisu eksplicitno vezani za sam kvalitet signala, nego su potpuno druge prirode – česta preklapanja govornika, nepotpune rečenice, reči ili čak i veće rečenične celine koje su slabo artikulisane, a čija simulacija nije obuhvaćena predloženim augmentacijama.

Tabela 1. WER za različite modele dobijen na standardnim javno dostupnim bazama

	Osnovni model	**Doobuka**	**Doobuka+ augmentacija**
CommonVoice	85.6	8.07	7.24
Fleurs	44.9	10.61	10.12

Tabela 2 WER za različite modele dobijen na internim test-podacima

	Doobuka	**Doobuka+ augmentacija**
Sastanci	23.53	24.89
Emisije	13.17	14.13
Tel. razgovori	7.83	5.8

V. ZAKLJUČAK

U radu je ispitivan uticaj augmentacije podataka na performanse doobučenog *Whisper* modela. Pokazano je da u slučaju javno dostupnih baza doobuke sa korišćenim augmentovanim materijalom utiču na poboljšanje performansi modela. Sa druge strane, doprinos predloženih tehnika augmentacije u slučaju snimaka dobijenih u realnim uslovima, kao što je transkripcija sastanaka ili TV emisija, nekonzistentan je i ograničen jer se kod ovakvih snimaka javljaju problemi koji nisu isključivo vezani za degradaciju kvaliteta signala. Na osnovu toga može se zaključiti da je potrebno ispitati i druge specifične tehnike augmentacije, koje bi, primera radi, uključivale preklapanje originalnog signala sa pozadinskim govorom ili simulaciju promene rastojanja od mikrofona.

ZAHVALNICA

Istraživanje sprovedeno uz podršku Fonda za nauku Republike Srbije, za projekat br. 7449 „Multimodalna višejezička komunikacija između čoveka i mašine, AISPEAK“.

LITERATURA

[1] V. Delić, D. Pekar, M. Sečujski, B. Popović, E. Pakoci and S. Suzić, „Development of speech technology for Serbian and its applications“, 1st Serbian International Conference on Applied Artificial Intelligence (SICAAI), Kragujevac, Srbija, Maj 2022

[2] A. Coy, P.S. Mohammed and P. Skerrit, “Inclusive Deaf Education Enabled by Artificial Intelligence: The Path to a Solution”, International Journal of Artificial Intelligence in Education, str.1-39, 2024

[3] H. Younis, and J.H. Hansen, “Challenges in real-time-embedded IoT command recognition”, 7th World Forum on Internet of Things (WF-IoT), pp. 848-851, jun 2021.

[4] S.E. Levinson, L.R. Rabiner and M.M. Sondhi, “An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition”, BellSystem Technical Journal 62, str. 1035–1074, 1983.

[5] F. Jelinek, “Statistical methods for speech recognition”, MIT press, 1998.

[6] G.E. Dahl, D. Yu, L. Deng, L. and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition”, IEEE Transactions on audio, speech, and language processing, 20(1), str. 30-42, 2011.

[7] T. Mikolov, M.Karafiát, L. Burget, J. Cernocký, J. and S. Khudanpur, “Recurrent neural network based language model”, 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, str. 1045-1048, Japan, Septembar 2010.

[8] R. Prabhavalkar, T. Hori, T.N. Sainath, R. Schlüter and S. Watanabe, “End-to-end speech recognition: A survey”. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32, pp.325-351, 2023

[9] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey and I. Sutskever, “Robust speech recognition via large-scale weak supervision”, International conference on machine learning, str. 28492-28518, Jul 2023.

[10] L. Wagner, B. Thallinger and M. Zusag, “CrisperWhisper: Accurate Timestamps on Verbatim Speech Transcriptions, arXiv preprint arXiv:2408.16589, 2024

[11] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.L. Han, Q.L. and Y. Tang, “A brief overview of ChatGPT: The history, status quo and potential future developmen”, IEEE/CAA Journal of Automatica Sinica, 10(5), pp.1122-1136, 2023

[12] M. P. Fernández-Gallego, and D. T. Toledano, “A Study of Data Augmentation for ASR Robustness in Low Bit Rate Contact Center Recordings Including Packet Losses”, Applied Sciences, 12(3), 1580, 2022.

[13] S. Tomar, “Converting video formats with ffmpeg”, Linux journal, 2006(146):10, 2006.

[14] H. Malik, and H. Farid, "Audio forensics from acoustic reverberation," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010), Dallas, TX, USA, 2010, pp. 1710-1713

[15] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), New Orleans, LA, USA, 2017, pp. 5220-5224.

[16] J. Malek,, and J. Zdansky, “On Practical Aspects of Multi-condition Training Based on Augmentation for Reverberation-/Noise-Robust Speech Recognition”. In: Ekštein, K. (eds) Text, Speech, and Dialogue. TSD 2019. Lecture Notes in Computer Science(), vol 11697. 2019, Springer, Cham.

[17] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Ng, “Deep Speech: Scaling up end-to-end speech recognition,” in arXiv, 2014.

[18] www.alfanum.co.rs

[19] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F.M. Tyers and G. Weber, “Common voice: A massively-multilingual speech corpus”. arXiv preprint arXiv:1912.06670, 2020

[20] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, “Fleurs: Few-shot learning evaluation of universal representations of speech”, IEEE Spoken Language Technology Workshop (SLT), str.. 798-805, 2023

The Impact of Training Data Augmentations on the Performance of a Fine-Tuned Whisper Model

Siniša Suzić, Tijana Nosek, Nikola Simić, Darko Pekar, Vlado Delić

ABSTRACT

Whisper is one of the most popular speech recognition models in recent times. Its popularity is based on the openness of the model itself and its support for recognition in a large number of languages. However, since the amount of data used for training varied significantly across languages, the recognition performance also differs greatly. This paper analyses the positive impact of fine-tuning the medium variant of the Whisper model on word error rate values for the Serbian language, which is an example of an under-represented language in the initial model's training data. The influence of adding new training data through augmentation by compression, noise addition, and reverberation was also examined. It was shown that adding augmented data in the case of publicly available datasets leads to WER improvement. On the other hand, when testing the fine-tuned model on real-world examples, the impact of data augmentation on model performance was not consistent, which highlights the need for additional testing using other augmentation techniques as well.

Прикупљање података и означавање у процесу формирања скупова података говора мржње на српском језику

Дражен Драшковић
Универзитет у Београду – Електротехнички факултет,
Београд, Србија
drazen.draskovic@etf.bg.ac.rs
ORCID: 0000-0003-2564-4526

Урош Раденковић
Универзитет у Београду – Електротехнички факултет,
Београд, Србија
uros.radenkovic@etf.bg.ac.rs
ORCID: 0000-0002-2440-0127

Марко Мићовић
Универзитет у Београду – Електротехнички факултет,
Београд, Србија
marko.micovic@etf.bg.ac.rs
ORCID: 0000-0001-7477-2503

Јелица Цинцовић
Универзитет у Београду – Електротехнички факултет,
Београд, Србија
jelica.cincovic@etf.bg.ac.rs
ORCID: 0000-0001-6440-9348

Адриан Милаковић
Универзитет у Београду – Електротехнички факултет,
Београд, Србија
adrian.milakovic@etf.bg.ac.rs
ORCID: 0000-0002-3005-3352

Владимир Јоцовић
Универзитет у Београду – Електротехнички факултет,
Београд, Србија
vladimir.jocovic@etf.bg.ac.rs
ORCID: 0000-0002-7140-5043

***Апстракт* - Говор мржње представља један од најозбиљнијих изазова савременог дигиталног друштва, нарочито услед масовне употребе друштвених мрежа и онлајн медија. Развој поузданих аутоматизованих система за његову детекцију захтева постојање квалитетних, пажљиво прикупљених и анотираних скупова података, што је посебно изазовно за језике са ограниченим дигиталним ресурсима, као што је српски језик. У овом раду описан је процес прикупљања, обраде и означавања текстуалних података са циљем формирања скупа података за детекцију говора мржње на српском језику. Подаци су прикупљани са различитих извора, укључујући друштвене мреже и онлајн медијске платформе, коришћењем аутоматизованих и ручних техника, као и веб индексирања и прикупљања. Формирани скуп садржи 4300 кратких текстова који су анотирани у три класе: текст без говора мржње, увредљив говор и говор мржње, при чему је говор мржње додатно категорисан према релевантним типовима дискриминације у складу са актима Европске уније. Поред описа процеса анотације, у раду је дата и детаљна анализа скупа података, укључујући дистрибуцију класа, дужину текстова и најчешће специфичне речи. Представљени резултати указују на сложеност и разноликост говора мржње у онлајн комуникацији на српском језику и представљају значајну основу за даљи развој и евалуацију система заснованих на вештачкој интелигенцији.**

Кључне речи – обрада природних језика, говор мржње, увредљиве речи, скуп података, веб индексирање.

I. Увод

Говор мржње представља облик јавног изражавања којим се подстичу, оправдавају или шире мржња, дискриминација и насиље према појединцима или групама на основу њихових личних или друштвених обележја. Ова обележја могу обухватати националну или етничку припадност, веру, расу, пол, сексуалну оријентацију, инвалидитет, језик или неко друго стварно или претпостављено својство. За разлику од увредљивог или непристојног говора, говор мржње има шири друштвени утицај јер директно угрожава људско достојанство и може довести до реалних облика дискриминације и насиља.

У савременом друштву, појам говора мржње посебно добија на значају услед експанзије дигиталних медија и друштвених мрежа. Интернет је омогућио да се информације шире изузетно брзо и без јасних уредничких баријера, што је створило простор за масовно ширење различитих облика штетног говора. Анонимност и осећај некажњивости додатно охрабрују појединце да изражавају ставове које у традиционалним медијима или јавном простору не би изнели.

Говор мржње је данас широко присутан у онлајн коментарима, на друштвеним мрежама, форумима и платформама за дељење видео-садржаја. Посебно је уочљив у контекстима политичких расправа, миграција, међунационалних односа и друштвених криза, где се често користи као средство манипулације јавним мњењем. Медији, иако имају значајну улогу у информисању јавности, понекад својим сензационалистичким приступом или недовољном контролом коментара доприносе нормализацији оваквог дискурса.

Идентификација говора мржње у дигиталном окружењу представља сложен задатак, јер значење поруке често зависи од контекста, језичких нијанси и културних специфичности. Исти израз може у једном контексту бити неутралан, а у другом имати јасну дискриминаторну или подстицајну функцију. Додатну сложеност уносе ирониja, сарказам и имплицитне поруке, које је тешко препознати без дубљег разумевања садржаја.

Традиционално, говор мржње се идентификовао мануелно, од стране модератора. Иако овај приступ омогућава високу прецизност, он је временски захтеван, скуп и тешко одржив у условима огромне количине садржаја који се свакодневно објављује на интернету. Због тога је у пракси често немогуће благовремено реаговати на све спорне садржаје.

Са развојем вештачке интелигенције и обраде природног језика, све више се примењују

аутоматизовани системи за детекцију говора мржње. Ови системи користе технике машинског учења и дубоког учења како би анализирали текстуалне податке и идентификовали потенцијално проблематичан садржај. Њихова предност лежи у брзини и могућности обраде великих количина података у реалном времену.

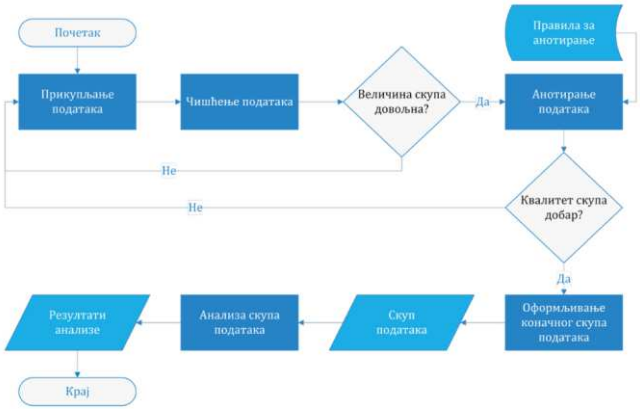
Ипак, аутоматска детекција говора мржње има значајна ограничења. Алгоритми често тешко разликују говор мржње од слободе изражавања, критике или сатире, што може довести до лажних позитивних или негативних резултата. Посебан изазов представљају језици са ограниченим дигиталним ресурсима, где недостаје довољно означених података за квалитетно тренирање модела.

Због тога се у савременим системима све чешће примењује хибридни приступ, који комбинује аутоматску анализу и људску експертизу. Аутоматски системи се користе за иницијално филтрирање и означавање сумњивог садржаја, док коначну одлуку доносе људски модератори. Овакав приступ омогућава бољу равнотежу између ефикасности и поузданости.

У овом раду описан је процес прикупљања података у оквиру националног истраживачког пројекта „Софтвер за превенцију текстуалних увреда на српском језику: Откривање говора мржње помоћу вештачке интелигенције“, као и анализе прикупљених података.

II. Прикупљање података

У овом поглављу дат је процес прикупљања података, њиховог процесуирања у сврху креирања анотираног скупа података, анализа података како комплетног скупа тако и подељених скупова за обучавање, валидацију и тестирање. На слици 1 приказан је поменути процес. Пре прикупљања текстова са говором мржње на српском језику прикупљен је скуп података у виду речника увредљивих, ружних и погрдних речи у српском језику помоћу више штампаних издања таквих речника, који су техникама оптичког препознавања карактера (енгл. *Optical Character Recognition*, скр. *OCR*) дигитализовани. Коришћењем тог скупа, аутори су реализовали и веб-екстензију за веб-прегледач *Google Chrome*, која цензурише овакве непримерене речи [1].



Слика 1. Приказ процеса прикупљања података са анализом.

Како би се постигла превенција говора мржње на интернету приликом обучавања модела вештачке интелигенције потребно је водити рачуна о скупу података који се користи за обучавање. Потребно је да скуп података буде довољно велик али и разноврстан како би се скупом покрио што већи број карактеристика говора мржње. Подаци прикупљени за потребе овог рада потичу из различитих извора као што су друштвене мреже - *Facebook* компаније Мета, *Instagram* компаније Мета, *X* (раније *Twitter*), *YouTube* и *Reddit*. Такође, поред поменутих извора, подаци су прикупљани и са медијских платформи са дневним вестима - *Blic*, *Informer*, *N1* и *Sportal*.

Извори су изабрани циљано у складу са потребама мастер рада, док је критеријум за избор извора вероватноћа постојања говора мржње на њему. Наведене друштвене мреже имају велики број корисника у Србији, широк спектар типова садржаја и могућност коментарисања истог. Са друге стране новинске платформе претежно деле садржај из домена политике, економије, друштва и спорта који често може садржати коментаре увредљиве природе. По потреби на поменутим платформама је таргетиран садржај за који постоји већа вероватноћа садржања говора мржње како би скуп података постао балансиранији.

Прикупљање података вршило се аутоматизовано и ручно. Аутоматизовано дохватање података било је могуће путем програмских интерфејса апликације (енгл. *Application Programming Interface*, скр. *API*) код оних друштвених мрежа где је то могуће, као што су *Facebook*, *YouTube* и раније *X*. За дневне вести са медијских сајтова писани су специјализовани мањи програми у виду скрипти, популарни веб индексери (енгл. *web crawler*) и веб прикупљачи (енгл. *web scraper*). Скрипте су писане у програмском језику *Python*. Мали део података настао је синтетичким путем коришћењем великих језичких модела и дотадашњег скупа података као референце. Велики језички модели, чак и уз различите видове навођења, нису били у стању да генеришу квалитетне податке. Веома мали број генерисаних података је употребљив из разлога што се подаци генеришу шаблонски. Такође, поједини велики језички модели попут модела *ChatGPT* не подржавају генерисање говора мржње у новијим верзијама, док старије верзије често не дају довољно добре резултате у општем случају. Последњи део података преузет је из скупа података *ReLDI-NormTagNER-sr 3.0* направљен од стране центра за проучавање јужнословенских језика *CLASSLA* (*CLARIN Knowledge Centre for South Slavic languages*)[2]. Овај скуп података није намењен за класификацију текста али садржи велик скуп коментара преузетих са друштвене мреже *X* (*Twitter*). Након спајања извршено је чишћење скупа података од коментара који не представљају смислене целине. Резултујући скуп података садржи 4300 коментара на српском језику.

III. Анотација података

За потребе анотирања усвојен је систем сличан систему представљеном у раду [3], у ком се подаци означавају једном од три класе: класа 0 – не садржи говор мржње; класа 1 – садржи увредљив говор; класа 2 – садржи говор мржње. Начини дискриминације који спадају у говор мржње су дискриминација на националној, етничкој, религијској, полној/сексуалној, политичкој и навијачкој основи као и било који други вид

дехуманизације и деградације на основу порекла, занимања, физичког изгледа, способности и инвалидитета. Списак категорија и поткатегорија које су обухваћене у овом истраживању дат је у табели 1, у складу са Кодексом понашања Европске Уније (ЕУ) о сузбијању незаконитог говора мржње на мрежи. У табели 2 дат је списак примера за текстове без говора мржње, текстове само са увредама, и текстове са елементима говора мржње у српском језику.

Може се десити да у једној реченици или једном кратком тексту са више реченица, буде изражено више категорија говора мржње. Са друге стране увредљив говор обухвата изразе попут псовки, вулгарности и других индивидуалних увреда, као и груб и агресиван тон на начин да може бити непријатан али не и дискриминаторски. Дакле, сваки говор мржње је увредљив говор, али није сваки увредљив говор заправо говор мржње:

- не садржи говор мржње – „Живео и здрав био у љубави и радости Никола Јокић“;
- увредљив говор – „Зна ли се име тог потпуковника? Дајте објавите нека му потомци знају који је олош био“;
- говор мржње – „А што пуно причаш са тим лажовима? Поубијај ту четничку ђубрад“.

Табела 1. Предложена класификација говора мржње према актима ЕУ.

Тип говора мржње	Назив категорије и поткатегорије говора мржње
К1) Расни, етички и национални говор мржње	**Раса / боја коже** → расизам, колоризам
	Етничка афилијација → етничка дискриминација, етничка мржња
	Националности / Порекло (миграционо, регионално) → ксенофобија
К2) Религиозни говор мржње	**Религија и веровања** → верска нетолерација, верска дискриминација
К3) Говор мржње заснован на полу и сексуалности	**Пол** → сексизам
	LGBTQ+ идентитет → хомофобија, трансфобија, квирфобија
К4) Говор мржње заснован на физичком изгледу и здрављу	**Физички изглед** → лукизам (дискриминација на основу изгледа)
	Болест / Инвалидитет → аблизам
К5) Говор мржње заснован на годинама	**Године (младост, старост)** → ејџизам
К6) Социо-економски статус	**Социоекономски статус / класа** → класизам
	Занимање / Професија→ стигматизација професија
К7) Спортска нетрпељивост	**Спортска припадност (идентитет навијача)** → спортска нетолерација, хулиганство

Табела 2. Примери категорија и поткатегорија које ће бити део детектора говора мржње на српском језику

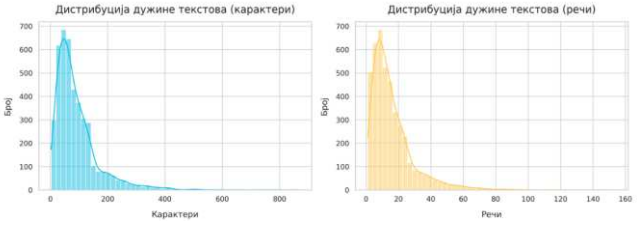
Улазни подаци (текст на српском језику)	Излазни подаци	
	К	Категорија/ Поткатегорија
Свиђа ми се што су Кинези радилице и што ће завршити пре рока.	**0**	/
Је л' ти то сестра? Изгледате као ретарди обоје.	**1**	/
Не желимо Цигане у нашој земљи, треба их уклонити.	**2**	**К1, расизам**
Ја нећу запошљавати дивље Албанце са Косова.	**2**	**К1, ксенофобија**
Све адвентистичке групе су лажови који нас јуре по улици.	**2**	**К2, религија**
Жене су преслабе за лидерске улоге.	**2**	**К3, сексизам**
Геј популација је одвратна.	**2**	**К3, хомофобија**
Ружни људи не заслужују пажњу.	**2**	**К4, лукизам (изглед)**
Особе са инвалидитетом су бескорисне.	**2**	**К4, аблизам (инвалидитет)**
Тинејџери су глупи и неодговорни.	**2**	**К5, ејџизам (старост)**
Новинари су лажови.	**2**	**К6, професија**
Сви навијачи Партизана су кретени.	**2**	**К7, навијачки идентитет**

IV. Анализа података

У овом поглављу приказан је процес анализе података у сврху бољег разумевања скупа. Процес обухвата статистичке анализе података попут дужине текстова, дистрибуције класа, прегледа најчешћих речи сваке од класа и других.

А. Анализа дужине података

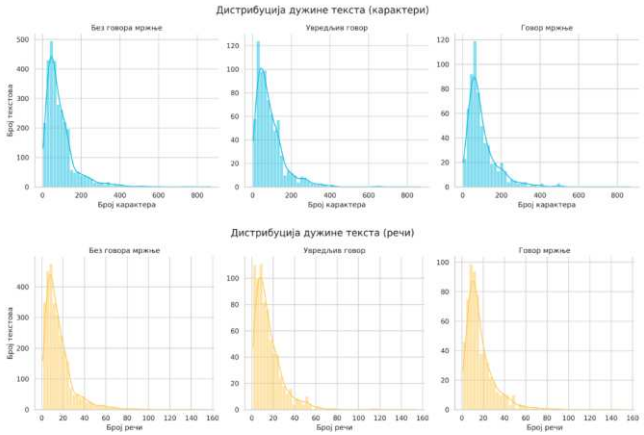
На скупу од 4300 коментара просечна дужина коментара је приближно 91 карактер тј. 16 речи. Најкраћи текст садржи само једну реч од 3 карактера, док је најдужи дугачак 868 карактера тј. 154 речи. Занимљиво је да испод 50% текстова има дужину мању од просечне. Статистика о распону дужине текстова значајна је због оптимизације процеса обучавања модела. На слици 2 види се да је највећи број текстова кратак и садржи мали број карактера/речи, док је број дугачких текстова јако мали.



Слика 2. Дистрибуција дужине текстова.

На слици 3 приказана је дистрибуција дужине текста по класама. Треба приметити да је број текстова класе „Без говора мржње“ већи од броја текстова друге две класе. Изглед графика за све три класе је уједначен и нема великих одступања у дужини текстова између класа

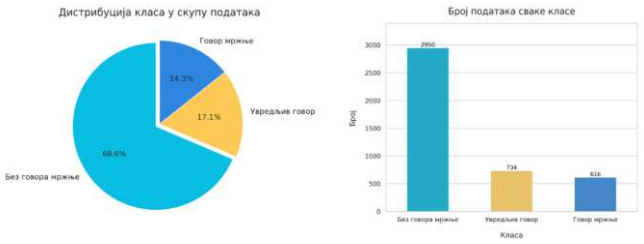
што значи да број текстова одређене класе не утиче на дистрибуцију дужина текстова у скупу.



Слика 3. Дистрибуција дужине текстова по класама.

B. Анализа класа

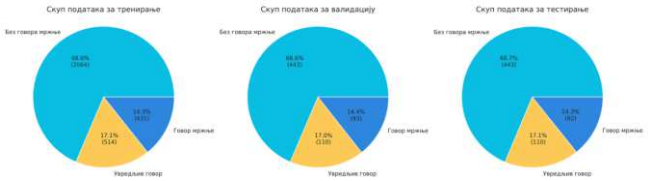
Једна од основних статистичких метрика за процену квалитета скупа података је балансираност скупа. Иако је често веома тешко постићи потпун баланс између броја података сваке класе, важно је тежити балансираном скупу и проценити у којој мери је скуп података балансиран. На слици 4 приказана је дистрибуција класа у скупу података. Скуп података који се користи за потребе овог истраживачког рада није балансиран и преко две трећине података припада класи „Без говора мржње“. Оваква неравнотежа је очекивана с обзиром на то да у комуникацији говор мржње није учестао колико нормалан говор. Изоловањем друге две класе види се да је њихов однос доста уравнотеженији што је добро.



Слика 4. Дистрибуција све три класе текстова – без говора мржње, увредљивог говора и са говором мржње.

C. Анализа скупова података за обучавање, валидацију и тестирање

За потребе обучавања, валидације и тестирања различитих модела вештачке интелигенције првобитни скуп података подељен је у три нова скупа и то: скуп за обучавање – 70% првобитног скупа; скуп за валидацију – 15% првобитног скупа; скуп за тестирање – 15% првобитног скупа. Важно је да сваки од нових скупова података задржи карактеристике првобитног скупа, попут дистрибуције класа. У супротном перформансе модела могу бити значајно умањене. На слици 5 приказана је расподела класа за сваки од нових скупова која је процентуално идентична оној из првобитног скупа.



Слика 5. Расподела класа по скуповима за тренирање, валидацију и тестирање.

D. Анализа специфичних речи

Откривањем најчешћих речи специфичних речи у сваком од скупова података могуће је стећи додатан увид у карактеристике говора мржње, увредљивог говора, као и интернет комуникације на српском језику. Из сваког од скупова, коришћењем лексикона најчешћих речи у српском језику (*stop words*), избачене су речи које не доприносе контексту текста. Тиме се добија скуп специфичних речи релевантних за врсту говора – класу. На слици 6 приказане су најчешће речи говора мржње.

Слика 6. Представа најчешћих речи говора мржње из целог скупа података кратких текстова на српском језику.

Коришћењем дефиниције говора мржње наведене у уводном поглављу, на датом графикону можемо уочити речи које се могу повезати са неким од облика мржње:

- мржњом на националној основи - „Срби, Усташе, Швабе, Шиптари,…“;
- мржњом на етничкој основи - „Цигани, Јевреји, …“;
- мржњом на верској основи - „Муслимани, Фрањевци, …“;
- мржњом на полној/сексуалној основи - „Геј, Педери, LGBT, …“;
- мржњом на политичкој основи – „СНС, НАТО,…“;
- подстицањем на насиље, агресивност, вређање – „Јебем, смрт, курац, …“.

Примећује се да су најуочљивије речи које се могу повезати са говором мржње на националној основи што је и очекивано узимајући у обзир друштвено-политичке аспекте и историју земаља на Балканском полуострву. Важно је напоменути да неке од ових речи могу представљати и други облик мржње у зависности од контекста, на пример реч „Цигани“ може се посматрати у контекстима мржње на етничкој, где су Цигани

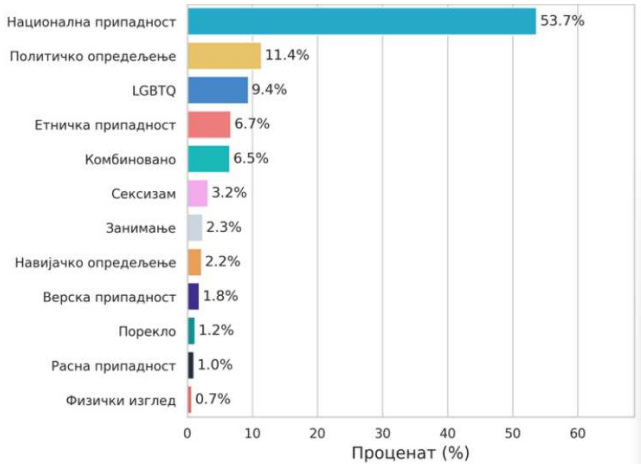
представници ромске националне мањине, и навијачкој основи, где су Цигани навијачи спортског клуба Црвена Звезда.

На слици 7 приказане су најчешће речи увредљивог говора. Занимљиво је да су на овој представи уочљиве речи које се односе на: политичке функционере и државне званичнике - Зеленски, Путин, Вучић, Бајден, Шмит,... и увреде - кретен, говно, дебил, глуп, слепац, ретард, итд.

Слика 7. Представа најчешћих увредљивих и ружних речи из целог скупа података кратких текстова на српском језику.

E. Анализа категорија говора мржње

Коришћењем раније поменуте дефиниције говора мржње текстови који садрже исти додатно су анотирани и према категорији и поткатегорији. На слици 8 приказана је дистрибуција категорија говора мржње за прикупљени скуп података. Као што је раније поменуто, мржња по националној припадности заиста јесте најучесталији облик мржње у говору на српском језику. Следе мржња на основу политичког опредељења, полног/сексуалног опредељења (LGBTQ), етничке припадности и говор који садржи комбиноване облике мржње. Од 12 анализираних категорија показало се да су прве три категорије појављују у чак око 75% кратких текстова са говором мржње. Категорија „Комбиновано“ означава више категорија говора мржње у одређеном кратком тексту.



Слика 8. Дистрибуција категорија говора мржње.

V. Закључак

Говор мржње у дигиталном окружењу представља сложен друштвени проблем који захтева систематичан и интердисциплинаран приступ. У овом раду приказан је детаљан процес прикупљања и анотације текстуалних података на српском језику, са посебним фокусом на разликовање увредљивог говора и говора мржње, као и на категоризацију различитих облика дискриминације. Анализа прикупљеног скупа података показује очекивану неравнотежу класа, али и довољну разноврсност садржаја неопходну за обучавање и тестирање модела вештачке интелигенције.

Посебан допринос рада огледа се у дефинисању јасног и применљивог система анотације прилагођеног језичким и друштвеним специфичностима српског језика, као и у идентификовању доминантних категорија говора мржње у онлајн комуникацији. Добијени скуп података представља вредан ресурс за даља истраживања у области обраде природних језика и аутоматске детекције штетног садржаја.

Будући рад може бити усмерен на проширење скупа података, унапређење балансираности класа и примену напреднијих великих језичких модела у процесу анотација података. Након завршетка припремања скупова података, у наставку пројекта аутори ће се бавити анализом модела који би били најпогоднији за проблем детекције говора мржње за српски језик. На тај начин могуће је допринети развоју ефикаснијих модела и система за превенцију говора мржње и очување здравијег и одговорнијег дигиталног јавног простора.

Етичка изјава

Сви примери говора мржње укључени у ову студију генерисани су искључиво у сврху истраживања и припремања скупа података за даљи развој и обучавање модела. Њихова употреба има за циљ да подржи развој аутоматизованих детектора заснованих на моделима вештачке интелигенције, способних да детектују, класификују и цензуришу говор мржње на српском језику. Аутори експлицитно осуђују говор мржње у свим облицима и циљ им је да промовишу свест и образовање о његовим штетним ефектима, како у Србији тако и на међународном нивоу.

Захвалница

Истраживање спроведено уз подршку Фонда за науку Републике Србије, бр. пројекта 11113, “*Software for Text Offences Prevention in Serbian: AI-driven Hate Speech Detection*” – STOP. Истраживање је спроведено у Палати науке – Задужбини Миодрага Костића, у Центру за примену вештачке интелигенције.

Литература

[1] A. Milaković, V. Jocović, J. Cincović, M. Mićović, U. Radenković, D. Drašković, Detecting ugly and derogatory words in Serbian language using a web browser extension, 32nd Telecommunication Forum TELFOR 2024, pp. 1 - 4, Telecommunications Society, University of Belgrade - School of Electrical Engineering, IEEE Region 8, IEEE Serbia & Montenegro Section, IEEE Serbia & Montenegro ComSoc Chapter, IEEE Serbia & Montenegro CIS Chapter, Belgrade, Serbia, Nov, 2024

[2] N. Ljubešić, T. Erjavec, V. Batanović, M. Miličević, and T. Samardžić, “Serbian Twitter training corpus ReLDI-NormTagNER-sr 3.0,” 2019. [Online]. Доступно на: http://hdl.handle.net/11356/1794 (приступано дана: 7.2.2025.)

[3] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated Hate Speech Detection and the Problem of Offensive Language”, arXiv:1703.04009, 2017. [Online]. Доступно на: https://arxiv.org/pdf/1703.04009 (приступано дана 16.1.2025.)

Data collection and annotation in the process of forming datasets hate speech in the Serbian language

Dražen Drašković, Uroš Radenković, Marko Mićović,
Jelica Cincović, Adrian Milaković, Vladimir Jocović

ABSTRACT

Hate speech represents a serious social and technological challenge in the contemporary digital environment, particularly due to its widespread presence on social networks, online forums, and media platforms. The rapid growth of user-generated content, combined with anonymity and the lack of effective moderation, has significantly contributed to the proliferation of discriminatory, offensive, and violent discourse. In this context, the development of high-quality datasets is a crucial prerequisite for building reliable automated systems for hate speech detection, especially for languages with limited digital resources, such as Serbian.

This paper presents a comprehensive overview of the data collection and annotation process used in the formation of a hate speech dataset in the Serbian language. The dataset was created within a national research project aimed at preventing textual offenses through artificial intelligence methods. Data were collected from multiple sources, including social networks, online news portals, and public discussion platforms, using both automated techniques (APIs, web crawlers, and scrapers) and manual procedures. The resulting corpus consists of 4,300 short text samples. A three-class annotation scheme was adopted, distinguishing between non-hateful content, offensive language, and hate speech, in accordance with European Union guidelines and relevant legal and ethical frameworks. In addition, hate speech instances were further categorized based on different forms of discrimination, such as national, ethnic, religious, gender-based, and political hatred. The paper also provides statistical and qualitative analyses of the dataset, including class distribution, text length characteristics, and the most frequent terms associated with hate speech. The presented dataset represents a valuable resource for future research in natural language processing and artificial intelligence, enabling the development and evaluation of automated hate speech detection models for the Serbian language.

Primena veštačke inteligencije u analizi i obradi govora kroz transkripciju,verifikaciju i evaluaciju izjava

Ana Mandić[1], Marko Jelović[1], Ana Bulatović[1], Filip Nikolić[1], Hana Mijatović[1], Nikola Vučićević[1], Ana Stanić[1], Natalija Bogdanović[1], Luka Lazović[1], Anja Mihajlov[1], Lana Popović[1], Adrian Milaković[1], Vladimir Jocović[1], Dražen Drašković[1]

[1] Univerzitet u Beogradu, Elektrotehnički fakultet, Beograd, Bulevar kralja Aleksandra 73, SRBIJA

Kontakt autor: Dražen Drašković (ORCID: 0000-0003-2564-4526), e-pošta: drazen.draskovic@etf.bg.ac.rs

Apstrakt - U svetu gde se informacije šire izuzetnom brzinom, često je izazovno razlikovati tačne od netačnih podataka koje dobijamo iz medija. Ručna analiza javnih govora zahteva mnogo vremena i resursa, što je čini nepraktičnom za obradu velikog broja informacija. Kako bi se odgovorilo na ovaj izazov, razvijen je sistem koji kombinuje moć velikih jezičkih modela i pretragu na internetu, omogućavajući automatsku proveru činjenica i pružanje pouzdanih rezultata. Ovakva tehnologija ima široku primenu – od medijskih kuća i istraživačkih timova, do svih onih koji žele da dobiju pouzdane informacije i bolje razumeju sadržaj koji im se plasira.

Ključne reči – tačnost informacija, javni govor, veliki jezički modeli, veštačka inteligencija, automatizovana provera

I. Uvod

Ovaj rad istražuje primenu velikih jezičkih modela (engl. *Large Language Model*, skr. LLM) u analizi govora javnih ličnosti na srpskom jeziku, sa posebnim fokusom na detekciju dezinformacija i nekonzistentnosti [1][2]. Rad se sastoji iz tri ključna segmenta: transkripcije, provere tačnosti tvrdnji i provere nekonzistentnosti.

Transkripcija se bazira na sistemu koji pretvara govor u tekst, dok provera tačnosti koristi napredne modele za verifikaciju izrečenih tvrdnji putem pretrage na internetu. Sistem za detekciju nekonzistentnosti identifikuje nesuglasice u izjavama koristeći specifične velike jezičke modele za upoređivanje trenutno posmatranog govora sa postojećim javnim govorima sačuvanim u bazi podataka. Rezultati istraživanja pokazuju efikasnost velikih jezičkih modela u prepoznavanju suštinskih razlika između tačnih i netačnih izjava, kao i sposobnost identifikovanja njihovih nedoslednosti. Kombinovanjem ovih komponenti, naš cilj je da unapredimo transparentnost informacija i podignemo svest o odgovornosti onih koji ih iznose.

U savremenom informacionom društvu, sve veći značaj pridaje se razvoju inteligentnih sistema za obradu i analizu javno dostupnih informacija. Sa pojavom naprednih tehnika obrade prirodnog jezika (engl. *Natural Language Processing*, skr. NLP) i velikih jezičkih modela, stvoreni su uslovi za automatizovano prepoznavanje, analizu i evaluaciju izgovorenih izjava, što otvara prostor za primenu u različitim domenima – od novinarstva i politike do akademskih i regulatornih okvira [3].

Polazeći od pretpostavke da pouzdana analiza govora mora obuhvatiti sve njegove ključne komponente, u ovom radu se istražuje integrisani pristup koji objedinjuje tri funkcionalna podsistema: transkripciju govora, verifikaciju tačnosti tvrdnji i analizu konzistentnosti izjava. Poseban akcenat stavljen je na javni govor na srpskom jeziku, koji zbog svoje jezičke specifičnosti i manje zastupljenosti u globalnim modelima, predstavlja izazovan, ali značajan istraživački kontekst.

Funkcionisanje celokupnog sistema zasniva se na hijerarhijskoj obradi podataka – proces započinje automatskom transkripcijom audio/video zapisa pomoću modela za pretvaranje govora u tekst, čime se formira ulaz za naredne faze analize. U sledećim koracima, sistem koristi velike jezičke modele za identifikaciju potencijalno netačnih tvrdnji kroz upoređivanje sa relevantnim izvorima informacija, kao i za detekciju nekonzistentnosti u iskazima analizom prethodno datih javnih izjava iz odgovarajuće baze podataka.

U okviru ovog rada biće predstavljena arhitektura razvijenog sistema, primenjeni modeli i alati, kao i evaluacija efikasnosti svakog od podsistema. Kroz analizu dobijenih rezultata, biće razmotren potencijal ovakvog pristupa za unapređenje transparentnosti i odgovornosti u javnom govoru, kao i mogućnosti za dalji razvoj sistema u kontekstu šireg jezičkog okruženja.

II. Transkripcija i dijarizacija govora

Automatska transkripcija govora predstavlja ključni prvi korak u procesu analize audio i video zapisa, jer omogućava pretvaranje govora u tekstualni format, što je preduslov za sve dalje obrade. U okviru ovog projekta, poseban akcenat stavljen je na tačnost transkripcije i jasno razgraničenje među govornicima.

Za automatsku transkripciju korišćen je *Whisper* model. Prvobitno je testiran model *Whisper Medium*, zbog njegove brzine i nižih zahteva za resursima [4]. Međutim, ispitivanja su pokazala da ovaj model ne pruža zadovoljavajuću tačnost na srpskom jeziku, naročito u segmentima koji uključuju politički govor, zbog čega je odlučeno da se pređe na *Whisper Large* model, koji je omogućio značajno preciznije rezultate

[5].

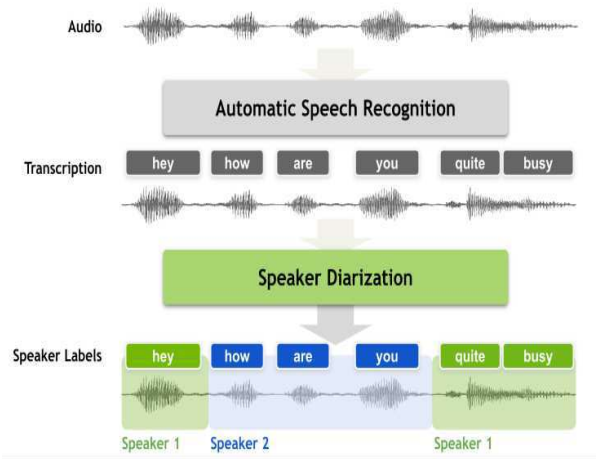
Nakon transkripcije, sprovedena je dijarizacija, odnosno proces identifikacije i odvajanja govornika u audio zapisu. Za ovu fazu korišćen je *Whisper-Diarization* alat, koji kombinuje automatsku transkripciju sa dodatnim modelima za prepoznavanje govornika, čime je omogućeno precizno označavanje govora svakog učesnika u transkriptu.

S obzirom na to da sirova transkripcija nije uvek optimalna za dalje analize, bilo je neophodno dodatno obraditi transkripte [6]. Za ovaj zadatak testirani su modeli *DeepSeek* i *GPT-4o Mini*, kao što je prikazano u Tabeli 1. Model *DeepSeek* je pokazao zadovoljavajuću tačnost, ali nedovoljnu brzinu za potrebe projekta. Sa druge strane, *GPT-4o Mini* je omogućio i visoku tačnost i znatno veću brzinu obrade, što je dovelo do njegovog izbora za završno uređivanje transkripta.

Tabela 1. Analizirani modeli za transkripciju

Model	Accuracy	Speed
DeepSeek	Dobro	Slabo
GPT-4o Mini	Dobro	Dobro

Radi dodatnog unapređenja strukture transkripta, sproveden je proces segmentacije teksta na logičke celine (chunking). Testirane su različite metode segmentacije, uključujući semantičko grupisanje na osnovu tematskih promena, kao i jednostavnije pristupe poput segmentacije na osnovu dužine rečenica i pauza u govoru [7]. Na slici 1 prikazana je ilustracija procesa transkripcije i dijarizacije govora.



Slika 1: Ilustracija procesa automatske transkripcije i dijarizacije govora.

Konačni rezultat transkripcije i dijarizacije predstavlja strukturirani JSON izlaz, u kojem su za svaki segment definisani govornik i pripadajuća tema razgovora prikazano na slici 2. Ovaj format omogućava jednostavno prosleđivanje podataka sistemima zaduženim za proveru činjenica i konzistentnosti izjava, čime se značajno ubrzava analiza i smanjuje potreba za manuelnim radom.


```
"date": "2/3/2025",
"topics": [
  {
    "topic_name": "Rušenje Srbije",
    "description": "Diskusija o političkoj i društvenoj situaciji u Srbiji, uključujući
                    kritiku vlasti, prelaznu vladu i rušenje same države.",
    "speaker_segments": [
      {
        "speaker_name": "Interviewers",
        "start_time": 0.00,
        "end_time": 2.596,
        "text": "Da li razmišljate o ostavci ili ekspertskoj prelaznoj vladi?"
      },
      {
        "speaker_name": "Speakers",
        "start_time": 3.403,
        "end_time": 35.827,
        "text": "Nismo se ovde skupili da čuvamo vlast bilo čiju, da čuvamo nečiju
                 fotelju, nismo došli da govorimo o tome koliko žele izbore..."
      },
      {
        "speaker_name": "Speakers",
        "start_time": 36.231,
        "end_time": 62.084,
        "text": "A onda je došlo vreme za njihov pokušaj da Srbiju sruše. Naš
                 je posao da svim tim ljudima pružimo ruku, da pokušamo..."
      }]
  }]
```


Slika 2. Izgled JSON izlaza

Proces transkripcije i dijarizacije može se predstaviti sledećim koracima:

A. ***Ulazni podaci**: audio/video zapis*

B. ***Transkripcija** korišćenjem Whisper modela*

C. ***Dijarizacija** i razdvajanje govornika*

D. ***Klasifikacija govornika***

E. ***Poliranje i uređivanje transkripta***

F. ***Chunking** (podela na logičke celine)*

G. ***Generisanje JSON izlaza***

Automatizacijom ovog procesa uspostavljen je čvrst i pouzdan temelj za dalje korake u analizi govora.

III. PROCES PROVERE ČINJENICA (FACT CHECKING)

U ovom projektu, automatski sistem za proveru činjenica koristi velike jezičke modele za obradu prirodnog jezika i pretragu na internetu kako bi se verifikovale tvrdnje izrečene u audio i video zapisima.

Početak procesa provere činjenica je prijem podataka u JSON formatu od sistema za transkripciju. Ovaj format sadrži tekstualne transkripte govora, uključujući tvrdnje koje je potrebno proveriti, kao i relevantne informacije o kontekstu, poput datuma govora i imena govornika. Na osnovu ovih podataka, veliki jezički model analizira sadržaj i izdvaja ključne činjenice iz svake tvrdnje. Na ovaj način, sistem generiše listu činjenica koje je potrebno proveriti putem pretrage na internetu.

Nakon što se činjenice izdvoje, sistem generiše odgovarajuće upite za pretragu na internetu. Upiti se šalju pretraživaču korišćenjem SerpAPI servisa, koji omogućava prikupljanje rezultata sa pouzdanih i kredibilnih izvora, koji uključuju članke iz vesti, naučne publikacije i vladine izvore, dok se izbegavaju tabloidi i druge nesigurne platforme [8].

Prikupljeni podaci se potom preuzimaju sa relevantnih veb stranica koristeći Jina API servis, koji omogućava ekstrakciju sadržaja [9].

Kada su podaci prikupljeni, sistem koristi LLM za analizu i upoređivanje sa originalnim tvrdnjama, kako bi procenio tačnost informacija. Sistem generiše izveštaj, koji pruža detaljan pregled svih činjenica koje su bile predmet provere, zajedno sa oznakama koje ukazuju na njihov status tačnosti (tačne, netačne ili neodređene). Pored toga, za svaku netačnu tvrdnju, izveštaj sadrži tačne informacije, kao i izvore sa kojih su tačne tvrdnje preuzete. Ovaj izveštaj se generiše u CSV formatu, čime se omogućava brzo pregledanje i dalja obrada rezultata.

Ovaj proces omogućava efikasnu i brzu proveru činjenica, smanjujući potrebu za manuelnim radom i povećavajući preciznost u analizi. Automatizacija ovog procesa čini ga pogodnim za velike količine podataka, kao što su transkripti govora političara i drugih javnih ličnosti, čime se omogućava brza detekcija netačnih ili manipulativnih tvrdnji. Konačan rezultat ovog procesa je tačna, objektivna i verifikovana lista činjenica koja je dostupna za dalje analize ili javnu upotrebu. Proces provere činjenica može se predstaviti sledećim koracima:

A. ***Ulazni podaci**: JSON format sa transkriptima govora*
B. ***Ekstrakcija ključnih činjenica** korišćenjem LLM modela*
C. ***Generisanje upita za pretragu** koji su optimizovani za pronalaženje relevantnih informacija.*
D. ***Pretraga na internetu** korišćenjem SERPAPI servisa.*
E. ***Preuzimanje sadržaja** sa relevantnih veb stranica koristeći Jina API.*
F. ***Analiza i procena tačnosti** prikupljenih podataka sa LLM-om.*
G. ***Generisanje izveštaja u CSV formatu**, koji uključuje verifikovane činjenice i njihove izvore.*

Automatizacija ovog procesa omogućava bržu, precizniju i efikasniju analizu govora, posebno u kontekstu političkih i društvenih debata, gde tačnost informacija ima ključnu ulogu u oblikovanju javnog mišljenja.

IV. PROCES PROVERE KONZISTENTNOSTI (CONSISTENCY CHECKING)

Consistency checking segment, zadužen za automatsku proveru konzistentnosti javnih govora, temelji se na arhitekturi generisanja proširenim pretraživanjem (engl. *Retrieval-Augmented Generation,* skr. RAG) [10]. U izradi celog sistema pa tako i segmenta provere konzistentnosti korišćen je Python kao osnovni programski jezik, prvenstveno zbog svoje čitljivosti, jednostavne sintakse i širokog spektra biblioteka koje omogućavaju efikasan rad sa podacima, implementaciju modela veštačke inteligencije, kao i jednostavnu integraciju sa različitim API servisima.

Izlaz sistema na kojem je rađeno u početnom delu sistema transkripcije je upravo ulaz sistema zaduženog za proveru konzistentnosti govora, to je JSON fajl koji sadrži transkribovani tekst izdeljen na segmente, kao i sve neophodne informacije o sagovornicima, dužini trajanja razgovora itd.

I faza - klasifikacija

U početnoj fazi rada, korišćenjem velikog jezičkog modela izvršena je klasifikacija segmenata JSON fajla kako bismo utvrdili da li transkript govora sadrži informacije za koje je potrebno ispitati konzistentnost ili ne. Na primer, izjave poput „dobar dan“ ili „kako ste“ ne podležu daljoj analizi, jer ne sadrže suštinske informacije koje bi mogle biti predmet provere.

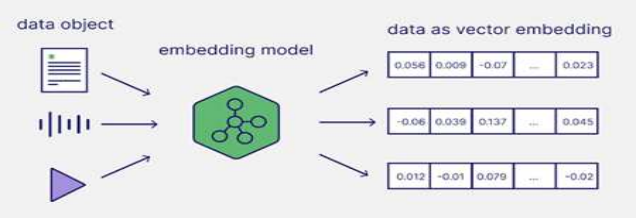
Pre samog procesa klasifikacije neophodno je podeliti svaki segment JSON fajla na manje celine, kako bi se obezbedio što efikasniji dalji rad. Za ovo je korišćena tehnika segmentiranje fiksne dužine (engl *fixed length chunking*) sa preklapanjem, pri čemu smo u početku fiksnu dužinu celine podesili na 200 karaktera i 50 karaktera za preklapanje. Na samom kraju impementacije sistema smo podešavali ove dužine kako bismo dobili što tačnije rezultate. Za potrebe klasifikacije segmenata iz transkribovanih govora korišćen je *OpenRouterAPI*, koji omogućava fleksibilno povezivanje sistema sa više velikih jezičkih modela.

U okviru ovog sistema, primenjen je model Claude 3 Haiku kompanije Anthropic, poznat po efikasnosti u obradi i razumevanju prirodnog jezika.

Klasifikacija je izvršena prompt-based metodom, u zero-shot režimu, pri čemu model na osnovu jasno definisanih uputstava (engl. *prompt*) odlučuje da li određeni segment sadrži proverivu tvrdnju relevantnu za političku funkciju.Ova procedura omogućava visoku fleksibilnost i skalabilnost bez potrebe za dodatnim treniranjem modela na domenskim podacima.

II faza – embedovanje

Ukoliko je neophodno ispitati konzistentnost segmenta vršimo njegovo embedovanje.



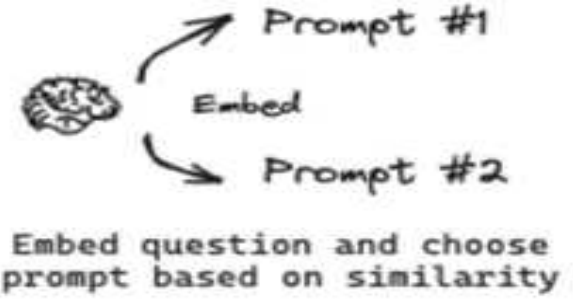
Slika 3. Prikaz procesa embedovanja

Embedovanje je ključni proces pretvaranja originalnih podataka, u našem slučaju pretvaranje transkribovanog govora, u numeričke vektore u višedimenzionalnom prostoru, kao što je prikazano na Slici 3. Ovi vektori su od suštinskog značaja, jer sažimaju semantičko značenje originalnih podataka na način koji omogućava računarskim modelima da lakše uporede i analiziraju sadržaj.

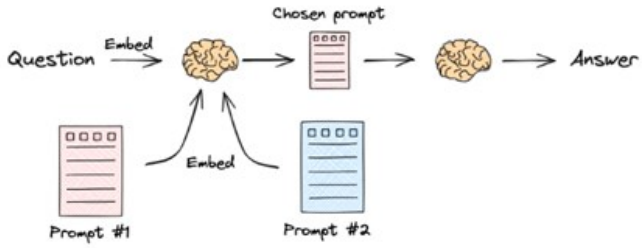
Za proces ugradnje (engl. *embedding*) tekstova u vektorski prostor korišćen je COHERE API, sa modelom *embed-multilingual-v3.0* [11]. Izabrali smo ovaj model jer je posebno specijalizovan za rad sa srpskim jezikom i time omogućava očuvanje semantičkog značaja tekstova prilikom vektorisanja, što je korisno za dalji rad.

III faza - Stvaranje baze, pretraga, odgovori

Sledeći korak podrazumeva pretraživanje baze podataka koja u sebi sadrži embedinge svih prethodnih govora neophodnih za proveru konzistentnosti. U ovom koraku je cilj pronaći prvih k embedinga koji su najrelevantniji za ispitivanje posmatranog govora korišćenjem K-najbližih suseda (engl. *K nearest neighbor search*, skr. KNN). Odabrani najrelevantniji embedinzi će se potom koristiti za generisanje odgovora kojim se popunjava kolona CSV fajla .



Slika 4. Izbor prompta na osnovu semantike sličnosti



Slika 5. Mehanizam odgovora pomoću najbližeg prompta

Ovaj odgovor može biti:

- „Konzistentno" / „Kontradiktorno" uz prethodno datu izjavu i datum kada je to rečeno
- „Ne znam" uz dodatno objašnjenje u slučaju kada nema relevantnih podataka za proveru

	A	B	C
1	Sentence;	"Response";	"Relevant";
2	Rekli ste da Srbija g	"Kontradiktorno";	"Srbija ima najvece
3	Vise od pola miliona	"Ne znam. Nema r	"Nema relevantnih
4	Tokom 2018. godin	"Konzistentno";	"Da li znate da smo

Slika 6. Prikaz generisanog odgovora u alatu Excel

U određenim delovima sistema, gde je bilo potrebno dodatno zaključivanje i generisanje odgovora, korišćen je i model GPT-4o-mini, poznat po brzini i efikasnosti u obradi prirodnog jezika. Kombinacijom ovih servisa i modela razvijen je pouzdan sistem sposoban da automatski detektuje kontradiktorne tvrdnje u tekstovima, uz mogućnost proširenja i primene u različitim domenima.

Jedna od glavnih poteškoća tokom razvoja bila je vezana za korišćenje OpenRouter API, budući da su API ključevi brzo trošeni zbog ograničenog broja poziva, što je zahtevalo pažljivo planiranje i optimizaciju korišćenja modela tokom testiranja i razvoja [12].

V. Zaključak

Istraživanje predstavljeno u ovom radu omogućilo je dublje razumevanje potencijala velikih jezičkih modela u oblasti analize javnog govora, sa posebnim fokusom na automatsku transkripciju, proveru tačnosti i detekciju nekonzistentnosti izjava. Kroz implementaciju savremenih modela poput Whisper, GPT-4o Mini i dodatnih API servisa, razvijen je funkcionalan sistem sposoban za obradu složenih audio i video zapisa, njihovu preciznu segmentaciju, kao i semantičku analizu izrečenih tvrdnji.

U radu su obrađeni izazovi transkripcije i dijarizacije, generisanja upita, evaluacije sadržaja sa interneta i izvođenja zaključaka o tačnosti izjava, a predstavljeni rezultati ukazuju na visoku efikasnost sistema u realnim scenarijima. Iako je evaluacija sprovedena na ograničenom broju govora, dobijeni nalazi pružaju vredne uvide za buduća istraživanja u oblasti političke analitike, medijske pismenosti i automatizovanog alata za proveru činjenica.

Dalji razvoj mogao bi da uključi rad sa većim korpusima, primenu modela koji podržavaju više jezika, kao i integraciju sa bazama podataka zvaničnih izvora informacija, čime bi se dodatno povećala pouzdanost i tačnost rezultata u analizi javnog diskursa.

Zahvalnica

Istraživanje sprovedeno uz podršku Fonda za nauku Republike Srbije, br. projekta 11113, "*Software for Text Offences Prevention in Serbian: AI-driven Hate Speech Detection*" – STOP. Istraživanje je sprovedeno u Palati nauke - Zadužbini Miodraga Kostića, u Centru za primenu veštačke inteligencije.

Literatura

[1] Vykopal, I., Pikuliak, M., Srba, I., Moro, R., Macko, D. and Bielikova, M., 2024, August. Disinformation capabilities of large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 14830-14847).

[2] Pavlyshenko, B.M., 2023. Analysis of disinformation and fake news detection using fine-tuned large language model. *arXiv preprint arXiv:2309.04704*.

[3] Kuntur, S., Wróblewska, A., Paprzycki, M. and Ganzha, M., 2024. Under the influence: A survey of large language models in fake news detection. *IEEE Transactions on Artificial Intelligence*.

[4] Stanojev, V., Nosek, T., Suzić, S., Pekar, D., Delić, V. and Sečujski, M., 2025, October. Improving Whisper-Based Serbian ASR Using Synthetic Speech. In *International Conference on Speech and Computer* (pp. 118-129). Cham: Springer Nature Switzerland.

[5] Macháček, D., Dabre, R. and Bojar, O., 2023. Turning whisper into real-time transcription system. *arXiv preprint arXiv:2307.14743*.

[6] Spiller, T.R., Rabe, F., Ben-Zion, Z., Korem, N., Burrer, A., Homan, P., Harpaz-Rotem, I. and Duek, O., 2023. Efficient and accurate transcription in mental health research-A tutorial on using whisper AI for audio file transcription. *OSF Preprints*.

[7] Zhao, J., Ji, Z., Fan, Z., Wang, H., Niu, S., Tang, B., Xiong, F. and Li, Z., 2025. MoC: Mixtures of Text Chunking Learners for Retrieval-Augmented Generation System. *arXiv preprint arXiv:2503.09600*.

[8] SerpAPI [Online]. Available: https://serpapi.com/ (accessed on January, 20th, 2025).

[9] JinaAPI [Online]. Available: https://jina.ai/ (accessed on January, 27th, 2025).

[10] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H. and Wang, H., 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, *2*(1).

[11] Kamalloo, E., Zhang, X., Ogundepo, O., Thakur, N., Alfonso-Hermelo, D., Rezagholizadeh, M. and Lin, J., 2023. Evaluating embedding APIs for information retrieval. *arXiv preprint arXiv:2305.06300*.

[12] Vijayakumar, P., Pyingkodi, M. and Devi, S., 2025, July. Comparative Analysis of AI Chatbot for Assessing Gemini AI, DeepSeek AI, and Qwen AI via OpenRouter API Integration. In *2025 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)* (pp. 01-08). IEEE.

Application of artificial intelligence in speech analysis and processing through transcription, verification and evaluation of statements

Ana Mandić, Marko Jelović, Ana Bulatović, Filip Nikolić, Hana Mijatović, Nikola Vučićević, Ana Stanić, Natalija Bogdanović, Luka Lazović, Anja Mihajlov, Lana Popović, Adrian Milaković, Vladimir Jocović, Dražen Drašković
(University of Belgrade, School of Electrical Engineering, Serbia)

ABSTRACT

In a world where information spreads at an exceptional speed, it is often challenging to distinguish accurate from inaccurate data obtained from the media. Manual analysis of public statements requires a significant amount of time and resources, making it impractical for processing large volumes of information. To address this challenge, a system has been developed that combines the power of large language models with web search, enabling automated fact-checking and the delivery of reliable results. Such technology has a wide range of applications—from media organizations and research teams to anyone seeking trustworthy information and a better understanding of the content presented to them.

YU #4: Sesija 4

Softverski sistemi i savremene informacione tehnologije

Vremenska analiza procesa integracije primenom Simpsonovog pravila na procesorima sa više jezgara u Javi

Marko Smilić
Univerzitet u Prištini sa privremenim sedištem u Kosovskoj Mitrovici, Prirodno-matematički Fakultet
Kosovska Mitrovica, Srbija
marko.smilic@pr.ac.rs
0000-0003-2227-2536

Časlav Stefanović
Univerzitet u Prištini sa privremenim sedištem u Kosovskoj Mitrovici, Prirodno-matematički Fakultet
Kosovska Mitrovica, Srbija
caslav.stefanovic@pr.ac.rs
0000-0003-4385-3356

Dejan Milić
Univerzitet u Nišu, Elektronski Fakultet
Niš, Srbija
dejan.milic@elfak.ni.ac.rs
0000-0001-6472-2027

Sanja Đurović
Univerzitet u Prištini sa privremenim sedištem u Kosovskoj Mitrovici, Prirodno-matematički Fakultet
Kosovska Mitrovica, Srbija
sanja.djurovic@pr.ac.rs
0000-0001-7847-9953

Danijel Đošić
Univerzitet u Prištini sa privremenim sedištem u Kosovskoj Mitrovici, Prirodno-matematički Fakultet
Kosovska Mitrovica, Srbija
danijel.djosic@pr.ac.rs
0000-0002-0144-5795

***Abstrak* – U ovom radu razmatramo vremensku analizu procesa integracije korišćenjem Simpsonovog pravila na procesorima sa više jezgara u programskom jeziku Java. Kako se računarska snaga povećava, paralelizam na procesorima sa više jezgara postao je ključan u ubrzavanju matematičkih operacija. Istražujemo kako se Simpsonovo pravilo, široko korišćeni metod za numeričku integraciju, može optimizovati za paralelno izvršavanje. Merimo vreme procesa integracije Frenelovog integrala, transcendentalne matematičke funkcije, na procesorima sa više jezgara i upoređujemo ga sa implementacijom sa jednim jezgrom. Takođe, grafički prikazujemo rezultate implementacije Simpsonovog pravila za određeni hardver (broj jezgara i specifikacije procesora), određen softver (Java verzija i Java biblioteke) i tačnost rezultata u zavisnosti od praga. Rezultati ukazuju na značajno smanjenje vremena računanja sa povećanjem broja dostupnih jezgara i bolju preciznost izborom optimalne vrednosti za prag. Pored toga, razmatramo implikacije za računarstvo visokih performansi u Javi**.

***Ključne reči* – Numerička integracija, Simpsonovo pravilo, Procesori sa više jezgara, Paralelno računanje, Java programiranje.**

I. Uvod

Potreba za implementacijom bržih i efikasnijih numeričkih metoda u savremenim računarskim zadacima dovela je do značajnog interesovanja za paralelnim računarstvom. Kako se povećanje računarske snage nastavilo sa razvojem procesora sa više jezgara, paralelno izvršavanje matematičkih algoritama postalo je ključno za postizanje računarskih performansi visokog nivoa. [1]. Numeričko rešavanje integral predstavlja ključnu operaciju u oblastima kao što su fizika, matematika, inženjerstvo i ekonomija. Potrebe ovih oblasti često uključuju zahtevne i obimne proračune u kojima se primenjuje proces paralelizacije [2] - [5]. Jedan takav metod, Simpsonovo pravilo, se često koristi za proces numeričke integracije zbog svoje jednostavnosti, preciznosti rezultata i efikasnosti u aproksimaciji funkcija određenih integrala [6]. Međutim, kako matematički modeli i simulacije postaju sve kompleksiniji, tradicionalna obrada zadataka na jednom jezgru može postati usko grlo u pogledu performansi, pre svega zbog velikih količina podataka ili složenosti zadataka koji se rešavaju [7].

Paralelno računanje, posebno korišćenjem procesora sa više jezgara, nudi idealno rešenje za ove računarske izazove. Podelom radnog opterećenja na više procesorskih jezgara, paralelno izvršavanje može značajno smanjiti vreme potrebno za izvođenje računarskih zadataka, omogućavajući obradu složenijih problema u kraćem vremenskom intervalu [8] - [10]. Korišćenje procesora sa više jezgara izazvalo je revolucionarne promene u mnogim oblastima, ubrzavajući računarski zahtevne zadatke kao što su numerička integracija, operacije sa matricama i razvoj algoritama za simulaciju [11]. U ovom kontekstu, tehnike paralelizacije mogu pomoći u optimizaciji metoda poput Simpsonovog pravila, koji se ranije obično izvršavao na jednom jezgru, tako što će se proces integracije raspodeliti na više jezgara.

Ovaj rad se fokusira na analizu vremena izvršenja Simpsonovog pravila – klasične metode numeričke integracije – kada se izvodi paralelno na procesorima sa više jezgara koristeći programski jezik Java. Iako je paralelizam dobro utemeljen u računarskim zadacima [12], optimizacija umeričkih metoda poput Simpsonovog pravila za sisteme sa više jezgara predstavlja jedinstvene izazove i prilike za poboljšanje računarske efikasnosti [13], [14]. Sa svojim razvijenim ekosistemom i podrškom za višenitno programiranje i paralelne računarske okvire poput ForkJoinPool i ExecutorService [15], programski jezik Java pruža svestranu platform za implementaciju ovakvih paralelnih procesa na način koji je istovremeno efikasan i prenosiv na različite hardverske konfiguracije.

Nedavne studije ističu napredak u tehnikama numeričke integracije, fokusirajući se na tačnost i efikasnost. Abdulhameed i Memon [16] predstavili su poboljšano trapezoidno pravilo sa smanjenom greškom, dok su komparativne analize Karpagam i Vijayalakshmi [17] i Dhali i dr. [18] naglasile veću preciznost Simpsonovog

pravila za glatke funkcije, posebno kod ravnomerno raspoređenih podataka. Vinsensia i dr. [19] naglasili su važnost kombinovanja metoda poput Gauss-Legendre pravila i Simpsonovog pravila za složene integrande, pokazujući značaj izbora metode u zavisnosti od primene. Nowicki i dr. [20], [21] demonstrirali su efikasnost Java biblioteke PCJ u radu sa HPC (high-performance computing), Big Data i AI radnim opterećenjima, kao i njenim performansama u oblaku za paralelne algoritme. Bhojwani i Singh [22] istraživali su paralelizaciju u određivanju integrala, dok su Mehrabi i dr. [23] predstavili @PT, Java okvir sa oznakama koji pojednostavljuje paralelno programiranje, naglašavajući nenametljivu implementaciju i poboljšanu konkurentnost.

U ovom radu ćemo istražiti kako Simpsonovo pravilo, kada se prilagodi paralelnom izvršavanju, koristi prednosti arhitektura sa više jezgara u smislu vremena izvršavanja. Analizirajući proces integracije Frenelovog integrala sinusne funkcije, demonstriraćemo kako paralelizacija može značajno smanjiti vreme računanja u poređenju sa implementacijom na jednom jezgru. Dalje analiziraćemo uticaj različitih hardverskih konfiguracija (broj jezgara i specifikacije procesora) i softverskih podešavanja (verzija Jave i biblioteke za paralelizaciju) na performanse, pružajući uvid u praktične aspekte implementacije efikasnih paralelnih algoritama u Javi.

Rezultati ovog istraživajna pružaju sveobuhvatno razumevanje potencijalnih prednosti korišćenja procesora sa više jezgara u zadacima numeričke integracije. Ovi rezultati doprinose optimizaciji matematičkih algoritama u HPC, posebno u okruženjima koja se u velikoj meri oslanjaju na Java aplikacije.

U narednim sekcijama ćemo predstaviti detaljan opis metodologije korišćene za paralelizaciju Simpsonovog pravila, analizu rezultata i diskusiju o implikacijama za buduća istraživanja u oblasti HPC i dizajna algoritama za paralelno izračunavanje.

II. Simpsonovo Pravilo

Simpsonovo pravilo se zasniva na aproksimaciji površine ispod krive prilagođavanjem paraboličnih segmenata funkciji. Umesto korišćenja pravolinijskih segmenata (kao kod trapezoidnog pravila), Simpsonovo pravilo koristi kvadratne funkcije za modelovanje funkcije preko malih intervala. Primenom kvadratne aproksimacije na svaki mali segment intervala integracije, Simpsonovo pravilo može preciznije proceniti površinu ispod krive, posebno za glatke i neprekidne funkcije [24]. Simpsonovo pravilo je definisano formulom:

$$\int_a^b f(x)dx \approx \frac{h}{3}\big(f(a) + 4(a+h) + f(b)\big) \qquad (1)$$

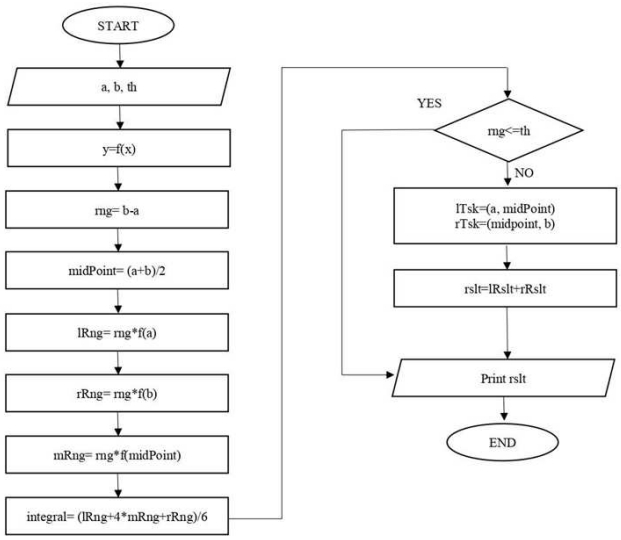
gde a i b predstavljaju donju i gornju granicu integrala, a $h = \frac{b-a}{n}$ predstavlja korak sa n podintervala [25], [26]. Za $n = 2,$ formula Simpsonovog pravila postaje:

$$\int_a^b f(x)dx \approx \frac{(b-a)}{6}\left(f(a) + 4\left(\frac{a+b}{2}\right) + f(b)\right) \qquad (2)$$

Simpsonovo pravilo podrazumeva evaluaciju funkcije u više tačaka unutar intervala integracije. Ove evaluacije mogu se podeliti na manje, nezavisne zadatke koji ne zavise jedan od drugog, što Simpsonovo pravilo čini inherentno pogodnim za paralelizaciju. Podelom intervala integracije, Simpsonovo pravilo može biti efikasno prilagođeno za iskorišćavanje pune snage procesora sa više jezgara [27].

Podela domena integracije na manje podintervale i dodeljivanje tih podintervala različitim jezgrima predstavlja ključni aspekt paralelnog izvršavanja numeričke integracije. Ovaj process uključuje podelu domena, raspodelu podintervala među jezgrima, omogućavanje svakom jezgru da nezavisno obavlja proračune, a zatim sinhronizaciju i agregaciju rezultata [28].

Ovakav pristup koristi računarsku snagu procesora sa više jezgara kako bi se značajno smanjilo vreme potrebno za numeričku integraciju velikih zadataka. Svakom jezgru se dodeljuje približno jednak deo ukupnog domena integracije. Ako je broj podintervala n deljiv brojem jezgara p, tada svako jezgro može obradjivati n/p podintervala. Ako n nije savršeno deljiv sa p, neka jezgra mogu obraditi jedan podinterval više od drugih. Idealno, podela treba da osigura ravnomernu raspodelu računarskog opterećenja među jezgrima.



Slika 1. Implementacija algoritma Simpsonovog pravila

Java pruža skup alata za konkurentno programiranje u okviru paketa java.util.concurrent, omogućavajući programerima da iskoriste više procesorskih jezgara i izvršavaju zadatke paralelno. Klase ključne za upravljanje paralelizmom uključuju ForkJoinPool, ExecutorService, Executor i ThreadPoolExecutor. Ove klase su deo Java Concurrency API-ja, koji nudi apstrakcije visokog nivoa za upravljanje nitima, raspoređivanje zadataka i paralelne proračune [29], [30].

U ovom radu je prikazana vremenska analiza numeričke integracije nad Frenelovim integralom primenom Simpsonovog pravila na procesorima sa više jezgara u programskom jeziku Java [31].

$$S(x) = \frac{2}{\sqrt{2\pi}}\int_0^x (\sin(t^2)\,dt) \qquad (3)$$

Analizirali smo vreme izvršavanja na tri različita računara i uporedili smo rezultate uzimajući u obzir uticaj verzija Jave

i biblioteka za paralelizaciju. Pored toga, analizirali smo tačnost rezultata u zavisnosti od praga.

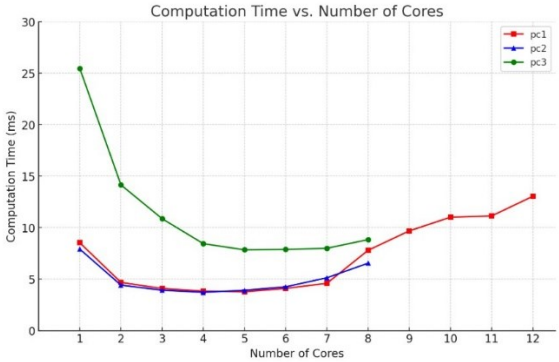
III. Rezultati i diskusija

U ovom odeljku predstavljamo rezultat analize vremena procesa integracije koristeći Simpsonovo pravilo na procesorima sa više jezgara. Takođe, prikazujemo tačnost rezultata u zavisnosti od vrednosti praga. Tabela 1 prikazuje hardverske specifikacije korišćenih računara. PC1 koristi verziju Jave 23, PC2 koristi verziju Jave 11, a PC3 koristi verziju Jave 8. Na svim računarima koristimo ForkJoinPool za upravljanje nitima.

Tabela 1. Specifikacija procesora korišćenih računara

#	Tip procesora	Frekvencija	Osnovni takt procesora	Turbo takt procesora	Jezgra	Niti
PC 1	13th Intel® Core™ i7, 1355U	1.7GHz	100MHz	5GHz	10	12
PC 2	11th Intel® Core™ i7,1165G7	2.8GHz	100MHz	4.7GHz	4	8
PC 3	Intel® Core™ i7-930 Processor	2.8GHz	133MHz	3.1GHz	4	8

Slika 2 prikazuje odnos između vremena računanja i broja procesorskih jezgara na tri računarska sistema (PC1, PC2 i PC3). X-osa predstavlja broj jezgara, dok Y-osa označava vreme računanja u milisekundama..



Slika 2. Vremenska analiza procesa integracije primenom Simpsonovog pravila za različit broj jezgara (niti)

Rezultati ukazuju na značajno smanjenje vremena računanja kako se broj procesorskih jezgara povećava. Ovo ponašanje je u skladu sa očekivanjima, jer paralelizacija numeričke integracije omogućava istovremeno izvršavanje više računarskih zadataka. Međutim, ubrzanje nije striktno linearno, što sugeriše prisustvo overhead-a i ograničenja resursa.

Na svim testiranim sistemima (PC1, PC2 i PC3) primetan je brz pad vremena izvršenja pri prelasku sa jednog jezgra na konfiguracije sa više jezgara. Međutim, analiza grafikona pokazuje da se najznačajnije poboljšanje performansi dešava do pet jezgara, nakon čega smanjenje vremena izvršenja postaje manje izraženo. Ovaj efekat je naročito očigledan u konfiguraciji sa više od osam jezgara, gde dalja povećanja broja jezgara ne donose značajne dobitke u performansama.

Upoređivanje testiranih sistema otkriva razlike u apsolutnim vremenima izvršenja. PC2 dosledno postiže niža vremena računanja u odnosu na PC3, dok PC1 pokazuje najduže vreme izvršenja među testiranim sistemima. Ove varijacije mogu se pripisati hardverskim specifikacijama, arhitekturi procesora, širini memorijskog interfejsa i razlikama u efikasnosti algoritma.

Tabela 2. Tačnost rezultata u zavisnosti od vrednosti praga.

#	Vrednost praga	Rezultat direktne integracije	Rezultat primene Simpsonovog pravila
1	0.01	0.097913140298856	0.097912862499545
2	0.001	0.097913140298856	0.097913140231474
3	0.0001	0.097913140298856	0.097913140298855
4	0.00001	0.097913140298856	0.097913140298856
5	0.000001	0.097913140298856	0.097913140298856

Tabela 2 prikazuje tačnost rezultata u zavisnosti od vrednosti praga. Rezultati pokazuju da metode direktne integracije i Simpsonovo pravilo za numeričku integraciju proizvode gotovo identične rezultate kako se vrednost praga smanjuje. Pri višim vrednostima praga, kao što je 0.01, postoji mala razlika između ove dve metode, ali kako se prag smanjuje, razlike brzo nestaju. Kada prag dostigne vrednost 0.0001, rezultati obe metode se podudaraju do nekoliko decimalnih mesta, što ukazuje na to da obe metode konvergiraju ka istom rešenju.

Kako prag prilazi minimalnim vrednostima (ispod 0.0001), obe numeričke tehnike integracije pokazuju izuzetnu doslednost u svojim rezultatima. Ovo ukazuje na to da je Simpsonovo pravilo, kao metoda aproksimacije, gotovo isto precizno kao direktna integracija za ovaj problem, naročito kada je potrebna visoka preciznost. S obzirom na blisku usklađenost rezultata pri minimalnim pragovima, može se zaključiti da je Simpsonovo pravilo pouzdana i efikasna metoda za integraciju, naročito kada su računarski troškovi ili vreme bitni faktori.

IV. Zaključak

U paralelnom računarstvu u Javi, istraživači nastavljaju sa istraživanjima i inovacijama kako bi poboljšali efikasnost i skalabilnost algoritama u različitim računarskim zadacima. Studije su se fokusirale na optimizaciju paralelnih implementacija metoda numeričke integracije, kao što su trapezoidno i Simpsonovo pravilo, kako bi se iskoristili procesori sa više jezgarai smanjilo vreme izvršavanja za probleme velikih zadataka.

U ovom radu, prikazane su prednosti paralelnog računanja u Javi korišćenjem Simpsonovog pravila za rešavanje Frenelovog integrala. Rezultati dobijeni u ovom radu pokazali su značajno smanjenje vremena izračunavanja integrala primenom paralelizacije, nezavisno od arhitekture procesora. Programsko okruženje Java i njegova ForkJoinPool biblioteka omogućili su efikasno upravljanje nitima. Vrednost praga ima značajnu ulogu u određivanju preciznosti rezultata numeričke integracije. Suštinski, ona deluje kao mera prihvatljive margine greške u aproksimaciji.

Rezultati prikazani u ovom radu otvaraju mogućnosti za dalja istraživanja u oblasti HPC, matematičkog modelovanja,

aproksimacije i optimizacije korišćenjem Simpsonovog pravila i paralelnog računanja..

REFERENCE

[1] R. Robey and Y. Zamora, *Parallel and High-Performance Computing*. New York: Manning Publications Co., 2021.

[2] Y. Huang, "Parallel computing and its applications," *2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pp. 715–718, 2022, doi: 10.1109/ICAICA54878.2022.9844487.

[3] S. Tavara, "Parallel Computing of Support Vector Machines," *ACM Computing Surveys (CSUR)*, vol. 51, pp. 1–38, 2019, doi: 10.1145/3280989.

[4] S. Lloyd, R. Irani, and M. Ahmadi, "Using Neural Networks for Fast Numerical Integration and Optimization," *IEEE Access*, vol. 8, pp. 84519–84531, 2020, doi: 10.1109/ACCESS.2020.2991966.

[5] M. K. Shahvandi, "Applications of numerical integration in geodesy and geophysics," *Acta Geophysica*, pp. 1–17, 2021, doi: 10.1007/s11600-020-00525-x.

[6] W. Guo, "Solving problems involving numerical integration (II): Modified Simpson's methods for equal intervals of odd numbers," *STEM Education*, p., 2023, doi: 10.3934/steme.2023011.

[7] D. Datta and M. Gordon, "A Massively Parallel Implementation of the CCSD(T) Method Using the Resolution-of-the-Identity Approximation and a Hybrid Distributed/Shared Memory Parallelization Model.," *J Chem Theory Comput*, p., 2021, doi: 10.1021/acs.jctc.1c00389.

[8] S. Kurgalin and S. Borzunov, "Fundamentals of Parallel Computing," in *A Practical Approach to High-Performance Computing*, S. Kurgalin and S. Borzunov, Eds., Cham: Springer International Publishing, 2019, pp. 17–35. doi: 10.1007/978-3-030-27558-7_3.

[9] S. Kurgalin and S. Borzunov, "Implementation of Parallel Algorithms," in *A Practical Approach to High-Performance Computing*, S. Kurgalin and S. Borzunov, Eds., Cham: Springer International Publishing, 2019, pp. 93–115. doi: 10.1007/978-3-030-27558-7_6.

[10] B. Ong and J. Schroder, "Applications of time parallelization," *Comput Vis Sci*, vol. 23, pp. 1–15, 2020, doi: 10.1007/s00791-020-00331-4.

[11] P. D. Michailidis and K. G. Margaritis, "Scientific computations on multi-core systems using different programming frameworks," *Applied Numerical Mathematics*, vol. 104, pp. 62–80, 2016, doi: 10.1016/j.apnum.2014.12.008.

[12] S. Kurgalin and S. Borzunov, "Fundamentals of Parallel Computing," in *A Practical Approach to High-Performance Computing*, S. Kurgalin and S. Borzunov, Eds., Cham: Springer International Publishing, 2019, pp. 17–35. doi: 10.1007/978-3-030-27558-7_3.

[13] O. Allamov, J. Yusupova, M. Davronov, M. Matyakubov, O. Chuponov, and S. Omonov, "Analysis of Parallel Computing Methods and Algorithms," *2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE)*, pp. 1710–1713, 2023, doi: 10.1109/APEIE59731.2023.10347601.

[14] A. Ivutin, A. Troshina, and A. Novikov, "Optimization Strategies for Automated Parallelization for Multicore Architectures," *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, pp. 1–6, 2019, doi: 10.1109/RADIOELEK.2019.8733489.

[15] L. Nigro, "Parallel Theatre: An actor framework in Java for high-performance computing," *Simul. Model. Pract. Theory*, vol. 106, p. 102189, 2021, doi: 10.1016/j.simpat.2020.102189.

[16] A. Abdulhameed and Q. Memon, "An improved Trapezoidal rule for numerical integration," *J Phys Conf Ser*, vol. 2090, p., 2021, doi: 10.1088/1742-6596/2090/1/012104.

[17] A. Karpagam and V. Vijayalakshmi, "Comparison Results of Trapezoidal, Simpson's 13 rule, Simpson's 38 rule, and Weddle's rule," *Materials Science Educator: Courses*, p., 2018, [Online]. Available: https://consensus.app/papers/comparison-results-trapezoidal-simpson-rule-simpson-rule-karpagam/c9c50ee19bfd55dba06c139228b213aa/

[18] M. Dhali, M. F. Bulbul, and U. Sadiya, "Comparison on Trapezoidal and Simpson's Rule for Unequal Data Space," *International Journal of Mathematical Sciences and Computing*, p., 2019, doi: 10.5815/ijmsc.2019.04.04.

[19] D. Vinsensia, Y. Utami, and A. Fitra, "Application of the Trapezoid Method, the Gauss Legendre Method, and the Simpson Method in Numerical Integration Solutions," vol. 9, pp. 58–66, 2021, doi: 10.35337/SCIENTIA.VOL9.PP58-66.

[20] M. Nowicki, L. Górski, and P. Bała, "Performance Evaluation of Java/PCJ Implementation of Parallel Algorithms on the Cloud," *Euro-Par 2020: Parallel Processing Workshops*, vol. 12480, pp. 213–224, 2021, doi: 10.1007/978-3-030-71593-9_17.

[21] M. Nowicki, L. Górski, and P. Bała, "PCJ Java library as a solution to integrate HPC, Big Data and Artificial Intelligence workloads," *J Big Data*, vol. 8, pp. 1–21, 2021, doi: 10.1186/s40537-021-00454-6.

[22] Y. Bhojwani and R. Singh, "Parallelization of Definite Integration," *International Research Journal of Engineering and Technology*, 2019, [Online]. Available: www.irjet.net

[23] M. Mehrabi, N. Giacaman, and O. Sinnen, "@PT: Unobtrusive parallel programming with Java annotations," *Concurr Comput*, vol. 31, p., 2018, doi: 10.1002/cpe.4831.

[24] P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, 2nd Edition. New York: Dover Publications, 2007.

[25] G. V. Milovanović, *Numericka analiza II*, 2nd ed. Beograd: Naučna knjiga, 1988.

[26] M. K. Jain, S. R. K. Iyengar, and R. K. Jain, *Numerical Methods: Problems and Solutions*, 1st ed. New Delhi: NEW AGE INTERNATIONAL, 2020.

[27] S. K. Sharma, "Performance Analysis of Parallel Algorithms on Multi-core System using OpenMP," *International Journal of Computer Science, Engineering and Information Technology*, vol. 2, no. 5, pp. 55–64, Oct. 2012, doi: 10.5121/ijcseit.2012.2506.

[28] N. Melab, J. Gmys, M. Mezmaz, and D. Tuyttens, "Multi-core versus many-core computing for many-task Branch-and-Bound applied to big optimization problems," *Future Generation Computer Systems*, vol. 82, pp. 472–481, 2018, doi: https://doi.org/10.1016/j.future.2016.12.039.

[29] "Java Script Data Transformation Library using Fork Join Pool and Web Workers Technology," *Int J Eng Adv Technol*, p., 2019, doi: 10.35940/ijeat.b2954.129219.

[30] M. S. Ayub, M. Adnan, and M. Y. Shafi, "Design and Development of a Java Parallel I/O Library," *ArXiv*, vol. abs/2305.07414, p., 2023, doi: 10.48550/arXiv.2305.07414.

[31] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 5th edition. San Diego: Academic Press, Inc., 1994.

Time analysis of the integration process using Simpson's Rule on multicore-processors in Java

Marko Smilić, Časlav Stefanović, Dejan Milić, Sanja Đurović, Danijel Đošić

ABSTRACT

This paper considers the time analysis of the integration process using Simpson's Rule on multi-core processors in Java programming language. As computational power increases, parallelism on multi-core processors has become crucial in speeding up mathematical operations. We investigate how Simpson's Rule, a widely used method for numerical integration, can be optimized for parallel execution. We measure the time of the integration process on multi-core processors for Fresnel integral, transcendental mathematical functions and compare it to a single-core implementation. In addition, we graphically present the results of Simpson's Rule implementation for specific hardware setup (number of cores and processor specifications), software setup (Java version and Java libraries), and accuracy of the results depending on the threshold. The results indicate a significant reduction in computation time with increased available cores and better accuracy by choosing an optimal value for the threshold. In addition, we discuss the implications for high-performance computing in Java.

Uporedna analiza radnih okvira Angular, React i Flutter na osnovu implementacije aplikacije za poslastičarnicu

Janko Tufegdžić
Elektrotehnički fakultet, Univerzitet u Beogradu
janko@etf.bg.ac.rs
0009-0001-7439-7034

Luka Hrvačević
Elektrotehnički fakultet, Univerzitet u Beogradu
lukah@etf.bg.ac.rs
0009-0008-6456-9390

Iva Potkonjak
Elektrotehnički fakultet, Univerzitet u Beogradu
potkonjak.iva@gmail.com

Uroš Mutavdžić
Elektrotehnički fakultet, Univerzitet u Beogradu
urosm2000@gmail.com

Jelica Cincović
Elektrotehnički fakultet, Univerzitet u Beogradu
jelica@etf.bg.ac.rs
0000-0001-6440-9348

Marija Punt
Elektrotehnički fakultet, Univerzitet u Beogradu
marija.punt@etf.bg.ac.rs
0000-0002-1944-7086

***Apstrakt* - Savremeni razvoj internet aplikacija nezamisliv je bez radnih okvira, koji olakšavaju implementaciju, povećavaju sigurnost i doprinose konzistentnosti koda. Pravilnim odabirom radnog okvira utiče se na efikasnost, skalabilnost i dugoročnu održivost aplikacije. Ovaj rad pruža uporednu analizu tri radna okvira – Angular, React i Flutter – kroz implementaciju aplikacije za poslastičarnicu sa identičnim funkcionalnim zahtevima. Poređenje se vrši na osnovu performansi, složenosti koda, vremena razvoja, podrške za testiranje i sigurnosnih aspekata. Rezultati analize ističu prednosti i nedostatke svakog okvira, pružajući smernice za optimalan izbor u zavisnosti od potreba projekta.**

Ključne reči – radni okviri, internet aplikacije, uporedna analiza, performanse

I. Uvod

U savremenom razvoju internet aplikacija, radni okviri postali su neizostavan alat koji olakšava, ubrzava i obezbeđuje konzistentnost koda [1]. Pravilan odabir radnog okvira ključan je za razvoj kvalitetnih, efikasnih i sigurnih aplikacija, budući da radni okvir aktivno učestvuje u svim fazama razvoja, od planiranja i dizajna, preko implementacije i testiranja, do održavanja i nadogradnje. Ovaj izbor takođe značajno utiče na skalabilnost i dugoročnu održivost rešenja.

Ovaj rad pruža komparativnu analizu tri popularna radna okvira: Angular, React i Flutter, sa ciljem pružanja smernica za odabir najprikladnijeg okvira u zavisnosti od specifičnih potreba i zahteva projekta. Angular[2], razvijen od strane kompanije Google, često se koristi za kompleksne i skalabilne aplikacije zbog svoje modularnosti. React[3], poznat po fleksibilnosti i jednostavnosti, idealan je za interaktivne korisničke interfejse. Flutter[4], takođe proizvod kompanije Google, omogućava razvoj višeplatformskih aplikacija sa jedinstvenom bazom koda.

Analiza radnih okvira sprovedena je kroz implementaciju aplikacije za poslastičarnicu, gde su iste funkcionalnosti i sličan korisnički interfejs implementirani u svakom od navedenih okvira. Komparacija je obuhvatila performanse, složenost koda, vreme razvoja, podršku za testiranje i sigurnosne aspekte. Rezultati ove analize ističu specifičnosti, prednosti i nedostatke svakog okvira, sa ciljem olakšavanja informisanog odabira u zavisnosti od potreba korisnika.

II. Metodologija

Prilikom uporedne analize radnih okvira, ključno je odabrati relevantne aspekte za poređenje kako bi se donela informisana odluka o najprikladnijem rešenju za određeni projekat. Temeljna analiza zasnovana na jasno definisanim parametrima omogućava realno sagledavanje prednosti i ograničenja svakog okvira, olakšavajući tako proces donošenja odluka u odabiru radnog okvira. Parametri po kojima su poređeni izabrani radni okvir su:

- **Performanse** – zauzeće RAM memorije, optimizacija učitavanja i renderovanja, veličina generisanih paketa.
- **Složenost koda** - jednostavnost sintakse, čitljivost i održivost koda, modularnost i mogućnost ponovne upotrebe komponenti.
- **Vreme razvoja** - brzina implementacije funkcionalnosti, dostupnost biblioteka i alata, stepen automatizacije procesa razvoja.
- **Podrška za testiranje** - dostupnost i integracija alata za testiranje, jednostavnost pisanja testova, podrška za unit i end-to-end testiranje.
- **Sigurnosni aspekti** - zaštita od najčešćih ranjivosti (XSS, CSRF, SQL injection), podrška za autentifikaciju i autorizaciju, bezbednost prenosa podataka.

III. Pregled radnih okvira

U ovoj sekciji su analizirana tri popularna radna okvira: Angular, React i Flutter, i pružiti kratak pregled njihovih ključnih osobina, prednosti i ograničenja. Ovi okviri se široko koriste u modernom razvoju aplikacija, a svaki od njih nudi jedinstven pristup razvoju internet i mobilnih aplikacija.

A. Angular

Angular je robustan i sveobuhvatan radni okvir razvijen i podržan od strane kompanije Google. Angular koristi TypeScript [5], nadskup jezika JavaScript, i omogućava statičku analizu koda, što dovodi do bolje detekcije grešaka i lakšeg održavanja koda. U svom osnovnom dizajnu, Angular

se temelji na arhitekturi MVC (*Model-View-Controller*), što omogućava podelu aplikacije na jasno definisane komponente, čime se pojednostavljuje razvoj, održavanje i testiranje. Angular koristi dvostrano povezivanje podataka (*two-way data binding*), što znači da se podaci u modelu i prikazu automatski sinhronizuju, olakšavajući rad sa formama i podacima. Međutim, dvostrano povezivanje podataka može negativno uticati na performanse u složenijim aplikacijama. Angular je takođe veoma modularan, što omogućava ponovnu upotrebu komponenata, servisa i direktiva. Ovaj okvir dolazi sa bogatom podrškom za testiranje, uključujući alate kao što su *Jasmine* [6] i *Karma* [7] za jednostavno testiranje komponenti i servisa, kao i podršku za *unit* i *end-to-end* testiranje. Ipak, zbog svoje kompleksnosti, Angular može biti izazovan za početnike i zahteva veće resurse za inicijalnu konfiguraciju, što može biti prepreka za brzi razvoj, naročito u manjim projektima.

B. React

React je biblioteka razvijena od strane kompanije Meta koja se koristi za izgradnju dinamičkih i interaktivnih korisničkih interfejsa. React se fokusira samo na *view* (pogon za korisnički interfejs) u arhitekturi aplikacije, dok ostatak funkcionalnosti, kao što je upravljanje podacima i rutiranje, zahteva integraciju sa drugim bibliotekama. Jedna od najsnažnijih karakteristika radnog okvira React je upotreba virtualnog DOM-a, koji omogućava brže i efikasnije renderovanje stranica smanjenjem broja potrebnih stvarnih renderovanja na ekranu. React se temelji na komponentama, što znači da aplikacija može biti podeljena u male, ponovo upotrebljive delove koda. Komponente u ovom radnom okviru mogu biti kombinovane kako bi se izgradili složeniji interfejsi, što olakšava održavanje i testiranje. React je veoma fleksibilan, jer omogućava korišćenje samo one komponente i biblioteke koje su im potrebne za projekat, ali ta fleksibilnost može biti i mana, jer često zahteva korišćenje dodatnih alata kao što su *Redux* za upravljanje stanjem i *React Router* za upravljanje rutama. Iako je React vrlo popularan, njegova fleksibilnost može biti izazov za nove programere, naročito jer se stalno razvija, što može dovesti do brze zastare nekih alata i biblioteka.

C. Flutter

Flutter je relativno nov radni okvir razvijen od strane kompanije Google, koji koristi Dart [8] programski jezik i omogućava razvoj aplikacija za više platformi, uključujući iOS, Android i internet, iz jednog zajedničkog koda. Flutter se ističe upotrebom vlastitog grafičkog engine-a, što znači da ne zavisi od tradicionalnog DOM-a, kao što je slučaj sa radnim okvirima Angular i React. Ovaj pristup omogućava visok nivo kontrole nad korisničkim interfejsom i konzistentnost prikaza na svim platformama. Flutter nudi izuzetne performanse, jer direktno renderuje UI koristeći *native* komponente, što omogućava aplikacijama razvijenim u radnom okviru Flutter da rade glatko, čak i na starijim uređajima. Takođe, *hot reload* funkcionalnost radnog okvira Flutter omogućava programerima da odmah vide promene u aplikaciji bez potrebe za ponovnim pokretanjem, čime se značajno ubrzava razvoj, naročito u dizajnu korisničkog interfejsa. Flutter je najpoznatiji po uspehu u razvoju mobilnih aplikacija, ali njegova podrška za internet aplikacije još uvek nije na istom nivou kao za mobilne aplikacije. Flutter zahteva upotrebu programskog jezika Dart, jezika koji nije toliko popularan kao JavaScript, što može biti prepreka za timove koji nemaju iskustva sa njim. Iako je Flutter izuzetno moćan za razvoj aplikacija za više platformi, njegova podrška za internet razvoj nije još uvek potpuno razvijena, pa može biti manje zreo u poređenju sa radnim okvirima Angular i React za internet aplikacije.

IV. REZULTATI ANALIZE

U ovoj sekciji predstavljeni su rezultati uporedne analize tri izabrana radna okvira: Angular, React i Flutter bazirana na implementaciji aplikacije za poslastičarnicu. Aplikacija obuhvata dva tipa korisnika, kupce i zaposlene. Kupcima je omogućeno da pregledaju proizvode i promocije, ostavljaju komentare, naručuju proizvode i prate status svojih porudžbina. Zaposleni imaju pristup alatima za upravljanje proizvodima i odobravanjem narudžbina. Sistem se, pored poslastičarnica, može koristiti i za restorane, internet prodavnice, salone lepote i slične delatnosti. Dovoljno je izvršiti manje izmene u sadržaju i funkcionalnostima specifičnim za konkretnu oblast, dok osnovna arhitektura ostaje primenljiva bez značajnih promena. Analiza je zasnovana na prethodno pomenutim parametrima.

A. Performanse

Performanse radnih okvira su poređene po četiri odabrana kriterijuma vidljiva u tabeli 1. Vreme učitavanja (*FCP*) meri vreme potrebno da se prvi sadržaj učita i postane vidljiv korisnicima. Merenje *FCP* je izvršeno pomoću alata *Google Lighthouse* [9]. Brzina renderovanja DOM elemenata ocenjuje koliko brzo radni okvir renderuje promene na ekranu. Može se izmeriti pomoću *Chrome DevTools Performance* alata, koji omogućava praćenje i analizu procesa renderovanja. Veličina generisanog paketa meri veličinu datoteka koje se preuzimaju prilikom učitavanja aplikacije. Veličina paketa je merena koristeći *Source Map Explorer*. Zauzeće RAM memorije meri količinu RAM-a koju aplikacija koristi tokom rada. Merenje je obavljeno pomoću *Chrome DevTools Memory* alata, koji omogućava praćenje potrošnje memorije tokom rada aplikacije.

Kriterijum	Angular	React	Flutter
Vreme učitavanja (FCP)	360ms	200ms	270ms
Brzina renderovanja DOM elemenata	7ms	6ms	8ms
Veličina generisanog paketa (Bundle size)	598 KB	309 B	2.3 MB
Zauzeće RAM memorije	108 MB	53 MB	706 MB

Tabela 1 – Rezultati poređenja performansi

Prvi kriterijum, vreme učitavanja (*First Contentful Paint* – FCP), meri vreme potrebno da se prvi sadržaj prikaže korisniku. Rezultati pokazuju da React ostvaruje najbolje performanse sa vremenom od 200 ms, dok Angular zaostaje sa 360 ms, a Flutter postiže 270 ms. Brže učitavanje aplikacije doprinosi poboljšanom korisničkom iskustvu, naročito u aplikacijama sa velikim brojem interaktivnih elemenata. Kod drugog kriterijuma, brzine renderovanja DOM elemenata, razlike su minimalne, ali React postiže najnižu vrednost od 6 ms, dok Angular i Flutter imaju nešto veće vrednosti od 7 ms i 8 ms. Ovo sugeriše da je React optimizovaniji za rad sa dinamičkim podacima i učestalim promenama interfejsa. Treći kriterijum, veličina generisanog paketa (*bundle size*), direktno utiče na brzinu učitavanja aplikacije i mrežnu efikasnost. React je najoptimizovaniji sa paketom od 309 B, dok je Angular drugi sa veličinom od 598 KB. Flutter generiše znatno veći paket od 2.3 MB, što može negativno uticati na učitavanje i potrošnju resursa, posebno u aplikacijama koje ciljaju sporije mrežne konekcije. U pogledu zauzeća RAM memorije, React pokazuje najveću efikasnost sa 53 MB, dok Angular koristi 108 MB. S druge strane, Flutter zahteva značajno više memorije – čak 706 MB, što može predstavljati izazov za uređaje sa ograničenim hardverskim resursima, naročito u mobilnim aplikacijama.

Na osnovu dobijenih rezultata može se zaključiti da React pokazuje najviše optimizacije u pogledu vremena učitavanja, efikasnosti memorije i brzine renderovanja, čime se potvrđuje njegova pogodnost za manje aplikacije koje zahtevaju visoke performanse. Angular nudi uravnotežen pristup između funkcionalnosti i performansi, dok je Flutter resursno zahtevniji.

B. Složenost koda

Složenost koda direktno utiče na brzinu razvoja, održavanje aplikacije i mogućnost ponovne upotrebe koda. Radni okviri se razlikuju po načinu strukturiranja aplikacija, sintaksi, modularnosti i lakoći učenja. Angular donosi strogu organizaciju sa jasno definisanim pravilima, React nudi fleksibilniji pristup razvoju, dok Flutter koristi specifičnu arhitekturu zasnovanu na widgetima. Kako bi se sagledale razlike u složenosti koda, sprovedeno je poređenje prema nekoliko relevantnih kriterijuma, prikazano u tabeli ispod.

Kriterijum	Angular	React	Flutter
Struktura i organizacija	Stroga struktura (moduli, komponente, servisi)	Fleksibilna, nije striktno definisana	*Widget-based* arhitektura
Modularnost	Visoka, ali sa više konfiguracija	Visoka, ali zavisi od eksternih biblioteka	Vrlo visoka, UI baziran na *widgetima*
Ponovna upotrebljivost	Dobra, ali zahteva strukturalni pristup	Odlična, lakša ponovna upotreba komponenti	Odlična, svi elementi su *widgeti*
Ekosistem i biblioteke	Bogat, ali često preopširan	Fleksibilan, ali zahteva dodatne biblioteke	Sve dolazi ugrađeno u SDK

Tabela 2 – Pregled kriterijuma za složenost koda

Iz tabele se može zaključiti da je Angular najkompleksniji zbog svoje stroge strukture, TypeScript sintakse i velikog broja konfiguracija, ali pruža visoku modularnost i skalabilnost. React je fleksibilniji i jednostavniji za početnike zahvaljujući JSX sintaksi i lakšoj ponovnoj upotrebi komponenti, ali zahteva dodatne biblioteke za punu funkcionalnost. Flutter donosi specifičan razvojni model zasnovan na *widgetima*, što omogućava visok nivo modularnosti i ponovne upotrebljivosti, ali zahteva poznavanje Dart jezika i prilagođavanje njegovoj arhitekturi.

C. Vreme razvoja

U ovoj sekciji biće predstavljeno poređenje vreme razvoja aplikacije za poslatičarnicu u radnim okvirima koji se porede. Sve 3 aplikacije su implementirane od strane timova koji imaju isti nivo znanja i iskustva u radu sa korišćenim radnim okvirima.

Flutter je najbrži za razvoj aplikacija, zahvaljujući svojoj funkcionalnosti "*Hot Reload*", koja omogućava brza testiranja i promene bez ponovnog pokretanja aplikacije. Pored toga, Flutter omogućava razvoj za više platformi koristeći jedan kod, što značajno smanjuje vreme potrebno za implementaciju aplikacija na različitim uređajima. Ovaj okvir omogućava brzo prototipisanje i završavanje aplikacija u vrlo kratkom vremenskom periodu. Razvoj pomenute aplikacije u Flutteru trajao je 3 dana. React, s druge strane, takođe omogućava brzi razvoj, ali zbog svoje jednostavne strukture i bogatog ekosistema, može da postane složeniji u većim aplikacijama. Aplikacija u Reactu bila je implentirana za 5 dana. Angular je najzahtevniji od tri okvira, pre svega zbog svoje kompleksne arhitekture koja zahteva više vremena za postavljanje i implementaciju svih potrebnih komponenti i struktura. Iako pruža robusnu i skalabilnu platformu, njegov složeniji pristup razvojnom procesu može značajno usporiti tempo rada, posebno za manje projekte, pa je razvoj aplikacije u radnom okviru Angular trajao 6 dana.

D. Podrška za testiranje

Testiranje je jedna od ključnih komponenti razvoja aplikacija, a svaki okvir nudi različite alate i pristupe koji mogu olakšati ili otežati implementaciju testova.

Flutter pruža odličnu podršku za testiranje, sa ugrađenim alatima kao što su *flutter_test* paket za jedinično testiranje i *widget_test* za testiranje korisničkog interfejsa [10]. Flutter takođe omogućava automatizovano testiranje aplikacija na različitim uređajima i platformama, što je posebno korisno kada se razvija za više platformi. Flutter nudi i alate za integracione i funkcionalne testove, čineći testiranje jednostavnim i efikasnim procesom. React takođe nudi solidnu podršku za testiranje, sa popularnim bibliotekama kao što su *Jest* i *React Testing Library* [11][12]. *Jest* [13] je široko korišćen za jedinično testiranje, dok *React Testing Library* olakšava testiranje interakcija korisničkog interfejsa. React omogućava brzo pisanje testova, a ekosistem oko njega pruža mnogo alata za testiranje komponenata, integraciju, pa čak i *end-to-end* testove pomoću alata kao što je *Cypress* [13]. Iako React ima odličnu podršku za testiranje, složenost testova može rasti kako aplikacija postaje veća, zbog potrebe za obuhvatanjem različitih interakcija između komponenti. Angular ima najkompletniju i najintegrisaniju podršku za testiranje među ova tri radna okvira. Angular dolazi sa ugrađenim testnim alatima, kao što su *Karma* i *Jasmine* za jedinično testiranje, zajedno sa *Protractor* za *end-to-end*

testove. *Karma* je testni pokretač koji omogućava izvršavanje testova u različitim pretraživačima, dok *Jasmine* pruža sintaksu za pisanje testova. Angularova integracija testiranja sa samim okvirom čini implementaciju testova jednostavnom, ali može biti zahtevnija u većim aplikacijama zbog same kompleksnosti Angulara. Takođe, kao i u Reactu, testovi mogu postati složeniji kako aplikacija raste, ali Angular ima robustan set alata koji omogućava duboko testiranje svih aspekata aplikacije.

E. Sigurnosni aspekti

Sigurnost aplikacija je bitan faktor u izboru radnog okvira, posebno kada se razvijaju aplikacije koje obrađuju osetljive podatke, poput aplikacije za poslastičarnicu koja može uključivati online naručivanje i plaćanje. Svaki od analiziranih okvira – Angular, React i Flutter – pruža različite mehanizme za zaštitu aplikacije, pri čemu se razlikuju po ugrađenim bezbednosnim funkcijama i mogućnostima zaštite od uobičajenih ranjivosti [15].

Angular se ističe po visokom nivou sigurnosti, zahvaljujući ugrađenim mehanizmima zaštite od uobičajenih napada kao što su XSS (*Cross-Site Scripting*) i CSRF (*Cross-Site Request Forgery*). Automatski se saniraju potencijalno opasni podaci kada se ubacuju u DOM, čime se smanjuje rizik od XSS napada. Takođe, podržava OAuth i JWT autentifikaciju, što ga čini pogodnim za implementaciju sigurnih prijava i upravljanja korisničkim sesijama. React nema ugrađene bezbednosne mehanizme poput Angulara, što znači da je odgovornost za implementaciju sigurnosnih praksi u velikoj meri na programeru. Na primer, React ne pruža automatsku zaštitu od XSS napada, ali preporučuje korišćenje bezbednog umetanja podataka pomoću JSX sintakse, gde se podrazumevano sprečava umetanje nebezbednog HTML-a. Za dodatnu sigurnost, React aplikacije često koriste *helmet.js* za postavljanje HTTP zaglavlja i povećanje zaštite. Takođe podržava sigurne metode autentifikacije, uključujući OAuth i JWT, ali njihova implementacija zavisi od dodatnih biblioteka i podešavanja. Flutter, kao okvir za razvoj nativnih aplikacija, nudi drugačiji pristup sigurnosti u poređenju sa Angularom i Reactom. Budući da Flutter renderuje UI unutar sopstvenog grafičkog okruženja, smanjuje izloženost uobičajenim pretnjama koje pogađaju web aplikacije, kao što su XSS i injekcioni napadi. Pored toga, Flutter nudi ugrađene mehanizme za sigurno skladištenje podataka, poput *Flutter Secure Storage* koji koristi enkripciju za zaštitu poverljivih informacija. Za autentifikaciju, Flutter takođe podržava OAuth, JWT kao i biometrijsku autentifikaciju kroz različite dodatke.

V. ZAKLJUČAK

Ovaj rad pružio je detaljnu uporednu analizu tri popularna radna okvira – Angular, React i Flutter – kroz implementaciju aplikacije za poslastičarnicu u sva tri radna okvira sa identičnim funkcionalnim zahtevima. Analiza je obuhvatila aspekte kao što su performanse, složenost koda, vreme razvoja, podrška za testiranje i sigurnosni aspekti, omogućavajući objektivno sagledavanje prednosti i nedostataka svakog okvira.

Rezultati pokazuju da je React najefikasniji u pogledu performansi, zauzimajući najmanje memorije i pružajući najbrže vreme učitavanja i renderovanja DOM elemenata. Takođe, njegova fleksibilnost i jednostavnost strukture čine ga pogodnim za aplikacije koje zahtevaju visoke performanse uz brzi razvoj. Angular se istakao po svojoj robusnoj strukturi i bogatoj podršci za testiranje, ali njegova kompleksnost može predstavljati izazov za manje iskusne programere. Njegova stroga organizacija čini ga pogodnim za velike i dugoročne projekte kojima je potrebna modularnost i skalabilnost.

Flutter se pokazao kao najbrži za razvoj aplikacija, zahvaljujući funkcionalnosti "Hot Reload" i mogućnosti višestruke platformske implementacije iz jedinstvenog koda. Međutim, njegovi veći zahtevi za memorijskim resursima i manja zrelost u domenu internet aplikacija mogu predstavljati izazove u određenim scenarijima.

Na osnovu dobijenih rezultata, izbor radnog okvira treba biti vođen specifičnim potrebama projekta. React je najpogodniji za dinamične i visoko interaktivne aplikacije, Angular je idealan za velike i kompleksne projekte sa dugoročnom održivošću, dok je Flutter najbolji izbor kada je brzina razvoja i višeplatformska kompatibilnost ključni faktor.

LITERATURA

[1] Cincović, J., & Punt, M. (2025). Comparison: Angular vs. React vs. Vue. Which framework is the best choice? *ICIST 2025 Proceedings*, 123-130.

[2] Angular, https://angular.io / , [pristupljeno 29.03.2025.].

[3] React, https://reactjs.org/ , [pristupljeno 29.03.2025.].

[4] Flutter, https://flutter.dev/ , [pristupljeno 29.03.2025.].

[5] Typescript, https://www.typescriptlang.org/ , [pristupljeno 30.03.2025.].

[6] "Jasmine: Behavior-driven JavaScript testing framework,", https://jasmine.github.io/ , [pristupljeno: 30.03.2025.].

[7] "Karma: Spectacular test runner for JavaScript," dostupno na: https://karma-runner.github.io/ , [pristupljeno: 30.03.2025.].

[8] Dart, https://dart.dev/ , [pristupljeno: 30.03.2025.].

[9] "Lighthouse: Open-source, automated tool for improving web quality", https://developer.chrome.com/docs/lighthouse/ [pristupljeno: 30.03.2025.].

[10] M. Işıtan and M. Koklu, "Comparison and Evaluation of Cross Platform Mobile Application Development Tools", International Journal of Applied Mathematics Electronics and Computers, vol. 8, no. 4, pp. 273–281, 2020, doi: 10.18100/ijamec.832673.

[11] E. Gülcüoğlu, A. B. Üstün, i N. Seyhan, "Comparison of Flutter and React Native Platforms," *Istanbul University Journal of the School of Business*, vol. 50, no. 2, pp. 129-143

[12] "Jest: Delightful JavaScript testing framework",https://jestjs.io/ [pristupljeno: 30.03.2025.].

[13] Denko, Blaž & Pecnik, Spela & Fister jr, Iztok. (2021). A Comprehensive Comparison of Hybrid Mobile Application Development Frameworks. International Journal of Security and Privacy in Pervasive Computing. 13. 78-90. 10.4018/IJSPPC.2021010105.

[14] "Cypress: JavaScript end-to-end testing framework" https://www.cypress.io/, [pristupljeno: 30.03.2025.].

[15] Singh, M., & Shobha, G. (2021). Comparative analysis of hybrid mobile app development frameworks. *International Journal of Soft Computing and Engineering (IJSCE)*, 10(6), 21-27.

A comparative analysis of Angular, React and Flutter frameworks based on the implementation of a pastry shop application

Janko Tufegdžić

Luka Hrvačević

Iva Potkonjak

Uroš Mutavdžić

Jelica Cincović

Marija Punt

ABSTRACT

Modern development of Internet applications is unthinkable without frameworks, which facilitate implementation, increase security and contribute to code consistency. Choosing the right framework affects the efficiency, scalability and long-term sustainability of implemented applications. This paper provides a comparative analysis of three frameworks - Angular, React and Flutter - through the implementation of a pastry shop application with identical functional requirements. The comparison is made based on performance, code complexity, development time, testing support and security aspects. The results of the analysis highlight the advantages and disadvantages of each framework, providing guidelines for the optimal choice depending on the needs of the project.

Pregled pozicionih algoritama u proširenoj i virtuelnoj stvarnosti

Mihajlo Ogrizović
Katedra za Računarsku tehniku i informatiku
Elektrotehnički fakultet, Univerzitet u Beogradu
Beograd, Srbija
ogrizovic@etf.bg.ac.rs

Filip Janković
Katedra za Računarsku tehniku i informatiku
Elektrotehnički fakultet, Univerzitet u Beogradu
Beograd, Srbija
jf160343d@student.etf.bg.ac.rs

Aleksa Ilić
Katedra za Računarsku tehniku i informatiku
Elektrotehnički fakultet, Univerzitet u Beogradu
Beograd, Srbija
aleksa.d.ilic@gmail.com

Dražen Drašković
Katedra za Računarsku tehniku i informatiku
Elektrotehnički fakultet, Univerzitet u Beogradu
Beograd, Srbija
drazen.draskovic@etf.bg.ac.rs

***Sažetak*—Virtuelna i proširena stvarnost danas imaju veliku primenu u različitim oblastima poput edukacije, industrijskih obuka, virtuelnih obilazaka znamenitosti i drugim. Za potrebe implementacije simulacija obilazaka ovakvih prostora u virtuelnoj realnosti neophodno je imati predstavu o poziciji u tom prostoru. Različiti uređaji proširene i virtuelne stvarnosti (VR) pružaju različite mehanizme i algoritme detekcije pozicije u prostoru. U ovom radu biće prikazan pregled i analiza različitih algoritama za određivanje pozicije u prostoru i detekcije ivica i granica u tom prostoru. Pregled će obuhvatiti veći broj različitih modela VR uređaja, dok će sam eksperiment biti primenjen korišćenjem uređaja Meta Quest 3. Eksperiment će obuhvatiti nekoliko različitih prostora i objekata unutar njih. Posmatraće se preciznost pozicije referentno u odnosu na druge, ali i u apsolutnom prostoru.**

***Index Terms*—softverski sistemi, virtuelna realnost, augmentovana realnost**

I. Uvod

Virtuelna (VR) i proširena stvarnost (AR) predstavljaju revolucionarne tehnologije koje ubrzano menjaju način na koji ljudi komuniciraju, uče i rade. Njihova primena se širi izvan zabave i igara, ulazeći u oblasti obrazovanja, medicine, industrije, trgovine i mnoge druge. Razvoj VR i AR tehnologija poslednjih godina beleži značajan rast zahvaljujući napretku hardvera i softvera, poboljšanoj dostupnosti i smanjenju troškova. Brži procesori, napredni algoritmi i efikasnije grafičke kartice omogućili su realističnije vizuelne prikaze, dok su bolji senzori i kamere unapredili interaktivnost i preciznost AR sistema. Podaci pokazuju da tržište VR i AR tehnologija beleži konstantan rast. Prema istraživanjima, vrednost globalnog VR i AR tržišta mogla bi premašiti stotine milijardi dolara u narednim godinama, što potvrđuje njihov sve veći značaj u različitim sektorima **precedence2024arvr**. Za potrebe implementacije simulacija obilazaka ovakvih prostora u virtuelnoj realnosti neophodno je imati predstavu o poziciji u tom prostoru. Različiti uređaji proširene i virtuelne stvarnosti pružaju različite mehanizme i algoritme detekcije pozicije u prostoru. U ovom radu biće prikazan pregled i opis različitih algoritama za određivanje pozicije u prostoru i detekcije ivica i granica u tom prostoru. Poseban fokus će biti dat uređaju Meta Quest 3, za koji će biti sproveden eksperiment zarad utvrđivanja preciznosti algoritama određivanja njegove pozicije. Sekcija 2 će prikazati šta je bilo istraživano na ovu temu do sada i kako istraživanje pokazano ovim radom prevazilazi ista. Sekcija 3 će služiti kao pregled i opis različitih metoda detekcije pozicije uređaja u prostoru, sa posebnim fokusom na uređaju Meta Quest 3. U sekciji 4 će biti opisani detalji i uslovi eksperimentima u kojima će se odrediti preciznost uređaja Meta Quest 3. U sekciji 5 će biti prikazani rezultati eksperimenta sa propratnom diskusijom. U poslednjoj sekciji će biti prikazani predlozi daljih istraživanja.

II. Pregled literature

Na temu tačnosti i istraživanja pozicionih algoritama jesu rađena različita istraživanja, mada primarno za druge VR uređaje. U radu **bauer2021accuracy**, rađena je analiza ponašanja VR sistema HTC Vive Pro i tačnosti njegovog pozicionog sistema, gde je zaključeno da u svim eksperimentima dobijena milimetarska tačnost uređaja. U 2019. godini, rađena su istraživanja tačnosti pozicioniranja HTC Vive Tracker-a u dinamičkom 3D okruženju **van2019agreement**. U njihovoj testnoj postavci, tracker je bio postavljen na pokretnu robotsku ruku. Istraživanja sa baznim stanicama (lighthouses) druge generacije pokazala su u ovom slučaju veća odstupanja nego što je Niehorster **niehorster2017accuracy** objavio za svoja statička istraživanja prve generacije. Konačan zaključak se ne može doneti, jer su korišćeni različiti hardver i testne postavke. Što se tiče Meta Quest 3 uređaja, u literaturi nije nađeno mnogo na temu provere tačnosti istog. Na primer, u **cheng2024comparing** je rađena uporedna analiza tačnosti pozicija različitih uređaja, među kojima je bio i Meta Quest 3. Tamo je pokazano kroz različite eksperimente da je imao makar decimetarsku tačnost uređaj (tačnost je opadala

u noćnim i otvorenim uslovima merenja) Još jedno istraživanje **eger2020measuring**, na Oculus Quest uređaju prikazuje pozicionu tačnost ispod 6.86mm. Eksperiment je bio rađen tako što je uređaj bio povezan za robosku ruku. Zaključeno je da njegova tačnost varira dosta od vremenskih uslova kao što su osvetljenje.

III. Pregled tehnologija

Postoje različiti sistemi detekcije pozicije uređaja u prostoru, gde se klasifikacija radi na osnovu korišćene tehnologije zarad detekcije. U nastavku sekcije se nalazi pregled istih, sa opisima respektivno.

A. Bžično praćenje

Bežični sistemi za praćenje koriste skup ankera koji su postavljeni duž perimetra prostora, dok se tagovi prate kako bi se odredila njihova pozicija. Ovaj sistem je sličan GPS-u, ali je dizajniran da funkcioniše i u zatvorenom i na otvorenom prostoru, pa se često naziva i "unutrašnji GPS". Tagovi određuju svoju 3D poziciju putem triangulacije, koristeći ankere postavljene na perimetru. Bežična tehnologija Ultra Wideband (UWB) omogućava praćenje pozicije sa preciznošću manjom od 100 mm. Korišćenjem fuzije senzora i brzih algoritama, preciznost može da dostigne 5 mm uz brzinu osvežavanja od 200 Hz i latenciju od 5 ms. **shi2016survey**

B. Magnetno praćenje

Magnetno praćenje funkcioniše na principu merenja intenziteta magnetnog polja u različitim pravcima. Obično se koristi bazna stanica koja generiše naizmenični (AC), jednosmerni (DC) ili impulsni jednosmerni napon. Kako se povećava udaljenost između mernog mesta i bazne stanice, intenzitet magnetnog polja opada. Rotacija merne tačke menja raspodelu magnetnog polja duž različitih osa, što omogućava određivanje orijentacije. Magnetno praćenje je implementirano u nekoliko proizvoda, uključujući Razer Hydra. U kontrolisanim uslovima, preciznost magnetnog praćenja može biti veoma visoka (prema specifikacijama Hydre, preciznost pozicioniranja je 1 mm, a preciznost rotacije 1 stepen). Međutim, ovaj sistem može biti podložan smetnjama uzrokovanim provodnim materijalima u blizini emitera ili senzora, magnetnim poljima drugih elektronskih uređaja, kao i prisustvom feromagnetnih materijala u prostoru za praćenje. **PREIM2014625**

C. Akustičko praćenje

Akustičko praćenje se zasniva na merenju vremena koje je potrebno da akustični signal stigne do poznatih prijemnika. Uobičajeno, više predajnika je postavljeno u prostoru koji se prati, dok su prijemnici (mikrofoni) postavljeni na objektima koji se prate. Ako su prijemnici upoznati sa vremenom kada je signal poslat, vreme putovanja signala može se koristiti za izračunavanje udaljenosti od predajnika. Kada su na čvrstom objektu postavljeni prijemnici/mikrofoni na poznatim pozicijama, razlika u vremenu prijema signala između njih može otkriti orijentaciju objekta u odnosu na predajnike. Kompanije poput Intersense-a uspešno su implementirale akustička rešenja za praćenje. **CHRIST2014561**

Međutim, akustički sistemi za praćenje zahtevaju dugotrajnu kalibraciju, podložni su greškama izazvanim ambijentalnom bukom i često ne omogućavaju visoke stope osvežavanja podataka. Zbog tih ograničenja, akustičke tehnologije se često kombinuju sa drugim vrstama senzora, kao što su inercijalni senzori, čime se postiže veća preciznost kroz „fuziju senzora“. **abramson2007recursive**

D. Inercijsko praćenje

Inercijalno praćenje koristi akcelerometre i žiroskope. Akcelerometri mere linearno ubrzanje. Pošto je izvod položaja po vremenu brzina, a izvod brzine ubrzanje, izlaz iz akcelerometra može se integrisati da bi se dobila brzina, a zatim ponovo integrisati da bi se dobio položaj (preciznije, položaj u odnosu na neku početnu tačku). Žiroskopi mere uglovnu brzinu. Današnji žiroskopi su komponente bazirane na MEMS tehnologiji, ali funkcionišu po istom principu kao i mehanički žiroskopi. Kao i kod akcelerometra, uglovna brzina se može integrisati da bi se odredio uglovni položaj (odnosno položaj u odnosu na početni ugao). Prednost inercijalnog praćenja je u tome što je vrlo jeftino, omogućava visoke frekvencije osvežavanja i nisku latenciju. Međutim, zbog procesa integracije i dvostruke integracije dolazi do značajnog drifta, naročito kod pozicionih informacija, pa se zato inercijalno praćenje teško može koristiti kao pouzdan izvor za određivanje položaja. **BLESER200959**

E. Optičko praćenje

Optičko praćenje koristi kamere postavljene na ili oko headset-a kako bi se odredila pozicija i orijentacija na osnovu algoritama računarskog vida. Ova metoda zasniva se na istom principu kao i stereoskopski ljudski vid. Kada osoba gleda objekat pomoću binokularnog vida, može da proceni približnu udaljenost objekta zbog razlike u perspektivi između dva oka. U optičkom praćenju, kamere se kalibrišu kako bi odredile udaljenost do objekta i njegov položaj u prostoru. Optički sistemi su pouzdani i relativno jeftini, ali mogu biti teški za kalibraciju. Pored toga, sistem zahteva direktnu svetlosnu liniju bez prepreka, u suprotnom će dobijati pogrešne podatke. Optičko praćenje se može izvoditi sa markerima ili bez njih. Praćenje sa markerima podrazumeva ciljeve sa poznatim šablonima koji služe kao referentne tačke, a kamere konstantno traže te markere i koriste različite algoritme (na primer, POSIT algoritam) kako bi izračunale položaj objekta. Markeri mogu biti vidljivi, kao što su štampani QR kodovi, ali mnogi koriste infracrvenu (IR) svetlost koju kamere mogu da detektuju. Aktivne implementacije koriste markere sa ugrađenim IR LED svetlima koji se mogu uključivati i isključivati radi sinhronizacije sa kamerom, što olakšava eliminisanje drugih IR izvora svetlosti u zoni praćenja. Pasivne implementacije koriste retroreflektore koji reflektuju IR svetlost nazad ka izvoru sa minimalnim rasipanjem. Praćenje bez markera ne zahteva unapred postavljene ciljeve, već koristi prirodne karakteristike okruženja kako bi odredilo poziciju i orijentaciju. **roadtoVR2014 techcrunch2019**

1) Inside-out praćenje: U ovoj metodi, kamera je postavljena na uređaj koji se prati i usmerena je ka spolja kako bi odredila njegovu lokaciju u okruženju. Headset-ovi koji koriste ovu tehnologiju imaju više kamera okrenutih u različitim pravcima kako bi dobili prikaz cele okoline. Ova metoda može funkcionisati sa markerima ili bez njih. Lighthouse sistem koji koristi HTC Vive je primer aktivnih markera. Svaki spoljašnji Lighthouse modul sadrži IR LED diode kao i laserski niz koji se kreće horizontalno i vertikalno, dok senzori na headset-u i kontrolerima mogu detektovati ove zrake i koristiti vreme prijema za određivanje pozicije. **techcrunch2019** Praćenje bez markera, kao kod Oculus Quest-a, ne zahteva ništa postavljeno u spoljašnjem okruženju. Ono koristi kamere na headset-u za proces nazvan SLAM (simultano određivanje lokacije i mapiranje), gde se trodimenzionalna mapa okruženja generiše u realnom vremenu. Algoritmi mašinskog učenja zatim određuju gde se headset nalazi unutar te 3D mape, koristeći detekciju karakterističnih tačaka za rekonstrukciju i analizu okoline. Ova tehnologija omogućava da vrhunski headset-ovi poput Microsoft HoloLens-a budu samostalni uređaji, ali takođe otvara mogućnosti za jeftinije mobilne headset-ove koji ne zahtevaju povezivanje sa spoljnim računarima ili senzorima. **ferroneTracking**

2) Outside-in praćenje: U ovoj metodi, kamere se postavljaju na fiksne pozicije u okruženju kako bi pratile položaj markera na uređaju koji se prati, kao što su HMD (head-mounted display) ili kontroleri. Više kamera omogućava različite poglede na iste markere, a to preklapanje omogućava precizna očitavanja pozicije uređaja. Originalni Oculus Rift koristi ovu tehniku, postavljajući konstelaciju IR LED dioda na svom headset-u i kontrolerima, što omogućava spoljašnjim kamerama u okruženju da očitavaju njihove pozicije. Ova metoda je najzrelija i ima primenu ne samo u VR-u već i u tehnologiji snimanja pokreta za filmsku industriju. Međutim, ovo rešenje je prostorno ograničeno, jer zahteva da se spoljašnji senzori stalno nalaze u vidokrugu uređaja. **techcrunch2015 Pustka2012**

F. Fuzija senzora

Česta situacija je da se ne primenjuje isključivo jedan algoritam za detekciju pozicija već simultano veći broj njih u kombinaciji zarad što tačnije procene pozicije. Algoritmi u različitim udelima primenjuju onda različite algoritme zarad što veće preciznosti. Česta takva kombinacija je korišćenje optičkog i inercijalnog praćenja, kao što je to slučaj kod većine VR HMD.

G. Pregled korišćenih algoritama kod različitih uređaja

U nastavku se nalazi tabelarni prikaz korišćenih tehnologija po različitim komercijalnim rešenjima za uređaje (HMD). Kao što se može primetiti iz prikaza tabele, većina uređaja koristi Inside-out praćenje. Razlozi za ovo su što je praktičnost korišćenja znatno veća u odnosu na Outside-in sisteme, koji zahtevaju više prostora da bi se postavili markeri za pozicioniranje i generalno nisu toliko zastupljeni. Iako su to primarmo korišćeni optički sistemi za pozicioniranje, oni zapravo koriste fuziju senzora, tako što kombinuju više sistema za praćenje (najčešće inercijsko sa optičkim).

Tabela I
PREGLED UREĐAJA SA ALGORITMIMA POZICIONIRANJA

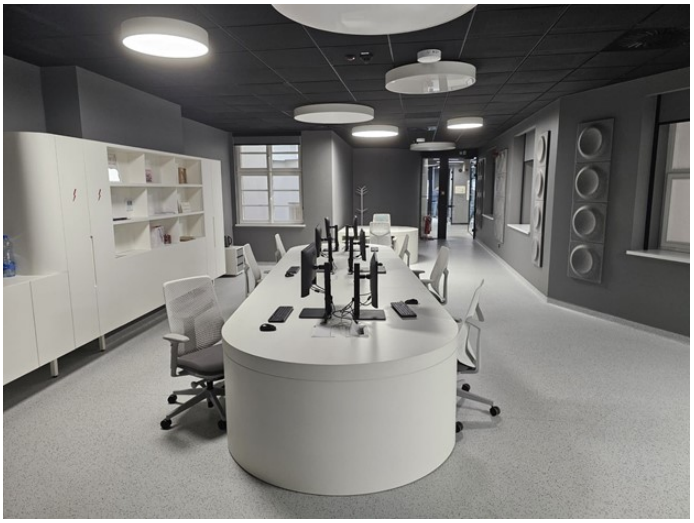
Uređaj	Primarni metod pozicioniranja
Meta Quest 3	Inside-out bez markera
Meta Quest 2	Inside-out bez markera
Valve Index	Inside-out (Lighthouse)
Oculus Rift S	Outside-in
PlayStation VR	Outside-in
HTC Vive Pro 2	Inside-out (Lighthouse)
Apple Vision Pro	Inside-out bez markera
Varjo XR-3	Inside-out bez markera
Microsoft HoloLens 2	Inside-out bez markera

IV. OPIS EKSPERIMENTA

Kao što je bilo napomenuto u uvodnoj sekciji, jedan od ciljeva ovog rada je izvođenje eksperimenta gde se istražuje preciznost određivanja pozicije Meta Quest 3 uređaja. Eksperiment se izvodi u kancelarijama Palate Nauke, u dve različite prostorije, (1) VR sobi koja je prazna i pravilnog oblika, i druga koja je (2) kancelarija sa nameštajem. Na slikama ispod se mogu videti obe prostorije. Prilikom izrade eksperimenta koristio se Meta Quest 3 Boundary sistem za određivanje dimenzija prostorija **mq3**. U slučaju (2) nije bilo moguće pokriti celu prostoriju Boundary opsegom (maksimalne dimenzije prostora su 25x25 stopa **mq3boundary**), pa je uzet samo deo sobe u kojoj se izvodio eksperiment.

Slika 1. Prva prostorija za snimanje lokacija (u nastavku P1).

Zarad implementacije aplikacije u kojoj se dohvata i ispisuje trenutna lokacija koristilo se razvojno okruženje Unity, zbog ogromne podrške i lakoće korišćenja. Za implementaciju XR dela aplikacije koristili su se *Meta XR Interaction SDK* i *Meta XR Core SDK*. Meta SDK ima podrazumevano upaljeno računanje koordinata referentno od mesta pokretanja aplikacije (*Floor*) pa je bilo neophodno da se odradi podešavanje koordinata da se računaju u odnosu na centar prostora koji koristi Boundarije snimljene prostorije (*Stage*). Eksperiment se izvodio po sledećem principu: u obe prostorije su odabrane dve nasumične tačke odakle će se meriti pozicija u odnosu na centar prostora (određen Boundarijem) u uslovima svetlosti i mraka. Kao referencu za nasumične pozicije su se koristili objekti koji su postavljeni u Unity koji će biti vidljivi u



Slika 2. Druga prostorija za snimanje lokacija (u nastavku P2).

```
        if (OVRManager.instance != null)
        {
            OVRManager.instance.trackingOriginType =
OVRManager.TrackingOrigin.Stage;
        }
        OVRPlugin.SetTrackingOriginType(OVRPlugin.TrackingOrigin.Stage);
```

Slika 3. Parče koda koje podešava način računanja centra prostora.

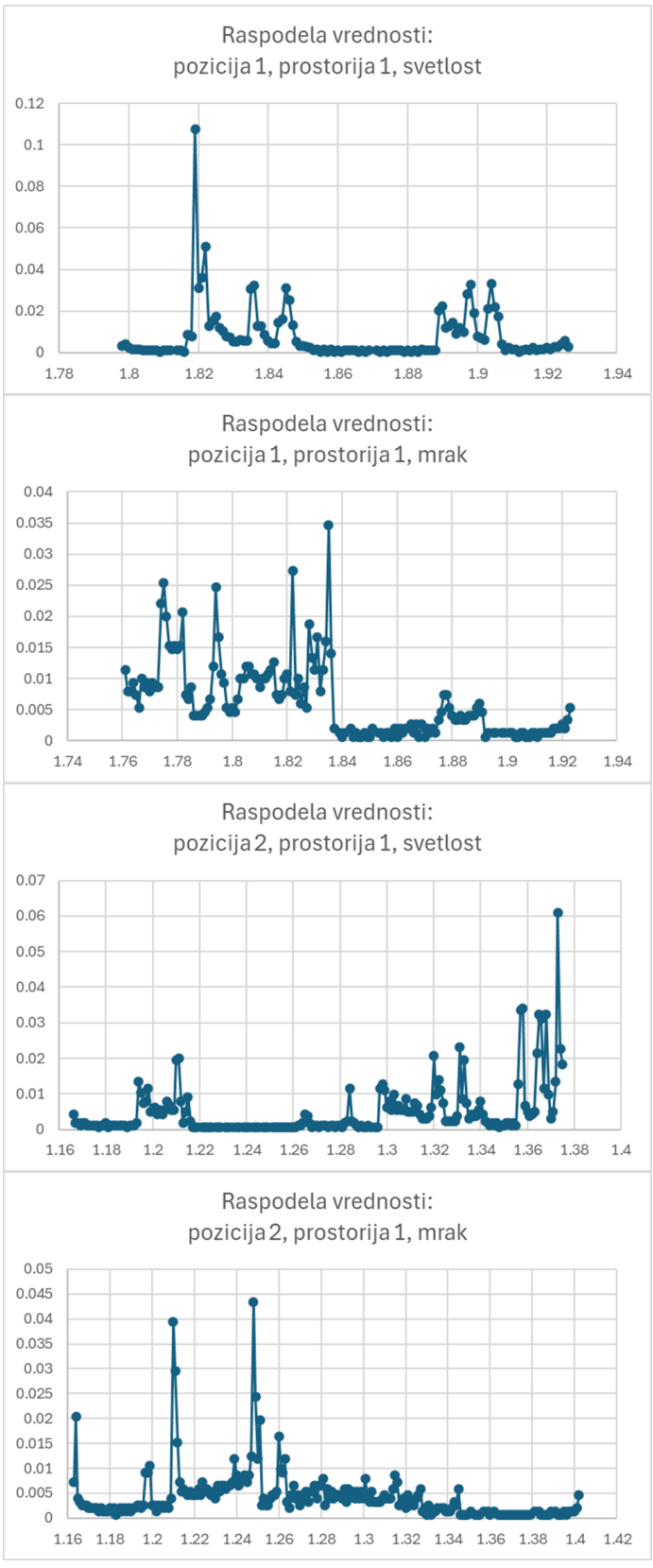
prostoriji (čija će tačnost takođe biti poređena). Merenje se radilo putem Meta Quest 3 HMD na kome se ispisuje distanca u odnosu na centar prostorije (prikazan Unity objektom). Kao referentnu vrednost tačnosti koristilo se merenje laserskim metrom (uzeta je srednja vrednost razdaljine nakon 4 merenja). Temperatura u kancelarijama tokom izvođenja eksperimenta je bila 23 stepeni. Aplikacija ima režim za snimanje, pa su se koordinate pozicija snimale u dokument samo ukoliko je taj režim upaljen.

V. Rezultati eksperimenta

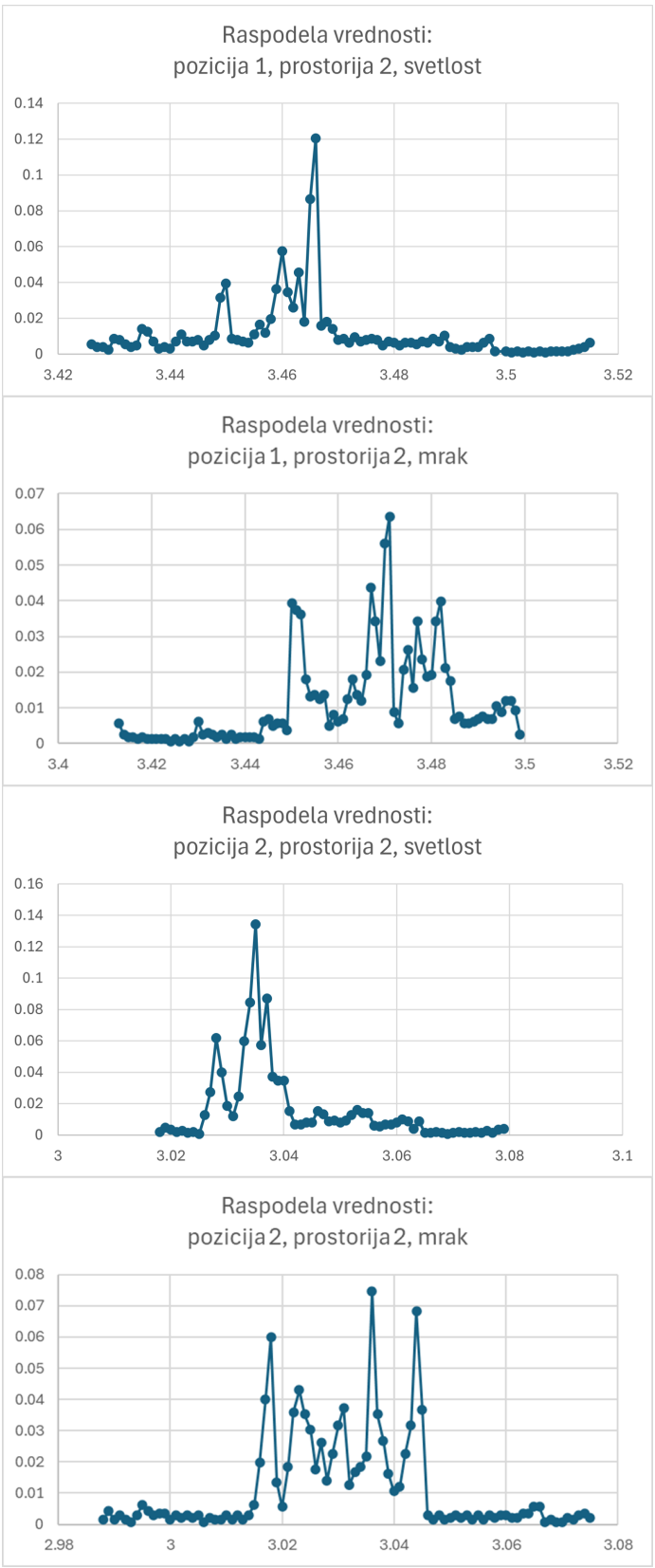
U ovoj sekciji će biti predstavljeni rezultati eksperimenta, bazirani na snimljenim vrednostima distance. Prvo (u podsekciji 5.1) će biti prikazani spektri vrednosti distance izmerenih sa Meta Quest 3 uređajem zajedno sa njihovom učestalošću, a zatim (u podsekciji 5.2) će biti prikazana poređenja prosečnih vrednosti distanci dobijenih preko Meta Quest 3 uređaja, laserskog metra, kao i predviđenih vrednosti objekata u samom Unity alatu.

A. Prikaz izmerenih vrednosti distanci sa Meta Quest 3 uređajem

Kao što je bilo pomenuto u opisu eksperimenta, merenja su bila rađena u dve različite prostorije na dve odabrane pozicije u svakoj, u uslovima svetlosti i mraka. U skladu sa time, u nastavku na slikama 4 i 5 se mogu videti rezultati merenja u svim navedenim uslovima.



Slika 4. Rezutati merenja pozicija u prostoriji 1.



Slika 5. Rezutati merenja pozicija u prostoriji 2.

B. Uporedba vrednosti distanci dobijenih putem različitih metoda

U nastavku na tabelama 2 i 3 se mogu videti u prostorijama 1 i 2 (R1 i R2 oznake su korišćene u tabeli) poređenja vrednosti za istu koordinatu u istim uslovima svetlosti. Vrednosti koje su bile poređene su redom: prosečna vrednost distance za odgovarajuću koordinatu dobijena putem vrednosti dobijene preko Meta Quest 3 uređaja, originalne distance izračunate na osnovu distance vrednosti koordinata unetih za odgovarajući objekat odakle se radi merenje u samom Unity razvojnom okruženju i na kraju merenje sa odgovarajuće pozicije dobijeno laserskim metrom, gde je uzeta prosečna vrednost četiri merenja zarad što veće preciznosti i smanjenja ljudske greške.

Tabela II
POREĐENJA PROSEKA RAZLIČITIH MERENJA U PROSTORIJI 1.

	R1, P1, S	**R1, P1, M**	**R1, P2, S**	**R1, P2, M**
Meta Quest 3 merenja	1.855470081	1.814214143	1.306806098	1.255630263
Unete lokacije u Unity	1.802775638	1.802775638	1.220655562	1.220655562
Merenje metrom	1.844	1.844	1.31225	1.31225

Tabela III
POREĐENJA PROSEKA RAZLIČITIH MERENJA U PROSTORIJI 2.

	R2, P1, S	**R2, P1, M**	**R2, P2, S**	**R2, P2, M**
Meta Quest 3 merenja	3.467665627	3.462897638	3.038845847	3.03089507
Unete lokacije u Unity	3.44818793	3.44818793	3.008321791	3.008321791
Merenje metrom	3.54075	3.54075	3.0485	3.0485

VI. DISKUSIJA I PLANOVI DALJIH ISTRAŽIVANJA

Iz prikazanog istraživanja u prethodnoj sekciji se može videti da Meta Quest 3 daje solidne rezultate tačnosti. Sama distanca izmerena uređajem je davala dovoljno tačne rezultate u poređenju sa merenjem laserskim metrom, gde je tačnost uvek makar na decimetarskom nivou. Primećuje se da u obe prostorije tačnost distance se smanjuje u uslovima mraka. To se takođe može primetiti po tome što su veće devijacije vrednosti u uslovima mraka, dok se u uslovima svetlosti izdvaja približno jednom maksimumu. Interesantno, na osnovu prikazanih merenja, primećuje se da najnetačnije vrednost daju objekti izgenerisani Unity okruženjem. Iz ovoga se može zaključiti da Meta Quest 3 ima slabiju procenu pozicioniranja virtuelnih objekata u odnosu na tačnu procenu samog uređaja i korisnika. Postoji nekoliko smerova u kojima se može dalje proširiti istraživanje. Jedan je da se isto istraživanje sprovede u otvorenim uslovima, gde se očekuje na osnovu rada SLAM algoritama Meta Quest 3 daje slabije procene distance. Drugi smer je da se isto istraživanje sprovede u kontrolisanijim uslovima sa preciznijom referentnom vrednošću (umesto laserskog metra da se koristi precizniji uređaj za merenje lokacije, poput Leica Viva TS15). Istraživanje se može proširiti tako

što se meri tačnost pozicije ne samo HMD, već i propratnih ručki. Moguće je odraditi istraživanje sa većim distancama i lokacijama.

ZAHVALNICA

Istraživanje je realizovano u prostorijama Palate nauke, Zadužbini Miodraga Kostića. Ovaj rad je finansijski podržalo Ministarstvo za nauku, tehnološki razvoj i inovacije Republike Srbije po ugovoru broj: 451-03-137/2025-03/200103

Overview of Positioning Algorithms in Augmented and Virtual Reality

Mihajlo Ogrizović, Filip Janković, Aleksa Ilić, Dražen Drašković

ABSTRACT

Virtual and augmented reality are widely used today in various fields such as education, industrial training, virtual tours of landmarks, and more. For the purpose of implementing simulations of such spaces in virtual reality, it is necessary to have an understanding of one's position within that space. Different augmented and virtual reality (VR) devices provide various mechanisms and algorithms for spatial position detection. This paper presents an overview and analysis of different algorithms used for determining spatial position and detecting edges and boundaries within that space. The review will cover a wide range of VR device models, while the experiment itself will be conducted using the Meta Quest 3 device. The experiment will include several different environments and objects within them. The accuracy of the positioning will be observed both relative to other positions and in absolute space.

Analiza performansi različitih algoritama broadcast komunikacije u Open MPI biblioteci

Miloš Nastić
Univerzitet u Beogradu, Elektrotehnički fakultet
Beograd, Srbija
nm235045p@student.etf.bg.ac.rs
0009-0009-9902-1912

Marko Mišić
Univerzitet u Beogradu, Elektrotehnički fakultet
Beograd, Srbija
marko.misic@etf.bg.ac.rs
0000-0002-7369-4010

Pavle Vuletić
Univerzitet u Beogradu, Elektrotehnički fakultet
Beograd, Srbija
pavle.vuletic@etf.bg.ac.rs
0000-0001-5600-2652

Jelica Protić
Univerzitet u Beogradu, Elektrotehnički fakultet
Beograd, Srbija
jelica.protic@etf.bg.ac.rs
0000-0003-0846-0290

***Apstrakt* - Message Passing Interface (MPI) je de fakto standard za pisanje paralelnih programa korišćenjem programske paradigme razmene poruka. Razmena poruka se može obavljati korišćenjem zajedničke memorije, ali i slanjem podataka putem mreže u zavisnosti od toga da li su MPI procesi pokrenuti na istom računaru ili ne. Ovaj rad se bavi analizom MPI_BCAST primitive za kolektivnu komunikaciju kojom se vrši slanje poruke od jednog do svih specificiranih procesa. Rad analizira konkretne algoritme kojima je implementirana ova primitiva u okviru Open MPI biblioteke, prikazuje matematičke modele, za svaku od implementacija, kojima se može vršiti predikcija vremena izvršavanja i upoređuje ih sa rezultatima merenja u Mininet mrežnom emulatoru. Utvrđeno je da je način odabira algoritma komunikacije značajno promenjen u verziji 4.1 Open MPI biblioteke i da u nekim situacijama vreme izvršavanja može biti više od 30 puta kraće u odnosu na verziju 4.0.**

Ključne reči – Message Passing Interface (MPI), paralelno programiranje, mrežna komunikacija, modelovanje i merenje performansi, Mininet

I. Uvod

Message Passing Interface (MPI) je standardizovana, prenosiva specifikacija biblioteke za razmenu poruka na programskim jezicima C i Fortran. Koristi se u računarstvu visokih performansi (eng. *high-performance computing*) i aplikacijama koje koriste tehnike paralelnog procesiranja [1]. MPI procesi izvršavaju MPI program i povremeno, po potrebi, međusobno komuniciraju. Komunikacija se može obavljati korišćenjem deljene memorije npr. kada su učesnici u komunikaciji procesori u nekom *UMA* (eng. *Uniform memory access*) sistemu. Češće, komunikacija se obavlja slanjem paketa putem mreže ukoliko su MPI procesi pokrenuti na fizički različitim uređajima. Komunikacija se obavlja korišćenjem namenskih objekata koji se nazivaju komunikatori i koji mogu obuhvatiti određeni broj procesa. Komunikacione primitive MPI standarda mogu da se prema nameni i broju učesnika podele u tri grupe kojima se vrši:

- Komunikacija između tačno dva procesa (eng. *point-to-point* komunikacija),
- Kolektivna komunikacija, u kojoj učestvuje više procesa,
- Sinhronizacija procesa koji učestvuju u komunikaciji.

Komunikacija između tačno dva procesa obuhvata različite implementacije operacija za slanje (MPI_SEND) i prijem (MPI_RECV) koje mogu uključiti blokiranje, sinhrono i asinhrono slanje i prijem, upotrebu bafera i sl. Kolektivna komunikacija obuhvata operacije u kojima učestvuje više procesa. Tipične operacije uključuju slanje istog podatka svima u grupi (eng. *broadcast*, MPI_BCAST), podelu paketa podataka unutar grupe (eng. *scatter*, MPI_SCATTER), objedinjavanje podataka od svih u grupi (eng. *gather*, MPI_GATHER), kao i redukcione operacije sa smeštanjem rezultata u jednom od procesa ili svim procesima (eng, *reduce*, MPI_REDUCE i MPI_ALL_REDUCE).

U procesu profilisanja izvršavanja različitih MPI programa uočeno je da 99% vremena odlazi na izvršavanje samo 19 MPI rutina. Pokazano je da je MPI_BCAST druga najčešće korišćena kolektivna operacija, posle operacije MPI_ALL_REDUCE [2]. Utvrđeno je da kod tipičnih MPI programa može biti utrošeno i više od 80% vremena na izvršavanje kolektivnih operacija [3].

MPI specifikacija [4] definiše semantiku komunikacionih operacija, ali ne i njihovu implementaciju koja je ostavljena konkretnim bibliotekama kao što su Open MPI i MPICH. Ovaj rad se bavi analizom performansi različitih implementacija MPI_BCAST primitive za kolektivnu komunikaciju u situaciji kada se koristi mrežna komunikacija u okviru Open MPI biblioteke. Mrežna komunikacija je nekoliko redova veličine sporija od komunikacije korišćenjem deljene memorije i benefiti ostvareni njenim ubrzavanjem značajno doprinose smanjenu ukupnog vremena izvršavanja MPI programa.

Rad je organizovan na sledeći način. U poglavlju 2 dat je pregled radova iz oblasti. Poglavlje 3 opisuje način rada Open MPI biblioteke [5] u navedenom slučaju. Poglavlje 4 upoređuje matematičke formule za određivanje očekivanog vremena izvršavanja različitih algoritama pomoću kojih je realizovana *broadcast* komunikacija. Poglavlje 5 prikazuje test okruženje i opisuje postupak merenja. U poglavlju 6 dati su rezultati merenja i njihovo tumačenje. U poglavlju 7 dat je zaključak razmatranja.

II. Pregled oblasti istraživanja

U situacijama kada je obrada vremenski i memorijski složena da jedan računar nije dovoljan, već se mora koristi više njih povezanih mrežom u klaster, brzina mrežne komunikacija ima veliki uticaj na ukupno vreme izvršavanja programa. Najviše prostora za unapređivanje ostavljaju primitive za kolektivnu komunikaciju kao što su MPI_BCAST, MPI_GATHER, MPI_SCATTER i sl, jer su uobičajeno realizovane korišćenjem različitih šema *point-to-point* komunikacije.

U radu [6] su upoređene različite implementacije primitiva kolektivne komunikacije i njihova očekivana vremena izvršavanja korišćenjem tri različita matematička modela.

Doktorska disertacija [7] ističe uticaj sinhronizacije između procesa na performanse kolektivne komunikacije. U navedenom radu prezentuje se novi radni okvir „HAN“ [8] za Open MPI uz pomoć kog se ostvaruje ubrzanje do 4,86 puta u odnosu na dotadašnju brzinu izvršavanja MPI_BCAST primitive.

Rad [9] ističe da odabir odgovarajućeg algoritma može zavisiti od različitih parametara i predlaže radni okvir za automatsko odabiranje najbolje implementacije za određenu MPI primitivu korišćenjem tehnika mašinskog učenja.

Autori rada [3] su predložili i uporedili sa realnim merenjima pristup u kome analitičke modele definišu na osnovu procene vremena izvršavanja konkretnog koda kojim je implementiran neki algoritam, a ne na osnovu njegovog matematičkog opisa. U većini slučajeva njihov model je odabrao najefikasniji algoritam, a u najgorem slučaju algoritam koji je do 10% sporiji od najefikasnijeg algoritma za testiranu veličinu poruke.

III. Način rada Open MPI biblioteke

Open MPI je biblioteka otvorenog koda koja implementira primitive definisane MPI standardom [4]. U nastavku rada će akcenat biti stavljen na implementaciju MPI_BCAST primitive.

A. Semantika MPI_BCAST primitive

MPI standard definiše listu parametara za svaku MPI primitivu. Kôd 1 prikazuje definiciju MPI_BCAST primitive.

```
MPI_BCAST(buffer, count, datatype, root, comm)
```

Kod 1. Definicija interfejsa MPI_BCAST primitive prema MPI standardu[4]

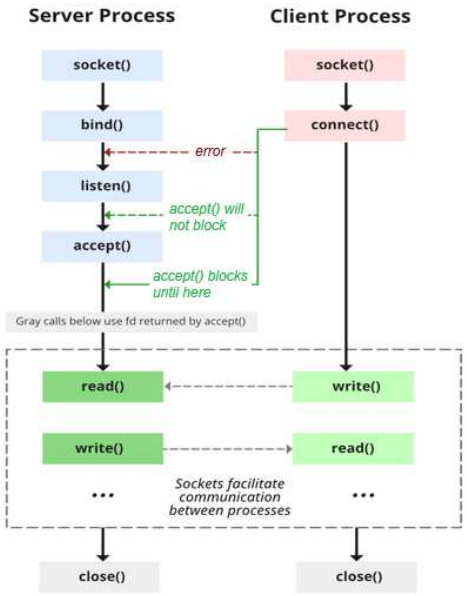
Značenje parametara je sledeće:

- *buffer* – Kod pošiljaoca je to adresa bafera u kome se nalaze podaci koje treba poslati. Kod primaoca je to adresa bafera u koji treba smestiti primljene podatke.
- *count* – Broj elemenata tipa *datatype* koji se prenosi.
- *datatype* – MPI tip podataka u baferu.
- *root* – Identifikator procesa pošiljaoca. S obzirom da se navedena primitiva poziva na identičan način kod svih MPI procesa, ovaj parametar služi biblioteci da utvrdi ko šalje podatke, a ko treba da ih primi.
- *comm* – Objekat komunikatora koji predstavlja grupu procesa koji učestvuju u komunikaciji.

B. Implementacija unutar Open MPI biblioteke

1) Obezbeđivanje mrežne komunikacije u Linux operativnom sistemu

U situacijama kada se komunikacija obavlja putem mreže na Linux operativnom sistemu koristi se *SOCK_STREAM* [10] koji predstavlja *socket* za TCP protokol. Glavni benefit ovog pristupa je taj što se time garantuje pouzdan prenos i prenos paketa u redosledu slanja. Obe stvari su neophodne za ispravan rad MPI programa. Svaki učesnik u komunikaciji mora dobiti sve podatke koji su njemu namenjeni i mora ih dobijati u redosledu u kom ih je pošiljalac poslao. *Linux socket* API razdvaja onoga ko šalje podatke (klijent) od onoga ko prima podatke (server). Bitno je navesti da je, zbog prirode TCP protokola, neophodno da se server startuje pre pokretanja klijenta. U suprotnom dobija se greška koja je prikazana na dijagramu na Slici 1 kada klijentski proces pozove *connect()* funkciju pre nego što je serverski proces pozvao *bind()* funkciju.



Slika 1. Dijagram stanja klijentskog i serverskog procesa u programskom jeziku C [11]

2) Implementacija kolektivne komunikacije

TCP protokol podržava samo *point-to-point* komunikaciju, pa se semantika MPI primitiva za kolektivnu komunikaciju mora prilagoditi ovom ograničenju. Najjednostavniji način bi bio da pošiljalac (proces sa rangom jednakim vrednosti *root*) pošalje *unicast* pakete ka svakom primaocu. Ovo rešenje ima vremensku složenost *O(n)*. Iz tog razloga takvo rešenje nije skalabilno za veći broj učesnika u komunikatoru i implementacije MPI standarda koriste naprednije algoritme. Uvidom u izvorni kod Open MPI biblioteke u datoteci *coll_base_bcast.c* uočava se 9 osnovih implementacija. U Tabeli 1 su navedeni ti algoritmi sa rednim brojevima koji su im dodeljeni u datoteci *coll_tuned_decision_fixed.c*.

Tabela 1. Šabloni komunikacije za ostvarivanje broadcast komunikacije unutar Open MPI biblioteke

Open MPI v4.1 broj	Algoritam	Koristi se u v4.0 i ranijim	Koristi se u v4.1 i kasnijim
1	*basic_linear*	NE	DA
2	*chain*	NE	DA
3	*pipeline*	DA	DA
4	*split_bintree*	DA	DA
5	*bintree*	NE	DA
6	*binomial*	DA	DA
7	*knomial*	NE	DA
8	*scatter_allgather*	NE	DA
9	*scatter_allgather_ring*	NE	NE

Izbor algoritma se vrši na osnovu veličine komunikatora

(broja učesnika u komunikaciji) i količine podataka koja se šalje. Ukupan broj broadcast algoritama koji se koristi, kao i algoritam odluke su se značajno menjali tokom vremena. Kôd 2 predstavlja isečak koda za odabir algoritma slanja u verzijama zaključno sa 4.0, dok Kod 3 predstavlja isečak koda za odabir algoritma slanja u verzijama počev od 4.1 Open MPI biblioteke.

```
/* Decision function based on MX results for messages up
to 36MB and communicator sizes up to 64 nodes */
const size_t small_message_size = 2048;
const size_t intermediate_message_size = 370728;
const double a_p16  = 3.2118e-6; /* [1 / byte] */
const double b_p16  = 8.7936;
const double a_p64  = 2.3679e-6; /* [1 / byte] */
const double b_p64  = 1.1787;
const double a_p128 = 1.6134e-6; /* [1 / byte] */
const double b_p128 = 2.1102;
...
if ((message_size < small_message_size) || (count <= 1)){
  /* Binomial without segmentation */
  segsize = 0;
  return  ompi_coll_base_bcast_intra_binomial(buff,
        count, datatype, root, comm, module, segsize);
} ...
else if (communicator_size < (a_p128 * message_size
                              + b_p128))
  /* Pipeline with 128KB segments */
  segsize = 1024  << 7;
  return ompi_coll_base_bcast_intra_pipeline(buff, count,
         datatype, root, comm, module,segsize);
}
```

Kod 2. Isečak koda za odabir algoritma slanja u Open MPI v4.0

```
if (communicator_size < 4) {
  if (total_dsize < 32) {
    alg = 3;
  } else if (total_dsize < 256) {
    alg = 5;
  } else if (total_dsize < 512) {
    alg = 3;
  } ...
} else if (communicator_size < 8) {
  if (total_dsize < 64) {
    alg = 5;
  } else if (total_dsize < 128) {
    alg = 6;
  } else if (total_dsize < 2048) {
    alg = 5;
  } ...
} ...
```

Kod 3. Isečak koda za odabir algoritma slanja u Open MPI v4.1

Odluke u verzijama zaključno sa 4.0 se donose na osnovu logičkih izraza u kojima figurišu vrednosti "a_" i "b_" konstanti koje su empirijski određene na osnovu rezultata merenja performansi tokom testiranja *benchmark* programima.

Algoritmi su numerisani u kodu počev od verzije 4.1. Mapiranje između broja i njemu odgovarajućeg algoritma je prikazano u Tabeli 1. Odluke u verzijama počev od 4.1 su suštinski po istom principu (u funkciji veličine komunikatora i količine podataka koja se šalje), ali postoje i određene razlike. Više se ne koriste "a_" i "b_" konstante. If-else konstrukt u v4.0 je imao 8 mogućih ishoda, dok u v4.1 je taj broj 44. Poslednja razlika koja se uočava je da su u v4.0 korišćena samo 3 algoritma, a u verziji 4.1 njih 8. Ova promena govori o značaju odabira najadekvatnijeg algoritma.

U trenutku pisanja ovog rada postoje problemi gubitka konekcije sa čvorom [12] i beskonačnog blokiranja [13] u v4.1 i v5.0 Open MPI biblioteke koji se odnose na izvršavanje kolektivnih operacija. Iz navedenog razloga testiranja su rađena na Open MPI v4.0. Samim tim nije bilo moguće testirati sve algoritme iz Tabele 1. Testirani su algoritmi koji su označeni u odgovarajućoj koloni Tabele 1 uz dodatak algoritma *basic_linear*, *bin_tree* i *knomial*. Ovo je bilo moguće uraditi zato što implementacije svih algoritama iz Tabele 1 već postoje i unutar koda v4.0, ali se još uvek nisu koristile. Navedeni algoritmi su se uspešno izvršavali nakon pokretanja iz modifikovane *coll_tuned_decision_fixed.c* datoteke. To nije bio slučaj sa preostala 3 algoritma (*chain*, *scatter_allgather*, *scatter_allgather_ring*) koji su se beskonačno blokirali ili generisali greške pri pokušaju pokretanja u v4.0 tako da su oni izostavljeni iz dalje analize.

IV. MATEMATIČKO MODELOVANJE ALGORITAMA KOMUNIKACIJE

Za modelovanje komunikacionih aspekata algoritama za kolektivnu komunikaciju često se koriste *Hockney* i *LogP* model (i njegova proširenja) [3]. Tabela 2 poredi kako različiti modeli definišu vreme slanja poruke između dva učesnika i ograničenja svakog modela. Simboli su sledeći:

- α – kašnjenje poruke (eng. *message latency*)
- $\beta = 1$ / *bandwidth*
- m – veličina poruke
- m_s – veličina segmenta (deo ukupne poruke)
- L – maksimalna vrednost kašnjenja
- o – režijski troškovi (eng. *overhead*)
- G – pauza između slanja dva uzastopna bajta
- g – pauza između slanja dve uzastopne poruke
- n_s – broj segmenata koji se šalju
- P – broj učesnika (čvorova) u komunikaciji

Tabela 2. Poređenje modela za analitičko opisivanje algoritama mrežne komunikacije [6]

Model	Vreme slanja između dva učesnika	Ograničenje
Hockney	$\alpha + m \times \beta$	Zagušenje u mreži nije moguće modelovati.
LogP	$L + 2o$	Pretpostavlja da se šalju samo male poruke konstantne veličine.
LogGP	$L + 2o + (m - 1)G$	Dozvoljava slanje najviše $\lfloor \frac{L}{g} \rfloor$ poruka istovremeno.
PLogP	$L + g(m)$	Za određivanje funkcije *g(m)* najčešće se koriste modeli mašinskog učenja što zahteva veliku količinu trening podataka.

S obzirom da se ovaj rad bavi idealnom situacijom kada je pokrenut samo jedan MPI program i MPI procesi jedini komuniciraju u mreži za očekivati je da zagušenje neće postojati, pa su u Tabeli 3 prikazane samo jednačine po *Hockney* modelu. U jednačinama se koristi simbol m_s koji označava veličinu segmenta koji se šalje. Vrednost m_s je manja ili jednaka veličini poruke u zavisnosti od toga da li se vrši segmentacija poruke ili ne.

Najjednostavniji i u većini slučajeva najsporiji algoritam je *basic_linear* zato što pošiljalac svaki segment šalje svakom

primaocu. Kod *pipeline* algoritma učesnici su poređani tako da svaki, sem prvog i poslednjeg, ima jedan čvor od kojeg prima i jedan kome šalje podatke. Ubrzanje potiče od činjenice da nakon što je primio paket učesnik može dalje da ga prosledi sledećem čvoru u paraleli sa primanjem narednog paketa i to važi za sve učesnike. Preostala četiri algoritma (*binomial*, *knomial*, *bintree* i *split_bintree*) komunikaciju između čvorova obavljaju u formi stabla od korenog čvora ka listovima. Najjednostavnija varijanta je slanje u šemi binarnog stabla (*bintree*) gde čvorovi dobijaju pakete od pretka i prosleđuju potomcima u *log2(n)* koraka.

Tabela 3. Hockney modeli različitih implementacija broadcast komunikacije [6]

Algoritam	Ukupno vreme slanja
basic_linear	$T = n_s \times (P-1) \times (\alpha(m_s) + m_s \times \beta(m_s))$
pipeline	$T = (P + n_s - 2) \times (\alpha(m_s) + m_s \times \beta(m_s))$
binomial	$T = \lceil \log_2(P) \rceil \times n_s \times (\alpha(m_s) + m_s \times \beta(m_s))$
knomial	$T = \lceil \log_k(P) \rceil \times n_s \times (\alpha(m_s) + m_s \times \beta(m_s))$
bintree	$T = (\lceil \log_2(P+1) \rceil + n_s - 2) \times (2 \times \alpha(m_s) + m_s \times \beta(m_s))$
split_bintree	$T = (\lceil \log_2(P+1) \rceil + n_s - 2) \times (2 \times \alpha(m_s) + m_s \times \beta(m_s)) + \alpha\left(\frac{m_s}{2}\right) + \frac{m_s}{2} \times \beta\left(\frac{m_s}{2}\right)$

Modifikacija ovog algoritma je *split_bintree* kod kojeg koreni čvor deli poruku na dva dela pre daljeg slanja. Jednu polovinu dobijaju potomci u levom podstablu, a drugu polovinu potomci u desnom. Potomci kasnije međusobno razmenjuju primljene sadržaje, sa odgovarajućim čvorom iz drugog podstabla, da bi kompletirali celu poruku. Ovaj algoritam pomaže u situacijama kada postoji zagušenje i kada je vreme za prijem svih podataka, pre slanja dalje, veliko. U drugim situacijama može imati i negativne efekte.

Binomno stablo se definiše rekurzivno.

- Binomno stablo reda nula predstavlja samo koreni čvor.
- Binomno stablo reda *n* ima koreni čvor čiji su potomci koreni čvorovi binomnih stabla reda *n-1*, *n*-2, … , 2, 1, 0.

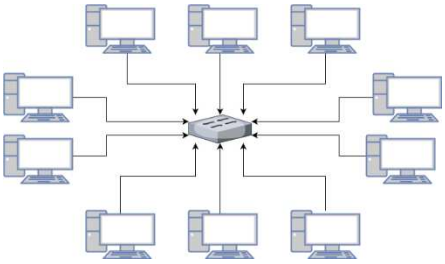
Slanje u šemi binomnog stabla (*binomial* algoritam) se obavlja takođe u $log_2(n)$ koraka. *Binomial* je vrsta *knomial* stabla za *k=2*. Kod *knomial* algoritma broj koraka iznosi $log_k(n)$.

S obzirom da su očekivana vremena izvršavanja prikazana u Tabeli 3 iskazana u odnosu na iste parametre (broj učesnika, kašnjenje, propusni opseg i veličinu poruke) moguće je i bez određivanja konkretnih vrednosti tih parametara izvršiti predikciju koji algoritam bi trebao da bude brži od kog. Za slučaj kada posmatramo manji broj učesnika u komunikaciji (npr. 10), kada ne očekujemo zagušenja i kada veličina segmenata koji se šalju nije tolika da značajno utiče na vreme slanja poruke očekujemo odnos prikazan u (1).

$$T_{split-bintree} \leq T_{basic-linear} \leq T_{pipeline} \leq T_{bintree} \leq T_{binomial} \leq T_{knomial} \quad (1)$$

V. Postupak testiranja

Radi provere teorijskih pretpostavki i utvrđivanja da li i u kojoj meri odabir konkretnog algoritma slanja podataka, pri pozivu MPI_BCAST primitive, utiče na ukupno vreme slanja podataka izvršena su merenja u *Mininet* mrežnom emulatoru. Korišćena je jednostavna topologija sa 10 čvorova (*host*) povezanih na jedan svič prikazana na Slici 2.



Slika 2. Topologija korišćena za testiranje

Svaki *host* je imao omogućenu SSH komunikaciju sa ostalim *host*-ovima, što je ključno za pokretanje MPI procesa pomoću komandi *mpi_exec* ili *mpi_run*. Ove komande koriste *hostfile*, čiji je primer prikazan u kodu 4. U konkretnom slučaju, pokreće se po jedan MPI proces na svakom *host*-u (*slots=1*). Na osnovu hostfile-a, određuju se učesnici komunikacije, a *host* sa kojeg je pokrenuta komanda pokušava da uspostavi SSH vezu sa svakim od njih i prosledi im parametre za pokretanje MPI programa.

```
10.0.0.1 slots=1
10.0.0.2 slots=1
...
```

Kod 4. Primer hostfile datoteke

Testiranje je obavljeno sa test programom čija pojednostavljena verzija je prikazana u Kodu 5. Program je za svaki od 5 algoritama pokretan za različite veličine bafera tj. vrednosti *num_elements* promenljive (125, 500, 2000 i 8000). Najmanja vrednost 125 odgovara slanju 500B podataka (125x4B) što predstavlja manje od MTU (eng. *Maximum Transmission Unit*) i ceo bafer se može poslati u jednom paketu. Ostale vrednost su nastale množenjem sa 4 tako da je najveća testiranja vrednost ekvivalenta baferu veličine 32KB. Slanje je obavljano u petlji od 10000 iteracija.

```
int main(int argc, char** argv) {
  ...
  if (world_rank == 0) {
    // Inicijalizacija bafera podacima i pocetak merenja
    total_mpi_bcast_time -= MPI_Wtime();
  }
  for (int i = 0; i < num_trials; i++) {
    MPI_Bcast(data, num_elements, MPI_INT, 0,
              MPI_COMM_WORLD);
  }
  if (world_rank == 0) {
    total_mpi_bcast_time += MPI_Wtime();
    printf("Ukupno vreme slanja = %lf\n",
           total_mpi_bcast_time);
  }
  ...
}
```

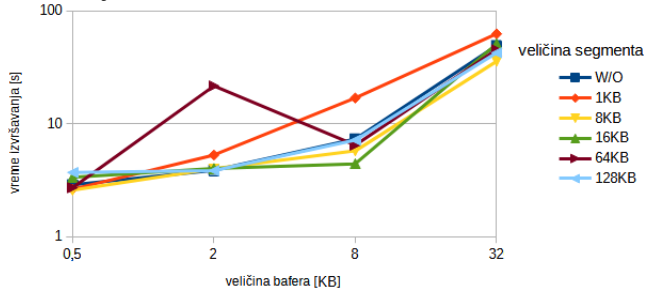
Kod 5. Test program korišćen za merenje brzine izvršavanja MPI_BCAST primitive

Sam test predstavlja sintetičko opterećenje koje se ne viđa u realnim primenama, ali najbolje prikazuje i najmanje razlike u vremenu izvršavanja različitih implementacija.

VI. Rezultati

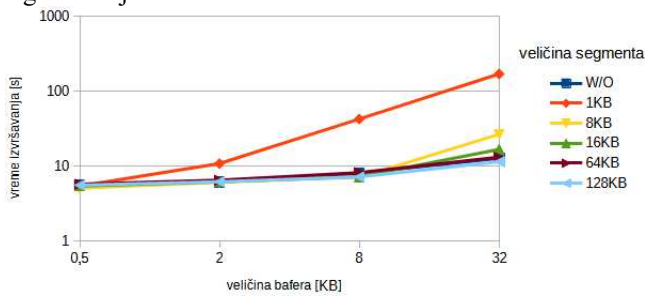
Analizom izvornog koda Open MPI biblioteke utvrđeno je da se pored odabira algoritma slanja vrši i odabir veličine segmenta koji se šalje. Uočene su vrednosti segmenta od 0 (bez segmentacije – šalje se ceo bafer), 1KB, 8KB, 16KB, 64KB i 128KB. Svaki algoritam je testiran sa navedenim veličinama segmenata i dobijeni su grafici prikazani na Slikama 3-6. *Basic_linear* algoritam ne vrši segmentaciju paketa, pa za njega ne postoji grafik zavisnosti od veličine segmenta. S obzirom da se *knomial* i *binomial* algoritmi razlikuju samo u vrednosti osnove (eng. *radix*) i da su na njihovim graficima relativni odnosi vrednosti za različite veličine segmenata skoro identični prikazan je samo grafik za *binomial* (što je ekvivalentno sa *knomial* za *radix* = *2*).

Slika 3 govori da kod *pipeline* algoritma veličine segmenta od 8KB i 16KB daju najbolje rezultate za testirane veličine bafera. Veličina segmenta od 64KB ima loš uticaj na vreme izvršavanja za veličinu bafera od 2KB.

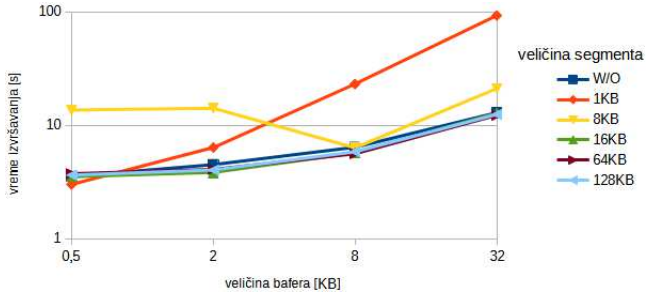


Slika 3. Grafik zavisnosti ukupnog vremena izvršavanja *pipeline* algoritma od veličine segmenta i veličine bafera koji se šalje.

Na Slici 4 vidimo da povećanje veličine segmenta kod *binomial* algoritma utiče na smanjenje vremena slanja. Veličine segmenta od 64KB, 128KB i slanje bez segmentacije imaju praktično ista vremena izvršavanja. To je i očekivano s obzirom da je najveća testirana veličina bafera 32KB, što je manje od navedenih veličina pa svakako nema segmentacije bafera.



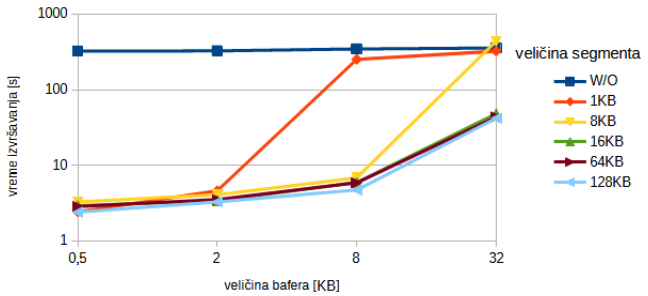
Slika 4. Grafik zavisnosti ukupnog vremena izvršavanja *binomial* algoritma od veličine segmenta i veličine bafera koji se šalje.



Slika 5. Grafik zavisnosti ukupnog vremena izvršavanja *bintree* algoritma od veličine segmenta i veličine bafera koji se šalje.

Na osnovu grafika na Slici 5 može se zaključiti da je veličina segmenta od 8KB nepovoljna za male veličine bafera. U svim ostalim situacijama veća veličina segmenta dovodi do manjeg vremena slanja paketa.

Za razliku od prethodno posmatranih algoritama, na Slici 6 se može uočiti da *split_bintree* algoritam ima najgore performanse kada se ne vrši segmentacija bafera. Ovo je očekivano, jer *split_bintree* je modifikacija *bintree* algoritma koja ubrzanje ostvaruje upravo tako što podatke koji treba da se pošalju deli u manje celine i šalje u paraleli. Kada toga nema šema slanja ovog algoritma je manje efikasna od običnog *bintree* algoritma i zbog toga se dobijaju lošiji rezultati. Što se tiče veličine segmenta i kod ovog algoritma uočava se smanjenje vremena izvršavanja sa povećanjem veličine segmenta.



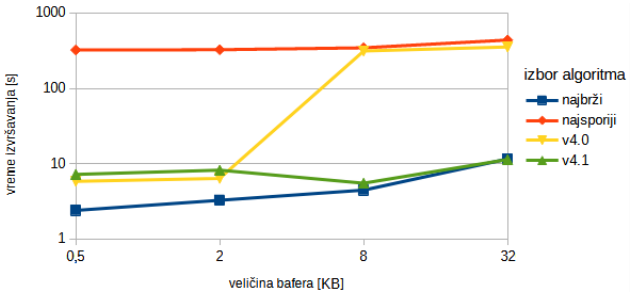
Slika 6. Grafik zavisnosti ukupnog vremena izvršavanja *split_bintree* algoritma od veličine segmenta i veličine bafera koji se šalje.

Na osnovu prethodnih merenja su izdvojeni najbrži i najsporiji rezultati za svaku od testiranih veličina bafera i upoređeni sa vrednostima koje se dobijaju u v4.0 i v4.1 Open MPI biblioteke bez modifikacije koda. Rezultati su prikazani u Tabeli 4 i na grafiku na Slici 7. Može se uočiti da za male vrednosti bafera (500B i 2KB) v4.0 i v4.1 imaju približne vrednosti, koje su 2-3 puta sporije od najbržih mogućih za te veličine bafera. Za vrednosti bafera od 8KB i 32KB v4.0 ima vremena slanja koja su približna najsporijim mogućim, dok v4.1 ima vrednosti koje su vrlo približne najbržim mogućim. To dovodi do zaključka da nova heuristika odabira implementacije MPI_BCAST primitive ima bolje performanse od one korišćene u starijim verzijama biblioteke.

Vrednosti za v4.1 usled problema [12] nije bilo moguće testirati direktno već su određene posrednim putem. Utvrđeno je koju implementaciju iz Tabele 3 bi ta verzija koristila (prikazano u Tabeli 5) za svaku od veličina bafera i onda su te implementacije forsirane unutar v4.0 koda i izvršena merenja. Iz tog razloga stoji zvezdica u Tabeli 4.

Tabela 4. Poređenje najbržeg, najsporijeg i vremena izvršavanja v4.0 i v4.1 implementacija bez modifikacije za različite veličine bafera.

Izbor algoritma	**Veličina bafera [KB]**			
	0,5	***2***	***8***	***32***
najsporiji	322,79 s	325,55 s	345,14 s	437,58 s
najbrži	2,39 s	3,26 s	4,43 s	11,54 s
v4.0	5,82 s	6,37 s	312,85 s	353,05 s
v4.1	*7,19 s	*8,16 s	*5,49 s	*11,30 s



Slika 7. Poređenje najbržeg, najsporijeg i vremena izvršavanja v4.0 i v4.1 implementacija bez modifikacije za različite veličine bafera.

Tabela 5 pokazuje da za veličine bafera od 8KB i 32KB verzija 4.1 koristi algoritme koji su u (1) na kraju poglavlja IV, na osnovu matematičkih modela, navedeni kao optimalni i merenja u *Mininet* emulatoru su to potvrdila.

Tabela 5. Algoritmi koji se koriste u različitim verzijama Open MPI biblioteke u zavisnosti od veličine bafera koji se šalje

Verzija Open MPI	**Veličina bafera [KB]**			
	0,5	***2***	***8***	***32***
v4.0	*Binomial, segsize=0*	*Binomial, segsize=0*	*Split_bintree, segsize=1KB*	*Split_bintree, segsize=1KB*
v4.1	*Knomial, segsize=0, radix=4*	*Knomial, segsize=0, radix=4*	*Binree, segsize=0*	*Binomial, segsize=0*

VII. Zaključak

Ovaj rad se bavi analitičkom i simulacionom analizom performansi dostupnih implementacija MPI_BCAST primitive u Open MPI biblioteci. Objašnjena je razlika u semantici između kolektivne komunikacije jedan prema svima (*broadcast*) i *unicast* komunikacije kao jedinog načina komunikacije korišćenjem TCP protokola. Objašnjena je potreba za korišćenjem efikasnijih komunikacionih algoritama od običnog slanja od izvora svakom primaocu zasebno. Prikazani su algoritmi koji se koriste zaključno sa verzijom 4.0 kao i algoritmi koji se koriste počev od verzije 4.1 Open MPI biblioteke. Date su jednačine za računanje očekivanog ukupnog vremena izvršavanja koristeći *Hockney* model. Izvršena su merenja sa jednostavnim sintetičkim test programom. Na kraju je prikazan učinak implementacija v4.0 i v4.1 u odnosu na najbrže i najsporije dobijene rezultate. Došlo se do zaključka da je brzina slanja za bafere od oko 8KB i veće značajno poboljšana počev od v4.1. Prikazano je da odabir odgovarajućeg algoritma slanja i veličine segmenta značajno utiče na performanse izvršavanja MPI_BCAST primitive i da u najekstremnijim slučajevima odnos između najbržeg i najsporijeg izvršavanja može biti veći od 30 puta.

Zahvalnica

Ovaj rad je delimično finansiran od strane Ministarstva nauke, tehnološkog razvoja i inovacija pod ugovorom 451-03-451-03-65/2024-03/200103. Autori su zahvalni na finansijskoj podrški.

Literatura

[1] W. Gropp, E. Lusk, and A. Skjellum, “Using MPI: portable parallel programming with the message-passing interface,” 3rd ed. MIT Press, Cambridge, MA, USA, 2014.

[2] R. Rabenseifner, “Automatic MPI counter profiling of all users: First results on a CRAY T3E 900-512,” Proceedings of the Message Passing Interface Developer’s and User’s Conference, 1999, pp. 77–85.

[3] E. Nuriyev and A. Lastovetsky, "Efficient and Accurate Selection of Optimal Collective Communication Algorithms Using Analytical Performance Modeling," in IEEE Access, vol. 9, pp. 109355-109373, 2021, doi: 10.1109/ACCESS.2021.3101689.

[4] “MPI Documents,” [Online] dostupno na: https://www.mpi-forum.org/docs/ [pristupljeno dana 9.12.2024.].

[5] “Open MPI: Open Source High Performance Computing,” [Online] dostupno na: https://www.open-mpi.org/ [pristupljeno dana 11.12.2024.].

[6] J. Pjesivac-Grbovic, T. Angskun, G. Bosilca, G. E. Fagg, E. Gabriel i J. J. Dongarra, “Performance analysis of MPI collective operations, “ 19th IEEE International Parallel and Distributed Processing Symposium, Denver, CO, USA, 2005, pp. 8 pp.-, doi: 10.1109/IPDPS.2005.335.

[7] X. Luo “Optimization of MPI Collective Communication Operations, “ PhD diss., University of Tennessee, 2020. [Online] dostupno na: https://trace.tennessee.edu/utk_graddiss/5818 [pristupljeno dana 11.12.2024.].

[8] X. Luo et al., “HAN: a Hierarchical AutotuNed Collective Communication Framework,” 2020 IEEE International Conference on Cluster Computing (CLUSTER), Kobe, Japan, 2020, pp. 23-34, doi: 10.1109/CLUSTER49012.2020.00013.

[9] S. Hunold, A. Bhatele, G. Bosilca and P. Knees, “Predicting MPI Collective Communication Performance Using Machine Learning,” 2020 IEEE International Conference on Cluster Computing (CLUSTER), Kobe, Japan, 2020, pp. 259-269, doi: 10.1109/CLUSTER49012.2020.00036.

[10] “socket(2) — Linux manual page,” [Online] dostupno na: https://man7.org/linux/man-pages/man2/socket.2.html [pristupljeno dana 14.12.2024.].

[11] “Internet Relay Chat (IRC),” [Online] dostupno na: https://github.com/sensoyyasin/ft_irc?tab=readme-ov-file [pristupljeno dana 14.12.2024.].

[12] “[4.1.5] ORTE has lost communication with a remote daemon #11830,” [Online] dostupno na: https://github.com/open-mpi/ompi/issues/11830 [pristupljeno dana 15.12.2024.].

[13] “openmpi-5.0.5 won't spawn #12916,” [Online] dostupno na: https://github.com/open-mpi/ompi/issues/12916 [pristupljeno dana 15.12.2024.].

Performance analysis of different broadcast communication algorithms in the Open MPI library

Miloš Nastić, Marko Mišić, Pavle Vuletić, Jelica Protić

ABSTRACT

The Message Passing Interface (MPI) is the de facto standard for writing parallel programs using the message passing programming paradigm. Message passing can be done using shared memory, but also by sending data over the network, depending on whether MPI processes are running on the same device or not. This paper deals with the analysis of the MPI_BCAST primitive for collective communication, which is used to send a message from one to all specified processes. The paper analyzes the specific algorithms that implement this primitive within the Open MPI library, shows mathematical models, for each of the implementations, that can be used to predict the execution time, and compares them with the measurement results in the Mininet network emulator. It was found that the way the communication algorithm is selected has been significantly changed in version 4.1 of the Open MPI library and that in some situations the execution time can be more than 30 times shorter compared to version 4.0.

Analiza algoritama dokaza bez znanja u blokčejn tehnologiji

1st Miloš Obradović
Elektrotehnički fakultet
Univerzitet u Beogradu
Beograd, Srbija
milos.obradovic@etf.bg.ac.rs
https://orcid.org/0000-0001-9225-3213

2nd Pavle Vuletić
Elektrotehnički fakultet
Univerzitet u Beogradu
Beograd, Srbija
pavle.vuletic@etf.bg.ac.rs
https://orcid.org/0000-0001-5600-2652

***Abstrakt*—Dokazi bez znanja (eng. *Zero-Knowledge Proofs - ZKP*) su familija kriptografskih protokola koji omogućavaju jednom entitetu - Dokazivaču (eng. *Prover*) da dokaže drugom entitetu - Verifikatoru (eng. *Verifier*) validnost skupa informacija bez otkrivanja samih informacija. U poslednjih nekoliko godina ZKP nalaze velike primene u blokčejnu (BČ) jer pomažu u rešavanju dva velika problema BČ tehnologije - privatnost podataka i skalabilnost mreže. Uprkos širokoj primeni, zbog brzog razvoja i kompleksnosti ZKP tehnologije, u literaturi nedostaje rad koji predstavlja dobru polaznu osnovu rada ZKP, koja obuhvata sve aktuelne pravce razvoja ZKP u BČ i suštinu rada glavnih ZKP protokola. Cilj ovog rada je da napravi pregled različitih vrsta ZKP protokola, prikaže osnovne karakteristike, prednosti i mane svake vrste i da buduće pravce istraživanja u ovoj oblasti.**

***Ključne reči*—dokazi bez znanja, blokčejn, kriptografija, ZK-SNARK, ZK-STARK, BulletProof**

I. Uvod i Motivacija

Kriptografski protokoli koji omogućavaju jednom entitetu - Dokazivaču (eng. *Prover*) da uveri drugi entitet - Verifikatora (eng. *Verifier*) da poseduje neke informacije, a da pritom ne otkrije same informacije nazivaju se dokazi bez znanja (eng. *Zero-Knowledge Proofs - ZKP*) [1]. Tri osnovne osobine koje treba da ispune ZKP su:

1) Kompletnost (eng. *Completness*) - ukoliko Dokazivač prati protokol, Verifikator će prihvatiti dokaz,
2) Pouzdanost (eng. *Soundness*) - zlonamerni Dokazivač ne može da prevari Verifikatora (osim sa izuzetno malom verovatnoćom),
3) Nulto znanje (eng. *Zero-Knowledge*) - Verifikator ne može da sazna ništa o skrivenoj informaciji osim dokaza da je ona tačna.

U blokčejn (BČ) tehnologiji, ZKP se koriste da obezbede skalabilnost sistema i privatnost informacija. Jedna od osnovnih inženjerskih hipoteza razvoja BČ tvrdi da nije moguće u isto vreme postići decentralizaciju, sigurnost i skalabilnost sistema [2], iako postoje primeri [3] koji govore da je lako dizajnirati sistem koji ima dve od tri navedene osobine. Jedan od modernih pristupa razvoja BČ je da se mreža podeli u 2 sloja, gde prvi sloj služiti da obezbedi decentralizovanost i sigurnost sistema, dok drugi sloj koristi sigurnost prvog sloja i obezbeđuje skalabilnost sistema, odnosno brže i jeftinije izvršenje transakcija [4].

Rešenja drugog sloja koja su zadužena za obradu i smeštanje podataka van prvog sloja, ali za svaku transakciju ostavljaju trag na prvom sloju i tako nasleđuju sigurnost prvog sloja rade po mehanizmu grupisanja transakcija (eng. *Rollups*) [5]. Prema načinu rada ova rešenja se dele na optimistična grupisanja transakcija (eng. *Optimistic Rollups*) i grupisanja transakcija bez znanja (eng. *Zero-Knowledge Rollups)*. Rad prvih zasniva se na dokazima o nepravilnosti (eng. *Fraud proofs*), dok se rad drugih zasniva na ZKP. Kako sama tehnologija dokaza o nepravilnosti donosi limit u propagaciji informacija između drugog i prvog sloja BČ, smatra se da će dugoročno svi algoritmi drugog sloja koristiti ZKP tehnologiju [6].

Pored upotrebe ZKP u razvoju skalabilnosti BČ, ova tehnologija nalazi i veliku primenu u očuvanju privatnosti korisnika BČ. ZKP protokoli omogućavaju postojanje anonimnih transakcija [7], sakrivanje ukupne količine novca koju korisnik poseduje [8], verifikaciju dela informacija iz digitalnog identiteta [9] i druge aspekte zaštite ličnih informacija.

U nastavku ovog rada odrađen je pregled rada ZKP tehnologije koji se koriste u BČ u 2025. Drugo poglavlje rada opisuje poželjne karakteristike ZKP za upotrebu u BČ i prikazuje kategorije ZKP algoritama. Naredno poglavlje rada analizira sve osnovne osobine rada ZKP kroz konstrukciju 1 protokola. Četvrto poglavlje poredi različite kategorije ZKP, uključujući bezbednosne elemente i efikasnost svakog. Završno poglavlje rada daje zaključak i opisuje buduće pravce istraživanja.

II. Dokazi bez znanja na blokčejnu

Pored neophodnih karakteristika navedenih u prethodnom poglavlju (Kompletnost, Pouzdanost, Nulto znanje) koje poseduju svi ZKP, na osnovu pregleda literature [10] [11] prepoznaju se još neke poželjne osobine koje ZKP treba da poseduju da bi bili efikasni za korišćenje na BČ:

1) Skalabilnost (eng. *Scalability*)- verifikacija dokaza treba biti što manje kompjuterski zahtevna i primenljiva za kompleksne tvrdnje,
2) Kompaktnost (eng. *Succinctness*) - veličina dokaza treba biti što manja,

3) Neinteraktivnost (eng. *Non-Interactivity*) - komunikacija se sastoji iz 1 poruke koju Dokazivač šalje Verifikatoru.

U BČ ulogu Verifikatora obično izvršava neki pametni ugovor (eng. *smart contract*). Kako se računarske operacije pametnih ugovora izvršavaju na BČ, njihove operacije su skupe i zbog toga je važno da ZKP poseduju što bolju Skalabilnost. Pored toga, osobine Kompaktnosti i Neinteraktivnosti značajno smanjuju cenu distribucije i smeštanja poruka na BČ i dosta utiču na primenljivost ZKP. Dodatna pogodnost je što dokaze koji poseduju osobinu Skalabilnost i Neinteraktivnosti može svako da proveri [12].

Pored navedenih, postoje i dodatne poželjne osobine ZKP kao što je bezbednost od kvantnih računara. Međutim, iako treba analizirati ove karakteristike, one nisu glavna prepreka za primenu ZKP u 2025 i zbog toga nisu detaljno izložene u ovom radu.

Na osnovu pregleda repozitorijuma otvorenog pristupa (eng. *open source*) [13] i naučne literature [12], uočavaju se 3 velike grupe ZKP protokola koji se koriste u BČ:

1) **Z**ero-**K**nowledge **S**uccinct **N**on-Interactive **AR**guments of **K**nowledge (ZK-SNARK),
2) **Z**ero-**K**nowledge **S**calable **T**ransparent **AR**guments of **K**nowledge (ZK-STARK),
3) **B**ullet**P**roofs (BP),

Pored navedenih poznatih grupa algoritama, postoje i druge implementacije koje se zasnivaju na korišćenju rekurzije i sličnih koncepata u cilju poboljšanja performansi algoritma. Iako ovi protokoli mogu da imaju i veće izmene u odnosu na osnovnu implementaciju algoritma, dok god su matematički principi na kojima se zasniva rad algoritma isti, ove promene nisu dovoljne da bi se algoritmi svrstali u zasebnu kategoriju.

III. Komponente dokaza bez znanja na primeru ZK-SNARK algoritma

Demonstracija ZKP protokola je praktičnija ukoliko su polazni uslovi zadati u matematički pogodnom obliku. Skup problema za koje ima smisla konstruisati ZKP čine problemi koji nisu lako rešivi, a za koje se može lako proveriti zadato rešenje, odnosno NP skup problema. NP-kompletni problemi [14] predstavljaju podskup NP problema koji ima osobinu da svaki NP problem može da se svede na zadati NP-kompletni problem. Na osnovu ove teoreme rad dokaza bez znanja je dovoljno demonstrirati na jednom NP-kompletnom problemu, računajući da onda bilo koji problem može da se svede na taj oblik.

Svođenje [15] originalnog skupa problema na pogodan NP-kompletan problem za rad ZK-SNARK algoritma [16] [17] odvija se u 3 koraka. Najpre se formuliše aritmetički izraz koji modeluje originalni problem, zatim se aritmetički izraz transformiše u sistem ograničenja prvog reda (eng. *Rank-1 Constraint System* - R1CS), nakon čega se R1CS redukuje na kvadratni aritmetički program (eng. *Quadratic Arithmetic Program* - QAP).

Format R1CS [18] prikazan je kroz formulu 1, gde a_i predstavlja elemente vektora veličine m koji predstavlja tajnu informaciju (eng. *witness*) koju Dokazivač zna i ne želi da otkrije; U, V, W predstavljaju matrice dimenzija $n \times m$ koje modeluju originalni skup ograničenja, gde je n broj ograničenja nakon transformacije u R1CS; $*$ predstavlja množenje matrica, a $\bullet$ Hadamardov proizvod.

$$(U * a) \bullet (V * a) = W * a \tag{1}$$

Naredni korak redukcije je primena Langražove interpolacije na kolone matrica U, V i W prikazane kroz jednačinu 1 čime se polazni skup uslova pretvori u jednakost vektora u tačkama interpolacije $1, 2, \ldots, n$. Novodobijena jednakost naziva se QAP [19] i prikazana je kroz formulu 2, gde je a vektor tajnih informacija, $u_1(x) \ldots u_m(x), v_1(x) \ldots v_m(x), w_1(x) \ldots w_m(x)$ su polinomi stepena n koji modeluju originalni skup ograničenja, $z(x)$ je najjednostavniji polinom koji ima korene u tačkama $1 \ldots n$, odnosno $(x-1) * (x-2) * \cdots * (x-n)$, a $h(x)$ je polinom koji Dokazivač mora da izračuna na osnovu svih ostalih podataka u jednačini i koji će moći da se odredi samo kada leva strana jednačine ima nule u tačkama $1, 2, \ldots, n$, odnosno samo kada je R1CS ispunjen.

$$\sum_{i=1}^{m} a_i u_i(x) \sum_{i=1}^{m} a_i v_i(x) - \sum_{i=1}^{m} a_i w_i(x) = h(x) z(x) \tag{2}$$

Nakon što je originalni skup ograničenja transformiran u jednakost polinoma oblika datog u jednačini 2, potrebno je proveriti validnost ove jednakosti bez da Dokazivač javno objavi vektor tajnih informacija a. Da bi se ovo postiglo na vektor a se primenjuje standardna enrkipcija pomoću eliptičnih krivih (eng. *Elliptic Curve Cryptography*), kao i funkcija bilinearnog uparivanja (eng *Bilinear Pairing*) [20]. Enkripcija korišćenjem eliptičnih krivih sastoji se iz toga da se broj p pomnoži javno dostupnom tačkom G koja pripada eliptičnoj krivi i tako dobije enkriptovana tačka $P = p * G$, koja pripada istoj toj krivoj. Pokazano je da kompleksnost inverzne operacije kod diskretnih eliptičnih krivih može da se svede na problem pronalaženja diskretnog logaritma (eng. *Dicrete logarithm*), na čemu počiva asimetrična kriptografija. Funkcija bilinearnog uparivanja e omogućava da za 2 tačke sa različitih eliptičnih krivih G_1 i G_2 i proizvoljno izabrani broj p važi jednakost $e(p * G_1, G_2) = e(G_1, p * G_2)$.

Na osnovu kriptografije eliptičnih krivih i bilinearnog uparivanja moguće je proveriti jednakost polinoma prikazanog u jednačini 2, ali ne na efikasan način. Na osnovu Švarc-Zipel leme [21] dovoljno je proveriti jednakost polinoma u jednoj nasumično izabranoj tački jer bi jednakost u toj tački implicirala ili jednakost polinoma u svim tačkama ili da je nasumično odabrana tačka jedna od najviše n tačaka u kojima se polinomi seku. Kako je stepen polinoma drastično manji od mogućih vrednosti za evaluaciju polinoma, verovatnoća da je druga situacija ispunjena je zanemarljiva i jednakost polinoma u jednoj tački evaluacije implicira jednakost polinoma generalno, što znatno olakšava proveru ispunjena jednačine 2.

Korišćenjem enkripcije pomoću eliptičnih krivih, funkcije bilinearnog uparivanja i činjenice da je za proveru jednakosti

polinoma dovoljna provera jednakosti u jednoj tački, dolazi se do osnovne varijante protokola za proveru ispunjenosti jednačine 2. Algoritam verifikacije se sastoji iz 3 koraka:

1) Verifikator generiše jednu nasumičnu tačku τ u kojoj se vrši evaluacija polinoma i pošalje je Dokazivaču;
2) Dokazivač izračuna i pošalje Verifikatoru vrednosti $A_1 = \sum_{i=1}^{m} a_i u_i(\tau) * G_1$, $B_2 = \sum_{i=1}^{m} a_i v_i(\tau) * G_2$ i $C_1 = (\sum_{i=1}^{m} a_i w_i(\tau) + h(\tau)z(\tau)) * G_1$;
3) Verifikator proveri da li važi da je $e(A_1, B_2) = e(C_1, G_2)$.

Prethodno opisani algoritam predstavlja suštinu rada ZK-SNARK protokola. Kako će treći korak uvek biti ispunjen za ispravno izračunate vrednosti A_1, B_2 i C_1, ovaj algoritam ima osobinu Kompletnosti. Pored toga, sam dokaz se sastoji iz 3 tračke na eliptičnim krivima, pa ovaj dokaz ima osobinu Kompaktnosti. Da bi algoritam pripadao kategoriji ZKP potrebno je još da ima osobine Pouzdanost i Nulto znanje. Dodatno, da bi algoritam bio primenljiv u BČ poželjno je da poseduje i svojstvo Neinteraktivnosti. Nastavak ovog poglavlja podeljen je u 3 potpoglavlja od kojih svako opisuje kako ispuniti jednu od ovih osobina.

A. Pouzdanost ZK-SNARK protokola

Osnovni problem prethodno prikazanog algoritma je što zlonamerni Dokazivač može da iskoristi 2 nasumična broja a i b i na osnovu njih izračuna broj c kao $c = a * b$. Koristeći brojeve a, b i c zlonamerni Dokazivač može da izračuna tačke na eliptičnim krivima $A_1 = a * G_1$, $B_2 = b * G_2$ i $C_1 = c * G_1$ koje pošalje Verifikatoru. Kako ovako generisane tačke zadovoljavaju jednakost koju Verifikator proverava i kako Verifikator nema načina da proveri da li tačke dobijene rešavanjem jednakosti polinoma koja se dokazuje, ovaj algoritam ne zadovoljava osobinu Pouzdanosti.

Kako bi se omogućilo da Verifikator proveri da li je Dokazivač dostavio tačke A_1, B_2, C_1 i H_1 koje odgovaraju delovima jednačine 2, neophodno je transformisati datu jednačinu tako da ona bude proširena parametrima α i β. Nakon što se jednačina proširi, Dokazivač računa tačke A_1, B_2 i C_1 na način koji je prikazan kroz jednačine 3.

$$\begin{aligned} A_1 =& \alpha * G_1 + \sum_{i=1}^{m} a_i u_i(\tau) * G_1 \\ B_2 =& \beta * G_2 + \sum_{i=1}^{m} a_i v_i(\tau) * G_2 \\ C_1 =& (\sum_{i=1}^{m} a_i * (\alpha * v_i(\tau) + \beta * u_i(\tau) + w_i(\tau)) \\ & + h(\tau)z(\tau)) * G_1 \end{aligned} \tag{3}$$

Nakon što Dokazivač izračuna tačke na eliptičnim krivima na način prikazan kroz formulu 3, Verifikator može da proveri validnost dokaza tako što evaluira jednakost prikazanu kroz jednačinu 4.

$$e(A_1, B_2) = e(\alpha * G_1, \beta * G_2) * e(C_1, G_2) \tag{4}$$

Ključna činjenica na kojoj se zasniva mogućnost da Validator potvrdi da je Dokazivač stvarno izračunao tačke A_1, B_2 i C_1 na ispravan način zasniva se na tome da Dokazivaču nije neophodno da zna vrednosti α i β kako bi izračunao odgovarajuće tačke. Umesto toga, Dokazivaču je dovoljno proslediti niz enkriptovanih vrednosti $\Psi_i = (\alpha * v_i(\tau) + \beta * u_i(\tau) + w_i(\tau)) * G_1$ za računanje takče C, kao i vrednosti $\alpha_1 = \alpha * G_1$ i $\beta_2 = \beta * G_2$ za računanje tačaka A i B.

Kako dokazivač nema informacije o tačnim vrednostima za α i β i kako mu je za računanje tačke C_1 dostupan niz tačaka na eliptičnoj krivoj gde za svaku važi da je α u nekoj vezi sa polinomom v, a β u nekoj vezi sa polinomom u, smatra se da je jedini način da Dokazivač generiše tačke koje zadovoljavaju formulu 4 onaj koji je prikazan u jednačinama 3. Ovo verovanje se dodatno zasniva na bilinearnoj Difi-Helman pretpostavci (eng. *Bilinear Diffie-Hellman Assumption*) koja tvrdi da nije moguće za nasumično odabranu vrednost c izračunati $x = e(\alpha_1, \beta_2) + e(c * G_1, G_2)$, a zatim pronaći tačke A_1 i B_2 za koje važi da je $e(A_1, B_2) = x$.

Drugi problem koji narušava Pouzdanost u trenutnoj postavci algoritma dešava se zbog toga što se Dokazivač ne obavezuje na neko rešenje a koje zadovoljava jednačinu 2 pre nego što Verifikator pošalje tačku evaluacije. Zbog toga Dokazivač može da konstruiše vektor a koji zadovoljava polaznu jednačinu za konkretnu vrednost τ koju je poslao Verifikator, a koji nije validno rešenje inače.

Da bi se sprečilo navedeno zlonamerno ponašanje, tačka evaluacije τ treba da bude sakrivena od Dokazivača, kao što su sakriveni i parametri α i β. Da bi Dokazivač mogao da izračuna odgovarajuće tačke dokaza, potrebno mu je dostaviti vektor $G_1, \tau * G_1, \ldots, \tau^{n-1} * G_1$ za računanje tačke A_1, $G_2, \tau * G_2, \ldots, \tau^{n-1} * G_2$ za računanje tačke B_2 i $t(\tau) * G_1, \tau * t(\tau) * G_1, \ldots, \tau^{n-1} * t(\tau) * G_1$ za računanje tačke C_1.

B. Nulto znanje ZK-SNARK protokola

Iako u opisanoj verziji ZK-SNARK algoritma, nije moguće na osnovu dokaza rekonstruisati vektor tajnih informacija koji je Dokazivač koristio, Verifikator može da za neki konkretni vektor tajnih informacija proveri da li je na osnovu tog vektora generisan dokaz. Ovo je posebno problematično u primenama gde postoji mali broj tajnih informacija koje imaju poznati skup mogućnosti, kao što je tajno glasanje manjeg broja predodređenih ljudi i zbog toga se ne može smatrati da prethodno opisani algoritmom poseduje osobinu Nultog znanja.

Da bi protokol imao osobinu Nultog znanja, Dokazivač prilikom objavljivanja dokaza generiše 2 nasumična parametra p i q i menja formule za generisanje tačaka A_1, B_2 i C_1 prikazane kroz jednačine 3 tako što će tački A_1 dodatno dodati vrednost $p * G_1$, tački B_2 dodatno dodati vrednost $q * G_2$, a tački C_1 dodatno dodati vrednost $A_1 * q + B_2 * p - p * q$. Verifikator i dalje može da proveri validnost dokaza korišćenjem jednačine 4, ali da bi neko proverio da li je neka vrednost korišćena za dokaz mora u isto vreme da pogodi vektor a i tajne parametre p i q.

C. Neinteraktivni ZK-SNARK protokol

Trenutna postavka ZK-SNARK algoritma se sastoji iz 3 koraka: 1) Verifikator izabere 3 nasumična broja τ, α i β i generiše odgovarajuće tačke koje je neophodno proslediti Dokazivaču; 2) Dokazivač izračuna potrebne tačke na eliptičnim krivama i pošalje ih Verifikatoru; 3) Verifikator proveri da li je zadovoljena odgovarajuća jednakost. Uočava se da je korak 1 mnogo računarski zahtevniji za Verifikatora od koraka 3, kao i da je korak 1 dovoljno odraditi jednom, nakon čega je moguće više puta raditi dokaz i verifikaciju odgovarajuće formule. Zbog ovoga se u protokol uvodi treći entitet koji se naziva Verodostojna postavka (eng. *Trusted setup*), koji služi da za zadatu jednačinu oblika prikazanog kroz jednačinu 2 generiše sve potrebne parametre za kasniju komunikaciju Dokazivača i Verifikatora.

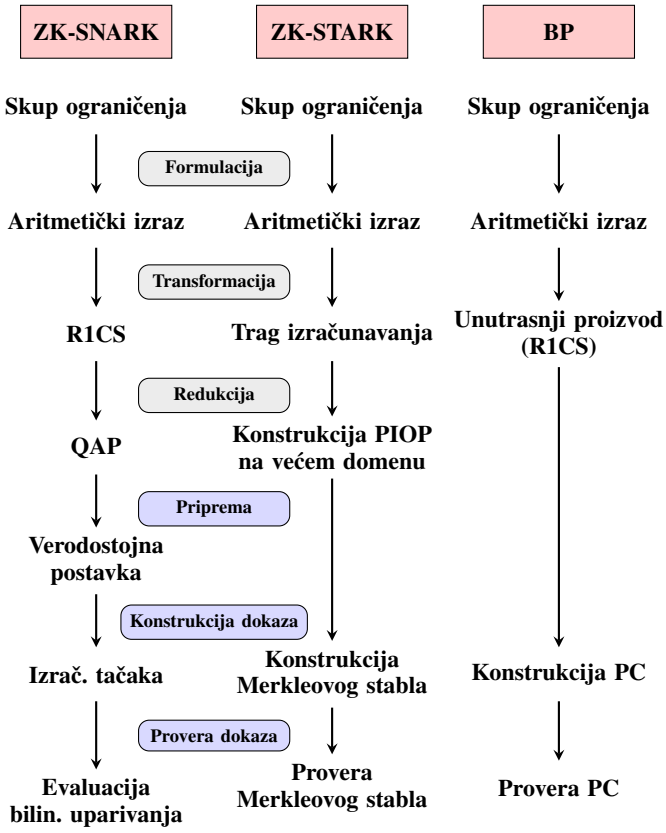
Prednost postojanja Verodostojne postavke je da nakon inicijalnog generisanja skrivenih parametara τ, α i β, i objavljivanja enkriptovanih vrednosti neophodnih za rad algoritama, ZK-SNARK algoritam ima malu veličinu dokaza i jednostavnu verifikaciju. Dodatno, dokazivanje konkretnih činjenica se dalje sastoji samo iz koraka 2 i 3 što čini ovaj protokol neinteraktivnim i time koristim za upotrebu u BČ.

Mana Verodostojne postavke je što se mora verovati da su nakon objavljivanja svih javnih parametara potrebnih za rad ZK-SNARK algoritma trajno obrisani skriveni parametri τ, α i β. Ovo je dodatna pretpostavka koja je neophodna za rad ZK-SNARK algoritma i narušavanje ove pretpostavke bi poništilo sve prethodno generisane dokaze. Zbog toga postoje veliki napori da se smanji šansa da skriveni parametri ostanu sačuvani, tako što u generisanju skrivenih parametara Verodostojne postavke učestvuje više entiteta i parametri ostaju sakriveni sve dok barem 1 entitet ne otkrije svoje parametre [22].

Kombinacijom svih tehnika opisanih u ovom poglavlju dobija se funkcionalni ZK-SNARK protokol. U praksi se mora voditi računa o nekim dodatnim stvarima kao što su javni ulazni parametri dokaza [23], a Verifikator mora proveriti još nekoliko jednakosti [24].

IV. Poređenje ZK-SNARK, ZK-STARK i BP

Na slici 1 prikazani su osnovni koraci rada ZK-SNARK, ZK-STARK i BP algoritama. Može se uočiti da ZK-SNARK i ZK-STARK algoritmi imaju sličnu strukturu, pri čemu ZK-STARK algoritam nema korak pripreme u kome Verodostojna postavka generiše potrebne parametre za konstrukciju dokaza na osnovu QAP formule. Zbog ove sličnosti se u literaturi pominje da je ZK-STARK algoritam zapravo vrsta ZK-SNARK algoritma koji eliminiše komponentu Verodostojne postavke [13]. Mi tvrdimo da ovo nije adekvatno poređenje i da je ZK-STARK dokaz suštinski različit od ZK-SNARK dokaza, jer se zasniva na drugim matematičkim principima, kriptografskim pretpostavkama i podložan je drugoj vrsti napada i zbog toga se ne može smatrati podskupom ZK-SNARK protokola. BP protokol takođe nema korak pripreme (što dodatno naglašava trud da se eliminiše potreba za Verodostojnom postavkom), a ima i pojednostavljenu redukciju u čemu se suštinski razlikuje od druga 2 algoritma [25].



Slika 1. Pregeld rada ZK-SNARK, ZK-STARK i BP protokola

Pre početka konstrukcije dokaza u svim navedenim algoritma se originalni skup ograničenja najpre svede na odgovarajuću jednačinu polinoma za koju se dokazuje određena relacija. Kod ZK-SNARK i ZK-STARK protokola, ova transformacija je nešto kompleksnija i unosi dodatno računarsko opterećenje za Dokazivača. Sa druge strane kod BP algoritma se originalni skup problema može direktno modelovati kao unutrašnji proizvod (ili se može pretvoriti u R1CS nad kojim se direktno može primeniti unutrašnji proizvod bez potrebe da se redukuje na QAP) [26], nakon čega se pomoću Pedersonove šeme obavezivanja (eng. *Pedersen Commitments - PC*) direktno konstruiše dokaz zasnovan na unutrašnjem proizvodu (eng. *Inner Product Arguments - IPA*).

Rad ZK-STARK algoritma se zasniva na ideji da je moguće uočiti zavisnost između bliskih vrednosti evaluacije (ili konstruisati takvu zavisnost ukoliko ona ne postoji prirodno), a zatim takvu zavisnost evaluirati na mnogo većem domenu od polaznog. Osnovna jednačina koju ZK-STARK algoritam dokazuje naziva se interaktivni proročki dokaz polinoma (eng. *polynomial interactive oracle proof* - PIOP) i modeluje korelaciju u tragu izračunavanja (eng. *execution trace*) između bliskih vrednosti evaluacije.

Nakon što se formira polinom koji modeluje korelaciju (eng. *constraint checking polynomial*) u tragu izračunavanja na manjem domenu, konstruiše se Merkleovo stablo (eng. Merkle tree) nad evaluacijom ovog polinoma na mnogo većem

Tabela I
PREGLED KARAKTERISTIKA ZK-SNARK, ZK-STARK I BP PROTOKOLA

Protokol / Osobina	ZK-SNARK	ZK-STARK	BP
Algoritam	Evaluacija QAP	Evaluacija PIOP na velikom domenu, FRI	IPA, PC
Algoritamska kompleksnost Dokazivač	$O(N * log(N))$	$O(N * poly\text{-}log(N))$	$O(N * log(N))$
Algoritamska kompleksnost Verifikator	$O(1)$[1]	$O(poly\text{-}log(N))$	$O(N)$
Memorijska kompleksnost dokaza	$O(1)$	$O(poly\text{-}log(N))$	$O(log(N))$
Kriptografske pretpostavke	Discrete Logarithm Hardness, Bilinear Diffie-Hellman, Computational Diffie-Hellman, Decisional Diffie-Hellman, Knowledge of Exponent	ireverzibilnost Heš funkcija	Discrete Logarithm Hardness
Druge bezbednosne pretpostavke	Verodostojna postavka	Heš funk. otporne na koliziju Pouzdanost Fiat-Šamir transf.	Pouzdanost Fiat-Šamir transf.
Sigurnost od kvantnih računara	Ne	Da	Ne

[1]Konstantan broj operacija bilinearnog uparivanja

domenu. Evaluacija polinoma na velikom domenu, uz korišćenje činjenice da se različiti polinomi čiji je stepen znatno manji od domena evaluacije razlikuju u skoro svim tačkama, predstavlja specijalan slučaj Rid-Salmon kodova za ispravljanje grešaka (eng. *Reed-Solomon error correction*) [27].

Provera konstruisanog ZK-STARK dokaza se sastoji iz provere evaluacije u jednoj tački i provere Merkleovog stabla od tačke evaluacije do korena stabla. Međutim, napad kojem je podložan ZK-STARK algoritam [28] proilazi iz činjenice da Verifikator nema kontrolu nad stepenom polinoma iz PIOP relacije i nema garanciju da je on drastično manji od domena evaluacije.

Inovativni pristup koji koriste ZK-STARK algoritmi da bi obezbedili da polinomi budu malog stepena zasniva se na dokazu o bliskosti polinoma za Rid-Salmon interaktivni proročki dokaz (eng. *Fast Reed-Solomon Interactive Oracle Proofs of Proximity* - FRI) [29]. Nekoliko radova [30] [31] [32] bave se utvrđivanjem da protokoli koji koriste FRI algoritam zaista imaju osobinu Pouzdanosti, sa zadovoljavajućom verovatnoćom.

Korišćenjem Fiat-Šamir heuristike (eng. *Fiat-Shamir heuristic*) ZK-STARK i BP protokole je moguće učiniti neinteraktivnim. Ova optimizacija se zasniva na pretpostavci da za postizanje osobine Pouzdanosti protokola nije neophodno da Verifikator stvarno dostavi potpuno nasumične vrednosti entropije, već je dovoljno da ove vrednosti budu nepredvidive u trenutku kada Dokazivač konstruiše dokaz.

U tabeli I prikazane su osnovne karakteristike ZK-SNARK, ZK-STARK i BP protokola. Iz tabele se jasno vidi da se ova 3 protokola suštinski razlikuju u radu algoritma, kriptografskim i drugim bezbednosnim pretpostavkama, ali i u veličini dokaza i tome koliko su računarski zahtevni za izvršavanje.

Na osnovu pregleda tabele, moguće je zaključiti da su ZK-SNARK algoritmi bolji i od ZK-STARK i od BP algoritama i po računarskom opterećenju Dokazivača i Verifikatora i po komunikacionom opterećenju mreže, odnosno veličini dokaza. Međutim, ovo nije slučaj zbog činjenice da iako u ZK-SNARK protokolu Verifikator treba da izvrši konstantan broj operacija, svaka od ovih operacija ja dosta računarski zahtevna jer se radi o operacijama bilinearnog uparivanja. Ove operacije su toliko računarski zahtevne da njihovo izvršenje u pametnim ugovorima na Eterijum (eng. *Ethereum*) BČ nije bilo moguće sve dok na sistemskom nivou nije implementirana funkcija koja omogućava njihov rad [33].

Zbog računarski zahtevnih operacija bilinearnog uparivanja, ZK-STARK algoritam su u praksi efikasniji za izvršavanje od ZK-SNARK algoritama. Efikasnost BP algoritama za Verifikatora je specifična jer mogućnost za svođenja originalnog skupa ograničenja direktno na IPA problem čini BP algoritme jako pogodnim za korišćenje kod problema gde je ova redukcija prirodna, kao što su dokazi opsega (eng. *Range Proof*). Međutim, u opštem slučaju, BP protokol ima lošije performanse i u odnosu na ZK-SNARK i u odnosu na ZK-STARK algoritam.

Pored računarskog opterećenja Verifikatora, jedna od najbitnijih karakteristika ZKP algoritama tiče se bezbednosti. Iako svi navedeni algoritmi imaju i kriptografske i druge bezbednosne pretpostavke, poznato je da su kriptografske pretpostavke na kojima se zasniva rad ZK-SNARK i BP algoritama ranjive na napade kvantnih računara. Zbog toga se smatra da je rad ZK-STARK protokola dugoročno bezbedniji i da budući razvoj ZKP tehnologije treba da se zasniva na ovim i njima sličnim algoritmima.

Poslednja jako bitna osobina navedenih algoritama tiče se njihove Kompaktnosti. Iako ZK-STARK protokol bolje koristi računarske resurse Verifikatora i dugoročno pruža bolju bezbednost, veličina dokaza ovog algoritma je previše velika za realnu upotrebu u BČ u 2025 [12]. Da je veličina dokaza glavni problem za primenu ZK-STARK algoritama pokazuje

i najnovija lista istraživačkih tema Eterijum fondacije (eng. *Ethereum Foundation*) [34].

V. Zaključak

U radu su prikazani osnovni koncepti rada ZKP na primeru ZK-SNARK algoritma, kao i način na koji se postižu neophodne i poželjne osobine ovih algoritama. Rad sadrži detaljnu uporednu analizu 3 velike grupe ZKP protokola, prednosti mane svake, kao i detaljnu analizu sigurnosti navedenih algoritama.

U radu su analizirane osnovne implementacije rada ZK-SNARK, ZK-STARK i BP algoritama, a buduće istraživanje bi trebalo da obuhvati više implementacija svakog od algoritma, uključujući najnovije teorijske pomake [35] i algoritme čiji se principi rada ne mapiraju direktno na neku od prethodno opisanih kategorija [36], kao i evaluaciju svake od ovih implementacija. Dodatno, potrebno je analizirati efikasnost različitih algoritama u različitim scenarijima upotrebe.

Pravci daljeg poboljšanja same ZKP tehnologije se pre svega sastoje iz poboljšanja performansi i otklanjanju problema postojećih algoritama, ali i dizajnu novih rešenja. Pored ovog, pregledi primene ZKP [12] pokazuju da na ovom polju ima prostora za doprinos, pogotovu za primenu ZK-STARK algoritama.

Literatura

[1] O. Goldreich and Y. Oren, "Definitions and properties of zero-knowledge proof systems," *Journal of Cryptology*, vol. 7, no. 1, pp. 1–32, 1994.

[2] J. Werth, M. H. Berenjestanaki, H. R. Barzegar, N. El Ioini, and C. Pahl, "A review of blockchain platforms based on the scalability, security and decentralization trilemma.," *ICEIS (1)*, pp. 146–155, 2023.

[3] V. Buterin, "The data availability problem." Ethereum Meetup, Silicon Valley, 9 2017. Available at https://www.youtube.com/watch?v=OJT_fR7wexw.

[4] "Ethereum roadmap." [Online; accessed 2025-03-03].

[5] V. Buterin, "An incomplete guide to rollups," 1 2021. [Online; accessed 2025-03-03].

[6] J. Agbo, "Optimistic vs. zero knowledge rollups: Which layer 2 is better?," 4 2024. [Online; accessed 2025-03-03].

[7] M. Nadler and F. Schär, "Tornado cash and blockchain privacy: a primer for economists and policymakers," *Federal Reserve Bank of St. Louis Review*, 2023.

[8] M. Möser, K. Soska, E. Heilman, K. Lee, H. Heffan, S. Srivastava, K. Hogan, J. Hennessey, A. Miller, A. Narayanan, *et al.*, "An empirical analysis of traceability in the monero blockchain," *arXiv preprint arXiv:1704.04299*, 2017.

[9] X. Yang and W. Li, "A zero-knowledge-proof-based digital identity management scheme in blockchain," *Computers & Security*, vol. 99, p. 102050, 2020.

[10] J. Groth, "On the size of pairing-based non-interactive arguments," in *Advances in Cryptology–EUROCRYPT 2016: 35th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Vienna, Austria, May 8-12, 2016, Proceedings, Part II 35*, pp. 305–326, Springer, 2016.

[11] E. Ben-Sasson, I. Bentov, Y. Horesh, and M. Riabzev, "Scalable, transparent, and post-quantum secure computational integrity," *Cryptology ePrint Archive*, 2018.

[12] B. Oude Roelink, M. El-Hajj, and D. Sarmah, "Systematic review: Comparing zk-snark, zk-stark, and bulletproof protocols for privacy-preserving authentication," *Security and Privacy*, vol. 7, no. 5, p. e401, 2024.

[13] matter labs, "Github - matter-labs/awesome-zero-knowledge-proofs: A curated list of awesome things related to learning zero-knowledge proofs (zkp).." [Online; accessed 2025-03-04].

[14] O. Goldreich, *P, NP, and NP-Completeness: The basics of computational complexity*. Cambridge University Press, 2010.

[15] V. Buterin, "Quadratic arithmetic programs: from zero to hero," 12 2016. [Online; accessed 2025-03-05].

[16] R. Gennaro, C. Gentry, B. Parno, and M. Raykova, "Quadratic span programs and succinct nizks without pcps," in *Advances in Cryptology–EUROCRYPT 2013: 32nd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Athens, Greece, May 26-30, 2013. Proceedings 32*, pp. 626–645, Springer, 2013.

[17] B. Parno, J. Howell, C. Gentry, and M. Raykova, "Pinocchio: Nearly practical verifiable computation," *Communications of the ACM*, vol. 59, no. 2, pp. 103–112, 2016.

[18] R. Skills, "Converting algebraic circuits to r1cs (rank one constraint system)," 7 2023. [Online; accessed 2025-03-05].

[19] R. Skills, "Converting algebraic circuits to r1cs (rank one constraint system)," 8 2023. [Online; accessed 2025-03-05].

[20] V. Buterin, "Exploring elliptic curve pairings," 1 2017. [Online; accessed 2025-03-05].

[21] R. Skills, "The schwartz-zippel lemma and its application to zero knowledge proofs," 8 2024. [Online; accessed 2025-03-05].

[22] R. Skills, "Trusted setup," 8 2023. [Online; accessed 2025-03-05].

[23] R. Skills, "Groth16 explained," 8 2023. [Online; accessed 2025-03-05].

[24] C. Reitwiessner, "zksnarks in a nutshell," *Ethereum blog*, vol. 6, pp. 1–15, 2016.

[25] B. Bünz, J. Bootle, D. Boneh, A. Poelstra, P. Wuille, and G. Maxwell, "Bulletproofs: Short proofs for confidential transactions and more," in *2018 IEEE symposium on security and privacy (SP)*, pp. 315–334, IEEE, 2018.

[26] R. Skills, "Bulletproofs explained," 10 2024. [Online; accessed 2025-03-5].

[27] S. B. Wicker and V. K. Bhargava, *Reed-Solomon codes and their applications*. John Wiley & Sons, 1999.

[28] V. Buterin, "Starks, part i: Proofs with polynomials," 11 2017. [Online; accessed 2025-03-07].

[29] StarkWare, "Low degree testing. the secret sauce of succinctness | by starkware | starkware | medium," 3 2019. [Online; accessed 2025-03-07].

[30] E. Ben-Sasson, I. Bentov, Y. Horesh, and M. Riabzev, "Fast reed-solomon interactive oracle proofs of proximity," in *45th international colloquium on automata, languages, and programming (icalp 2018)*, pp. 14–1, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018.

[31] E. Ben-Sasson, L. Goldberg, S. Kopparty, and S. Saraf, "Deep-fri: sampling outside the box improves soundness," *arXiv preprint arXiv:1903.12243*, 2019.

[32] E. Ben-Sasson, D. Carmon, Y. Ishai, S. Kopparty, and S. Saraf, "Proximity gaps for reed–solomon codes," *Journal of the ACM*, vol. 70, no. 5, pp. 1–57, 2023.

[33] A. Network, "Ethereum support for zk-snarks. the zk-snarks series continues with… | by aventus network | coinmonks | medium," 2 2019. [Online; accessed 2025-03-07].

[34] E. Foundation, "Academic grants round 2025 wishlist." [Online; accessed 2025-03-07].

[35] H. Lipmaa, "Polymath: Groth16 is not the limit," in *Annual International Cryptology Conference*, pp. 170–206, Springer, 2024.

[36] A. Gabizon, Z. J. Williamson, and O. Ciobotaru, "Plonk: Permutations over lagrange-bases for oecumenical noninteractive arguments of knowledge," *Cryptology ePrint Archive*, 2019.

Fiksni bežični pristup (FWA) – tehnologije, primena i perspektive

Dejan Nemec
Fakultet tehničkih nauka, Univerzitet u Novom Sadu
Novi Sad, Srbija
denem@uns.ac.rs

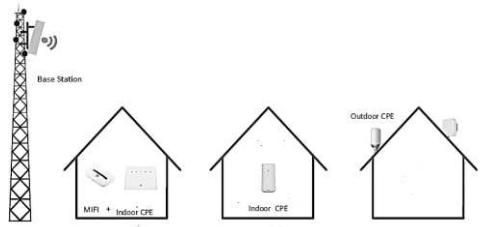
***Apstrakt* – Fiksni bežični pristup (FWA – *Fixed Wireless Access*) predstavlja tehnologiju koja omogućava bežično povezivanje korisnika na širokopojasne mreže, čime se eliminiše potreba za fizičkim kablovima. Ovaj rad istražuje razvoj, tehnološke aspekte, primenu i perspektivu FWA tehnologije, sa posebnim osvrtom na njenu ulogu u pružanju širokopojasnog Interneta u ruralnim i urbanim područjima. Analiziraju se različite tehnologije koje se koriste u FWA, uključujući WiMAX, satelitski pristup, 4G LTE i 5G, kao i njihove prednosti i izazovi. Takođe, rad razmatra ključne tehnološke aspekte FWA i perspektivu FWA u kontekstu globalne digitalne transformacije.**

Ključne reči – FWA, WLL, 4G LTE, 5G.

I. UVOD

Za fiksni bežični pristup FWA (*Fixed Wireless Access*) koriste još i termini bežična lokalna petlja WLL (*Wireless Local Loop*), fiksni radio pristup FRA (*Fixed Radio Access*), radio u pretplatničkoj petlji RITL (*Radio In The Loop*) i širokopojasni bežični pristup BWA (*Broadband Wireless Access*). FWA se odnosi na rešenja koja koriste radio talase za povezivanje krajnjeg korisnika na mrežu telekom operatora. Ta rešenja karakteristična su za ruralna područja gde žični pristup nije pogodan, ali i u urbanim sredinama, naročito u novije vreme gde se koristi za neka IoT (*Internet of Things*) rešenja ili kao *backup* pristup Internetu.

Uobičajeno je da se korisnički uređaj CPE (*Customer Premise Equipment*) radio signalom poveže za fiksnom pristupnom antenom bazne stanice koja je dalje žičnim putem (najčešće optikom) povezana na ostali deo mreže (Sl. 1). Bazna stanica sa svojim antenama omogućava povezivanje većeg broja korisnika, a to zavisi od tehnologije koja se u FWA primenjuje [1]-[7].



Slika 1. Pozicije uređaja u fiksnom bežičnom pristupu [2]

FWA tehnologija se širi globalno, posebno u regijama gde je infrastruktura ograničena, kao što su delovi Afrike, Azije i Južne Amerike. Operatori sve više investiraju u FWA kao alternativu ili dopunu tradicionalnim širokopojasnim uslugama.

II. ISTORIJSKI RAZVOJ FWA TEHNOLOGIJE

FWA tehnologija je prošla dug put od svojih skromnih početaka u mikrotalasnoj komunikaciji do modernih 5G mreža koje omogućavaju brzine i performanse koje su prethodno bile nezamislive. S nastavkom razvoja 5G i budućih tehnologija, FWA će verovatno ostati ključna komponenta u globalnoj širokopojasnoj infrastrukturi, posebno u područjima gde je izgradnja fiksnih mreža ekonomski ili tehnički izazovna.

FWA tehnologija ima svoje korene u ranim danima bežične komunikacije, a njen razvoj prati napredak u telekomunikacijama i bežičnim tehnologijama. Postoji nekoliko ključnih faza u istorijskom razvoju FWA tehnologije [8]-[11]:

- Rani bežični sistemi (1940-1970) – Prvi bežični sistemi bili su namenjeni za vojne i vladine potrebe, posebno za komunikaciju na daljinu. U ovom periodu razvijeni su prvi mikrotalasni linkovi, koji su omogućavali prenos podataka na velike udaljenosti. Ovi sistemi bili su preteča modernih FWA tehnologija.
- Pojava mobilnih mreža (1980-1990) – Sa razvojem mobilnih mreža (1G i 2G), fokus je bio na mobilnoj telefoniji, ali su se pojavile i prve primene fiksne bežične komunikacije. FWA tehnologija počela se koristiti za pružanje telefonskih usluga u ruralnim područjima gdje je postavljanje fiksnih linija bilo skupo ili tehnički izazovno. U ovom periodu, bežična tehnologija bila je ograničena na uske frekvencijske opsege i niske brzine prenosa.
- Prva generacija FWA (1990-e) – Tokom 1990-ih, FWA je počela da se koristi za pružanje fiksnih bežičnih usluga, posebno u ruralnim područjima gde postavljanje optičkih kablova nije bilo isplativo. Rane FWA mreže koristile su tehnologije kao što su LMDS (*Local Multipoint Distribution Service*) i MMDS (*Multichannel Multipoint Distribution Service*). Ove mreže su omogućavale brzine od nekoliko Mbit/s, što je u to vreme bilo dovoljno za osnovne usluge poput pristupa Internetu i telefonskih veza.
- Širokopojasni pristup (2000-2010) – S pojavom 3G mreža i povećanjem potražnje za širokopojasnim internetom, FWA tehnologija je dobila novi zamah. FWA je postala popularna opcija za pristup Internetu u područjima gde DSL (*Digital Subscriber Line*) ili kablovski distributivni sistemi nisu bili dostupni. Za pružanje bežičnog pristupa koristile su se frekvencije mikrotalasnog spektra (npr. 2,4 GHz i 5 GHz). Sa razvojem WiMAX (*Worldwide Interoperability for Microwave Access*) tehnologije sredinom 2000-ih, FWA je doživela značajan napredak. WiMAX je omogućio veće brzine prenosa (do 70 Mbit/s) i bolju pokrivenost, što ga je činilo atraktivnim za pružanje širokopojasnih usluga u ruralnim i urbanim područjima.
- 4G LTE (*Long Term Evolution*) i povećana brzina (2010-2018) – Uvođenje 4G LTE tehnologije

omogućilo je značajno povećanje brzina i pouzdanosti bežičnih mreža. FWA usluge postale su konkurentne fiksnim širokopojasnim uslugama, posebno u područjima s ograničenom infrastrukturom. Operatori su počeli koristiti LTE mreže za pružanje FWA usluga, što je omogućilo brzine od nekoliko desetina Mbit/s.

- 5G i budućnost FWA (2018-danas) – Uvođenje 5G tehnologije donelo je revolucionarne promene u FWA sektoru. 5G nudi znatno veće brzine (do nekoliko Gbit/s), niže kašnjenje i veći kapacitet. FWA postaje ključna komponenta u strategijama operatora za pružanje širokopojasnog pristupa u urbanim, prigradskim i ruralnim područjima. 5G FWA omogućava korisnicima brzine koje su konkurentne ili čak bolje od tradicionalnih fiksnih usluga, što je posebno važno u područjima gdje je optička infrastruktura nedostupna. Očekuje se da će FWA tehnologija nastaviti da evoluira sa daljim razvojem 5G i uvođenjem 6G mreža. FWA će verovatno postati ključna komponenta u pružanju širokopojasnih usluga u ruralnim područjima i u kontekstu IoT aplikacija.

III. KLJUČNI TEHNOLOŠKI ASPEKTI FWA

FWA je tehnologija koja pruža pouzdane Internet usluge koristeći bežičnu komunikaciju između fiksne antene na stubu provajdera usluga i fiksne krajnje tačke rezidencijalnog ili poslovnog korisnika. Naziv „fiksni" poseduje jer se bežična mobilna mreža povezuje sa fiksnom mrežom na određenoj lokaciji. Veza je bežična nekoliko stotina metara pre nego što se ponovo pretvori u žičnu odnosno optičku vezu. Dakle, za razliku od mobilnih bežičnih usluga, FWA je namenjen za stacionarnu upotrebu, pružajući pouzdanu alternativu tradicionalnim žičnim širokopojasnim vezama.

FWA tehnologija koristi napredne tehnike poput MIMO (*Multiple Input Multiple Output*), *beamforming*-a i visokofrekventnih spektara (mmWave) za poboljšanje performansi. Razvoj pametnih antena i sofisticiranih algoritama za upravljanje mrežom omogućava efikasniju upotrebu spektra i bolju pokrivenost.

Ključni tehnološki aspekti FWA jesu [12], [13]:

- Radio frekvencije i spektar – FWA koristi različite radio frekvencije, licencirane i nelicencirane opsege (2,4 GHz, 5 GHz, mmWave), uključujući LTE, 5G, Wi-Fi i frekvencijske spektre drugih bežičnih sistema. Upotrebom naprednih tehnologija kao što su MIMO i *beamforming* poboljšava se pokrivanje i kapacitet mreže.
- Arhitektura mreže – FWA sistemi se oslanjaju na centralne bazne stanice koje šalju i primaju signale do i od korisničkih uređaja. Ova arhitektura omogućava efikasno upravljanje mrežom i podršku velikom broju korisnika. Snaga signala koju može da emituje bazna stanica određuje oblast pokrivenosti. Veća snaga prenosa može pokriti veće površine, ali može biti podložna regulatornim ograničenjima. Korisnički uređaj, CPE, karakteriše usmerena ili jagi antena koja fokusira signal prema anteni bazne stanice povećavajući snagu i kvalitet signala.
- Tehnologije prenosa – Danas se najčešće korišćene tehnologije jesu 4G LTE i 5G koje nude visoke brzine prenosa podataka i malo kašnjenje, što je ključno za podršku aplikacija koje zahtevaju visok kvalitet usluge, kao što su video striming i *online* igre.
- Sigurnost i kvalitet usluga (QoS – *Quality of Service*) – FWA sistemi moraju osigurati visoki nivo sigurnosti i kvalitet usluge kako bi podržali kritične aplikacije i usluge.
- Integracija sa drugim tehnologijama – FWA se često koristi u kombinaciji sa drugim tehnologijama, kao što su Wi-Fi i optička infrastruktura, kako bi se poboljšala pokrivenost i kvalitet usluge.
- Inovativne primene – Upotreba *TV White Space* tehnologije (TVWS) omogućava efikasan pristup u ruralnim područjima, gdje postoje ograničenja u korišćenju tradicionalnih frekvencija.
- Fleksibilnost i adaptacija – FWA sistemi moraju biti fleksibilni i sposobni prilagoditi se različitim zahtevima korisnika i okolini, što uključuje podršku za različite vrste aplikacija i usluga.

IV. FWA TEHNOLOGIJE I ARHITEKTURE

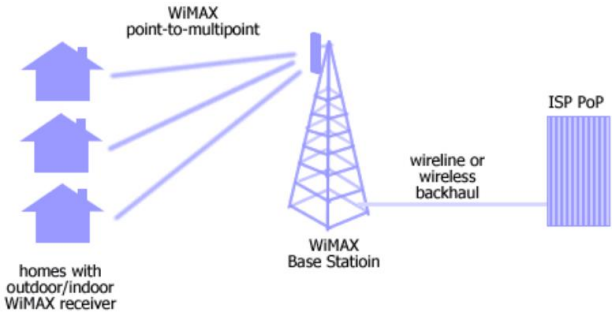
Ključne tehnologije i arhitekture koje omogućavaju FWA jesu:

- WiMAX,
- TVWS (*TeleVision White Space*),
- satelitski pristup,
- 4G LTE,
- 5G.

WiMAX je familija standarda za bežične širokopojasne komunikacije koje su zasnovane na skupu standarda IEEE 802.16, u okviru kojih su date specifikacije fizičkog sloja (PHY – *Physical Layer*) i kontrole pristupa medijima (MAC – *Medium Access Control*). Originalni IEEE 802.16 standard („Fiksni WiMAX") objavljen je 2001. god. i tada je omogućavao maksimalne brzine 30-40 Mbit/s, dok je standard IEEE 802.16-2004 omogućio brzine i do 75 Mbit/s. Nakon toga su sledila i druga unapređenja.

Konačna specifikacija „Mobilnog WiMAX-a", koja je podrazumevala neki stepen mobilnosti korisničkih uređaja i maksimalne brzine do 1 Gbit/s do fiksnih terminala, usvojena je 2011. god. u standardu IEEE 802.16m-2011. WiMAX Release 2.1, popularno nazvan i WiMAX 2+ obezbeđuje kompatibilnost sa LTE sistemima. Novije verzije standarda su WiMAX Release 2.2 (2014. god.) i WiMAX Release 3 (2021. god), gde poslednja podržava i kompatibilnost sa 5G.

WiMAX može da radi na frekvencijama 2-66 GHz i podržava različite tipove modulacija. Jedna bazna stanica može da poveže veći broj korisnika, a maksimalan domet između bazne stanice i terminala je 50 km (Sl. 2). U prvim verzijama standarda, za ostvarivanje veze između bazne stanice i terminala poželjno je bilo obezbediti optičku vidljivost [2], [13]-[15].



Slika 2. Osnovna arhitektura WiMAX mreže [15]

Iako nije u potpunosti zaživela na tržištu jer su druge tehnologije postale dostupne, poslednjih 10 i više godina, u cilju ostvarivanja FWA upotreba TVWS je privlačila mnogo pažnje zbog svojih prednosti, te je vredna pomena [16].

Frekvencijski spektar ispod 1 GHz je veoma pogodan za primene koje zahtevaju dobre uslove propagacije. Nakon implementacije sistema digitalne televizije (DVB-T, DVB-T2), značajan deo TV spektra ostao je slobodan u pojedinim zemljama i pojedinim regionima.

Imajući u vidu ekonomski i tehnološki značaj radio-spektra, veliki broj regulatora i komercijalnih organizacija započeo je istraživanje o načinima najboljeg iskorišćenja spektra koji je ostao neiskorišćen nakon uvođenja digitalne televizije. Naime, planiranje digitalne televizije zasnovano je na zonama raspodele i odgovarajućim dodelama, pa iz tog razloga preostaju oblasti koje su nepokrivene televizijskim signalom na određenim frekvencijama. *White Spaces* (WS) su slobodni delovi spektra koji se ne koriste na određenoj geografskoj lokaciji u određenom trenutku (vremenskom intervalu) od strane korisnika koji poseduju licence ili dozvole za korišćenje kanala. TVWS je slobodan deo spektra unutar opsega 470-790 MHz, primarno namenjenog za digitalno terestričko emitovanje televizijskog signala [17].

Razni predlozi, uključujući IEEE 802.11af, IEEE 802.22 i one od *White Spaces Coalition*, zagovarali su korišćenje *White Space*-a za obezbeđivanje bežičnog širokopojasnog pristupa Internetu. Uređaj namenjen za korišćenje ovih dostupnih kanala nazvan je *White Space Device* (WSD). Takvi uređaji projektovani su da detektuju prisustvo postojećih, ali i neiskorišćenih područja spektra koja su bila rezervisana za analognu televiziju, i da ih koriste za pristup Internetu [18], [19].

Kako je TVWS poređen sa Wi-Fi, on ima svoje prednosti i nedostatke. Jedna od prednosti je pokrivanje većeg područja, preko 10 km, a nedostatak predstavlja veličina antene koja je potrebna na nižim frekvencijama [20].

Za FWA mogu se koristiti i satelitske komunikacije. U zavisnosti od visine na kojima se sateliti nalaze, signal može biti oslabljen, a izraženo je i kašnjenje signala. Komunikacija sa satelitima niske i srednje orbite uglavnom daje prihvatljive rezultate.

Jedan od primera kada se satelitske veze koriste za pristup jesu VSAT (*Very-Small-Aperture Terminal*) sistemi. Antene koje se nalaze kod korisnika su prečnika 0,7-1,2 m i odatle je i naziv ovih sistema. Antene i primopredajnik mogu biti stacionarni (u okviru objekta) ili prenosni, odnosno sporo pokretni kada se nalaze na velikim brodovima ili u reportažnim kolima (Sl. 3).

Slika 3. VSAT terminali [18], [22], [23]

U novije vreme popularni su sistemi koji koriste satelite niske orbite, LEO (*Low Earth Orbit*). Poznat je sistem Starlink koji poseduje 7.000, a planira da postavi ukupno 12.000 satelita. Korisnicima omogućuje brzine 25-220 Mbit/s po pristupačnim cenama. U Evropi jednokratno se plaća oprema oko 400 evra, a mesečna naknada je 40-60 evra. Terminali su još manji od VSAT terminala, veličine nekoliko desetina cm (Sl. 4) [24]-[28].

Slika 4. Starlink antena [28]

FWA može se ostvariti i upotrebom 4G LTE tehnologije. 4G FWA koristi prednosti brzih performansi 4G mobilne komunikacione tehnologije. Uvođenje 4G LTE tehnologije omogućilo je značajno povećanje brzina (reda nekoliko desetina Mbit/s) i povećanje pouzdanosti bežičnih mreža. FWA usluge postale su konkurentne fiksnim širokopojasnim uslugama, posebno u područjima s ograničenom infrastrukturom.

4G LTE mreža koja povezuje mobilne terminale obično koristi makroćelije koje istovremeno opslužuju mnogo korisnika. Svaka ćelija obično može da šalje signale na maksimalnu udaljenost npr. 500 m u gustim urbanim sredinama, a u ruralnim područjima domet ide sve do 10-15 km. 4G koristi relativno niske frekvencije koje imaju dosta dobro prostiranje signala, ali poseduju manji deo spektra, pa brzina može biti problematična. 4G može da ostvari brzine do 100 Mbit/s, i to samo pod idealnim uslovima. Radio signali se dele između više korisnika i degradiraju sa razdaljinom, tako da su stvarne održive brzine u proseku daleko niže. Stoga, kada se bira 4G FWA kao fiksni širokopojasni pristup, treba ga proceniti u skladu sa sopstvenim potrebama i stvarnom situacijom. Ipak, 4G FWA je dobar alat za obezbeđivanje pokrivenosti širokopojasnim Internetom u teško dostupnim područjima, posebno ruralnim [29].

4G pristup je brži za postavljanje i lakši za instalaciju od tradicionalnog žičnog širokopojasnog pristupa. Korisnicima je potreban samo 4G uređaj i odgovarajuća SIM (*Subscriber Identity Module*) kartica [5], [28]-[31].

Sada smo ušli u eru 5G. Performanse 5G su daleko bolje od 4G, a propusni opseg je 10-100 puta veći od 4G. To znači da će mogućnosti FWA biti znatno poboljšane, donoseći korisnicima bolje iskustvo.

5G u srednjem opsegu spektra (2-6 GHz) se smatra savršenim za FWA. U odnosu na 4G, 5G ima mogućnost korišćenja većeg kanala za prenos podataka (do 100 MHz) i bolju spektralnu efikasnost. To se pretvara u više podataka i veće brzine. Kao takav, srednji opseg je idealan za FWA, nudeći 10 puta veći kapacitet od 4G. Operatori širom sveta nadograđuju svoje mreže na 5G za mobilne usluge i takođe koriste povećani kapacitet za FWA usluge. Postoji još jedna vrsta 5G koja koristi milimetarske talasni (mmWave) spektar, npr. od 24 GHz do 39 GHz. Visoke frekvencije omogućavaju daleko veći kapacitet i brzine reda Gbit/s. Međutim, signali na ovim frekvencijama su podložniji gubicima usled prepreka od objekata poput drveća i zgrada, pa čak i stakla.

Svakako, 5G FWA će se u budućnosti sve više koristiti, kako u ruralnim, tako u i urbanim područjima, kao i za povezivanje uređaja u IoT konceptu [5], [28]-[31].

V. PREDNOSTI I NEDOSTACI FWA

U odnosu na žične pristupe, FWA ima sledeće prednosti [32]-[35]:

- Brzo uspostavljanje servisa – FWA eliminiše potrebu za opsežnim i dugotrajnim instalacijama kablova, omogućavajući dobavljačima usluga da brže dođu do kupaca, naročito u ruralnim ili nedovoljno opsluženim područjima, gde bi postavljanje optičkih vlakana moglo biti ekonomski neisplativo.
- Jeftinije uspostavljanje servisa – Uspostavljanje fiksne bežične veze generalno podrazumeva niže troškove u poređenju sa polaganjem optičkih kablova na velike udaljenosti. Pored toga, FWA pomaže u smanjenju troškova održavanja povezanih sa održavanjem fizičkih kablova.
- Skalabilnost – FWA mreže su visoko skalabilne i lako mogu da se prilagode rastućoj potražnji za propusnim opsegom. Dodatne bazne stanice mogu se rasporediti kako bi se proširila područja pokrivenosti ili povećao kapacitet mreže, omogućavajući dobavljačima usluga da skaliraju svoje mreže prema zahtevima korisnika.
- Fleksibilnost – FWA nudi fleksibilnost u pogledu implementacije i mrežne arhitekture. Može se implementirati kao samostalno rešenje ili integrisati sa postojećom žičnom infrastrukturom kako bi se proširila pokrivenost ili poboljšala otpornost mreže.
- Pouzdanost – FWA bežične mreže su manje podložne fizičkim poremećajima kao što su prekidi kablova ili oštećenja usled vremenskih uslova, što obezbeđuje veću pouzdanost u poređenju sa tradicionalnim žičnim mrežama. Napredak u FWA tehnologiji poboljšao je stabilnost signala i smanjio smetnje, dodatno povećavajući pouzdanost.
- Pristupačnost – FWA omogućava pristup Internetu u područjima gde tradicionalne opcije širokopojasnog interneta mogu biti nedostupne ili neadekvatne. Korišćenjem bežične komunikacije, FWA proširuje povezanost na udaljene ili nedovoljno usluženе zajednice, premošćujući digitalni jaz i promovišući digitalnu inkluziju.
- Veza boljih karakteristika – FWA zasnovan na 5G mreži nudi znatno veće brzine prenosa podataka u poređenju sa tradicionalnim širokopojasnim vezama kao što su npr. DSL veze.
- Integracija sa IoT – FWA omogućava brzu, isplativu i bezbednu integraciju sa IoT sistemima u udaljenim preduzećima. FWA omogućava bržu komunikaciju sa IoT uređajima, integraciju sa trećim stranama, kontrolu, praćenje i ažuriranja u realnom vremenu.
- Interkomunikacija – FWA može koristiti preduzećima koja koriste VoIP (*Voice over IP*) za internu komunikaciju i u kontakt centirma. Neke od glavnih karakteristika FWA VoIP-a uključuju malo kašnjenje u prenosu govornih paketa, poboljšani audio i visokokvalitetne video konferencije.
- Bezbednost mreže – Baš kao što Wi-Fi omogućava korisnicima da podese zaštitu lozinkom, FWA u skladu sa propisima, korisnicima nudi opcije za obezbeđivanje privatnosti svojih mreža i zaštitu podataka. FWA obezbeđuje bezbedan prenos paketa podataka, autentifikaciju, kontrolu, zaštitu od napada i bezbednost mreže.

FWA nudi prednosti mobilnim operatorima i korisnicima, ali nije bez nedostataka. Provajderi FWA servisa bi trebalo da procene gde FWA zaostaje u odnosu na alternativne pristupe kako bi utvrdili da li je tehnologija vredna ulaganja.

Neki nedostaci FWA uključuju sledeće [35]:

- Pouzdanost – FWA može funkcionisati kao dobra širokopojasna veza, ali bežična veza i dalje može ponekad da ima smetnje i prekide usled vremenskih uslova ili uzrokovane preprekama (zgrade, drveće) na putanji prostiranja signala.
- Manje brzine u odnosu na optički pristup – FWA jeste jeftiniji i lakši za instalaciju od žičnih veza, ali optička vlakna i kablovi nude veće brzine, veću pouzdanost i bolje performanse.
- Ograničen domet – Termin „fiksni" u FWA odnosi se na činjenicu da širokopojasni pristup postoji samo u između definisane bazne stanice i korisničkog uređaja, što znači da FWA ne podržava bežični roming.
- Interferencija – Ukoliko ne postoji usklađenost, FWA je podložan smetnjama signala od bežičnih sistema koji se nalaze u okruženju.
- Ograničena pokrivenost – Iako može dosegnuti područja do kojih kablovi ne mogu, FWA pokrivenost i dalje zavisi od blizine najbliže bazne stanice.
- Promenljiva brzina – U zavisnosti od kvaliteta mreže, FWA brzine mogu varirati, posebno tokom vršnih perioda korišćenja. U nekim slučajevima ovo možda nije idealno za aplikacije koje zahtevaju veliki propusni opseg.

VI. PRIMENE FWA

FWA može biti podogan širokopojasni pristup u različitim oblastima i situacijama [33], [34]:

- Ruralna područja – FWA je posebno pogodan za ruralna i udaljena područja gde je postavljanje optičkih kablova logistički izazovno i skupo.
- Urbana područja – Čak i u gusto naseljenim urbanim područjima, FWA pronalazi svoje mesto kao efikasno rešenje za povezivanje „poslednjeg kilometra". Pomaže u ublažavanju zagušenja mreže i pruža alternativu/*backup* pristup Internetu kada za to postoji potreba.
- Poslovna i preduzetnička rešenja – Brzo postavljanje i isplativost FWA čine ga atraktivnim izborom za mnoga preduzeća koji ga koriste za namenske, brze Internet veze ili kao *backup* žičnom pristupu.
- Hitne situacije – U slučaju prirodnih katastrofa ili drugih situacija kada su fiksne mreže oštećene, FWA može pružiti brzu i efikasnu alternativu.
- Privremeni događaji – FWA se može koristiti za pristup Internetu na privremenim događajima kao što su festivali, sajmovi ili sportski događaji.
- Industrija 4.0 i IoT – FWA podržava povezivanje pametnih uređaja, senzora i automatskih sistema u fabrikama i poljoprivredi.

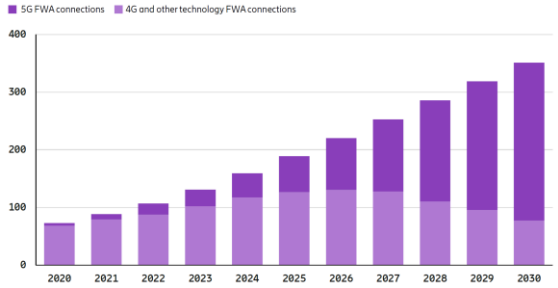
VII. PERSPEKTIVE FWA

Sve skorašnje prognoze govore da FWA ima veliku perspektivu, i da se očekuje rast tržišta, a naročito sa implementacijom 5G tehnologije na kojoj se uglavnom bazira budući FWA.

Kompanija Ericsson koja poseduje napredna rešenja za

FWA, sprovodila je više istraživanja o perspektivama FWA [37]-[38], a poslednje rezultate objavila je u seriji članaka pod nazivom *Fixed Wireless Access handbook 2025*:

- Na kraju 2024. broj FWA konekcija bio je oko 160 miliona.
- Predviđa se da će do 2030. god. broj FWA konekcija rasti širom sveta i da će dostići 350 miliona, a od toga će 280 miliona, ili 80%, biti 5G FWA veza (Sl. 5).
- Procenjuje se da će pružaocima usluga FWA donositi godišnje 74 milijarde dolara prihoda.

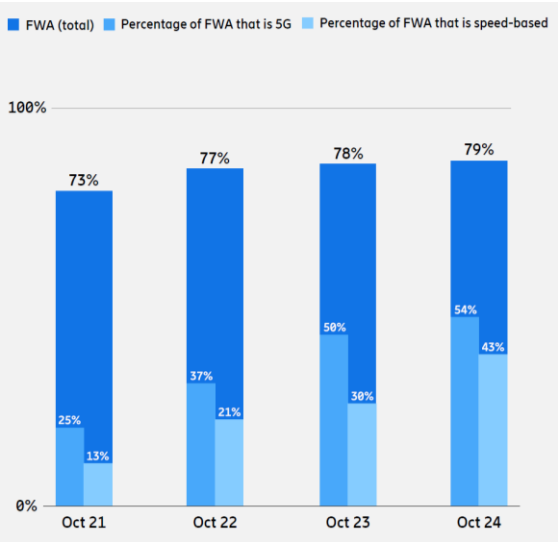


Slika 5. Trend rasta FWA korisnika do 2030. [38]

- Što se prenešenog saobraćaja tiče, na kraju 2024. god, FWA saobraćaj činio je 25% ukupnog saobraćaja globalno u mobilnim sistemima. Predviđa se da će FWA saobraćaj porasti za više od četiri puta do 2030. god. (skoro 170 EB), a procena je da će to biti 36% ukupnog saobraćaja.

Slično kao i u drugim sistemima, korisnicima su se do sada nudila generalno dva različita tarifna profila (Sl. 6):

- profil koji limitira ukupnu količinu saobraćaja koji se može preneti po korisniku po ugovorenoj brzini na mesečnom nivou i jeftiniji je od drugog tarifnog profila.
- *Flat-rate* pristup (*speed-based*), skuplji tarifni profil, ali sa svojim prednostima.

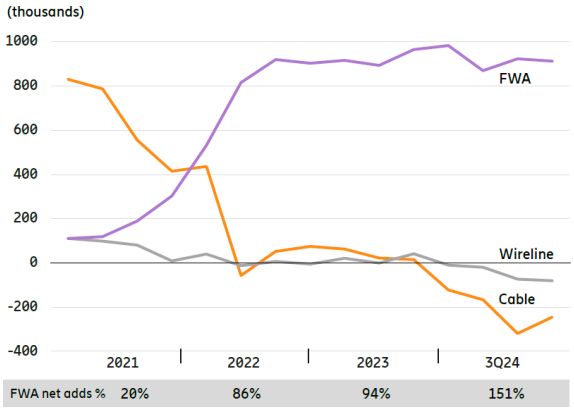


Slika 6. Procenat operatora na globalnom nivou koji nude FWA, procenat 5G FWA i procenat provajdera koji nude *flat-rate* (*speed-based*) [38]

Kao jedan od pokazatelja u kom pravcu ide implementacija FWA može se navesti da su do novembra 2024. god, tri najveća provajdera mobilnih usluga u SAD (AT&T, T-Mobile i Verizon), zajedno povezali oko 10,4 miliona novih korisnika u poslednje 4 godine. U slučaju T-Mobile i Verizon-a, većina su bili FWA korisnici, i to 6 miliona odnosno 4 miliona, respektivno, tokom posmatranog perioda. AT&T je FWA usluge počeo da nudi u avgustu 2023. i prijavio skoro 500.000 FWA konekcija do trećeg kvartala 2024. Tokom 2024. godine, segment FWA pristupa je najviše rastao na tržištu fiksnog širokopojasnog pristupa.

Sa druge strane, kablovski provajderi su doživeli pad udela u tržištu tokom ovog perioda. Ne mali pozitivni priraštaj korisnika tokom 2021. god, naglo pada do sredine 2022. god. čak prelazi u negativne vrednosti i praktično nastavlja sa padom. Do kraja 2023. godine, najveći kablovski provajderi (Comcast, Charter i Altice) počeli su da gube korisnike širokopojasnog pristupa (Sl. 7), iako su i dalje najveći širokopojasni segment u Americi [38].

Za segment žičnog pristupa karakterističan je prelazak širokopojasnih veza zasnovanih na bakarnim kablovima na veze zasnovane na optičkim vlaknima. Do 2023. ta tranzicija je rezultirala malim pozitivnim prirastom za žične segmente AT&T i Verizon, dok je 2024. pokazala pad, prvenstveno u AT&T (Sl. 7).



Slika 7. Priraštaj korisnika fiksnog širokopojasnog pristupa u tri najveća mobilna provajdera u SAD [38]

VIII. ZAKLJUČAK

Fiksni bežični pristup (FWA) predstavlja jedno od veoma korisnih rešenja za širokopojasni pristup. Postoji nekoliko tehnologija koje to omogućuju, ali se u poslednje vreme ističu 4G LTE i 5G. Za razliku od tradicionalnih žičnih pristupa, FWA eliminiše potrebu za fizičkim vezama, oslanjajući se umesto toga na bazne stanice i fiksne prijemnike za obezbeđivanje pristupa Internetu. Integracija naprednih tehnologija kao što su MIMO, *beamforming* i milimetarske frekvencije značajno je poboljšala njegovu brzinu, kapacitet i pokrivenost, čineći ga održivom alternativom optičkim mrežama. Pojava 5G je promenila pravila igre za FWA. Omogućava gigabitne brzine, nisku latenciju i pouzdane veze, pozicionirajući FWA kao ključno rešenje za premošćavanje digitalnog jaza i podršku novim aplikacijama kao što su IoT, pametni gradovi i *edge computing*.

ZAHVALNICA

Ovaj rad podržan je od strane Fakulteta tehničkih nauka u Novom Sadu, Departmana za energetiku elektroniku i telekomunikacije, u okviru projekta pod nazivom „Razvoj i primena savremenih alata i metoda istraživanja u energetici, elektronici i telekomunikacijama”.

LITERATURA

[1] White Paper, "Evolution of Fixed Access Networks", OFCOM, Published 13 September 2023.

[2] M. Victor-Ikoh, A. Moko, "Fixed Wireless Access: An Explorative Study of WIMAX FWA and 5G FWA Networks", IJCSMC, Vol. 10, Issue. 4, April 2021, pp. 99-107

[3] European Editors, "Wireless Local Loop", April 10, 2012, https://www.digikey.com/en/articles/wireless-local-loop, pristupljeno Februar 2025.

[4] "Fixed Wireless Access", Ericsson, https://www.ericsson.com/en/fixed-wireless-access, pristupljeno Februar 2025.

[5] "Fixed Wireless Access explained", Nokia, 28 Jun 2023, https://www.nokia.com/about-us/newsroom/articles/fixed-wireless-access-explained/, pristupljeno Februar 2025.

[6] "Wireless Local Loop", https://en.wikipedia.org/wiki/Wireless_local_loop, pristupljeno Februar 2025.

[7] "FWA", Open Fiber, https://openfiber.it/en/?glossary_new=fwa, pristupljeno Februar 2025.

[8] "Fixed Wireless Access, Handbook on Land Mobile (including Wireless Access)", Volume 1, Second Edition, International Telecommunication Union, Radiocommunication Bureau, 2001.

[9] Andrea Goldsmith, "Wireless Communications", Draft of Second Edition, Chapters 1-9, Feb. 18, 2020.

[10] C. Moozakis, S. M. Kerner, "What is wireless communications? Everything you need to know", TechTarget, January 2023.

[11] Bob Wallace, "Fixed Wireless Access: An Enterprise Broadband Alternative", www.networkcomputing.com, June 29, 2023, https://www.networkcomputing.com/wireless-networking/fixed-wireless-access-an-enterprise-broadband-alternative, pristupljeno Februar 2025.

[12] "The Technical Specifications of Fixed Wireless Access", Expero, October 10, 2024, https://www.expereo.com/blog/fixed-wireless-access-technical-specifications, pristupljeno Februar 2025.

[13] "WiMAX", Wikipedia, https://en.wikipedia.org/wiki/WiMAX, pristupljeno Februar 2025.

[14] E. Gregersen and other Editors of Encyclopædia Britannica, "WiMAX Technology", Britannica, Feb 13, 2025, https://www.britannica.com/technology/WiMax, pristupljeno Februar 2025.

[15] "Internet Access Guide: WiMAX", Conniq.com, https://www.conniq.com/InternetAccess_WiMAX-02.htm, pristupljeno Februar 2025.

[16] J. Engebratson, "Declaration Networks CEO: We're Shifting Focus from Fixed Wireless to Fiber", Benton Institute for Broadband & Society, July 24, 2023, https://www.benton.org/headlines/declaration-networks-ceo-we%E2%80%99re-shifting-focus-fixed-wireless-fiber, pristupljeno Februar 2025.

[17] D. Vujić, "Studija izvodljivosti uvođenja *White Space* uređaja u UHF opsegu (između 470-790 MHz)", Regulatorna agencija za elektronske komunikacije i poštanske usluge – RATEL, Beograd, Jun 2018.

[18] "White Spaces (radio)", Wikipedia, https://en.wikipedia.org/wiki/White_spaces_(radio), pristupljeno Februar 2025.

[19] "TV White Space Database", Wikipedia, https://en.wikipedia.org/wiki/TV_White_Space_Database, pristupljeno Februar 2025.

[20] "Television White Space", RFWEL, https://www.rfwel.com/us/index.php/tvws?srsltid=AfmBOopjqmR0e0tjFwfgHM3aFl9tjccQAxcEoyaI5eA3z-qnL5nyE4k8, pristupljeno Februar 2025.

[21] Product and Solution, "Residential & Enterprise Broadband", Global Invacom Group, https://globalinvacom.com/pages/new-markets, pristupljeno Februar 2025.

[22] "Professional Grade Satellite Internet Service", Ground Control, https://www.groundcontrol.com/products/vsat/, pristupljeno Februar 2025.

[23] Editorial Team, "What is VSAT (Very Small Aperture Terminal)?", EverythingRF, Apr 6, 2019, https://www.everythingrf.com/community/what-is-vsat-very-small-aperture-terminal?gad_source=1&gclid=Cj0KCQiA2oW-BhC2ARIsADSIAWqQ082mnj1rY4dwDJa2pdwAQPJxVncMXbGKZ9u18sDW9eV34-x9jqgaAlJ-EALw_wcB, pristupljeno Februar 2025.

[24] A. Božović, "Starlink Mini od sada dostupan u Evropi po vrlo pristupačnoj ceni, internet čak i u najzabačenijim mestima bez signala", 27.07.2024, Benchmark, https://benchmark.rs/vesti/uredaji/starlink-mini-od-sada-dostupan-u-evropi-po-vrlo-pristupacnoj-ceni-internet-cak-i-u-najzabacenijim-mestima-bez-signala/, pristupljeno Februar 2025.

[25] "Starlink", Wikipedia, https://en.wikipedia.org/wiki/Starlink, pristupljeno Februar 2025.

[26] "Satellite Technology", Starlink, https://www.starlink.com/technology, pristupljeno Februar 2025.

[27] J. Wallis, "Starlink Internet & Its Incredible Coverage – The Tech Behind Series", Intuji, December 12, 2022, https://intuji.com/starlink-tech-behind-internet-coverage/, pristupljeno Februar 2025.

[28] D. Anders, "Starlink Internet Review: Plans, Pricing, Speed and Availability", CNET, April 9, 2025, https://www.cnet.com/home/internet/starlink-internet-review/, pristupljeno Februar 2025.

[29] Blog, "What exactly is 4G FWA?", PUSR, September 11, 2023, https://www.pusr.com/blog/What-exactly-is-4G-FWA, pristupljeno Februar 2025.

[30] A. Mends Crentsil, "What is Fixed Wireless Access (FWA)?", Netscout, February 2, 2024, https://www.netscout.com/what-is/fixed-wireless-access-fwa, pristupljeno Februar 2025.

[31] FWA White Paper, "How fixed wireless access is erasing the digital divide", Quectel, https://www.quectel.com/blog/4g-5g-fwa-white-paper/, pristupljeno Februar 2025.

[32] "What is FWA (Fixed Wireless Access) Technology & How Does It Work?", Tata Play Fiber Blog, https://www.tataplayfiber.com/blog/what-fwa-fixed-wireless-access-technology-how-does-it-work, pristupljeno Februar 2025.

[33] "Fast and Flexible: Unleashing the Potential of Fixed Wireless Access, Droam Blog", https://droam.com/blog/fixed-wireless-access/, pristupljeno Februar 2025.

[34] V. Kohli, "FWA use cases for next-generation connectivity", TechTarget, 30 Aug 2023, https://www.techtarget.com/searchnetworking/tip/FWA-use-cases-for-next-generation-connectivity, pristupljeno Februar 2025.

[35] D. Darah, "Evaluate top 5G fixed wireless access benefits", TechTarget, 14 Jun 2023, https://www.techtarget.com/searchnetworking/feature/Evaluate-top-5G-fixed-wireless-access-benefits, pristupljeno Februar 2025.

[36] J. Rosenfeld, "Fixed Wireless Access (FWA): The future of connectivity?", September 26, 2024, Hologram, https://www.hologram.io/blog/fixed-wireless-access-connectivity/, pristupljeno Februar 2025.

[37] Report, "Capturing the 5G FWA opportunity: A household view", Ericsson, 2024, https://www.ericsson.com/en/reports-and-papers/consumerlab/reports/fixed-wireless-access-for-household, pristupljeno Februar 2025

[38] Fixed Wireless Access handbook 2025, "FWA momentum: Unlocking economies of scale", Insight 1 of 8, Ericsson, 2025.

[39] Laitsou, Eleni; Ioannou, Nikos; Katsianis, Dimitris: "5G Fixed Wireless Access for rural broadband", 31st European Conference of the International Telecommunications Society (ITS): "Reining in Digital Platforms? Challenging monopolies, promoting competition and developing regulatory regimes", Gothenburg, Sweden, 20th-21st June 2022.

Fixed Wireless Access (FWA) – Technologies, Applications And Perspectives

Dejan Nemec

ABSTRACT

Fixed Wireless Access (FWA) is a technology that enables wireless connection of users to broadband networks, thus eliminating the need for physical cables. This paper explores the development, technological aspects, application and outlook of FWA technology, with particular reference to its role in providing broadband Internet in rural and urban areas. The various technologies used in FWA, including WiMAX, satellite access, 4G LTE and 5G, are analyzed, as well as their advantages and challenges. Also, the paper discusses the key technological aspects of FWA and the perspective of FWA in the context of global digital transformation.

Evolucija RAN arhitektura: Od distribuiranih sistema do neprednih OPEN RAN modela

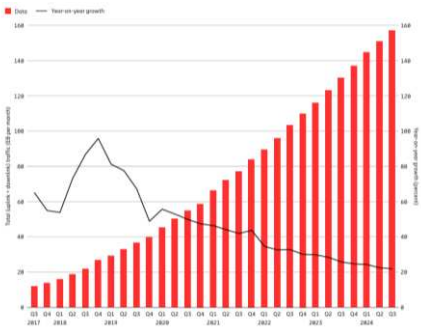
Katarina Stefanović
Telekom Srbija
Beograd, Srbija
katarinastef@telekom.rs
0009-0005-9847-8620

***Apstrakt* - Eksponencijalni rast mobilnog saobraćaja, uvođenje 5G servisa i sve strožiji zahtevi korisnika za QoE (Quality of Experience) nameću potrebu za inovativnim i efikasnim rešenjima u arhitekturama pristupnih mreža (RAN, *Radio Access Network*). Ovaj rad pruža sveobuhvatnu analizu evolucije RAN arhitektura, fokusirajući se na prelazak sa tradicionalnih distribuiranih modela na napredne koncepte kao što su C-RAN (*Cloud* RAN), H-CRAN (*Heterogeneous Cloud* RAN), F-RAN (*Fog*-RAN) i O-RAN (*Open* RAN). Detaljno su opisane njhove arhitekture, prednosti i izazovi u implementaciji, sa posebnim osvrtom na energetsku efikasnost, spektralnu efikasnost, troškove i alokaciju resursa. Kroz uporednu analizu, rad ističe kako je tranzicija ovih arhitektura obezbedila skalabilnosti i održivosti UDN (*Ultra-Dense Networks*) u 5G okruženju iz ugla operatora, kao i uvođenje veštačke inteligencije (AI, Artificial Intelligence). Na kraju rada se diskutuje ko će preuzeti vođstvo RAN tržišta, AI u RAN-u ili Open RAN, predlažući buduće pravce istraživanja.**

Ključne reči – 5G,C-RAN, H-CRAN, Open RAN, F-RAN

I. UVOD

Sa razvojem 5G mreža i inovativnih servisa trend eksponecijalnog rasta generisanh podataka sve je izraženiji u mobilnoj mreži. Kvartalni rast saobraćaja mobilnih mreža između Q2 2024. i Q3 2024. godine iznosio je oko 4 procenta, kao što je prikazano na slici 1, dok je ukupan mesečni globalni saobraćaj mobilnih mreža iznosio 157 EB ($157*10^{18}$ bajtova) [1].



Slika 1. Globalni paketski saobraćaj mobilnih korisnika [1]

Densifikacija ćelija *(cell densification)*, čija je strategija iz ugla operatora prikazana na slici 2, se predlaže za prevazilaženje problema za ratućim brojem generisanih podataka, korisnika i unapređenja spektralne efikasnosti postojećih mreža. Glavni nedostatak povećanja gustine baznih stanica (BS, *Base Station)* je to što ukupni interferencijski nivo u mreži takođe raste, što rezultira degradacijom velikog broja KPI (*Key Performance Indicator*), kao što su RSRP *(Reference Signal Received Power),* RSRQ *(Reference Signal Received Quality),* SINR *(Signal to Interference plus Noise Ratio),* BLER *(Block Error Rate)* i velikim broja dropova i neuspešnih hendovera. Pored toga, postavljanje više BS-ova nije ni ekonomski, a ni energetski efikasno jer RAN mreža troši veliku količinu energije. Ekonomska neefiksanost se ogleda kroz visoke troškove, kako za kupovinu same opreme, parcela na kojima se postavlja operma, tako i za održavanje odnosno omogućavanje nesmetanog i efikasnog funkcionisanja BS-ova. Trenutno, industrija informacionih i komunikacionih tehnologija odgovorna je za 5% globalnih emisija ugljen-dioksida (CO_2). Kako bi mobilni operatori mogli da odgovore na buduće zahteve 5G servisa, sve veći obim generisanih podataka i kako bi smanjili emisiju CO_2 i utrošak energije po baznoj stanici, moraju pronaći efikasna rešenja za poboljšanje kvaliteta servisa *(Quality of Service,* QoS), povećanje spektralne efikasnosti i održavanje stabilnih prihoda, uz istovremeno smanjenje troškova. Iz tog razloga su predložene brojne arhitekture i tehnologije za 5G mobilne mreže, posebno u oblasti RAN mreža. Inovacije odnosno strategije koje su operatori preduzeli kako bi bili u koraku sa zahtevima tržišta su:

- Povećanje spektralne efikasnosti kroz primenu MIMO *(Multiple Input Multiple Output), beamforming-a* i *mmWave* tehnologije,
- Primena piko/femto/mikro ćelija koje koegzistiraju sa novim tehnologijama, kao što su softverski definisane mreže (SDN, *Software-Defined Networking)* i vrtualizacija mrežnih funkcija (NFV, *Network Functions Virtualization),* i
- Optimizacija i rekonstrukcija RAN arhitektura. Inovativna renja za mrežne arhitekture poput C-RAN, H-CRAN, F-RAN i O-RAN-a, ne samo da poboljšavaju performanse sistema, već i energetsku efikasnost, a smanjuju kapitalne (CAPEX, *Capital Expenditures)* i operativne troškove (OPEX, *Operational Expenditures).*

Slika 2. Strategija operatora – densifikacija ćelija

Rad je organizovan na sledeći način. U prvom poglavlju je dat uvod i objašnjeno je zašto je došlo do evolucije RAN mreža, kao i zaito je to važno operatorima. U drugom poglavlju su objašnjeni tipovi RAN arhitektura, njihove prednosti i nedostaci. U trećem poglavlju je data uporedna analiza arhitektura. U četvrtom poglavlju su data zaključna razmatranja i budući pravci istraživanja.

II. Pregled evolucije bežičnih pristupnih mreža

A. Cloud RAN

Bežična pristupna mreža je interfejs mobilne mreže ka korisnicima i najskuplji deo mobilne mreže u smislu CAPEX-a i OPEX-a. Tehnički gledano, zbog ograničenja radio-interfejsa, RAN je izvor skoro 80% problema povezanih sa performansama mreže, koji utiču na kvalitet korisničkog iskustva (QoE, *Quality of Experience*). Iz tog razloga je RAN ključna oblast fokusa mobilnog operatora kako sa tehničkog, tako i sa finansijskog aspekta. Vendori i operatori stalno razvijaju nove pristupe za povećanje efikasnosti RAN-a kako bi zadovoljili sve veće zahteve, a istovremeno smanjili CAPEX, OPEX i ukupne troškove vlasništva (TCO, *Total Cost of Ownership*). U tradicionalnoj distribuiranoj arhitekturi (D-RAN, *Distributed Radio Access Network*), svaka BS ima svoju BBU (*Baseband Unit*) i RRU (*Remote Radio Unit*), koje su fizički na istoj lokaciji. Prednosti ove arhitekture su jednostavnost implementacije i brza lokalna obrada podataka. D-RAN arhitektura ima ograničenja u pogledu fleksibilnosti, smanjenja troškova, skalabilnosti jer je održavanje kompleksno usled hardvera i softvera, koji je vendorski zavistan (*vendor lock-in*) i efikasnosti performansi, kao i optimizovanje energetske i spektralne efikasnosti. Ispunjenje ovih ograničenja zahteva inovativniji pristup poput *cloud computing*-a, koji se koristi prilikom realizovanja C-RAN (*Cloud Radio Access Network*).

B. Cloud RAN

Koncept softverizacije i virtualizacije RAN resursa naziva se C-RAN. C-RAN podrazumeva premeštanje BBU jedinica sa baznih stanica na centralizovanu lokaciju izmedu ivice mreže (*edge*) i jezgra mreže (*core*), gde se formira BBU pool (data centri) radi povećanja skalabilnosti i pripreme RAN-a za virtualizaciju. Virtualizovane BBU jedinice (vBBU) se implementiraju na više NFV platformi koristeći cloud resurse koje operatori već imaju i/ili resurse cloud provajdera (Amazon Web Services, Microsoft Azure, Google Cloud Platform). Dok su udaljene radio jedinica (RRH, *Remote Radio Head*) smeštene na ivici mreže, ali su povezane sa vBBU jedinicama kroz proprijetarni (*vendor-specific*) fronthaul (eCPRl, *Enhanced Common Public Radio Interface*) interfejs. RRH jedinica vriši analogno-digitalnu konverziju, pojačanje signala, filtriranje i konverziju frekvencija. Dok BBU vrši (de)modulaciju, upravlja spektralnim i vremenskim resursima, vrši (de)kodiranje, sinhronizaciju i kontrolu RAN-a. Prednosti C- RAN arhitekture su sledeći [2-8]:

1) Smanjenje potrošnje energije i povećanje protoka. Za razliku od D-RAN-a, C-RAN koristi manji broj BBU-a, što rezultira značajnim smanjenjem potrošnje energije. C-RAN omogućava i prebacivanje energetski intenzivnih obrada podataka sa korisničkih uredaja (UE, *User Equipment*) i BS na obližnji cloud, čime se štedi energija. Zbog velike gustine RRH-ova, razdaljina izmedu radio jedinice i UE je smanjena kako bi se smanjila snaga emitovanja bez uticaja na ukupnu pokrivenost mreže. Niska snaga prenosa znači da će vek trajanja baterije UE biti duži, a potrošnja energije RAN-a smanjena. Tokom perioda niskog saobraćaja, nedovoljno iskorišćeni BBU pool-ovi/BS-ovi mogu biti isključeni, a njihov saobraćaj može biti redistribuiran na aktivne BBU pool-ove/BS-ove.

2) Povećanje kapaciteta i smanjenje ko-kanalske interferencije. U centralizovanom sistemu kao što je C-RAN, BBU pool koordinira rad više ćelija istovremeno, što omogućava optimalno raspoređivanje resursa i efikasnije upravljanje interferencijama. Drugim rečima, kada više ćelija koristi zajednički sistem za alokaciju resursa, može se precizno kontrolisati način na koji se signali prenose i primaju, čime se značajno smanjuje preklapanje i smetnje izmedu kanala. Ovo rezultira poboljšanjem kvaliteta signala, većim kapacitetom mreže i boljim QoE, posebno u gusto naseljenim oblastima ili na mestima sa visokim saobraćajnim opterećenjem. Pored toga, implementacijom C-RAN-a, brzine prenosa podataka u *downlink*-u na ivici ćelije mogu se poboljšati za 40%-70%, dok se brzine prenosa podataka u *uplink*-u na ivici ćelije mogu povećati i do 2-3 puta [4].

3) Smanjeni kapitalni i operativni troškovi. Raspodela L1, L2 i L3 funkcija izmedu BBU i RRH jedinica smanjuje CAPEX i OPEX u C-RAN arhitekturi. CMRI (*China Mobile Research Institute*) u radu [5] navodi da primenom C-RAN arhitekture može doći do smanjenja CAPEX-a do 15% i OPEX-a do 50%. Smanjenje OPEX za operatore je posledica manjeg broja poseta lokacijama, lakšeg nadogradnje i održavanja mreže, ali i nižih TCO. U C-RAN arhitekturi, broj lokacija za BBU može biti smanjen za jedan do dva reda veličine što je uzrok smanjenja CAPEX-a.

4) Poboljšanje mobilnosti. Među brojnim prednostima C-RAN-a, unapređeno upravljanje mobilnošću smatra se jednom od ključnih tema koje su intenzivno proučavane. C-RAN poboljšava upravljanje mobilnošću tako što integriše naprednu predikciju kretanja korisnika, inteligentinije donošenje odluka o handoveru usled poznavanja šire slike mreže tj. usled centralizovane obrade signala i alokacije resursa, grupisanje ćelija u klastere radi smanjenja hard hendovera i deljenja resursa. Time se broj nepotrebnih handover-a smanjuje za oko 20%, a QoE poboljšava.

Pored prednosti koje C-RAN donosi 5G sistemima, postoje i određena ograničenja, uključujući: sajber-napade koji su u stalnom porastu (MAC *spoofing*, IP *spoofing* i *hijacking*, TCP *flooding napadi,* kao i FTP napadi) čija je meta centralizovani sistem i ograničen kapacitet fronthaul linka. Ukoliko usled sajber-napada BBU pool otkaže dolazi i do otkaza većeg dela mreže što je glavni nedostataka ovog pristupa. Kako bi se prevazišao nedostatka fronthaul linka uvedene su tehnike deljenja podataka između BS-ova, tehnika kompresije podataka i razmatranje upotrebe FSO (*Free-Space Optics*) kao fronthaul rešenje. U poređenju sa postojećim fronthaul rešenjima (*mmWave*, optika), upotreba FSO kao novog fronthaul rešenja ima četiri glavne prednosti:

gigabitni protoci, usmeren laserski snop (otpornost na elektomagnetne inteferencije i visoka bezbednost) i brza implementacija. Medutim vremenski uslovi značajno degradiraju kvalitet signala. Gusti oblaci, gusta magla ili peščane oluje mogu ozbiljno degradirati performanse FSO. Na primer, u Pekingu, gde su peščane oluje česte tokom proleća, korišćenje FSO za fronthaul nije dobra opcija. Za postizanje visokog kapaciteta i dostupnosti linka, može se koristiti hibridni FSO/*mmWave* pristup koji je idalje u fazi istraživanja i koji potencijalno obezbeduje dostupnost linka na nivou od 99.999%. Ovaj hibridni sistem je robustan i za kišne i za maglovite uslove. Jedini vremenski uslovi, koji mogu uticati na performanse hibridnog FSO/mmWave sistema je kombinacija jake kiše i guste magle, što se retko dešava. Ni mmWave ni FSO ne mogu raditi u okruženju bez linije optičke vidljivosti. U ovakvim slučajevima mogu se koristiti optička vlakna ili sub 6 GHz (licencirani) opsezi [6,7,8].

C. Heterogeni RAN

Kako bi se prevazišla objašnjena ograničenja odnosno „usko grlo“ fronthaul linka sa rastućim brojem korisnika i kompleksnosti mreže došlo je do razvoja arhitekture H-CRAN. U tradicionalnim mrežama isključivo su se koristile makro ćelije za pokrivanje velikih oblast usled manje gustine korisnika. Sa razvojem 5G sistema dolazi do zagušenja mreže. Kako bi operatori postigli visoke protoke i opslužili sve veći broj povezanih UE na Internetu, uvedene su heterogene bežične mereže (HetNets) sa piko i femto ćelijama. Male ćelije omogućavaju ponovnu upotrebu istih frekvencija, što znači da više korisnika može koristiti mrežu bez zagušenja. H- CRAN kombinuje fleksibilnost i kapacitet HetNet-a sa efikasnošću obrade signala u C-RAN. Za razliku od C-RAN arhitekture, BBU pool u H-CRAN-u je dodatno povezan sa makro baznim stanicama (HPN, *High Power Node*), kako bi se smanjilo opterećenje na fronthaul linkovima. BBU pool u H-CRAN arhitekturi je dodatno povezan sa HPN-ovima pomoću Sl i X2 interfejsa, koji su preuzeti iz 3GPP standardizacije. Prednosti i novine koje se uvode sa ovom arhitekturom [9-16]:

1) Suzbijanje interfereneije između RRH jedinica korišćenjem naprednih kooperativnih tehnika u cloud-u. Istovremeno, rastuća veličina HetNets izaziva eksplozivni rast mobilnog saobraćaja što dovodi do pojave međuslojne (*inter-tier*) interferencije. Inter-tier interferencija između makro i femto ćelija najčešće se dešava kada korisnici koriste isti frekvencijski opseg, kada handover nije optimizovan - veliki SINR jer korisnik ulazi u zgradu i ne preuzima je femto ćelija koja ima slab signal, a makro proizvodi samo smetnje ili kada makro ćelije previše „curi“ u zatvorene prostore (*Coverage Leakage*). Rešenja uključuju korišćenje različitih frekevencijskih spektra izmedu marko i femto ćelije, kontrolu snage/optimizaciju vrednosti tilta i napredne tehnike smanjenja interferencije (eICIC, *Enhanced Inter-Cell Interference Coordination*) korišćenjem parametara kao što su ABS (*Almost Blank Subframes*), BO (*Bias Offset*), CRE (*Cell Range Expansion*), TPA (*Transmit Power Adjustment*) i CTV (*Cell Transmission Vector*). ABS je metoda koordinacije interferencije u vremenskom domenu koja ublažava interferenciju HPN ka malim ćelijama (*small cells*) tako što smanjuje snagu prenosa HPN tokom određenih podokvira (subframe). CRE omogućava rasterećenje saobraćaja HPN- ova tako što omogućava većem broju UE da se povežu na male ćelije, dodajući BO pomak na RSRP koji UE detektuje. TPA definite koliko će HPN smanjiti snagu prenosa tokom ABS podokvira kako bi se smanjila interferencija u downlink-u UE koji komuniciraju sa malim ćelijama. Dok CTV na bazi CQI (Channel Quality Indicator) povećava pokrivenost malih ćelija smanjenjem snage HPN-ova.

2) Upotreba kooperativnog upravljanja radio-resursima (CC-CRRM*, Cloud Computing-based Cooperative Radio Resource Management*) sa implementiranim samoorganizujućim mrežnim funkcijama (CC-SON, *Coverage and Capacity - Self-Organizing Network*) kako bi se postigla visoka efikasnost u korišćenju radio resursa, poboljšale performanse i smanjili operativni troškovi. CC-CRRM sistem dinamički rasporeduje radio resurse u mreži kako bi smanjio zagušenja i optimizovao performanse. CC-CRRM koristi globalne informacije *CSI* (*Channel State Information*) i QSI (Queue State Information) kako bi prioritetno prenosio saobraćaj osetljiv na kašnjenje. Ovaj pristup omogućava mreži da se prilagodi dinamičnom okruženju u 5G mrežama i optimizuje efikasnost upotrebe radio spektra. Mreža koristi samoorganizujuće funkcionalnosti kako bi automatski optimizovala parametre i minimizirala ljudsku intervenciju. CC-CRRM koristi adaptivne mehanizme za kompenzaciju prekida pokrivenosti kada korisnici prelaze između različitih ćelija (npr. izmedu RRH-ova i HPN-ova). To omogućava kontinuirani QoS, čak i u situacijama kada dolazi do gubitka signala ili preklapanja pokrivenosti. U poslednjem vreme, studije istražuju upotrebu algoritama mašinskog učenja za detekciju i kompenzaciju prekida rada ćelija (COD, *Cell Outage Detection*). Ovi algoritmi analiziraju podatke kao što su RSRP i RSRQ kako bi identifikovali problematične oblasti i automatski optimizovali mrežu. U nekim slučajevima koriste se i modeli kao što su HMM (*Hidden Markov Model*) ili KNN (*k-Nearest Neighbors*) algoritmi za precizno predviđanje i optimizaciju prekida u mreži.

3) Poboljšanje mobilnosti korisnika. Smanjenje broja neuspelih hendovera, niža stopa ping-pong efekta i niža stopa prekida veze za korisnike sa velikom brzinom kretanja ključni je cilj uvođenja HPN-ova. U H-CRAN mrežama, korisnici sa velikom brzinom kretanja se povezuju sa HPN jedinicama koje nude pouzdaniju vezu, dok se korisnici sa manjim brzinama kretanja preferencijalno povezuju sa RRH jedinicama.

4) Upotreba CC-CoMP (*Coordinated Clustered*-CoMP). CC-CoMP je tehnika koja koristi računarsku moć clouda da bi omogućila koordinisani rad više BS-ova (ili pristupnih tačaka) u mobilnoj mreži. Glavni cilj je smanjenje interferencije i poboljšanje protoka za korisnike na ivici ćelije. Koordinacija u CoMP-u može biti: JT (*Joint Transmission*), DPS (*Dynamic Point Selection*) i CS/CB (*Coordinated Scheduling/Beamforming*). JT podrazumeva da više BS-ova istovremeno šalje signal ka istom UE, povećavajući snagu signala i protok. DPS sistem bira najbolju BS u datom trenutku za slanje signala ka UE. Kod CS/CB sistema stanice zajedno planiraju resurse i usmeravaju snopove signala kako bi smanjile interferenciju.

U praksi su ograničeni kapacitet fronthaul-a i kašnjenje prenosa podataka neizbežni problemi. Osim toga, kako se mreža širi, postaje nepraktično pretpostaviti da su svi CSI idealno poznati i obrađeni za celu mrezu Zastarevanje CSI podataka zbog kašnjenja i visoke mobilnosti korisnika dovodi

do smanjenja preciznosti tehnika kao što je sparse beamforming i smanjuje ukupne performanse mreže. Ograničenja fronthaul linka utiču na performanse na nekoliko načina navedenih u narednom tekstu. Ukupna brzina prenosa podataka po RRH-u ne sme biti veća od kapaciteta njegovog fronthaul linka. Nedovoljan kapacitet fronthaul-a sprečava RRH da u potpunosti iskoristi dostupne radio resurse. Ovaj problem postaje još ozbiljniji kada se koriste napredne tehnike kao što su CC-CoMP i CC-CRRM, koje zahtevaju značajne količine podataka za koordinaciju. Prevelik broj HPN-ova u H-CRAN mrežama može smanjiti performanse RRH-ova, dok premali broj HPN-ova može pretvoriti H-CRAN u C-RAN. Zato je pronalaženje optimalne ravnoteže izmedu HPN-ova i RRH-ova ključno za efikasno funkcionisanje H-CRAN mreža. U budućnosti, potrebno je istražiti optimalnu gustinu i lokacije za postavljanje HPN-ova i RRH-ova kako bi se postigao najbolji kompromis u performansama [9, 14-16].

D. Fog RAN

Međunarodna korporacija za podatke (IDC*, International Data Corporation*) predviđa da će do 2020. godine biti 30 milijardi senzorskih uređaja povezanih na Internet. Međutim, trenutne tehnike računarstva u oblaku dovode do ograničenja koje idalje nije rešeno, a to je kašnjenje. Stoga je predložena nova arhitektura F-RAN, koja predstavlja poboljšanje u odnosu na H-CRAN jer smanjuje kašnjenje, zagušenje saobraćaja i troškove povezivanja premeštanjem obrade i skladištenja podataka bliže korisnicima. Ova arhitektura integriše prednosti cloud i fog računarstva, omogućavajući efikasniju raspodelu resursa i QoS. F-RAN koristi fog pristupne tačke (F-AP) i uređaje korisnika (F-UE) za lokalnu obradu podataka, čime se smanjuje opterećenje fronthaul-a i BBU pool-a. Povezivanjem F-AP-ova sa cloud-om putem optimizovanih fronthaul veza, F-RAN omogućava brži pristup sadržaju i poboljšava efikasnost mreže. F-RAN integriše edge keširanje i distribuiranu obradu podataka, smanjujući opterećenje fronthaul-a i poboljšavajući odziv sistema u realnom vremenu. Naglašavamo da, za razliku od C-RAN, cilj F-RAN arhitekture nije minimizacija troškova implementacije i operativnih troškova kroz smanjenu složenost čvorova na ivici mreže, već maksimizacija performansi sistema u smislu brzine isporuke podataka korišćenjem resursa u cloudu i na ivici mreže (keširanje). lako je bezbednost ključna za razvoj F-RAN sistema, malo se raspravlja o ovoj temi. Pomicanjem RAN resursa sa cloud na edge mreže, F-RAN mreže se suočavaju sa bezbednosnim pretnjama koje nisu prisutne u konvencionalnim, centralizovanim cloud sistemima. F-RAN mreže su podložnije napadima zlonamernih entiteta. U još nepovoljnijem scenariju, pošto se autentifikacija korisnika sprovodi na ivici mreže, umesto u cloud-u, napadač može lakše dobiti pristup mreži preko krajnjih pristupnih tačaka i ostati neotkriven od strane globalne cloud infrastrukture ili firewall sistema. Drugi izazov ovih arhitektura je nedostatak inicijativa i konzorcijuma. Do sada postoji samo OpenFog Consortium, koji je objavio referentnu arhitekturu za F-RAN (2017) kako bi ubrzao usvajanje IoT-a u preduzećima. Takode, nije razvijen nijedan standard specifičan za ove sisteme, dok u srodnim oblastima, poput MEC-a, postoje standardi za terminologiju, zahteve i okvire [17-21].

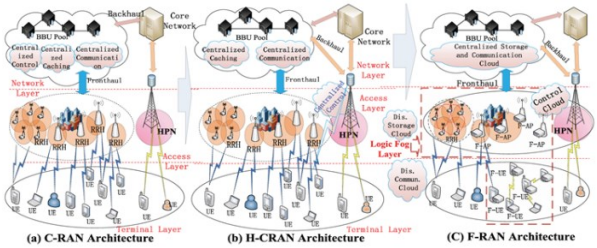
III. UPOREDNA ANALIZA

U narednom poglavlju je data tabela 1 i slika 3 sa uporednom analizom navedenih arhitektura, upoređujući odabrane parametre. Pri čemu su parametri odabrani po relevantnosti prilikom donošenja odluka za nove strategije i izmene RAN arhitektura iz ugla operatora. Uporedna analiza je urađena na osnovu sledećih parametra: mesto keširanja i kontrle, razdvajanje korisničke i kontorlne ravni, kao i opterećenja na svaki entitet u mreži, potrošnju energije, kašnjenja i troškova koje zahtevaju navedene arhitekture.

Tabela 1. Uporedna analiza arhitektura

Naziv arhiktekture	C-RAN	H-CRAN	F-RAN
Keširanje	Cetralizovana	Centralizovana	Centralizovana i distirbuirana
Kontrola mreže	Centralizovana	Centralizovana	Centralizovana i distriburiana
Razdvajanje korisničke i kontrolne ravni	Ne	Da	Da
Heterogenost	Srednja	Veoma visoka	Visoka
Edge mreža kompleksnot	Niska	Niska	Srednja
Procesiranje podataka	Cloud data centri	Cloud data Centri	Blizu korisnika (egde)
Opterećenje na fronthaul	Visoka	Srednja	Niska
Opterećenje na BBUpool	Visoka	Srednja	Niska
Inter-tier interfenrencija	Niska	Visoka	Srednja
Kašnjenje	Visoko	Nisko	Nisko
Potrošnja energije	Srednje	Visoko	Srednje
CAPEX/OPEX	Visoko	Veoma visoko	Srednje

Iz Tabele 1 možemo da zaključimo da je C-RAN zaslužan za centralizaciju procesa u RAN mrežama, ali i za velika kašnjenja i velika opterećenja na fronthaulu. Kako bi se prevazišlo "usko" grlo na fronthaulu uveden je H-CRAN koji je uveo heterogenost kao naredni problem, pri čemu kašnjenja idalje nisu rešena. Sa pojavom Fog RAN-a se rešava problem kašnjenja. Postavlja se zaključno razmatranje ove analize, a to je da su sve ove tranzicije arhitektura doprinele uvođenju AI u RAN mrežama, jer sada na centralizovanom mestu imamo veliku količinu podataka i sledeći cilj svakog opertora je da iskosriti centralizovanu veliku količinu podataka i generiše saznanja kako bi unapredio svoj budući rad.



Slika 3. Uporedna vizuelna analiza C- RAN, H-CRAN i F-RAN arhitektura

IV. ZAKLJUČAK

Sa dramatičnim porastom broja UE, ogromnom količinom podataka i novim zahtevima za izuzetno niskom latencijom i većom brzinom prenosa podataka, 5G zahteva duboko preispitivanje dizajna arhitekture mobilnih mreža, a posebno arhitekture RAN-a. Motivisani ovim, sproveden je pregled postojeće literature koja se odnosi na RAN arhitekture za 5G mobilne mreže, uključujući C-RAN, H-CRAN i F-RAN. Kako bi se izvršilo poboljašanje navedenih arhitektura sve veći broj istraživanja je usmeren na implementaciju ML

(*Machine Learning*) algoritama u RAN mreže. ML algoritmi su ključni za predviđanje budućih potreba za resursima, optimizaciju raspodele zadataka i poboljianje ukupne efikasnosti sistema. Nadzirano učenje (*Supervised Learning*), kao podgrupa ML algoritma, može da predvidi opterećenje sistema i zahteve za resursima na osnovu istorijskih podataka/prethodnih trendova, omogućavajući optimalnu raspodelu resursa. Nenadzirano učenje (*Unsupervised Learning*), kao druga podgrupa ML algoritma, identifikuje obrasce i strukture u neoznačenim podacima, što je posebno korisno za detekciju anomalija. Ova tehnika pomaže u identifikaciji neobičnih obrazaca koji mogu ukazivati na potencijalne problem, poput uskih grla u sistemu ili zlonamerne aktivnosti. DL (*Deep Learning*) tehnike su izuzetno efikasne u obradi nestrukturiranih podataka (slike, zvuk, tekst), što ih čini pogodnim za precizno predviđanje potreba za resursima uzimajući u obzir faktore poput ponašanja korisnika i aktivnosti IoT (*Internet of Things*) uredaja, čime se optimizuje dinamička dodela resursa. Dok RL (*Reinforcment Learning*) podrazumeva treniranje agenata da donose sekvencijalne odluke kroz nagrađivanje poželjnih ponašanja i kažnjavanje nepoželjnih. Ova tehnika je izuzetno efikasna u dinamičnim i složenim okruženjima kao što su autonomne strategije dodele resursa koje se prilagođavaju promenama u mrežnom okruženju bez ljudske intervencije. Implementacija AI modela koji mogu da obuhvate heterogenost u mreži je ključna jer univerzalni pristup nije dovoljan. Glavni izazov skaliranja AI u mrežnom okruženjima leži u koordinaciji velikog broja geografski disperzovanih čvorova i njihovim ograničenim računarskim kapacitetima. Potreba za otvorenošću i interoperabilnošću u RAN arhitekturi postaje sve izraženija. Open RAN se nameće kao jedno od najnovijih rešenja koje razdvaja mrežne komponente u standardizovane, otvorene interfejse, omogućavajući operatorima veću fleksibilnost i optimizaciju troškova [22]. Međutim nameće se pitanje da li je isplativije operatorima da uvode AI u već postojeće Cloud RAN infrastrukture ili da implementiraju Open RAN, što zavisi od mrežnog okruženja samog operatora. Iz tog razloga treba uzeti u obzir da li je u pitanju multivendorsko okruženje, da li je to mali ili veliki operator, koliko je razvijena ljudska ekspertiza za nove arhitekture, koja količina novčanog resursa se ulože, kao i da li Vendori nude odgovarajuća rešenja za nove arhitekture i da li postoje standardizacioni problemi. Usled postojanja niza pitanja i varijabli buduća istraživanja su usmerena na kalsifikaciju naučnih radova i pilot projekta u industiriji upoređujući efikasnost AI u Cloud RAN-u i Open RAN-a sa AI mogućnostima, kao i to ko će nositi veću tržišnu moć na telekomunikacionom tržištu.

LITERATURA

[1] F. Jejdling, „Ericsson Mobility Report ", 2024

[2] R. T. Rodoshi, T. Kim, W. Choi, "Resource Management in Cloud Radio Access Network: Conventional and New Approaches", *Sensors*, vol. 20, br. 9, 2020

[3] R. S. Alhumaima, M. Khan, H. S. Al-Raweshidy, "Component and parameterised power model for cloud radio access network", *IET Communications*, vol. 10, br. 7, str. 745-752, 2016

[4] V. Suryaprakash, P. Rost, G. Fettweis, " Are Heterogeneous Cloud-Based Radio Access Networks Cost Effective", *IEEE Journal on Selected Areas in Communications,* vol. 33, br. 10, 2015

[5] Y. Zhang, M. Chen, "Cloud Based 5G Wireless Networks", *Springer*, Chain, Switzerland, 2016

[6] J. Wu, Z. Zhang, Y. Hong, Y. Wen, "Cloud Radio Access Network (C-RAN): A Primer", *IEEE Network*, 2015, str. 35-41

[7] H. Zhang, Y. Dong, J. Cheng, J. Hossain, Member, V. Leung, "Fronthauling for 5G LTE-U Ultra Dense Cloud Small Cell Networks", *IEEE wireless communications*, 2016

[8] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, L. Dittman, "Cloud RAN for Mobile Networks Technology Overview", *IEEE Communications Surveys & Tutorials*, vo1. 17, br. 1, str. 405-426, 2015

[9] S. Pana, O. P. Babalola, V. Balyan, "5G radio access networks:A survey", Array, vol. 14, br. 1, 2022

[10] R. Sahu, V. Sahu, "An Energy-Efficient Algorithm for Resource Allocation in H-CRAN (EE H-CRAM) for 5G Networks", *Wireless Personal Communication,* vol. 138, str. 1483—1499, 2024

[11] M. A. Marotta, N. Kaminski, I. Gomez-Miguelez, L. Z. Granville, J. Rochol, L. DaSilva, C. B. Both, "Resource Sharing in Heterogeneous Cloud Radio Access Networks", *IEEE Wireless Conununications,* vo1. 22, br. 3, str. 74-82, 2015

[12] M. Peng, C. Wang, Y. Li, J. Jiang, J. Li, "Heterogeneous Cloud Radio Access Networks: A New Perspective for Enhancing Spectral and Energy Efficiencies", *IEEE Wireless Communications*, vo1. 21, br. 6, 2015

[13] Q. Liu, T. Han, N. Ansari, G. Wu, "On Designing Energy-Efficient Heterogeneous Cloud Radio Access Networks", *IEEE Transactions on Green Communications and Networking, vol.* 2, br. 3, str. 721-734, 2018

[14] P. Yu, F. Zhou, T. Zhang, W. Li, L. Feng, X. Qiu, "Self-Organized Cell Outage Detection Architecture and Approach for 5G H-CRAN", *Wireless Communications and Mobile Computing,* br. 4, pp.:1-11, 2018

[15] H. Choi, T. Kim, S. Lee, H. S. Choi, N. Yoo, "Energy-Efficient Dynamic Enhanced Inter-Cell Interference Coordination Scheme Based on Deep Reinforcement Learning in H-CRAN", *Sensors,* vol. 24, br. 24, 2024

[16] N. Chen, B. Rong, X. Zhang, M. Kadoch, "Scalable and flexible massive MIMO precoding for 5G H-CRAN", *IEEE Wireless Communications,* vol. 24, br. 1, str. 46-52, 2018

[17] K. Liang, L. Zhao, X. Chu, H. Chen "An Integrated Architecture for Software Defined and Virtualized Radio Access Networks with Fog Computing", *IEEE Network,* vol. 31, br. 1, str. 80-87, 2017

[18] C. Mouradian, D. Naboulsi, S. Yangui, R. H. Glitho, M. J. Morrow, P. Polakos, "A Comprehensive Survey on Fog Computing: State-of-the-Art and Research Challenges", *IEEE communications surveys & tutorials,* vo1. 20, br. 1, 2018

[19] Y. Ku, D. Lin, C. Lee, P. Hsieh, H. Wei, C. Chou, A. Pang, "5G Radio Access Network Design with the Fog Paradigm: Confluence of Communications and Computing", *IEEE Communications Magazine,* vol. 55, br. 4, str. 46-52, 2017

[20] D. Alsadie, "A Comprehensive Review of AI Techniques for Resource Management in Fog Computing: Trends, Challenges and Future Directions", *IEEE Access,* vol. 12, 2024

[21] M. Habibi, H. Schotteni, M. Binhan, "A Comprehensive Survey of RAN Architectures Toward 5G Mobile Communication System", IEEE Access, vol. 7, 2019

[22] K. Alain, M. Habibi, M. Saad, M. Renzo, T. Melodia, M. DebbaH, A. Schotten, "A Comprehensive Tutorial and Survey of O-RAN: Exploring Slicing-aware Architecture, Deployment Options, Use Cases, and Challenges", [Online] dostupno na: https://arxiv.org/abs/2405.03555

Evolution of RAN Architectures: From Distributed Systems to Advanced OPEN RAN Models

Katarina Stefanović

ABSTRACT

The exponential growth of mobile traffic, the introduction of 5G services, and user demands for Quality of Experience necessitate innovative and efficient solutions in RAN architectures. This paper provides a comprehensive analysis of the evolution of RAN architectures, focusing on the transition from traditional distributed models to advanced concepts such as Cloud RAN, Heterogeneous Cloud RAN, Fog RAN and Open RAN. Their architectures, advantages, and implementation challenges are described in detail, with a particular emphasis on energy efficiency, spectral efficiency, costs, and resource allocation. Through a comparative analysis, the paper highlights how the transition to these architectures has ensured the scalability and sustainability of Ultra-Dense Networks in the 5G environment from the

operator's perspective, as well as the integration of Artificial Intelligence in RAN. Finally, the paper discusses who will take the lead in the RAN market AI in RAN or Open RAN, while proposing future research directions.

Prednosti i izazovi primene prenosnih baznih stanica mobilne telefonije

Dejan Nemec
Fakultet tehničkih nauka, Univerzitet u Novom Sadu
Novi Sad, Srbija
denem@uns.ac.rs

***Apstrakt* – Prenosne Bazne Stanice (PBS) postale su ključni element u modernim mobilnim mrežama, omogućavajući brzu implementaciju mrežne infrastrukture kada za to postoji potreba. Ovaj rad istražuje scenarije primena prenosnih baznih stanica, kao što su veliki događaji, ruralna područja i u slučajevima katastrofe. Navode se karakteristike i tipovi prenosnih baznih stanica, odnosno mogućnosti prenosa PBS opreme na područje od interesa. Kroz rad navedene su prednosti koje PBS pružaju, a posebno su analizirani problemi i izazovi koji prate primenu PBS.**

Ključne reči – PBS, RDU, CoW, LAP, HAP.

I. UVOD

Mobilna povezanost je nešto što većina ljudi uzima zdravo za gotovo, ali to nekada može biti teško ostvarivo na mestima gde se privremeno okupi veliki broj ljudi, a mreža nije projektovana to da podrži ili u slučaju ekstremnih vremenskih prilika ili prirodne katastrofe kada se ošteti ili uništi infrastruktura i kada je od vitalnog značaja obezbediti komunikaciju spasilačkim ekipama i građanstvu.

Prenosne bazne stanice (PBS) su mobilne verzije tradicionalnih baznih stanica koje su postale ključni element u modernim mobilnim mrežama omogućavaju brzu implementaciju mrežne infrastrukture u područjima s ograničenim resursima, u ruralnim područjima, tokom velikih događaja (sportski događaji, sajmovi, koncerti), u hitnim situacijama i u slučaju prirodnih katastrofa. Njihova upotreba je strateška za brzo širenje mobilnih ćelijskih mreža.

PBS treba da podrži naglo povećanje mobilnog saobraćaja ili da obezbedi osnovne usluge kada one nisu dostupne. Postavljanje PBS, u principu, ne zahteva građevinske dozvole i građevinske radove.

Postoji više vrsta PBS, odnosno načina kako oprema može biti postavljena da pokrije određeno područje od interesa. Ove vrste se razlikuju po veličini i mogućnostima koje pružaju u pogledu broja korisnika koje mogu da povežu i kakve im usluge mogu ponuditi. Pored ovoga, PBS se razlikuju da li se postavljaju na zemlji ili je oprema u vazduhu prenošena različitim tipovima letelica. Dva bitna uslova se moraju ispuniti, a to je obezbeđivanje električne energije, iz mreže ili baterija i pristup jezgru telekomunikacione mreže, fizičkim vodovima kao što je optika ili bežičnim vezama kao što su radio-relejne ili satelitske veze.

Treba napomenuti da i državne organizacije za sprovođenje zakona (npr. policija, vojska) mogu koristiti PBS za prikupljanje obaveštajnih podataka.

Primarno, PBS su privremeno rešenje, ali po potrebi mogu se transformisati i u stalne bazne stanice [1]-[7].

II. SCENARIJI UPOTREBE PBS

Postoje dve primarne situacije kada su PBS od velike koristi za društvo [1]-[7]:

- Povećanje kapaciteta mreže kada se organizuju događaji sa većim brojem korisnika privremeno i sporadično na nekom području.
- Obezbeđivanje komunikacije u slučajevima katastrofa kada je osnovna infrastruktura oštećena ili uništena.

Poslednjih godina, pokrivenost područja i kapacitet mobilnih mreža je konstantno u porastu, posebno u gusto naseljenim sredinama. Međutim, u manje naseljenim područjima pristup brzim mobilnim mrežama ne prati ovaj trend. U stvari, operatori mobilnih mreža se sada suočavaju sa zadatkom da definišu ekonomski održive arhitekture mreže za primenu servisa velikih brzina u prigradskim ili ruralnim zonama. Takve arhitekture nisu nužno trajne i mogu se primeniti samo kada za njima postoji potreba. Kao rezultat toga, organizatorima masovnih događaja koji će se odvijati u područjima sa ograničenim kapacitetom mreže potrebna su fleksibilna rešenja koja bi omogućila povezivanje privremenim korisnicima, kao što su učesnici i sami organizatori. Takva privremena usluga se može koristiti za različite događaje, uključujući okupljanja na otvorenom, konferencije i seminare, gradilišta, festivale i sportske događaje. Komunikaciona rešenja za ove slučajeve mogu se oslanjati na privremene mobilne mreže, odnosno prenosne bazne stanice čime bi se zamenila ili dopunila već postojeća mrežna infrastruktura [2].

Nakon katastrofalnih događaja, kao što su vremenske nepogode (npr. poplava, oluja) ili katastrofe većih razmera (npr. zemljotres, tornado, cunami), različite spasilačke ekipe treba što pre da deluju kako samostalno tako i zajedno i sinhrono na ugroženom području. U cilju spasavanja, ključno je obezbediti pouzdane i interoperabilne komunikacione sisteme. U ovom kontekstu, kao bitne stvari koje treba obezbediti jesu komunikacije za one koji treba prvi da reaguju (FRC – *First Respoder Communication*), sistemi podrške za obnovu kritične infrastrukture, nadzor nakon katastrofe, mreže medicinskih usluga, itd. Međutim, postojeće mreže su možda oštećene ili uništene. Na primer, bazne stanice ćelijskih mobilnih mreža su možda bile pogođene zemljotresom ili cunamijem, kao i nestankom struje izazvanim višestrukim uzrocima kao što su teški vremenski događaji, uključujući poplave i uragane. Neke katastrofe ili vanredni događaji mogu se, u izvesnoj meri, predvideti. U svakom slučaju, oni koji prvi reaguju oslanjaju se na komunikacione uređaje, opremljene višestrukim senzorima i heterogenim primopredajnicima kako bi podržali

aplikacije koje zahtevaju sve veći propusni opseg, uključujući video strimovanje u realnom vremenu ili razmenu velikih količina podataka (npr. slike visoke rezolucije, medicinsko praćenje i praćenje uslova u životnoj sredini). Ovo postavlja snažan i izazovan zahtev službama spašavanja za pouzdanom i skalabilnom komunikacionom infrastrukturom kako bi se obezbedila pokrivenost mreže, malo kašnjenje i veliki kapacitet, kao i interoperabilnost sa drugim radio tehnologijama koje su dostupne u okruženju. U ciljanoj arhitekturi sistema, tipične aplikacije za javnu bezbednost mogu biti dvosmerna govorna komunikacija, poludupleks video konferencije, video strimovanje u realnom vremenu, masovni prenos podataka, elektronska pošta, pristup Internetu, LTE *Push-To-Talk* (*Long Term Evolution*), VoIP i 5G servisi [2].

III. KARAKTERISTIKE I TIPOVI PBS

PBS se sastoje od tri glavne komponente:

- Antenski sistem – omogućava pokrivanje signalom u određenom području.
- Oprema za obradu signala – uključuje radio opremu i kontrolne jedinice.
- Izvor napajanja – može iz energetske mreže ako je dostupna, baterijski, solarnih panela ili pomoću generatora.

PBS su dizajnirane da budu kompaktne, lako prenosive i energetski efikasne. Prenosne bazne stanice za povezivanje sa korisnicima koriste iste tehnologije kao i fiksne bazne stanice:

- 2G, 3G, 4G (LTE), 5G,
- mogu da koriste i MIMO (*Multiple-Input Multiple-Output*) i *beamforming* za poboljšanje jačine signala i efikasnosti,
- omogućavaju visokokvalitetni prenos podataka i podršku za veliki broj uređaja.

Postoji nekoliko mogućih rešenja za brzo postavljanje prenosnih baznih stanica [1]-[7]. Izbor optimalne mrežne konfiguracije, koja se sastoji od komunikacione platforme izabrane iz tri segmenta, zavisi od parametara koje diktira scenario u kojem se PBS implementira, kao što su veličina područja, pristupačnost, komunikacioni zahtevi i drugo. Takođe, različito se planiraju resursi u zavisnosti od vremena potrebnog za uspostavljanje komunikacije. Ukoliko se radi o događajima koji su unapred poznati (sajmovi, sportski događaji) tada je na neki način lakše postaviti prenosne bazne stanice i tada se uglavnom biraju prenosne stanice koje se instaliraju na zemlji. U slučaju vanrednih situacija PBS mogu biti postavljane na zemlji ili različitim tipovima vazduhoplova mogu biti transportovane iznad područja. Pored ova dva tipa PBS, za komunikaciju se može koristiti i komunikaciona oprema koja se nalazi na satelitima visokih i niskih orbita uz ograničenja koja satelitska komunikacija poseduje.

Zemaljske prenosne bazne stanice mogu se podeliti na [7]:

- *Rapid-Deployement Units* (RDU)
- *Cell on Wheels* (CoW)

RDU su prenosne bazne stanice većega gabarita (Sl. 1) koje se transportuju većim kamionima ili u okviru standardnih kontejnera koji vuče kamion [7]-[10].

Slika 1. RDU – izgled i postavljanje [7]

Osnovne karakteristike ovih PBS jesu:

- Njihova upotreba je strateška za brzo širenje ćelijskih mreža.
- Mogu imati velike mogućnosti jer su većega gabarita od ostalih.
- Poseduju rešetkasto montažno-demontažni stub sa merdevinama koji nosi jednu ili više antena – stub se višestrukim sajlama povezuje sa betonskim tegovima kako bi se osigurala stabilnost i povećala otpornost na udare vetra.
- Brzo i lako se montiraju u ograničenim prostorima – do 8 sati je potrebno za montažu.
- Ne zahtevaju se građevinske dozvole, ni građevinske radove i temelje – balastni betonski tegovi se koriste za ankerisanje kontejnera i stuba.
- U kontejneru se nalazi pribor za montažu, baterije, agregat, klima sistem, električna i radio oprema bazne stanice.
- Veza sa okosnicom telekomunikacione mreže se može ostvariti zemaljskim vodovima i zemaljskim i satelitskim radio linkovima.
- Može se postaviti više ovakvih PBS (Sl. 2) u situacijama kada se očekuje veliki broj korisnika [10].

Slika 2. Primena više RDU-a [10]

CoW predstavlja prenosnu baznu stanica koja se nalazi u manjem kamionu koji prevozi teret do 3.500 kg, kombi vozilu (Sl. 3) ili se transportuje na manjoj prikolici (Sl. 4) [7], [13]-[16]. Osnovne karakteristike CoW jesu:

- Garantuju pun rad u samo jednom danu u ograničenim prostorima – u većini slučajeva je potrebno do nekoliko sati, ne retko i manje od jednog sata za postavljanje.
- U odnosu na RDU, transport je brži i lakši u težim uslovima, kao što su katastrofe, imajući u vidu da se prenose mnogo manjim vozilima.
- Hidraulični sigurnosni stabilizatori, koji se brzo izvlače i uvlače, obezbeđuju stabilnost PBS.
- Aluminijumski hidraulični stub koji nosi antene je teleskopskog tipa, a sastoji se od više cevastih elemenata. Stub se podiže ručnim ili električnim vitlom.
- Hidraulični stub može biti različitih visina, do oko 20 m.
- Komunikacija sa jezgrom mreže ostvaruje se radio, satelitskim ili žičnim putem.

Slika 3. CoW u namenskom vozilu [14]

Slika 4. CoW koja se vuče ili transportuje na prikolici [16]

PBS koje se prenose vazduhoplovima mogu se podeliti na:

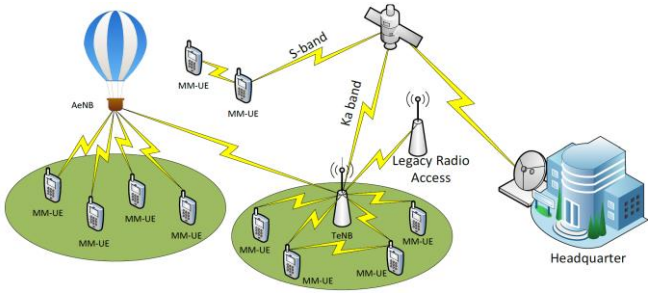
- PBS koje se prenose vazdušnim platformama – razlikuju se platforme na manjim (LAP – *Low Altitude Platform*) i platforme na većim visinama (HAP – *High Altitude Platform*). LAP platforme su na visinama do recimo 5 km, dok su HAP platforme koje se, prema IEEE (*Institute of Electrical and Electronics Engineers*), nalaze na visinama 20-50 km [17].
- PBS koje transportuju dronovi i bespilotne letelice.

PBS koju prenosi balon-zmaj (*kytoon – kite balloon)* punjen helijumom (*HeliKite*, Sl. 5) je jedan primer LAP platforme [2]. *HeliKite* predstavlja letelicu koja je privezana za zemlju. Letelica deo svog uzgona dobija dinamički pomoću dela koji ima oblik zmaja koji je teži od vazduha, a ostatak uzgona ostvaruje aerostatski pomoću balona punjenog helijumom koji je lakši od vazduha. Primarna prednost *HeliKite*-a je ta što ostaje u razumno stabilnom položaju iznad tačke vezivanja, bez obzira na jačinu vetra, dok su obični baloni i zmajevi manje stabilni. *HeliKite* se koristi u mnoge svrhe, i civilne i vojne, pa tako može biti korišćen i za prenos PBS. U zavisnosti od meteoroloških uslova može se nalaziti na visinama od 300 m do 4 km i može da nosi 10 kg tereta. U vazduhu može da ostane nekoliko dana, a na zemlji se od oštećenja čuva na svojoj bazi na kojoj se ujedno dopunjuje helijumom (Sl. 6) [18]-[22].

Slika 5. Balon-zmaj punjen helijumom [18]

Slika 6. Balon-zmaj i baza za čuvanje i punjenje helijumom [20]

Sl. 7 ilustruje predloženu arhitekturu implementacije za pružanje hitne i privremene komunikacije. Ova mreža treba da bude otporna, da podržava širokopojasne aplikacije i da bude bezbedna. Pored toga, mrežni servis će biti prilagođen vrsti katastrofe ili privremenog događaja. Štaviše, arhitektura mora da podržava brzu konfiguraciju i primenu širokopojasne mreže. Ovo se postiže dizajnom dva međusobno čvrsto povezana segmenta: vazdušnog segmenta, koji se sastoji od LAP-ova, i zemaljskog segmenta, omogućenog višestrukom kooperativnom zemaljskom opremom [2].

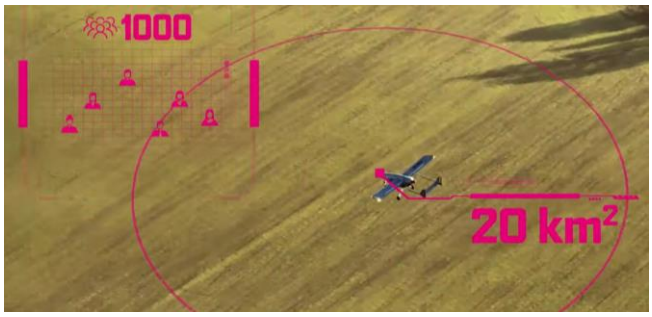


Slika 7. Arhitektura brze implementacije za pružanje širokopojasnog pristupa [2]

Kao primer PBS koja je instalirana na bespilotnoj letelici navešće se njena upotreba od strane Deutsche Telekom-a. Po prvi put, Deutsche Telekom je koristio bespilotnu letelicu (UAV – *Unmanned Aerial Vechicle*) da obezbedi pokrivenost mobilnom mrežom korisnicima na komercijalnoj mreži uživo. Sa visine od 2,3 km (maksimalna moguća visina letelice je 3 km), sa integrisanom mobilnom baznom stanicom obezbedila je pokrivenost na snežnim stazama na legendarnoj trci u skijaškom trčanju „Jizerska 50", u Češkoj, 2025. god. [23]-[25].

Letelica *Primoco One 150* (Sl. 8), koja prenosi baznu stanicu, dizajnirana je i proizvedena u Češkoj. Letelica je dugačka 3,65 m, visoka 1,25 m i ima raspon krila od 4,85 m i može da nosi 10 kg tereta, a u ovom primeru je transportovana PBS koja je težila nekoliko kg. Ugrađena bazna stanica na zemaljsku okosnicu može da se poveže radio vezom vazduh-zemlja, kao i preko satelita. Sistem omogućava *download* brzine do 200 Mbit/s i brzine do 75 Mbit/s u *upload* smeru. Istovremeno može da poveže 1000-1200 korisnika i da pokrije područje od 20 km^2 (Sl. 9).

Slika 8. Letelica *Primoco One 150* [23]



Slika 9. Područje pokrivanja signalom [25]

Letelica se kontroliše iz kontrolne stanice koja se nalazi u kombi vozilu (Sl. 10). Na krovu kombi vozila nalaze se dve antene, jedna za komunikaciju sa letelicom, a druga za satelitsku vezu za jezgrom mreže.

Slika 10. Kontrolna stanica u kombi vozilu [25], [23]

Kao još jedan primer korišćenja PBS može se navesti Norveška telekomunikaciona kompanija „Telia Norge AS" koja je 2024. god. testirala PBS tokom godišnje vežbe službi spašavanja [26]-[28]. Korišćene PBS bile su postavljene na prenosni stub. Stub i sva oprema mogu se preneti terenskim vozilom (Sl. 11 i Sl. 12). PBS je kompaktna, male težine i napaja se pomoću baterija, sa ostatkom mreže može da se poveže i pomoću satelita ukoliko drugi način nije dostupan. Podržava 4G i 5G, *network slicing* i višestruke virtuelne mreže. U drugom slučaju za potrebe pokrivanja područja drvne industrije (seča stabala u nepristupačnim područjima) Telia je koristila 5G PBS koja je prenošena dronom (Sl. 13) [29]-[31].

Slika 11. Vozila koja prenose PBS opremu kompanije Telia [27]

Slika 12. Prenosni stub koji nosi PBS [28]

Slika 13. PBS na dronu [29]

IV. PROBLEMI I IZAZOVI PRIMENE PBS

PBS imaju niz prednosti kao što su brza implementacija, fleksibilnost u različitim okruženjima.

Međutim implementacija PBS suočava se i sa izazovima i problemima među kojima su [32]-[39]:

- regulatorni izazovi,
- tehnički izazovi, i
- društveni izazovi.

A. *Regulatorni izazovi*

Postupak odobravanja postavljanja PBS nekada može biti kompleksan, ali je svakako mnogo lakši od postupka za dobijanje dozvole fiksnih baznih stanica. Kompleksnost može zavisiti od perioda primene PBS i lakše je ukoliko se radi o kratkotrajnoj primeni. Postupak i dokumentacija koja se zahteva zavisi od svake lokalne samouprave.

Neophodno je obezbediti saglasnost sa lokalnim vlastima za korišćenje frekventnog spektra kako ne bi dolazilo do interferencije sa drugim sistemima.

PBS i intenzivno korišćenje mobilnih telefona (npr. sajamska manifestacija) povećava elektromagnetno zračenje, te bi dobro bilo da postoji procena koliko je zračenje, odnosno da se obezbedi da se ne prelaze prihvatljive vrednosti. Nekada će se možda zahtevati procena uticaja na životnu sredinu. Iako zakoni ne zahtevaju uvek izradu studije o proceni uticaja na životnu sredinu za svaku baznu stanicu, PBS sa snagom većom od 250 W podložne su ovim pravilima, a odluka o potrebi izrade takve studije prepuštena je lokalnim vlastima.

U regulatorne izazove spadaju i urbanistička ograničenja. Ova ograničenja mogu uključivati minimalne udaljenosti od stambenih objekata, bolnica, škola, vrtića, što dodatno komplikuje proces postavljanja stanica u urbanim sredinama.

B. *Tehnički izazovi*

Kada je o tehničkim izazovima reč treba voditi računa o energetskoj efikasnosti i obezbeđivanju napajanja što može biti problematično u udaljenim ili ruralnim područjima gde infrastruktura nije razvijena.

Interoperabilnost baznih stanica može biti problem jer PBS moraju biti kompatibilne sa različitim mobilnim uređajima i mrežnim standardima, a to može zahtevati dodatna ulaganja i prilagođavanja.

PBS moraju biti projektovane za laku instalaciju i transport i neretko i pored frekventne usklađenosti sa lokalnom upravom treba voditi računa o izbegavanju interferencije sa postojećim sistemima.

PBS po pravilu imaju manji kapacitet u odnosu na fiksne sisteme, a troškovi održavanja mogu biti veliki u ekstremnim uslovima.

C. *Društveni izazovi*

Otpor javnosti i pogrešna percepcija rizika spadaju u društvene izazove primene PBS. Zabrinutost građana u vezi sa potencijalnim zdravstvenim rizicima od elektromagnetnog zračenja, nedovoljna informisanost i strah od nepoznatog može da dovode do otpora prema instalaciji PBS. Informacije o nejonizujućem zračenju često su pogrešno interpretirane, što dodatno doprinosi strahu i otporu lokalnog stanovništva prema novim tehnologijama.

V. ZAKLJUČAK

Prenosne bazne stanice predstavljaju ključno rešenje za brzu i fleksibilnu implementaciju mobilnih mreža. Njihova primena u hitnim situacijama, ruralnim područjima i na masovnim događajima pokazuje njihov ogroman potencijal. Sa daljim razvojem tehnologije, PBS će biti sve važnije u obezbeđivanju globalne povezanosti. Sa razvojem 5G tehnologije, PBS će postati još važnije zbog potrebe za većom gustinom mrežne infrastrukture. Takođe, integracija sa IoT (*Internet of Things*) i AI (*Artificial Intelligence*) otvara nove mogućnosti za primenu PBS.

ZAHVALNICA

Ovaj rad podržan je od strane Fakulteta tehničkih nauka u Novom Sadu, Departmana za energetiku elektroniku i telekomunikacije, u okviru projekta pod nazivom „Razvoj i primena savremenih alata i metoda istraživanja u energetici, elektronici i telekomunikacijama”.

LITERATURA

[1] Z. Zhang, R. Stanica, F. Valois, “Movable Base Stations in Mobile Networks for Emergency Communications”, PIMRC 2023 - IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications, Toronto, Canada, Sep 2023.

[2] A. Valcarce, T. Rasheed, K. Gomez, et al., “Airborne Base Stations for Emergency and Temporary Events”, Lecture Notes of the Institute for Computer Sciences, June 2013.

[3] J. Košmerl, A. Vilhar, “Base Stations Placement Optimization in Wireless Networks for Emergency Communications”, ICC'14 – W11: Advances in Public Safety and Emergency Communications, 2014.

[4] H. Shinbo, Y. Kunisawa, T. Sakai, Y. Kitatsuji, A. Endo, K. Tanaka, “Flying Base station for Temporary Mobile Communications in an Area Affected by a Disaster”, 5th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), 4-7 Dec. 2018.

[5] C. D. Gavrilovich, “Mobile communication system with moving base station”, Unitet States Patent, Patent No.: US 8,463,177 B2, Jun. 11, 2013.

[6] “Temporary Facilities During Emergencies”, Australian Goverment, Department of Infrastructure, Transport, Regional Development and Communications, October 2020.

[7] "Mobile cell sites", Wikipedia, https://en.wikipedia.org/wiki/Mobile_cell_sites, pristupljeno Februar 2025.

[8] Communication Trailers, Mini Cell on Wheels for Telstra, ICS Industries, EnerSys Company, 2022, http://www.icsindustries.com.au/projects/communication-trailers/mini-cell-on-wheels-for-telstra-emergency-response-fleet, pristupljeno Februar 2025.

[9] "RDU (Rapid Deployment Unit)", G-Enviro, http://www.g-enviro.com/Products_cat.php?id_sec_main=9&id_sec_sub=10&RDU (Rapid%20Deployment%20Unit), pristupljeno Februar 2025.

[10] "HDC (Heavy Duty Cow)", G-Enviro, http://www.g-enviro.com/Products_in.php?id_art=61&id_sec_main=9&id_sec_sub =23&G-env-HDC0-SH-35-00, pristupljeno Februar 2025.

[11] "Cell Site on Wheels (Pneumatic or Motorized Lattice): aka COWs With Shelters", Telecom Site Solutions, LLC, http://www.telecomsitesolutions.com/CellSiteonWheels.htm, pristipljeno Februar 2025.

[12] Brochure "Wireless connectivity that deploys fast and deploys just about anywhere", Temporary Wireless Solutions, Commscope, Ruckus, 2020.

[13] "Cell on Wheels", Engineered for all Environments, EnerSys Company, https://www.icsindustries.com.au/products/communication-trailers/cell-on-wheels, pristupljeno Februar 2025.

[14] "Cell Site on Light Trucks (aka COLTS)", Telecom Site Solutions, LLC, http://www.telecomsitesolutions.com/CellSiteonLightTrucks.htm, pristupljeno Februar 2025.

[15] "CommScope Introduces Wireless Rapid Deployment Unit for On-Demand Wireless Connectivity", CommScope, July 7, 2020, https://www.commscope.com/press-releases/2020/commscope-introduces-wireless-rapid-deployment-unit-for-on-demand-wireless-connectivity/, pristupljeno Februar 2025.

[16] Brochure "Communication Base Station", Engineered for all Environments, EnerSys Company

[17] "Aerial Base Station", Wikipedia, https://en.wikipedia.org/wiki/Aerial_base_station, pristupljeno Februar 2025.

[18] "Helikite Airborne Surveillance Systems", Allsopp Helikites LTD, https://www.helikites.com/aerial-surveillance-aerostats, pristupljeno Februar 2025.

[19] "Helikite Aerostats", Airborne Communications Ltd, https://www.airbornecomms.com/helikite-aerostats, pristupljeno Februar 2025.

[20] "Helikite training in very British weather", EPFL, https://www.epfl.ch/labs/eerl/helikite-training-in-very-british-weather/, pristupljeno Februar 2025.

[21] "Allsopp Helikite", Wikipedia, https://en.wikipedia.org/wiki/Allsopp_Helikite, pristupljeno Februar 2025.

[22] "Kytoon", Wikipedia, https://en.wikipedia.org/wiki/Kytoon, pristupljeno Februar 2025.

[23] A. Sanchez, "Deutsche Telekom uses drone as flying base station for temporary coverage", February 21, 2025, https://www.telekom.com/en/media/media-information/archive/deutsche-telekom-uses-drone-as-flying-base-station-for-temporary-coverage-1088440, pristupljeno Februar 2025.

[24] Primoco UAV, Wikipedia, https://en.wikipedia.org/wiki/Primoco_UAVa, pristupljeno Februar 2025.

[25] "First in EU: T-Mobile and Primoco UAV SE introduce mobile signal boost by UAV drone — to be deployed for the first time in the Jizerska 50 cross-country ski race", sUAS News, January 2025, https://www.suasnews.com/2025/01/first-in-eu-t-mobile-and-primoco-uav-se-introduce-mobile-signal-boost-by-uav-drone-to-be-deployed-for-the-first-time-in-the-jizerska-50-cross-country-ski-race/, pristupljeno Februar 2025.

[26] Srikapardhi, Telia, "Norwegian People's Aid Test Portable Mobile Base Stations via LEO Connectivity", March 25th 2024, TelecomTalk, https://telecomtalk.info/telia-test-portable-mobile-base-stations-leo/943550/, pristupljeno Februar 2025.

[27] "Norwegian People's Aid tests portable mobile base stations from Telia", Telia Company, March 25th 2024, https://www.teliacompany.com/en/news/norwegian-peoples-aid-tests-portable-mobile-base-stations-from-telia-2024-03-25-06-00-00, pristupljeno Februar 2025.

[28] "Telia tests portable mobile base station during annual rescue services exercises in Norway", June 2024, https://www.teliacompany.com/en/news/telia-tests-portable-mobile-base-station-during-annual-rescue-services-exercises-in-norway-2024-06-13-12-00-00, pristupljeno Februar 2025.

[29] N. Wood, "Telia puts 5G base station on drone and goes logging", June 2024, https://www.telecoms.com/5g-6g/telia-puts-5g-base-station-on-drone-and-goes-logging, pristupljeno Februar 2025.

[30] P. Lipscombe, "Telia uses drones to expand 5G network for logging industry", DCD, June 21, 2024, https://www.datacenterdynamics.com/en/news/telia-uses-drones-to-expand-5g-network-for-logging-industry/, pristupljeno Februar 2025.

[31] H. Kadia, "Drones and Telia's Private 5G Enable Remote Forestry Control in Sweden", TeckNexus LLC, June 23, 2024, https://tecknexus.com/5gusecase/drones-and-telias-private-5g-enable-remote-forestry-control-in-sweden/, pristupljeno Februar 2025.

[32] White Paper, C. Meyer, H. Basilier, "Ensuring critical communication with a secure national symbiotic network" Ericsson, December 2021.

[33] White Paper, "Radio waves and health – Base station", Ericsson, 2013.

[34] R. Ramirez-Vazqueza, J. Gonzalez-Rubiob, E. Arribasc, A. Najerad, "Personal RF-EMF exposure from mobile phone base stations during temporary events", Journal Environmental Research 175, Elsevier, pp. 266-273, 2019.

[35] D. Čukić, M. Matić, "Analiza regulatornog okvira i prakse za izgradnju mreže baznih stanica mobilne telefonije", NALED, jul 2021.

[36] Rezultati ankete sa lokalnim samoupravama, "Uslovi za postavljanje baznih stanica mobilne telefonije na lokalu", NALED, Savez za imovinu i investicije, Savez za e-upravu, mart 2021.

[37] S. H. Alnajjar, F. Malek, M. S. Razalli, M. S. Ahmad, "Low-Altitude Platform to Enhance Communications Reliability in Disaster Environments", Journal of Advances in Information Technology, vol. 5, no. 1, pp. 21-30, February 2014.

[38] K. Dervić, "RF zračenje baznih stanica", KesatNet, https://kesatnet.me/rf-zracenje-baznih-stanica/, pristupljeno Februar 2025.

[39] H. Wurzburg, "Simple Portable RTK Base Station", July 2022, https://discuss.ardupilot.org/t/simple-portable-rtk-base-station/87694, pristupljeno Februar 2025.

Advantages and Challenges of Using Portable Base Stations In Mobile Systems

Dejan Nemec

ABSTRACT

Portable Base Stations (PBS) have become a key element in modern mobile networks, enabling rapid deployment of network infrastructure when needed. This paper explores the application scenarios of portable base stations, such as major events, rural areas, and in disaster situations. The characteristics and types of portable base stations are listed, i.e. the possibilities of transporting PBS equipment to the area of interest. The paper outlines the advantages that PBS provides, and in particular, analyzes the problems and challenges that accompany the implementation of PBS.

YU #5: Sesija 5

Informacioni sistemi, računarske simulacije i edukacija

Миграција и конверзија SAP ЕРП-а на S/4 HANA технолошку платформу

Миодраг Гачић
ЕПС АД
Београд, Србија
miodrag.gacic@eps.rs

Наташа Ђокић
ЕПС АД
Београд, Србија
natasa.djokic@eps.rs

Милан Милојевић
ЕПС АД
Крагујевац, Србија
milan.milojevic@eps.rs

Циљ овог рада је да се компанијама које тек планирју да пређу на S/4HANA технолошку платформу упознају са корацима и проблемима са аспекта корисничког искуства. Постоји доста корисничке документације, али направити праве кораке у правом смеру није лако. Желели смо да са корисничког угла укажемо на искустава, изазове и технологије за тај нови корак. Није лако одабрати праве алате, ит стручњаке за овај посао. Потребан је и временски оквир за реализацију оваквих пројекта који може да буде и повећан због лоше анализе и припреме конверзије и миграције. Неко искуство говори да је за овакве пројекте потребно минимум 12 месеци ако су САП системи дуже у употреби, и ако систем поседује велики број развијеног корисничког кода. Прелазак на нова решења није једноставан и лак процес, јер укључује различите тимове специјалиста и пословних корисника. Мења се и пословна логика коју компанија мора да промени да би ефекат и предности нових технологија биле искоришћене. Све то повећава ефикаснист и модрнизацију компаније којима се добија и нова вредност, али и пословна сигурност компаније. Нова технологија омогућава компанијама доношење пословних одлука заснованих на подацима у реалном времену. Подаци су у меморији, што омогућава и неке BI анализе и извештаје, за које би требало развијати посебне извештајне системе.

Кључне речи – S/4 HANA , ЕРП систем изграђен на САП Хана платформи у меморији

1. УВОД

САП ЕРП као централизовано ЕРП решење великог броја компанија се састоји од више пословних модула: финансијско пословање ФИ,магацинско пословање ММ, контролинг ЦО, продаја СД, управљање људским ресурсима – ХЦМ. То је пословни пакет нове генерације који је дизајниран за једноставн рад у окружењу дигиталне технологије.

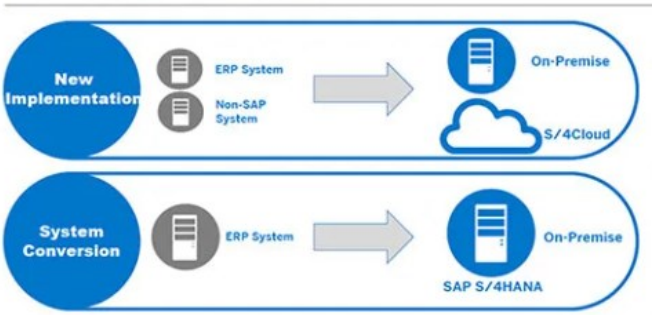
Разлози за прелазак на нову верију САП-а су вишеструки:

- Нове технологије
- Нови савременији кориснички интерфејси
- Нова правила безбедности и заштите података
- САП подршка за старе системе је доступна до 31.12.2027.године
- Нови поједностављен модел података са новим релацијама и везама
- Лакше интеграције са новим технологијама и новим системима
- Нове пословне функционалности које се развијају само за S/4 HANA, а не и за старе верзије САП-а

2. НОВА ТЕХНОЛОШКА ПЛАТФОТМА

За прелазак на нову технолошку платформу постоје два пута:

- Нова имплементације - где се подаци само мигрирају из старог система у нови систем
- Миграција и конверзија постојећег ЕРП система у нови S/4 HANA систем



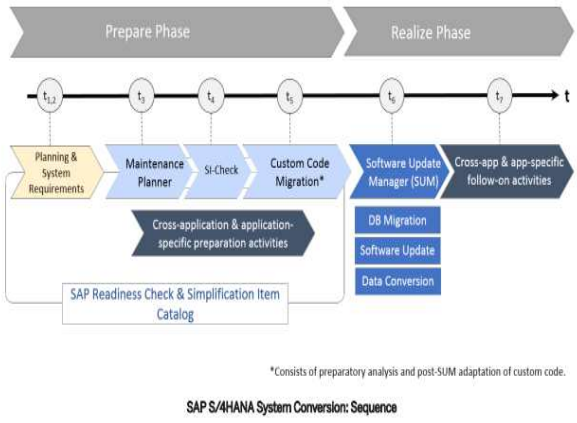
Слика: Приказ начина преласка на нову платформу

Концепт миграције и конверзије на нову технолошку платформу S/4 HANA обухвата конверзију и миграцију старог система у нови систем са новим паравилима. Мења се целокупан модел базе података, са свим релацијама и корисничким интерфејсом. То утиче на пословне процесе који треба да испрате ове промене.
У тај нови модел је потребно уклопити старе податке, и наставити са радом компаније. Потребно је променити пословну логику да би се пословање прилагодило новом моделу података. У старим системима постоје наслеђени програми развијени за потребе пословних корисника. Све те програме треба прилагодити новом моделу података. За то је потребан тим састављен од техничких лица, пословних корисника, консултаната и САП

подршке за евентуалне проблеме у конверзији. Сарадња различитих тимова је неопходна за успешан прелазак на нову технолошку платформу.
Захваљујући концепту табела у меморији, и чувању на начин на који то чине програми за компресију, нови систем омогућава обављање свих анализа у реалном времену. Резултат тога је да једна табела , која је у традиционалном смислу имала 100 MB података са припадајућим индексима, сада заузима свега 10 MB података доступних на великим брзинама RAM меморије.
Конверзија и миграција на нову технолошку платформу се обавља у три фазе:

- Припрему конверзије и миграције САП система
- Реализацију конверзије и миграције САП система
- Оптимизацију коначног решења

Графички приказ корака миграције и конверзије на нову технолошку платформу



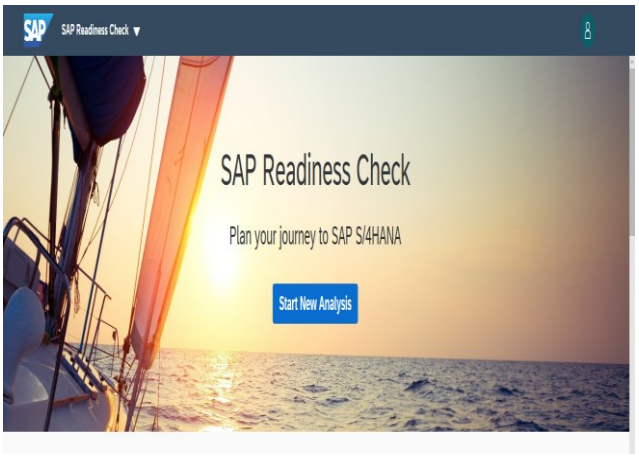
SAP S/4HANA System Conversion: Sequence

3. ПРИПРЕМЕ ПРОЈЕКТА И МИГРАЦИЈЕ НА НОВУ ТЕХНОЛОШКУ ПЛАТФОРМУ

Фаза припреме је најзначајнија, и најсвеобухватнија фаза у миграцији и конверзији на SAP S/4HANA технолошку платформу. Она захтева уску сарадњу различитих тимова и професија: аналитичари, програмери, техничка лица, пословни корисници, али и менаџмента компаније, с обзиром да се мењају и пословни процеси. Пословне одлуке морају да прате технолошке промене. Технолошке промене намећу и промене пословних процеса које је потребно ускладити. Компанија постаје ефикаснија и технолошки напреднија.
Планирање
У овој фази је потребно да се стекне слика о свим корацима и неопходним припремама за сам пројект миграције и конверзије. За планирање пројекта препорука САП-а је да се користи алат Roadmap Viewer где се виде svi кораци који су неопходни за прелазак на нову технолошку платформу, и алат SAP Readiness Check за SAP S/4HANA.

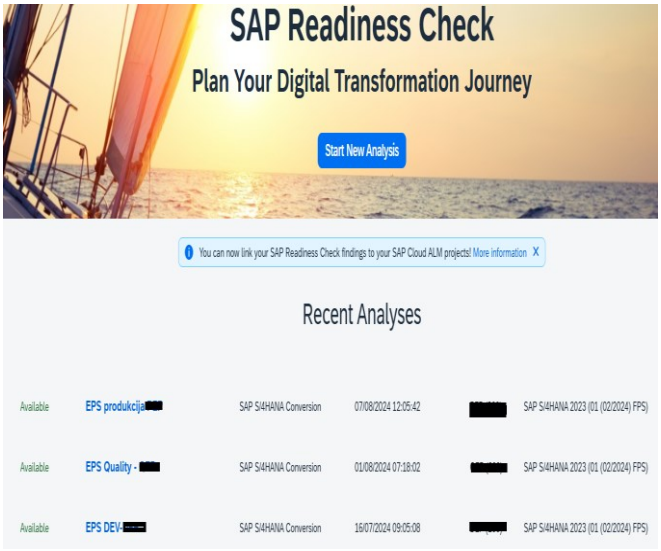
Захтеви хардвера
У зависности од величине базе и броја корисника одређује се величина хардвера. У овом случају је планирање ресурса у сопственом дата центру – привате цлоуд. Постоји и алат који за одређене улазне параметре може да препоручи будућу величину система.

Maintance planer
Maintance planer се користи као алат и први корак у фази конверзије. Алат проверава компоненте система, инсталиране add-in, као и пословне функције како би се осигурала компабилност старог са новим системом. Генерише stack fajl који се користи приликом стварне конверзије система.
Анализа развијеног кода
Фаза анализе развијеног (“home made”) кода је и кључна фаза да се осигура успешан прелазак на нову платформу. Овде се проверава код програма који је развијен од локалних програмера и консултаната независно од САП стандардних програма. У великим комапнијама има доста оваквог кода и то треба посебно мигрирати и тестирати. У новој верзији је промењена и пословна логика , па је потребно утвдити којим пакетима у новом систему припада развијен код, и јел има потребе за њим. Можда та решења постоје у новој верзији ЕРП-а. На тај начин се добија слика кода који није у складу са новим моделом базе података, и који је потребно редефинисати у складу са новим моделом. Ако тога има доста, то је захтеван и дуготрајан процес.
Припрема стандардних апликација
Ова фаза подразумева кораке на систему који морају да се обаве пре конверзије да би сама конверзија била успешна. Она обухвата брисање клијента 066 који у новом систему не постоји, инсталирање планера одговарајуће верзије, неке ноте за Солман систем, повезивање Солман система са старим и новим циљаним системом који су специфични за нову верзију САП-а. Након техничке конверзије са алатом СУМ потребно је конверовати и роле и ауторизације на новом систему. Препоручени концепт приступа функцијама SAP S/4HANA је преко SAP Fiori UKS са SAP Fiori Launchpad који се покрећу у претраживачу или SAP Business Client-u. Због тога је потребно прилагодити статри концепт ауторизација да подржи нови концепт менија концепту ауторизација SAP Fiori Lauchpad.
Data Transition Validation
Валидација транзиције података је алат који омогућава поређење пословних података пре и после конверзије из SAP ERP u SAP S/4HANA. Алат подржава и сценарије надоградње или конверзије за пословну валидацију података. Генерише пословне извештаје као инструменте поређења. У тим извештајима се виде подаци који су проблематични за сам ток миграције и конверзије.



Слика: SAP Readiness Check for SAP S/4HANA

На САП страни (САП Портал специфичан за корисника) се пушта посао који генерише документ, који се импортује на САП портал. Резултат је слика са генерисаним анализама и препорукама. Свакако је ово кључни део анализе, а самим ти и успешне миграције и конверзије на нови систем. Овај алат много помаже да се идентификују реални проблеми пре него сто се уђе у процес стварне миграције.
Потребно је напоменути да се мигрира и конвертује цело САП окружење (развој ,тест, продукција), а да се анализа ради за сваки систем посебно.



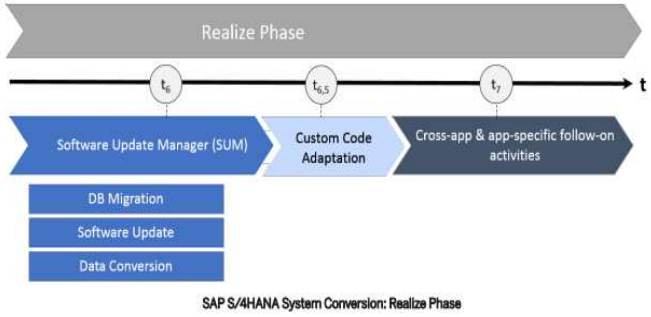
Слика: САП окружења sa које је урађена анализа

Слика: Резултати SAP Readiness Check-a

4. КОНВЕРЗИЈА И МИГРАЦИЈА НА НОВУ ТЕХНОЛОШКУ ПЛАТФОРМУ

За реализацију овог корака се користи алат Software Update Manager (SUM) који обавља три корака:

- Миграцију базе - опционо ако база није већ ХАНА већ неки други вендор
- Надоградњу софтвера – инсталација SAP S/4HANA софтвера
- Конверзију података - конверзија података у нову структуру коју користи koristi SAP S/4HANA



Слика: фазе реализације конверзије у нови систем

Након што је Softvare Update Manager (SUM) урадио техничку конверзију , може се почети конверзија кода који је направљен ван САП стандарда. Програмери и консултанти мењају код ван САП стандарда, као и правила пословних замена према новом пословном моделу базе података.
Потребно је да све модификације и надоградње користе стандардне САП трансакције SPDD, SPAU и SPAU_ENH. Морају се решити сви проблеми због промене модела базе података, и испоштовати препоруке из претходне анализе SAP Readiness Check for SAP S/4HANA.

The Obsolete Data Handling tool
Алат за обраду непотребних података нам омогућава да избришемо старе податке који могу остати након конверзије, а нема више потребе за тим подацима. То су случајеви који настају због промене модела подтака, и непотребних старих редуданси модела података – табела које се мењају приликом конверзије. Пословни подаци се мигрирају у нове структуре података. Ове промене су посебно изражене у САП ЕРП модулима управљању материјалима, финансијама, продаји и дистрибуцији. Изворни подаци се не бришу приликом конверзије, јер би то повећало временску дужину конверзије, и евентуално застој продукционих система. Зато се овај корак оставља за крај, када већ бизнис може нормално да функционише.

5. АЛАТИ ЗА МИГРАЦИЈУ И КОНВЕРЗИЈУ

Алати који су коришћени приликом миграције и конверзије на SAP S/4HANA технолошку платформу су:

- Roadmap Viewer

- SAP Readiness Check for SAP S/4HANA
- Maintance planer
- Data Transition Validation
- Software Update Manager (SUM)
- The Obsolete Data Handling tool

6. ЗАКЉУЧАК

S/4 HANA је пословни пакет следеће генерације САП-а. У потпуности је изграђен на најнапреднијој технолошкој платформи у меморији САП ХАНА. Користи модерне принципе дизајна са САП Фиори корисничким формама, као и нови модернији концепт корисничких форми заснованих на менијима. Са преласком на нову технолошку платформу компанија је добила следеће бенефите:

• Оптимизације апликација специфичне за САП ХАНА платформу

S/4 HANA, САП оптимизује апликацију како би на најбољи начин искористио могућности САП ХАНА-а база података у меморији. На пример, искључене су агрегације из модела података, и на тај начин је смањен и број података у бази

• Дизајн корисничког изгледа који одговара новим трендовима и технологијама

S/4 HANA, САП дизајнира апликацију са најновијим корисничким искуством заснованим на корисничким изгледима и формама (УКС).

• Обједињавање функционалности у моделу података

S/4 HANA, САП уклања редундантност претходног модела обезбеђујући једну функционалност за један пословни циљ компаније

S/4 HANA помаже компанијама да једноставно раде у дигиталној економији, укључујући такве алате као што су Internet of Things, Big Data, пословне мреже, Artificial Intelligence (AI) и кориснички интерфејси за мобилне апликације. Претходне верзије САП-а нису то дозвољавале. Омогућен је приступ корисничким формама са мобилних уређаја. Све то доприноси утиску модернизације и новом технолошком приступу целе компаније.

Литература:

Conversion Guide for SAP S/4HANA 202

SAP Readiness Check 2.0 - Setup & Execution https://help.sap.com/docs/SAP_S4HANA_ON-PREMISE?locale=en-US

Unapređenje performansi visokoškolskih institucija primenom analitike

Milica Škembarević, Marija Đukić, Gordana Savić, Nenad Aničić, Bisera Andrić Gušavac, Milena Popović, Danica Lečić-Cvetković
Fakultet organizacionih nauka, Univerzitet u Beogradu
Beograd, Srbija
milica.skembarevic, marija.djukic, gordana.savic, nenad.anicic, bisera.andric.gusavac, milena.popovic, danica.lecic-cvetkovic@fon.bg.ac.rs
0000-0003-0649-3005, 0000-0002-1136-4278, 0000-0002-6799-8691, 0000-0001-9438-5906, 0000-0002-2947-8054, 0000-0002-5016-9248

***Apstrakt* – Visokoškolske institucije se susreću sa novim izazovima u praćenju i analizi performansi učesnika u nastavnim i nenastavnim procesima. Tradicionalni informacioni sistemi visokoškolskih ustanova nemaju mogućnost direktne i sveobuhvatne analize performansi u okviru samog sistema niti definisane mehanizme za izračunavanje indikatora koji bi omogućili praćenje performansi kroz različite periode i na različitim nivoima. Cilj ovog rada jeste opis razvoja sistema za analitiku performansi, kroz definisanje i implementaciju mehanizama za praćenje ključnih indikatora za analitiku. Bazirano na analizi postojećih rešenja i identifikovanih potreba, napravljen je sistem koji podržava centralizovano učitavanje, ispitivanje i vizuelizaciju podataka uz pomoć Apache Superset alata. Razvijeno rešenje daje mogućnost analize performansi prema različitim kriterijumima i na različitim nivoima – od nivoa pojedinca do nivoa institucije i uz različite metode vizuelizacije adekvatno prilagođene podacima koji se prikazuju. Vizuelno predstavljene vrednosti indikatora u velikoj meri ukazuju na potencijalne pravce unapređenja kvaliteta obrazovanja i predstavljaju polaznu tačku pri donošenju dugoročnih strateških odluka.**

Ključne reči – akademska analitika, upravljanje nastavnim procesima, visoko obrazovanje, performanse studenata.

I. UVOD

Visokoškolske ustanove imaju veoma bitnu ulogu u razvoju i usavršavanju kadrova za tržište rada koje je zasnovano na znanju. Sa povećanjem upisnih kvota i rastućom potražnjom za određenim obrazovnim profilima, kao i usled naglog porasta raznovrsnosti zahteva na tržištu rada, obrazovne institucije se suočavaju sa izazovima u upravljanju svojim procesima. Poseban izazov predstavlja i praćenje indikatora efikasnosti koji proizilaze kao rezultat tih procesa.

Proces donošenja odluka, praćenje i predviđanje uspešnosti studenata i kvaliteta nastave koji se realizuju u okviru akademske institucije može biti olakšan korišćenjem alata akademske analitike [1, 2] kao i kreiranjem sistema za praćenje performansi. Akademska analitika, kao deo analitike obrazovanja, koristi kvantitativne metode deskriptivne, prediktivne i preskriptivne analitike pri kreiranju preporuka koje podržavaju donošenje odluka [3]. Postojeći sistem za evidenciju ne pruža mogućnost sveobuhvatnog praćenja i predviđanja trendova, kao ni analizu različitih faktora koji utiču na performanse studenata. Podaci su nestruktuirani, decentralizovani i raspoređeni u više različitih izvora. Dodatno, deo evidencije je dostupan isključivo u fizičkom obliku, jer još uvek nije digitalizovan. Uvođenje softverskog sistema za prikupljanje, obradu i analizu podataka može unaprediti upravljanje nastavnim procesima i donošenje strategija za poboljšanja kvaliteta obrazovanja [4]. Jasno predstavljeni i pravilno interpretirani podaci predstavljaju ključni resurs za akademske institucije koje teže inovativnosti i konkurentnosti. Indikatori performansi mogu pružiti uvid u rezultate rada nastavnika, strukturu i kvalitet studijskih programa, kao i uspešnost samih studenata. Korišćenjem istih indikatora, moguće je uporediti performanse i postaviti ciljeve razvoja na nivou predmeta, katedri, modula i programa. Izlazi iz ovakvog sistema korisni su različitim akterima u obrazovnom sektoru, uključujući ministarstva, akreditacione organizacije, univerzitete, rukovodstvo fakulteta, nastavnike i studente.

Cilj rada je unapređenje procesa praćenja performansi studenata kroz razvoj sistema za analitiku. Kroz rad treba odgovoriti na sledeća istraživačka pitanja:

1. Koji model analitike performansi je najpogodniji za praćenje uspešnosti studenata na različitim nivoima, od nivoa studenta do nivoa institucije?
2. Kako se performanse studenata mogu pratiti tokom vremena, upoređivati sa prethodnim rezultatima i koji su najefikasniji načini prikaza tih podataka?

II. PREGLED LITERATURE

Analitika je postala značajno prisutna u obrazovanju, što potvrđuje i razvoj posebne oblasti poznate kao akademska analitika. U [5] se ističe prelazak sa tradicionalnih metoda analize podataka ka naprednijim pristupima poput akademske analitike i rudarenja obrazovnih podataka (eng. *educational data mining*). Kao ključne prednosti ovih pristupa navode se unapređeno donošenje odluka, bolji ishodi učenja i efikasnija raspodela resursa. Međutim, autori ukazuju i na izazove poput zaštite privatnosti podataka, složenost integracije podataka, kao i potrebe za značajnim ulaganjima u infrastrukturu i stručni kadar. U tom kontekstu, predlažu uspostavljanje nove profesionalne uloge – Educational Data Scientist – koja podrazumeva spoj tehničkih znanja iz oblasti analize podataka i razumevanje obrazovnog sistema i njegovih specifičnosti. Autori [3] navode da je uspeh studenata ključan za ostvarenje akademskih ciljeva visokoškolskih institucija. Akademska analitika ima značajnu ulogu u ovom procesu, jer omogućava korišćenje podataka i prediktivno modeliranja za identifikaciju studenata koji su u riziku od akademskih poteškoća. Na taj način se otvara prostor za pravovremene intervencije koje mogu doprineti uspehu studenata, kao i povećanju stope diplomiranja. [4] navodi da su analitički projekti uglavnom usmereni na oblasti nastave i učenja, dok je njihova primena u administrativnom kontekstu znatno manje zastupljena. Autor predlaže proširenje obima prikupljanja podataka, kao i podsticanje korišćenja analitike

u administrativne kako bi se poboljšala institucionalna produktivnost i korišćenje resursa.

[6] su predložili razvoj inteligentnog sistema za pružanje povratnih informacija u vidu preporuka, sa ciljem unapređenja studentskih postignuća. Ovaj sistem automatski generiše intervencije zasnovane na analizi podataka o učinku studenata. Autori naglašavaju da je ključno da sistem bude u stanju da prepozna individualne potrebe studenata, ali i da omogući nastavnom osoblju zadržavanje kontrole nad predloženim preporukama.
Ovaj rad se oslanja na uvide i nalaze prethodnih istraživanja i prikazuje proces razvoja sistema za analitiku i praćenje performansi u visokoškolskom obrazovanju, na primeru Fakulteta organizacionih nauka. Razvoj sistema uključuje najpre definisanje relevantnih indikatora, a zatim i implementaciju samog sistema.

III. METODOLOGIJA

Razvoj sistema za analitiku i praćenje performansi visokoškolske institucije biće prikazan na primeru Univerziteta u Beogradu – Fakulteta organizacionih nauka. U početnoj fazi analizirani su analitički sistemi koji su razvijeni na drugim institucijama [7, 8, 9], kao i relevantna literatura [10, 11, 12]. Na osnovu sinteze prethodnih iskustava i specifičnih potreba fakulteta, definisan je početni okvir koji obuhvata šest modula. Za svaki modul definisane su osnovne grupe indikatora, opisan je način proračuna i praćenja definisanih indikatora, kao i koncept prikaza u okviru izveštaja. Ulazni podaci za sistem analitike Fakulteta organizacionih nauka potiču iz izveštaja studentske službe. Ovi izveštaji obuhvataju informacije o studentima, ispitnim rokovima, izlaznosti na ispite i ostvarenim rezultatima. Dodatno, izveštaju sadrže podatke o disciplinskim prijavama, kao i o povezanosti predmeta sa katedrama, odgovornim nastavnicima i akreditacionim periodima. Na osnovu ovih podataka izvedene su dodatne vrednosti relevantne za sistem analitike performansi, poput informacija o apsolventskim rokovima, dok su istovremeno kreirane procedure za proveru i prečišćavanje podataka. Za potrebe validacije postavljenih indikatora sprovedeno je istraživanje koristeći podatke o ispitnim rokovima iz dve školske godine, uključujući studente iz različitih akreditacionih perioda. Baza podataka obuhvata oko 160.000 zapisa. Razvijeni sistem omogućava pregled izveštaja i indikatora performansi, sa opcijama filtriranja i grupisanja prema nivou studija, demografskim karakteristikama, prethodnom obrazovanju i načinu finansiranja studija. Pored toga, sistem pruža mogućnost praćenja trendova i donošenja informisanih odluka zasnovanih na dostupnim podacima.

Podaci su skladišteni u PostgreSQL bazi podataka [13, 14], dok je za njihovu analizu i vizualizaciju korišćen alat Apache Superset [15]. Apache Superset je open-source alat namenjen analitičarima, koji omogućava naprednu analizu različitih indikatora na više nivoa, kao i kreiranje raznovrsnih grafikona i tabela. Zahvaljujući interaktivnom interfejsu i drag-and-drop funkcionalnosti prilikom izrade grafikona, alat je jednostavan za upotrebu i pristupačan i korisnicima bez tehničkog predznanja. U okviru komandnih tabli (eng. *dashboards*), moguće je prilagoditi prikaz podataka u skladu sa specifičnim potrebama analize. Za naprednije korisnike, Superset nudi i mogućnost direktnog pisanja složenijih SQL upita, čime se omogućava direktna pretraga baze podataka iz koje se učitavaju podaci [16].

IV. REZULTATI

U početnoj fazi razvoja informacionog sistema, na osnovu relevantne literature [17, 18, 19] definisan je opseg informacionog sistema koji će biti osnova za analitiku performansi. Definisano je četiri skupa indikatora performansi koji su u fokusu ovog istraživanja:

- Kvalitet upisa
- Kvalitet studenata
- Performanse nastavnika
- Performanse studijskog programa, modula ili predmeta

Za svaki od indikatora, opisano je njegovo izračunavanje, kao i način praćenja za svaki od indikatora. Primer definicije jednog od indikatora, stope odustajanja, dat je u nastavku:

$$\textit{Stopa odustajanja} = \frac{\textit{broj ispisanih kandidata posle prve godine studija}}{\textit{broj upisanih kandidata na prvu godinu studija}}$$

Kasnije je za svaki od indikatora definisano i u okviru kojih izveštaja i vizuelizacija se javlja i u kom obliku. U okviru naredne faze definisana je struktura baze podataka u koju će biti učitavani podaci dobijeni iz različitih izvora (studentske službe na različitim nivoima studija, komisije za upis, kadrovske službe, akreditacionih dokumenata itd.). Korišćena je PostgreSQL baza podataka. Podaci su podeljeni u okviru tabela koje se odnose na upis, studente, predmete, pripadnost katedrama i usklađivanje predmeta po različitim akreditacijama, polaganja u okviru ispitnih rokova, kao i disciplinske mere. Pre unosa u bazu podataka bilo je neophodno očistiti podatke (pretvoriti podatke iz izveštaja u odgovarajuće tipove, ukloniti ili zameniti nevažeće vrednosti, identifikovanje i razrešavanje situacija duplih unosa) i u nekim slučajevima spojiti podatke iz više izveštaja u jedan na osnovu identifikatora. Jedan deo dokumentacije nije bio digitalizovan (disciplinske mere) te je struktura izveštaja o disciplinskim merama osmišljena tako da sadrži informacije relevantne za prethodno definisane indikatore performansi.
Nakon kreiranja tabela u okviru šeme podataka i učitavanja podataka, javila se potreba za dodatnim poljima koje je moguće izvesti na osnovu postojećih vrednosti (kao što su indikatori o polaganjima u apsolventskim rokovima, pripadnost studenta generaciji), a koja su uvedena radi povećanja efikasnosti pretraživanja i filtriranja podataka u bazi. Za svaku od izvedenih vrednosti je definisan triger koji se poziva pri dodavanju ili izmeni odgovarajućih polja, iz kojih se vrednosti izvode, koji izračunava i upisuje vrednost u odgovarajuće kolone.
U okviru Apache Superset alata, povezane su tabele iz baze i kreirani odgovarajući izveštaji. Za neke od izveštaja je bilo neophodna definicija promenljivih u samom alatu kako bi bio omogućen prikaz prema dodatnim parametrima.

Glavni kriterijumi za filtriranje (koji se javljaju kao relevantni u većini definisanih indikatora) implementirani u okviru platforme su:

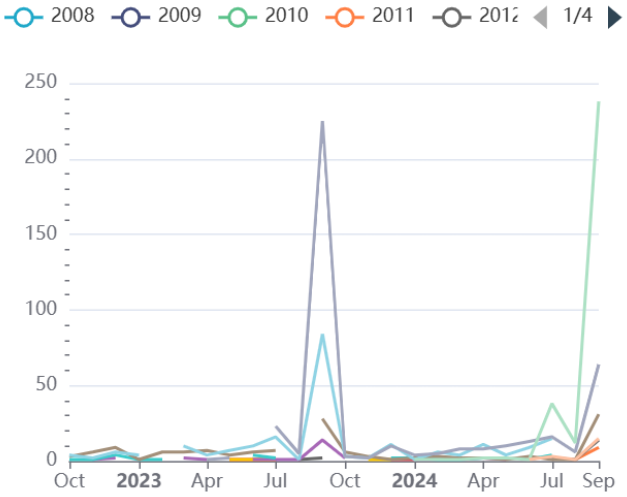
- Studijski program
- Status predmeta (obavezan, izborni i komisijska prijava)
- Predmet
- Ispitni rok
- Generacija
- Nastavnik
- Student

Prvi definisan izveštaj (Slika 1) odnosi se na broj ostvarenih ESP bodova tokom školske godine (može sadržati i više školskih godina u svrhu postizanja uporedivosti indikatora). Za svaku školsku godinu moguće je videti broj ESP bodova prema studijskim programima i generacijama, kao i međuzbirove. Uz pomoć ovog izveštaja moguće je pratiti trend uspeha studenata kroz vreme i poreditи apsolutne vrednosti što dovodi do lakšeg uočavanja pada ili rasta produktivnosti studenata i nosi važnu informaciju za upravu fakulteta.

			Метрика	Број студената				
			Остварено ЕСПБ	више од 60	између 37 і 60	мање од 37	Међузбир	Укупно
skolska_godina	studijski_program	Генерација						
2022/2023	ИСИТ09	2009/2010				1	1	1
		2010/2011			1	2	3	3
		2011/2012				7	7	7
		2012/2013				11	11	11
			Укупно	706	5182	3049	8937	8937

Slika 1 – Izveštaj o ostvarenim ESPB

Drugi definisan grafikon (Slika 2) je linijski dijagram koji se odnosi na datume i brojnost diplomiranja studenata. Linije koje povezuju tačke na dijagramu su označene različitim bojama za različite generacije pa je moguće pratiti i kad je koja generacija diplomirala i u kojoj meri. Ovaj grafikon, osim što sadrži trend diplomiranja, može igrati ulogu pri planiranju odbrana završnih radova gde bi se na osnovu podataka iz prošlih godina mogao uočiti trend opterećenosti određenih datuma i postupiti u skladu sa tim tokom pravljenja optimalnog rasporeda odbrana završnih radova, ali i ostalih nastavnih i nenastavnih aktivnosti.



Slika 2 - Linijski dijagram

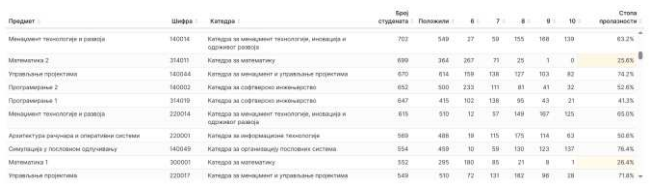
Naredni izveštaj se odnosi na pojedinačne ispite i rezultate studenata na tim ispitima. U izveštaju se prikazuje ukupan broj studenata koji su prijavili ispit, a nisu na isti izašli, a zatim i broj onih koji jesu ili nisu položili ispit. Nivoi prema kojima će ovaj broj biti prikazan mogu varirati i moguće je napraviti više ili manje modularan izveštaj. Relevantni parametri, osim naziva predmeta za koji se izveštaj generiše, su: školska godina, studijski program, studijski profil i ispitni rok. Na osnovu ovih podataka moguće je poreditи predmete među sobom prema stopi prolaznosti na različitim nivoima. Ovaj izveštaj takođe može služiti kao polazna tačka za planiranje ispita u budućnosti s obzirom na činjenicu da sadrži odnos između broja studenata koji su ispit prijavili, a na njega nisu izašli i broja studenata koji su polagali ispit, računajući i one koji su položili i one koji nisu. Modularnost izveštaja omogućava planiranje kako na makro (na nivou ispitnog roka npr.) tako i na mikro skali (na nivou pojedinačnog predmeta u ispitnom roku).

Prateći grafikon jeste stubičasti dijagram gde su na X osi označene ocene koje su studenti dobili na ispitu (računajući samo one koji su pristupili ispitu), a na Y osi je broj studenata koji je ostvario tu ocenu. Ovakvim vizuelnim prikazom se lakše donose zaključci o odnosu između ocena na samom predmetu, ali je moguće i poreditи predmete među sobom (na nivou katedre, profila, naučnoj oblasti itd.) i utvrditi korisne i uspešne prakse sa predmeta koji imaju dobre rezultate i primeniti ih kod predmeta sa lošijim performansama.

Još jedan prateći grafikon jeste kružni dijagram koji služi za lakše razumevanje i sagledavanje odnosa između studenata koji nisu izašli na ispit, onih koji su položili i onih koji nisu položili ispit. Sveobuhvatni prikaz sva tri dijagram prikazan je na slici 3.

Slika 3 - Dijagrami o polaganju ispita

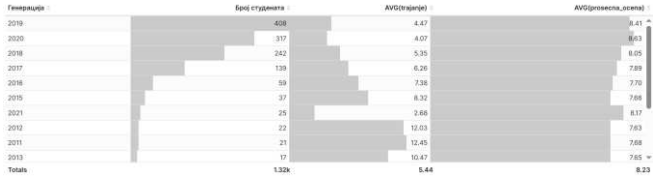
Naredna celina se odnosi na predmete i ostvarenu prolaznost (Slika 4). Prikazuje se broj studenata koji su polagali predmet, ukupan broj onih koji su položili, kao i pregled brojnosti prema ocenama. Za svaki predmet je data i stopa prolaznosti, a uvedena je i funkcionalnost koja određenom bojom označava neuobičajene vrednosti za stopu prolaznosti što predstavlja znak za dublju analizu tih predmeta. Kao dodatak ovom izveštaju se nalazi i kutijasti dijagram koji pokazuje raspodelu vrednosti ocena prema katedrama. Za svaku katedru su izračunate vrednosti o minimalnoj i maksimalnoj vrednosti, medijani i srednjoj vrednosti, broju primera u uzorku i vrednosti parametra za početak i kraj dva središnja kvartila. Na osnovu ovoga je moguće rangirati performanse na nivou katedri i preduzeti potrebne mere u slučaju da neke katedre imaju drastično različite rezultate u odnosu na ostale.



Slika 4 - Prolaznost po predmetima

Na osnovu podataka o diplomiranjima studenata napravljen je izveštaj o performansama studenata tokom studija prema

generacijama (Slika 5). Za svaku generaciju je izračunat broj studenta te generacije koji su diplomirali, prosečno trajanje studiranja izraženo u godinama, kao i prosečna ocena tokom celokupnih studija. U okviru tabele je dat i stubasti dijagram te je lakše uočiti razlike među generacijama i identifikovati naznake problema ili poteškoća pri studiranju na nivou generacije.



Slika 5 - Izveštaj o diplomiranju

V. ZAKLJUČAK

Pažljivo i detaljno izveštavanje o performansama visokoškolskih ustanova je preduslov za pravovremeno identifikovanje i rešavanje potencijalnih problema koji se mogu javiti tokom obrazovnog procesa. Na osnovu realnih potreba ustanove, odlukom da se koristi deskriptivni model analitike, razvijen je sistem koji pruža višeslojni pogled u stanje različitih procesa koristeći vrednosti definisanih indikatora i njihovu vizuelizaciju. Na osnovu postojećih podataka, moguće je uvideti šablone i pravilnosti koji se javljaju u okviru procesa na visokoškolskoj ustanovi i donositi odluke o daljim pravcima delovanja na osnovu toga. Kroz modularno izveštavanje, olakšana je interakcija korisnika sa sistemom i podešavanje parametara za izveštaj kroz intuitivan korisnički interfejs. Rezultati primene nad pravim podacima iz prethodne dve školske godine su pokazali da je moguće porediti indikatore performansi kroz vreme i prema različitim grupacijama što je osnova za pronalaženje faktora u sistemu koji utiču na razlike u performansama, kao i identifikovanje obrazaca koji mogu služiti za strateško ili operativno planiranje u budućnosti.

Neki od pravaca za buduća istraživanja se odnose na proširenje opsega indikatora, dodavanje podataka koji nisu obuhvaćeni ovim sistemom (npr. uspeh studenata nakon završetka studija), kao i integraciju ovog sistema sa ostalim, postojećim sistemima na fakultetu ili univerzitetu kako bi performanse mogle biti analizirane i na široj skali i donose zaključci na višem nivou. Takođe, moguće je dodati mapiranje predmeta koji obrađuju iste tematske celine u različitim akreditacionim periodima i posmatrati kako se to odražava na uspeh studenata.

ZAHVALNICA

Istraživanje sprovedeno u okviru inicijative „FON ideje – podrška projektima zaposlenih FON-a“.

LITERATURA

[1] F. J. Paz & S. C. Cazella (2019). Academic analytics: a systematic review of literature. International Journal of Development Research, 9(11), 31710–31716.

[2] F. C. Curran, S. Carlo & K. Harris-Walls (2024). Making the data visible: A systematic review of systems-level data dashboards for leadership and policy in education. Review of Educational Research, 00346543241288249.

[3] J. P. Campbell, P. B. DeBlois & D. G. Oblinger (2007). Academic analytics: A new tool for a new era. EDUCAUSE Review, 42(4), 40.

[4] Y. Y. Wong (2016). Academic analytics: a meta-analysis of its applications in higher education. International Journal of Services and Standards, 11(2), 176–192.

[5] T. Agasisti & A. J. Bowers (2017). Data analytics and decision making in education: towards the educational data scientist as a key actor in schools and higher education institutions. In Handbook of Contemporary Education Economics (pp. 184–210). Edward Elgar Publishing.

[6] U. Bin Mat, N. Buniyamin, P. M. Arsad & R. Kassim (2013, December). An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention. In 2013 IEEE 5th Conference on Engineering Education (ICEED) (pp. 126–130). IEEE.

[7] P. M. Hernandez-Diaz, J. A. Polanco & M. Escobar-Sierra (2021). Building a measurement system of higher education performance: Evidence from a Latin-American country. International Journal of Quality & Reliability Management, 38(6), 1278–1300.

[8] O. Adejo & T. Connolly (2017). An integrated system framework for predicting students' academic performance in higher educational institutions. International Journal of Computer Science and Information Technology, 9(3), 149–157.

[9] A. Giada, B. Giovanni & C. Vincenza (2014). A new indicator for higher education student performance. Higher Education, 68, 653–668.

[10] R. J. Shavelson, O. Zlatkin-Troitschanskaia & J. P. Mariño (2018). Performance indicators of learning in higher education institutions: An overview of the field. In Research Handbook on Quality, Performance and Accountability in Higher Education (pp. 249–263).

[11] M. Bratti, G. Barbato, D. Biancardi, C. Conti & M. Turri (2022). Degree-level determinants of university student performance. In Teaching, Research and Academic Careers: An Analysis of the Interrelations and Impacts (pp. 267–318). Cham: Springer International Publishing.

[12] A. M. Alsarmi & Z. A. Al-Hemyari (2014). Quantitative and qualitative statistical indicators to assess the quality of teaching and learning in higher education institutions. International Journal of Information and Decision Sciences, 6(4), 369–392.

[13] R. H. Obe & L. S. Hsu (2017). PostgreSQL: Up and Running: A Practical Guide to the Advanced Open Source Database (3rd ed.). O'Reilly Media.

[14] D. Fontaine (2019). Mastering PostgreSQL in Application Development (2nd ed.). Tapoueh.org Publishing.

[15] S. Shekhar (2019). Apache Superset Quick Start Guide: Learn How to Visualize and Explore Your Data Effectively with Apache Superset. Packt Publishing.

[16] Apache Software Foundation. (n.d.). Apache Superset Documentation. https://superset.apache.org/docs/

[17] S. Ahmad, M. A. El-Affendi, M. S. Anwar & R. Iqbal (2022). Potential future directions in optimization of students' performance prediction system. Computational Intelligence and Neuroscience, 2022(1), 6864955.

[18] G. Akçapınar, A. Altun & P. Aşkar (2019). Using learning analytics to develop early-warning system for at-risk students. International Journal of Educational Technology in Higher Education, 16(1), 1–20.

[19] H. Guruler, A. Istanbullu & M. Karahasan (2010). A new student performance analysing system using knowledge discovery in higher educational databases. Computers & Education, 55(1), 247–254.

Improving the performance of higher education institutions through the application of analytics

Milica Škembarević, Marija Đukić, Gordana Savić, Nenad Aničić, Bisera Andrić Gušavac, Milena Popović, Danica Lečić-Cvetković

ABSTRACT

Higher education institutions are facing new challenges in monitoring and analyzing the performance of participants in teaching and non-teaching processes. Traditional information systems of higher education institutions lack the ability to directly and comprehensively analyze performance within the system itself, nor do they have defined mechanisms for calculating indicators that would enable performance monitoring over different periods and at different levels. The aim of this paper is to describe the development of a performance analytics system through the definition and implementation of mechanisms for monitoring key analytics indicators. Based on the analysis of existing solutions and identified needs, a system was created that supports

centralized loading, testing, and visualization of data using the Apache Superset tool. The developed solution provides the ability to analyze performance according to different criteria and at different levels - from the individual level to the institution level and with different visualization methods adequately adapted to the data being displayed. Visually presented indicator values largely indicate potential directions for improving the quality of education and represent a starting point for making long-term strategic decisions.

Keywords: academic analytics, teaching process management, higher education, student performance

Analiza uticaja FONIS hakatona na razvoj karijere studenata

Dušan Kostić
Univerzitet u Beogradu – Fakultet organizacionih nauka
dk20243269@student.fon.bg.ac.rs
0009-0000-3432-6074

Miloš Jolović
Univerzitet u Beogradu – Fakultet organizacionih nauka
mj20243254@student.fon.bg.ac.rs
0009-0003-6580-2039

Aleksandar Joksimović
Univerzitet u Beogradu – Fakultet organizacionih nauka
aj20243251@student.fon.bg.ac.rs
0009-0008-5711-7636

Tamara Naumović
Univerzitet u Beogradu – Fakultet organizacionih nauka
tamara.naumovic@fon.bg.ac.rs
0000-0001-9849-7665

Petar Lukovac
Univerzitet u Beogradu – Fakultet organizacionih nauka
petar.lukovac@fon.bg.ac.rs
0000-0003-4561-8886

***Apstrakt* – Ovaj rad analizira uticaj učešća na hakatonima organizovanim od strane Udruženja studenata informatike FONIS na razvoj karijere studenata. Hakatoni su prepoznati kao značajan oblik neformalnog obrazovanja koji podstiče razvoj tehničkih i preduzetničkih veština. Cilj istraživanja je da se ispita na koji način i u kojoj meri učešće na ovim događajima doprinosi profesionalnom razvoju učesnika. Metodološki okvir istraživanja uključuje anketnu studiju sprovedenu na uzorku od 71 učesnika različitih hakatona organizovanih u periodu od 2013. do 2024. godine. Rezultati istraživanja ukazuju na značajan doprinos hakatona razvoju kritičkog mišljenja, preduzetničkog načina razmišljanja i profesionalnih veština učesnika. Zaključeno je da hakatoni predstavljaju efikasnu platformu za sticanje znanja i umrežavanje, ali da postoji prostor za unapređenje, naročito u domenu post-hakatonskih aktivnosti i podrške učesnicima u daljem razvoju njihovih projekata.**

Ključne reči – hakaton, otvorene inovacije, karijerni razvoj

I. UVOD

Ovaj rad predstavlja istraživanje i analizu efekata na razvoj profesionalne karijere studenata koji su nastali kao posledica učešća na hakatonima u organizaciji Udruženja studenata informatike FONIS. Obzirom da hakatoni predstavljaju važan pedagoški alat [1], a takođe podstiču razvoj profesionalnih kompetencija studenata[2]evidentan je disbalans između malog broja istraživanja uticaja učešća na hakatonu i velike važnosti, ali i popularnosti hakatona.

Cilj ovog istraživanja je da se identifikuju i detaljnije opišu različiti uticaji hakatona kao fenomena na lični i karijerni razvoj mladih. Istraživanje je sprovedeno na učesnicima hakatona u organizaciji Udruženja studenata informatike FONIS u prethodnih 10 godina. Uzorak je sačinjen od 71 ispitanika, koji su učestvovali na FON Hakatonu ili Hakatonu za srednjoškolce.

Teorijski okvir istraživanja čini ostvrt na definisanje i nastajanje hakatona kao fenomena. Takođe, hakatoni su opisani kao metod otvorenih inovacija, ali i u funkciji visokoškolskog obrazovanja, kao alat koji podstiče inovativno i kreativno razmišljanje kod studenata. U okviru metodologije istraživanja navedena su istraživačka pitanja i opisan je način prikupljanja podataka, kao i sam upitnik. Na samom kraju predstavljeni su rezultati, kao i njihova analiza. Predstavljeno je zadovoljstvo učesnika svojim učešćem, kao i kratkoročni i dugoročni efekti na njihov razvoj nakon hakatona. Takođe, analizirani su izvori motivacije, uticaj hakatona na unapređenje znanja i iskustva takmičara, ali i doprinos hakatona preduzetničkom razvoju.

II. TEORIJSKI OKVIR ISTRAŽIVANJA

A. Hakatoni

Prema Oksfordskom rečniku pojam hakaton se definiše kao "događaj, koji obično traje nekoliko dana, na kojem se veliki broj ljudi okuplja da učestvuje u zajedničkom programiranju". U pokušaju da konceptualizuju hakaton je definisan na sledeći način: „Hakaton je jedan tip organizovanog, ciljno orijentisanog takmičenja u inovacijama, kratkotrajan događaj sa vremenskim ograničenjem, u kojem se izazov rešava kreativno kroz kompeticiju i zajednički rad timova, a rezultati se prikazuju i prepoznaju tokom ceremonije na kraju događaja"[3].

Povećanjem popularnosti tokom godina, hakatoni su evoluirali od malih studentskih takmičenja do događaja koje sa svojim sredstvima organizuju mnoge organizacije, softverske kompanije, čak i vladine agencije sa ciljem ohrabrivanja digitalnih inovacija. Početkom 2000-ih godina, hakatoni su postali ugledni i velike kompanije, kao i investitori, počeli su ih smatrati sredstvom za brzi razvoj softvera i tehnologija, kao i za otkrivanje novih oblasti za inovacije i finansiranje [4][5].

U radu [6]ističe se da hakatoni predstavljeni u literaturi obično postoje van stabilnog organizacionog konteksta i okupljaju ljude koji uglavnom nisu ranije radili zajedno, pa čak se nisu ni upoznali, kao što je slučaj sa takmičenjima u digitalnim inovacijama ili hakatonima otvorenih podataka.

Između ostalog hakatoni su korišćeni u saradnji između univerziteta i industrije u svrhe visokog obrazovanja [7][8], čak i u virtuelnim događajima [9]. Ipak, ne postoji jedinstvena tipologija hakatona koja se koristi za različite svrhe. Hakatoni se mogu razlikovati i po dužini trajanja takmičenja, gde su najčešći primeri hakatona koji traju 24 ili 48 sati koji su manjeg obima, ali i dosta dinamičniji. Dok oni duži mogu trajati od nekoliko dana do par nedelja i podrazumevaju dosta komplikovanije zahteve. Prema [5] hakatoni se mogu podeliti u dve velike grupe, na tehnološki orijentisane (*Tech-centric*) i ciljno orijentisane (*Focus-centric*).

Uspeh hakatona zavisi od niza elemenata dizajna, kao što su naziv, datum, trajanje, sektor i vrsta inicijatora, ciljevi, pokretački faktori, teme, format , broj učesnika, rodna ravnoteža, način formiranja timova, veštine učesnika, definicija problema, generisanje ideja i struktura nagrada [1][10].

B. Hakatoni kao metod otvorenih inovacija

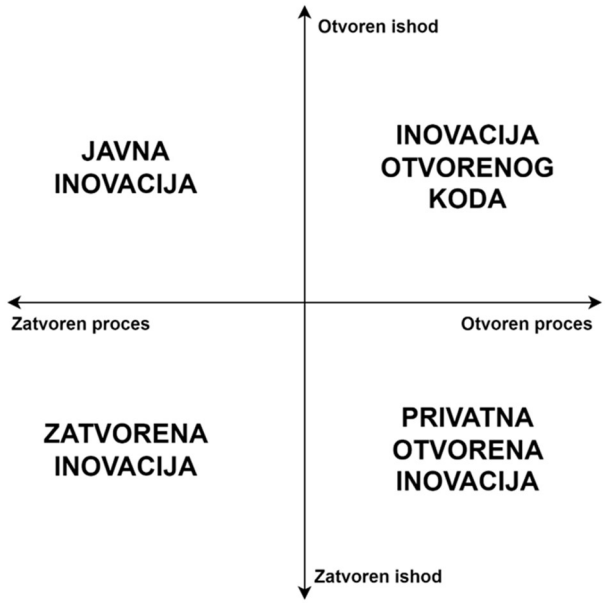
Termin "otvorena inovacija" pojavio se početkom 2000-ih i promovisao ga je profesor Henry Chesbrough. Koristio se za opis unapređenja koje organizacije traži izvan svojih internih krugova, koristeći razne spoljašnje izvore (kao što su povratne informacije korisnika, objavljeni patenti, konkurenti, spoljašnje agencije, javnost itd.) kako bi podstakla inovacije [4][11]. Prema [12] zaključuje se da je inovacija jedan od temeljnih stubova ekonomije zasnovane na znanju koja pokreće socioekonomski i društveni rast u razvijenom svetu.

Ranije korišćeni model generisanja novih ideja se može opisati kao model zatvorene inovacije. Kompanije i organizacije su se čvrsto oslanjale na interne resurse i instraživačko-razvojne centre i ulagale sredstva u njihov razvoj, verujući da uspešna inovacija zahteva strogu kontrolu, očekivajući da će im taj put doneti uspeh [4][11].

Model otvorenih inovacija je nastao kao posledica sve veće mobilnosti znanja zaposlenih i pojave rizičkog kapitala [11]. Ove promene proširile su granice inovacionih procesa unutar kompanija i stvorile su nov način istraživanja, omogućavajući kreiranje projekata iz unutrašnjih ili spoljašnjih izvora i omogućavajući da nova tehnologija uđe u bilo koju fazu (istraživanje, razvoj, plasman, itd.). Ovaj model omogućava kompanijama da smanje troškove tradicionalnih metoda istraživanja i razvoja, koristeći spoljne resurse za istraživanje i razvoj kako bi uštedeli vreme i novac, dok istovremeno povećavaju prihod licenciranjem internih tehnologija [4][13].

Huizingh je u radu [14] klasifikovao različite pristupe inovacijama na osnovu otvorenosti kako procesa, tako i inovacionog ishoda. On je razvio matricu sa četiri vrste inovacija:

- Zatvorena inovacija (Zatvoren proces - Zatvoren ishod),
- Privatna otvorena inovacija (Otvoren proces - Zatvoren ishod),
- Javna inovacija (Zatvoren proces - Otvoren ishod),
- Inovacija otvorenog koda (Otvoren proces - Otvoren ishod), koja je prikazana na Slici 1.



Slika 1. Matrica inovacija, adaptacija[14]

Hakatoni suštinski ubrzavaju proces otvorenih inovacija kroz podsticanje saradnje, rešavanje problema i postizanje opipljivih rezultata. Ovi događaji predstavljaju priliku za kreativnost i iskustveno učenje, promovišući vrednosti interdisciplinarne saradnje, brze izrade prototipova, praktičnog rešavanja problema i efektivnog umrežavanja [15][16].

Iako se hakatoni uspešno koriste kao novi oblik organizovanja inovacija proizvoda, dizajniranje hakatona zahteva pažljivo planiranje unapred i uzimanje u obzir ciljeva koje su organizatori i učesnici postavili za događaj, kao i za sebe [17]. Kompanije su veoma brzo prepoznale upotrebnu vrednost hakatona kao modela ulazne otvorene inovacije, gde studenti, ali i ostali učesnici, mogu kreirati ideje i prototipove koji pružaju korisna idejna rešenja [18].

C. Hakatoni u obrazovanju

Jedan od pristupa primeni koncepata otvorene inovacije u neformalnom obrazovanju su studentski hakatoni. U samom početku bili su organizovani s idejom da učesnici razvijaju prototipove softverskih rešenja u kratkim vremenskim intervalima intenzivnog programiranja, hakatoni su do danas evoluirali u različite modele studentskih takmičenja [15]. Glavna obrazovna korist hakatona je iskustveno učenje, gde studenti stiču veštine kroz aktivno angažovanje u projektima.

Visokoobrazovne ustanove imaju potrebu da razviju okruženje koje interno podržava inovacioni potencijal i unapređuje inovacione performanse, ali i da istraže na koji način aktivnosti poput hakatona mogu doprineti poboljšanju internih inovacionih performansi i potencijalu kroz ostvarene rezultate. Rastuća popularnost hakatona stvorila je potrebu za sve intenzivnijim istraživanjem ovog fenomena od strane stručnjaka [17][19].

Učešće na hakatonima, pomaže studentima da uče kroz preko potrebno praktično iskustvo [1]. Takođe, jedan od ciljeva hakatona je da proširi profesionalne kompetencije studenata računarstva ili softverskog inženjerstva. Pristup

rešavanju ovog problema zasniva se na primeni postojećeg teorijskog znanja i praktičnih veština studenata u stvarnim profesionalnim aktivnostima. Ovaj pristup može se realizovati kroz uključivanje studenata u različite vrste praksi u okviru obrazovnog procesa, kao i kroz angažovanje profesionalnih softverskih programera koji će sa studentima sarađivati van formalnog obrazovnog okvira [2].

Događaji ovog karaktera služe kao platforme za neformalna iskustva učenja [12], ali i kao dinamična okruženja za učenje, koristeći učenje zasnovano na projektima kako bi inspirisali studente da kreativno integrišu različite oblasti znanja kroz timski rad [1][20]. Hakatoni, između ostalog, predstavljaju vrlo praktičan koncept za prepoznavanje kvaliteta studenata, pri čemu balansiranje vrednosti koje učesnici mogu dobiti postaje centralni cilj hakatona, stvarajući obostrano korisne situacije [2][17].

Hakatoni u obrazovnim institucijama koriste moćne alate za postizanje pedagoških ciljeva, promovišući značajnu interakciju, timsku saradnju i angažovanje sa različitim saradnicima. Ovaj pristup aktivira ključne kompetencije, uključujući socijalne i komunikacione veštine, podstičući analitičko, kritičko i inventivno razmišljanje. Takođe, unapređuje sposobnosti socijalnog partnerstva i razvija prilagodljivost, koja je ključna za lični razvoj i poboljšanje kvaliteta života [1][21].

Postoji briga da dugotrajno angažovanje u ovakvim vannastavnim aktivnostima može škoditi regularnom nastavnom procesu, ali je prilikom analize studentskih ocena, utvrđeno je da je ova briga neopravdana. Analiza [22] je pokazala da studenti koji učestvuju u hakatonima imaju nešto viši prosek ocena (2–5%) u odnosu na studente koji ne učestvuju. Kada se uporedi uspeh na individualnom nivou, ne postoje značajni obrasci. Dakle, iako se ne može tvrditi da hakatoni uzrokuju ili utiču na bolje proseke, pretpostavka je da ovakve aktivnosti obično privlače grupu studenata sa većim akademskim postignućima.

Neformalno učenje i razvoj ovih veština pomažu studentima da se bolje pozicioniraju prilikom konkurisanja za prakse i poslove nakon završetka fakulteta [23].

III. METODOLOGIJA

Ovo istraživanje je imalo za cilj da ispita da li i na koji način učešće na hakatonu, u organizaciji Udruženja studenata informatike FONIS, utiče na razvoj profesionalne karijere učesnika.

Istraživačka pitanja u ovom istraživanju su:

• Koji su efekti na učesnike i njihovu karijeru nakon učešća na hakatonu?

• Kako učesnici procenjuju značaj celokupnog iskustva hakatona i njegov uticaj na razvoj njihovih znanja i iskustva nakon takmičenja?

• Kakav je odnos učesnika prema rešenjima koja su kreirali na hakatonu i u kojoj meri su ih dalje razvijali?

• Šta iskustvo hakatona donosi studentima u pogledu njihovih stavova o karijernom razvoju?

• Šta predstavlja najbitniji motivacioni faktor za učešće na hakatonu?

• Na koji način hakatoni, kao platforma za učenje, utiču na razvoj učesnika?

• Kakav je stav učesnika o organizaciji hakatona i aktivnostima nakon hakatona?

Udruženja studenata informatike FONIS koju čine studenti Fakulteta organizacionih nauka u Beogradu je dobrovoljno, nevladino i neprofitno udruženje koje postoji od 2000. godine i organizuje raznovrsne događaje na godišnjem nivou koji omogućavaju studentima i srednjoškolcima da uče, napreduju i stiču iskustvo koje će im biti neophodno za dalji razvoj karijere u industiji informacionih tehnologija. U periodu od 2013. godine do danas organizovano je preko 20 hakatona sa različitom tematikom. Hakatoni su se kroz iteracije razvijali, da bi na poslednjem održanom takmičenju postojalo više od 50 timskih prijava. Takođe, kroz godine menjao se i način održavanja samog hakatona. Bitno je naznačiti da tokom svih ovih godina nije postojalo sistematsko praćenje takmičara, kao i njihovih rezultata u daljoj karijeri, ali i da se ta praksa započinje ovim istraživanjem.

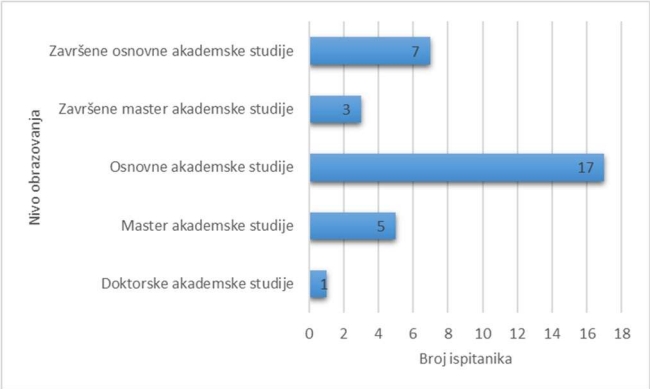
Po aktuelnom pravilniku takmičenja, timovi pre početka hakatona saznaju temu hakatona, nakon čega tokom diskusije mogu da postave dodatna pitanja u vezi iste. Takmičenje traje 24 časa, a takmičari tokom hakatona imaju dva termina konsultacija. Nakon završetka takmičenja timovi prezentuju svoju ideju i rešenje publici i članovima žirija, koji kasnije biraju pobednike.

Ciljani uzorak su predstavljali učesnici na haktonima u organizaciji FONIS-a u periodu od 2013. do 2024. godine. Ispitanici su konktaktirani putem email adresa koje se nalaze u bazama podataka koje su ostavili prilikom prijave za hakaton. Podaci o učesnicima su akumulirani sa 15 različitih hakatona, a poziv za učešće u istraživanju i upitnik su poslati na 433 različite email adrese. Ispitanici su anonimno popunjavali upitnik elektronskim putem, pomoću platforme Microsoft Forms. Nisu postojali delovi uzorka koji su bili uklonjeni, niti nepravilno popunjeni upitnici. Konačan uzorak istraživanja sačinjava ukupno 71 ispitanik.

Upitnik je činilo ukupno 30 pitanja od kojih je 7 pitanja bilo otvorenog tipa. Pitanja zatvorenog tipa bila su sa jednostrukim izborom i u obliku petostepene Likertove skale, osim jednog pitanja gde su ispitanici imali izbor između Da i Ne.

IV. ANALIZA REZULTATA

Ispitanici se mogu razvrstati u četiri starosne grupe: 16-20 godina ima 35% ispitanika, 20-23 ima 34%, 24-29 godina ima 24% ispitanika, a 30 ili više godina 7% ispitanika. U trenutku popunjavanja upitnika većinu uzorka su činili studenti (njih 38) dok je zaposlenih bilo 28. Ostali ispitanici su bili srednjoškolci. Obrazovna struktura uzorka prikazana je na Slici 2.



Slika 2. Obrazovna struktura uzorka

A. Efekti i zadovoljstvo nakon hakatona

U segmentu upitnika u kome je bilo potrebno ispitati koji su efekti na karijere učesnika nakon hakatona ispitanici su trebali da ocene u kojoj meri se slažu sa sledećim iskazima (1 – veoma se ne slažem, 5 – veoma se slažem). Rezultati su prikazani u Tabeli 1.

Tabela 1. Efekti nakon hakatona

Iskaz	AVG	STD	Interval poverenja
"Zaposlio/la sam se i bavim se projektima sličnim projektima sa hakatona."	2.24	1.292	0.30
"Zaposlio/la sam se, ali moj trenutni posao nema veze sa projektima kojima sam se bavio tokom hakatona."	2.44	1.490	0.35
"Krenuo/la sam u totalno drugom pravcu, ali i dalje gajim interesovanja prema tehnologijama korišćenim na hakatonu."	2.15	1.091	0.25
"Promenio/la sam fokus svoje karijere i više me ne interesuju tehnologije koje sam koristio/la na hakatonima."	1.73	1.028	0.24

Na pitanje da ocene jasnoću zahteva i zadataka koji su postavljeni na temu hakatona ispitanici su odgovorili sa prosečnom ocenom 3.79, ali je samo 13% ispitanika izrazilo nezadovoljstvo niskim ocenama za ovaj kriterijum, dok na pitanje da ocene u kojoj meri su zadovoljni sopstvenim učešćem i iskustvom na hakatonu, ispitanici su odgovorili sa prosečnom ocenom 4.31, a čak 89% ispitanika je odgovorilo sa veoma visokim ocenama.

B. Sticanje znanja i iskustva

Ispitanicima je postavljeno pitanje da na skali od 1 do 5 ocene sledeće kriterijume koji mogu imati uticaj na razvoj njihovih znanja, veština i iskustva. Rezultati su prikazani u Tabeli 2.

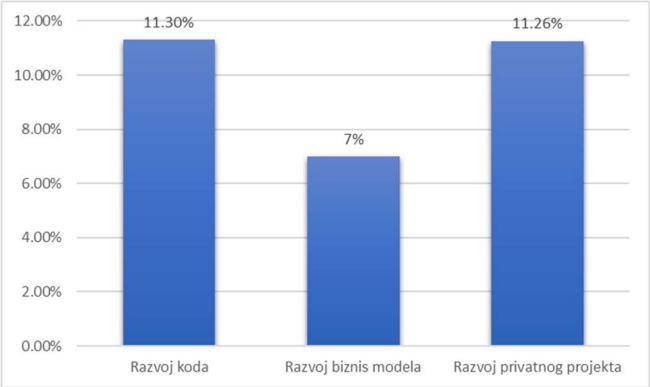
Tabela 2. Sticanje znanja i iskustva

Kriterijum	AVG	STD	Interval poverenja	Ocene 4 i 5
Efektnost prenosa znanja i mentorstva	3.44	1.097	0.26	50.7%
Uticaj na tehnička znanja	3.77	1.13	0.26	66.3%
Uticaj na razvoj mekih veština	3.96	1.01	0.24	74.6%
Uticaj na razvoj sposobnosti rada na projektnim zadacima	4.17	1.02	0.24	80.3%
Uticaj na razvoj kritičkog mišljenja, rešavanja problema, preduzetničkom duhu	4.14	1.04	0.24	77.5%
Značaj kontakta sa drugim takmičarima	3.61	1.28	0.30	64.8%

Iz ovih rezultata moguće je zaključiti da ispitanici prepoznaju da je najveći uticaj hakatona na razvoj sposobnosti i iskustva u radu sa projektim zadacima, kao i razvoj kritičkog mišljenja, rešavanja problema i prilagođavanju novim situacijama, koje predstavljaju neizostavni deo njihove buduće karijere. Interesantno je da ispitanici više prepoznaju uticaj hakatona na razvoj mekih veština, nego na razvoj tehničkih znanja, što govori da hakatoni nisu isključivo programerska takmičenja i da učesnici imaju fokus na više različitih aspekata razvoja jednog softverskog rešenja.

C. Preduzetnički razvoj

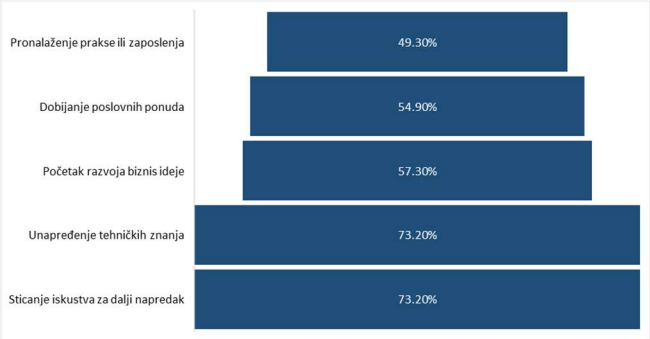
Prilikom popunjavanja upitnika, 76,1% ispitanika izjasnilo se da smatra da je njihovo rešenje na hakatonu rešilo postavljeni izazov u velikoj meri, dok 87.7% ispitanika tvrdi da je njihov tim u velikoj meri uzeo u obzir potrebe i preferencije potencijalnih korisnika prilikom razvoja rešenja, dolazi se do zaključka da je zadovoljstvo učesnika svojim rešenjem na visokom nivou. To potvrđuje i činjenica da 89% ispitanika izabralo visoku ocenu za zadovoljstvo svojim učešćem na hakatonu. Ipak, samo 11,3% ispitanika se izjasnilo da je u većoj meri nastavilo sa razvojem koda koji su kreirali tokom hakatona, a samo 7% ispitanika je nastavilo sa razvojem biznis modela. Takođe, bitno je izvojiti da dve trećine ispitanika nije izrazilo ambiciju da nastavi sa unapređenjem rešenja nakon završetka hakatona, kako sa timom, tako ni samostalno, što potvrđuje rezultate istraživanja da je samo 11,26% ispitanika je nastavilo da razvija ideju svog tima sa hakatona, od čega je velika većina pokrenula privatni projekat, što je prikazano na Slici 3.



Slika 3. Razvoj rešenja nakon hakatona

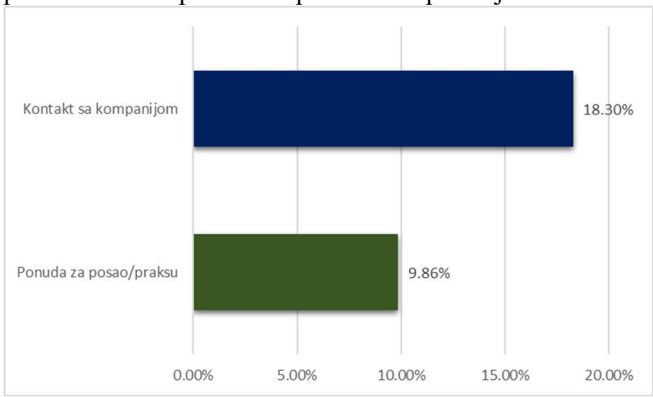
D. Karijerni razvoj

U pitanju u kome su ispitanici ocenjivali kao mogućnost za profesionalni razvoj, čak 73.2% ispitanika smatra da hakatoni pozitivno utiču na unapređenje tehničkih znanja, ali i na sticanje iskustva rada na projektima potrebnog za dalji napredak u karijeri. Takođe, 57.8% ispitanika smatra da hakatoni pružaju veliku mogućnost za početak razvoja biznis ideje. Kada su u pitanju poslovne prilike, 49.3% ispitanika smatra hakatone prilikom za pronalaženje prakse ili zaposlenja, dok 54.9% ispitanika kao priliku za dobijanje poslovnih ponuda. Rezultati su prikazani na Slici 4.



Slika 4. Karijerni razvoj

Na Slici 5. je prikazano da je od celokupnog uzorka, čak 18.3% ispitanika stupilo u bilo kakav kontakt sa kompanijama pokroviteljima hakatona, od kojih je više od polovine dobilo ponudu za praksu ili zaposlenje.



Slika 5. Saradnja sa kompanijama

Bitno je spomenuti da su predstavnici kompanija uglavnom pojedinačno kontaktirali učesnike za koje im se javilo interesovanje, nikada timove u celosti.

E. Motivacija

U pogledu motivacije učesnika za učešće na hakatonu, ispitanici su izvojili nagrade za pobednike, ali i stvaranje novih konekcija sa kolegama sličnih interesovanja kao najuticajniji faktor, čak 80.3% uzorka je ocenilo ove kategorije sa najvišim ocenama. Preostale kategorije ne zaostaju mnogo, mogućnost sticanja novog znanja i iskustva, sa 78.9% visokih ocena i povezivanje sa kompanijama i uglednim ljudima iz privrede, sa 76.1% visokih ocena. Ovi rezultati dovode do zaključka da je motivacija za učešće na hakatonima kod ispitanika veoma diverzifikovana.

Nagrade na hakatonima najčešće predstavljaju vaučere u novčanoj vrednosti za kupovinu tehnoloških proizvoda u unapred određenoj prodavnici, kao i dodatne poklone koje obezbeđuju organizatori. Upoznavanje kolega, članova svog tima, omogućava učesnicima da prepoznaju veštine i sposobnosti svojih saradnika za rad u timskom okruženju, ali i način fukcionisanja u stresnom i dinamičkom okruženju. Takođe, veoma su značajni i kontakti koji se mogu ostvariti sa članovima drugih timova, pošto hakatoni okupljaju mlade iz različitih okruženja, ali sa sličnim interesovanjima.

Hakatoni, pogotovo za one učesnike koji prvi put učesvuju, predstavljaju bitan izvor za sticanje tehničkih, ali i veština prezentovanja, liderstva i prilagođavanja nepredviđenim okolnostima. Pružaju učesnicima iskustvo koje predstavlja nešto najpribližnije onome što ih očekuje u realnom radnom okruženju.

Povezivanje sa kompanijama, pokroviteljima hakatona, učesnicima pruža mogućnost da se upoznaju sa stručnjacima iz različitih oblasti i na taj način ostvare konekcije koje im mogu biti veoma značajne u daljem razvoju karijere. Između ostalog, oni najbolji, često iskoriste priliku hakatona da pokažu svoje znanje i veštine i na taj način bivaju prepoznati od strane predstavnika iz industrije.

F. Hakatoni kao platforma za učenje

Na pitanje u kojoj meri dinamičnost hakatona stvara stresno okruženje za učenje ispitanici su odgovorili sa prosečnim odgovorom 3.55, a čak 57% njih je odgovorilo na ovo pitanje sa visokim ocenama. Iz ovih rezultata se izvodi takođe da je 85% učesnika istraživanja koji su ocenili hakatone kao visokostresno okruženje za učenje su iskazali veoma visoko zadovoljstvo svojim učešćem na hakatonu, što može dovesti do zaključka da je ovakva vrsta okruženja za učenje prijala čak 48% ispitanika.

Na pitanje, "Koliko smatrate da je bitna podrška i materijali koje ste dobili za celokupno iskustvo na hakatonu?" ispitanici su dali prosečnu ocenu 3.63, a čak 58% ispitanika smatra da je veoma bitan ovaj vid podrške za doživljaj učešća na hakatonu. Interesantno je da je ovo pitanje u visokoj korelaciji sa pitanjem o zadovoljstvu učešćem na hakatonu, pošto je 96% ispitanika koji smatraju da su podrška i mentorstvo veoma bitan faktor doživljaja i učenja tokom hakatonu, odgovorilo da je veoma zadovoljno sopstvenim iskustvom na hakatonu.

Na jedno od pitanja koje je najznačnije za odgovor na istraživačka pitanja, "u kojoj meri vam je hakaton pružio priliku za učenje i lični razvoj", ispitanici su u proseku dali ocenu 4.01, dok je 79% ispitanika ocenilo ovo pitanje sa

najvišim ocenama. Takođe, ovo pitanje su ispitanici koji smatraju da je veoma bitna podrška i mentorstvo, ocenili visokim ocenama sa preklapanjem od čak 93%.

Ovi rezultati lako dovode do zaključka da je učestvovanje na hakatonima koji sadrže mentorsku podršku kroz konsultacije i žiriranje blisko povezano sa omogućavanjem učesnicima da prihvate hakaton, ne samo kao takmičenje, nego i kao platformu za učenje i lični razvoj u različitim oblastima, ali i implicira stvaranje visokog zadovoljstva i pozitivnog osećaja učesnika nakon završetka hakatona.

G. Aktivnosti nakon hakatona

U sklopu poslednjeg pitanja pred ispitanicima je bila mogućnost da ocene potrebu za aktivnostima nakon hakatona, njihovu želju za ponovnim učestvovanjem na hakatonu, preporukama kolegama, ali i da napišu konkretne komentare na temu hakatona.

Prilikom ocenjivanja potrebe za organizovanjem post hakaton aktivnosti (u vidu okupljanja, druženja, networking-a, organizovanja programa za podršku startap-ima sa hakatona, itd...) ispitanici su dali prosečnu ocenu 3.68, a 60% ispitanika je smatralo da je veoma potrebno da se kreiraju dodatne aktivnosti nakon same realizacije hakatona. Bitno je napomenuti da je veliki procenat njih veoma visoko ocenilo sopstveno iskustvo učestvovanja na hakatonu, ali i hakaton kao mogućnost za učenje i lični razvoj. Dodatno, čak 83.1% ispitanika veoma rado preporučilo kolegama ili prijateljima da učestvuju na hakatonu.

V. ZAKLJUČAK

Ovaj rad je pružio detaljnu analizu uticaja FONIS hakatona na razvoj profesionalnih karijera studenata, koristeći kvantitativni metod istraživanja. Analizirane su različite kategorije, uključujući zadovoljstvo učesnika, sticanje znanja i iskustva, preduzetnički razvoj, motivaciju, poslovne prilike, učenje i networking. Prikazani rezultati istraživanja ukazuju na to da hakatoni pružaju učesnicima priliku za profesionalni napredak, sticanje novih veština i iskustava, ali i uspostavljanje važnih poslovnih kontakata.

Mogućnosti za dalji razvoj istraživanja uključuju proširenje uzorka kako bi se obuhvatili učesnici iz šireg spektra hakatona organizovanih u različitim industrijama i akademskim sektorima. Dodatno, buduća istraživanja mogla bi ispitati dugoročne efekte učešća na hakatonima na profesionalne uspehe, kao i analizirati mogućnosti za unapređenje organizacije hakatona, posebno u domenu aktivnosti nakon hakatona koje bi omogućile kontinuirani razvoj projekata i podršku učesnicima u daljoj karijeri. Takođe, bilo bi potrebno proširiti spektar pitanja i tako odrediti koje još dodatne vrednosti i apsekti hakatona doprinose pozitivnom utisku, i još važnije, ličnom i razvoju karijere učesnika.

LITERATURA

[1] A. Miličević, M. Despotović-Zrakić, D. Stojanović, M. Suvajžić, and A. Labus, "Academic performance indicators for the hackathon learning approach – The case of the blockchain hackathon," *Journal of Innovation & Knowledge*, vol. 9, no. 3, p. 100501, Jul. 2024, doi: 10.1016/J.JIK.2024.100501.

[2] Z. S. Seidametova, Z. S. Seidametova, Z. S. Abduramanov, and G. S. Seydametov, "Hackathons in computer science education: monitoring and evaluation of programming projects," *Educational Technology Quarterly*, vol. 2022, no. 1, pp. 20–34, Feb. 2022, doi: 10.55056/etq.5.

[3] S. Halvari, A. Suominen, J. Jussila, V. Jonsson, and J. Bäckman, "International Society for Professional Innovation Management. Conceptualization refinement of hackathon for innovation management," 2020.

[4] T. Naumović, B. Vajagić, L. Cvetković, and M. Proročić, "Open innovations and the role of hackathons," *E-business technologies conference proceedings*, vol. 2, no. 1, pp. 42–45, Jun. 2022, Accessed: Mar. 05, 2025. [Online]. Available: https://ebt.rs/journals/index.php/conf-proc/article/view/111

[5] G. Briscoe and C. Mulligan, "Digital Innovation: The Hackathon Phenomenon", Accessed: Mar. 05, 2025. [Online]. Available: http://qmro.qmul.ac.uk/jspui/handle/123456789/7682

[6] E. P. P. Pe-Than, A. Nolte, A. Filippova, C. Bird, S. Scallen, and J. Herbsleb, "Corporate hackathons, how and why? A multiple case study of motivation, projects proposal and selection, goal setting, coordination, and outcomes," *Hum Comput Interact*, vol. 37, no. 4, pp. 281–313, Jul. 2022, doi: 10.1080/07370024.2020.1760869.

[7] J. Jussila, A. H. Suominen, and T. Rainio, "Entrepreneurship Competence Using Educational Hackathons in Finland," *Journal of Finnish Studies*, vol. 23, no. 2, pp. 32–73, Dec. 2020, doi: 10.5406/28315081.23.2.05.

[8] A. H. Suominen, S. Halvari, and J. Jussila, "World Heritage meets Smart City in an Urban- Educational Hackathon in Rauma," *Technology Innovation Management Review*, vol. 9, no. 9, pp. 44–63, Sep. 2019, doi: 10.22215/TIMREVIEW/1268.

[9] J. Jussila, J. Raitanen, A. H. Suominen, and A. M. Järvenpää, "Virtual Hackathons—A Novel Approach for University-Industry Collaboration," *Springer Proceedings in Complexity*, pp. 247–257, 2021, doi: 10.1007/978-3-030-62066-0_19.

[10] E. P. P. Pe-Than, A. Nolte, A. Filippova, C. Bird, S. Scallen, and J. D. Herbsleb, "Designing Corporate Hackathons with a Purpose: The Future of Software Development," *IEEE Softw*, vol. 36, no. 1, pp. 15–22, Jan. 2019, doi: 10.1109/MS.2018.290110547.

[11] H. W. Chesbrough, "Open Innovation: The New Imperative for Creating and Profiting from Technology - Henry William Chesbrough - Google Books." Accessed: Mar. 05, 2025. [Online]. Available: https://books.google.rs/books/about/Open_Innovation.html?id=4hTRWStFhVgC&redir_esc=y

[12] M. B. Garcia, "Fostering an Innovation Culture in the Education Sector: A Scoping Review and Bibliometric Analysis of Hackathon Research," *Innov High Educ*, vol. 48, no. 4, pp. 739–762, Aug. 2023, doi: 10.1007/S10755-023-09651-Y/FIGURES/8.

[13] M. Elmquist, T. Fredberg, and S. Ollila, "Exploring the field of open innovation," *European Journal of Innovation Management*, vol. 12, no. 3, pp. 326–345, Jul. 2009, doi: 10.1108/14601060910974219/FULL/XML.

[14] E. K. R. E. Huizingh, "Open innovation: State of the art and future perspectives," *Technovation*, vol. 31, no. 1, pp. 2–9, Jan. 2011, doi: 10.1016/J.TECHNOVATION.2010.10.002.

[15] Z. Bogdanovic, A. Milicevic, D. Stojanovic, A. Labus, M. Despotovic-Zrakic, and B. Radenkovic, "Open Innovation Strategies in Engineering Education," *2023 IEEE 33rd International Conference on Microelectronics, MIEL 2023*, 2023, doi: 10.1109/MIEL58498.2023.10315923.

[16] D. Cobham, C. Gowen, K. Jacques, J. Laurel, and S. Ringham, "FROM APPFEST TO ENTREPRENEURS: USING A HACKATHON EVENT TO SEED A UNIVERSITY STUDENT-LED ENTERPRISE," *INTED2017 Proceedings*, vol. 1, pp. 522–529, Mar. 2017, doi: 10.21125/INTED.2017.0265.

[17] A. Miličević, M. Despotović-Zrakić, T. Naumović, M. Suvajdžić, and B. Radenković, "Measuring the performance of the innovative potential of the academy on the example of Algorand WEB 3.0 hackathon," *E-business technologies conference proceedings*, vol. 3, no. 1, pp. 217–223, Jun. 2023, Accessed: Mar. 05, 2025. [Online]. Available: https://ebt.rs/journals/index.php/conf-proc/article/view/195

[18] I. Attalah, P. A. Nylund, and A. Brem, “Who captures value from hackathons? Innovation contests with collective intelligence tools bridging creativity and coupled open innovation,” *Creativity and Innovation Management*, vol. 32, no. 2, pp. 266–280, Jun. 2023, doi: 10.1111/CAIM.12552.

[19] A. Nolte, E. P. P. Pe-Than, A. Filippova, C. Bird, S. Scallen, and J. D. Herbsleb, “You Hacked and Now What?,” *Proc ACM Hum Comput Interact*, vol. 2, no. CSCW, Nov. 2018, doi: 10.1145/3274398.

[20] C. Wallwey, M. M. Longmeier, D. Hayde, J. Armstrong, R. Kajfez, and R. Pelan, “Consider ‘HACKS’ when designing hackathon challenges: Hook, action, collaborative knowledge sharing,” *Front Educ (Lausanne)*, vol. 7, p. 954044, Sep. 2022, doi: 10.3389/FEDUC.2022.954044/BIBTEX.

[21] V. Kyrychenko and V. Necherda, “HACKATHON AS A TECHNOLOGY OF FORMING A SOCIALLY SUCCESSFUL PERSONALITY OF A PUPIL,” *Theoretical and Methodical Problems of Children and Youth Education*, vol. 26, no. 1, pp. 156–168, Oct. 2022, doi: 10.32405/2308-3778-2022-26-1-156-168.

[22] A. Nandi and M. Mandernach, “Hackathons as an informal learning platform,” *SIGCSE 2016 - Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, pp. 346–351, Feb. 2016, doi: 10.1145/2839509.2844590.

[23] D. C. Yen, S. Lee, and S. Koh, “Critical knowledge/skill sets required by industries: An empirical analysis,” *Industrial Management and Data Systems*, vol. 101, no. 8, pp. 432–442, 2001, doi: 10.1108/EUM0000000006173/FULL/XML.

Analysis of the impact of FONIS hackathons on students career development

Dušan Kostić, Miloš Jolović, Aleksandar Joksimović, Tamara Naumović, Petar Lukovac

ABSTRACT

This paper analyzes the impact of hackathons organized by the Association of Students of Information Science FONIS of the participating students on their career development. Hackathons are recognized as an important form of informal education that encourages the development of technical and entrepreneurial skills. Main goal of the research is to examine how and to what extent participation in these events contributes to the professional development of the participants. The methodological framework of the research includes a survey study conducted on a sample of 71 participants of various hackathons organized in the period from 2013 to 2024. The research results indicate a significant contribution of the hackathon to the development of critical thinking, entrepreneurial thinking and professional skills of the participants. It was concluded that hackathons represent an effective platform for acquiring knowledge and networking, but that there is room for improvement, especially in the domain of post-hackathon activities and support for participants in the further development of their projects.

Upravljački podsistem za praćenje, unapređenje i kontrolu materijalnih tokova u proizvodnji industrije prerade drveta

Ksenija Ćosić
Fakultet organizacionih nauka
Beograd, Republika Srbija
ksenijacosic0411@gmail.com
0009-0006-6134-9064

Slobodan Antić
Fakultet organizacionih nauka
Beograd, Republika Srbija
slobodan.antić@fon.bg.ac.rs
0000-0003-2726-0235

Nemanja Tulimirović
Fakultet organizacionih nauka
Beograd, Republika Srbija
nemanja.tulimirović@fon.bg.ac.rs
0009-0008-9722-6108

***Apstrakt* - U radu je analiziran proces proizvodnje preduzeća koje posluje u oblasti drvne industrije i bavi se proizvodnjom drvenih paleta za transport i skladištenje robe. Analizi proizvodnog procesa je pristupljeno u cilju unapređenja postojećeg procesa proizvodnje, tako da se poveća produktivnost, efikasnost, pojednostave procesi upravljanja i planiranja, kao i da se unaprede procesi praćenja i donošenja odluka. Tokom istraživanja, na osnovu prikupljenih informacija o preduzeću i procesu proizvodnje, kreiran je upravljački informacioni podsistem u vidu spredšit aplikacije, kojim je predstavljen upravljački podsistem za praćenje i kontrolu materijalnih tokova, koji se sastoji iz dva osnovna modula: upravljačkog modela za planiranje proizvodnje, nabavke i prodaje drvenih proizvoda i modula za praćenje informacija o robno-materijalnim tokovima. U radu će biti naveden uspešan primer razvoja i implementacije spredšit aplikacije u vidu praktičnog primera spredšit informacionog sistema, kao rezultata primene znanja iz oblasti spredšit inženjerstva.**

Ključne reči – upravljački informacioni podsistem, upravljački model, spredšit aplikacija, planiranje proizvodnje, optimizacija

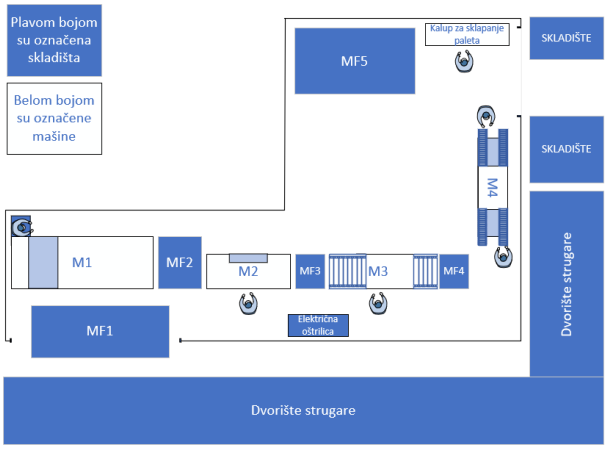
I. Uvod

U radu će biti proučavani rad i organizacija proizvodnje preduzeća „Rakić“. Preduzeće posluje u oblasti drvne industrije i bavi se proizvodnjom paleta za transport i skladištenje robe. Ideja rada je sagledavanje i analiza postojećeg stanja celokupnog poslovanja, a potom davanje predloga poboljšanja. Dati predlozi poboljšanja predstavljaju osnov za uvođenje upravljačkog informacionog podsistema. Cilj rada je razvoj aplikacije radi praćenja, unapređenja i kontrole materijalnih tokova. Spredšit aplikacija se sastoji iz dva osnovna modula: upravljačkog modela i modula za praćenje informacija o robno-materijalnim tokovima. Upravljački model služi za planiranje proizvodnje, nabavke i prodaje drvenih proizvoda. Materijalno praćenje tokova se vrši u delu aplikacije za robno-materijalno knjigovodstvo i služi za evidenciju poslovnih promena nad subjektima robno-materijalnih tokova. Razvijeni su radni listovi koji predstavljaju pomoćne evidencije koje će preduzeću biti od pomoći, s obzirom da preduzeće nema finansijsko knjigovodstvo. Aplikacija sadrži sve podatke iz proizvodnje, transformiše ih u informacije koje služe kao osnov za donošenje poslovnih odluka i tako omogućava efikasnije praćenje materijalnih tokova, a posledično unapređuje poslovanje i povećava profitabilnost preduzeća.

II. Opis postojećeg stanja

Preduzeće strugara „Rakić“ ima dugu istoriju poslovanja i svrstava se u domen malih porodičnih preduzeća. Osnovni proizvodi su dve vrste paleta, takozvane „obične“ i evro palete. Usled široke primene ovih proizvoda, tražnja je, uslovno rečeno, stalna i neograničena. Preduzeće može da proda onoliko paleta koliko može da proizvede. Postojeće stanje biće opisano kroz faze i zatim će biti prikazan dijagram robno-materijalnih tokova.

Strugara u procesu proizvodnje ima četiri mašine i dodatan alat. Mašine će biti imenovane redno: M1, M2, M3 i M4 (Slika 1). Mašine su postavljene linijski po redosledu operacija obrade materijala. Između svake mašine postoji prostor, koji služi kao međufazno skladište, takozvano bafer skladište. Bafer skladišta pomažu proizvođačima da pravovremeno zadovolje tražnju kupaca i poboljšaju fleksibilnost i produktivnost proizvodnog sistema [1]. Bafer skladišta su imenovana redno: MF1, MF2, MF3, MF4 i MF5.



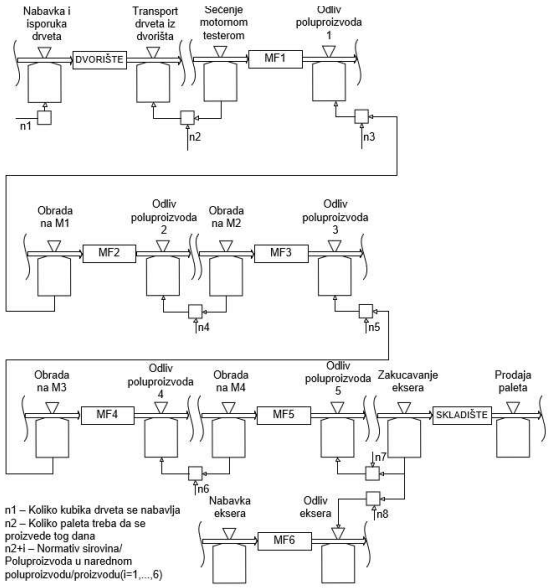
*Pod pojmom skladišta se podrazumevaju dvorište strugare, međufazna skladišta i skladište za odlaganje gotovih proizvoda

Slika 1. Tlocrt radionice

Pre nego što počne bilo koja operacija obrade drveta, potrebno je da se nabavi repromaterijal. Iako ne postoji fiksna količina koja se nabavlja, postoji utvrđen nivo sigurnosnih zaliha, koji kada se dostigne, naručuje se nova količina. Prilikom prodaje, dobavljači izdaju propratnicu kao prateću dokumentaciju. Izdavanje propratnice vrši se prilikom stavljanja posečenog drveta u promet i sadrži podatke o prodavcu i kupcu, kao i one vezane za specifikacije drveta (vrsta, obim, dužina, zapremina...) [2].

Materijalni tok počinje nabavkom drveta. Akcija nabavke drveta (prema normativu n1, koji određuje koliko je drveta potrebno nabaviti) prouzrokuje akumulaciju repromaterijala u dvorištu strugare. Na dnevnom nivou, radi proizvodnje, drvo se transportuje do prostora neposredno pored mašine M1 (akcija odliva iz dvorišta). Proizvodnja kreće tako što se viljuškarem prevozi potrebna količina drveta do međufaznog skladišta 1. Debla su duga po nekoliko metara, tako da se prvo motornom testerom krate na proizvoljnu dužinu (poluproizvod 1). Akcija skraćivanja drveta prouzrokuje akumulaciju poluproizvoda 1 na međufaznom skladištu 1 – MF1 (prema normativu n2 - koliko je potrebno paleta proizvesti taj dan). Odliv sa MF1 nastaje zbog uzimanja poluproizvoda radi obrade na mašini M1. Na mašini M1 se deblo seče po dužini, tako da rezultat ove faze budu daske sa neobrađenim stranama (poluproizvod 2). Prva i poslednja daska koje se iseku su daske koje imaju koru duž cele jedne strane, tako da se one odlažu ispred strugare kao otpad. Isečene daske sa dve neobrađene strane se ređaju na gomilu na podu između dve mašine (međufazno skladište 2) i spremne su za dalju obradu. Ova obrada (prema normativu n3 - koji je normativ utroška drveta za taj broj paleta) predstavlja akciju koja dovodi do akumulacije poluproizvoda 2 na međufaznom skladištu 2 – MF2. Odliv nastaje zbog uzimanja dasaka radi sledeće obrade na mašini M2. Daske se dalje obrađuju na mašini M2, koja seče samo jednu neobrađenu stranu. Sa međufaznog skladišta 2 se uzima daska, prislanja na graničnik i seče. Rezultat ove obrade su daske sa jednom obrađenom stranom (poluproizvod 3) i one se skladište između ove i naredne mašine (međufazno skladište 3). Akcija obrade na mašini M2, prema normativu n4 - koliko je poluproizvoda 2 potrebno za proizvodnju poluproizvoda 3, dovodi do akumulacije poluproizvoda na MF3. Odliv sa MF3 nastaje zbog uzimanja dasaka zbog naredne obrade na M3. Naredna obrada je na mašini M3 na kojoj se daske sa jednom obrađenom, ravnom stranom seku na dužinu 1.05 m ili 1.25 m. Uzima se daska sa međufaznog skladišta 3, postavlja prema graničniku i krati (poluproizvod 4). Ukoliko je potrebno odseći deo daske usled nezadovoljavajućeg kvaliteta, daske se seku i na dužine od pola metra. Skraćene daske se po završetku sečenja ređaju na gomilu neposredno pored mašine (međufazno skladište 4). Akcija obrade na mašini M3, kao i akumulacija na međufaznom skladištu MF4, su analogne akciji obrade na mašini M2 i akumulaciji na međufaznom skladištu MF3, sa pratećim normativima. Zatim se, sa MF4, uzima daska sa jednom obrađenom ravnom stranom i spušta na konvejer M4 okrenuta licem obrađene strane. Mašina ima dva konvejera, tako da kada testera iseče jednu letvu (poluproizvod 5), sa druge strane mašine se uzima daska koja dalje može da se krati i stavlja se na konvejer, koji odvodi dasku do druge strane mašine, gde se daska ponovo premešta na prvi konvejer. Ovo je poslednja mašina što se tiče obrade letvi potrebnih za sklapanje paleta. Nakon sečenja svake letve, ređaju se gotovi delovi na paletu. Letve dužine 1.05 m i 1.25 m ređaju se na jednu paletu, a letve dužine 0.5m na drugu paletu. Kada se natovare palete, ručnim viljuškarem se prevoze do mesta neposredno pored kalupa za sklapanje palete (međufazno skladište 5). Potom se uzimaju letve sa palete i ređaju u kalup. Akcija obrade na mašini M4, kao i akumulacija na međufaznom skladištu MF5, su analogne akciji obrade na mašini M3 i akumulaciji na međufaznom skladištu MF4, sa pratećim normativima. Osim osnovnog materijala drveta, za proizvodnju paleta potreban je još jedan pomoćni materijal – ekseri. Akcijom nabavke eksera dolazi do akumulacije na međufaznom skladištu 6 – MF6. Zbog naredne akcije, finalizacije paleta, postoji akcija odliva eksera. Kada su poređane letve u kalup i nakon obavljene kontrole, obavlja se zakucavanje eksera na predefinisana mesta. Gotove palete se odlažu na gomilu pored kalupa. Zalihe paleta se izvoze manuelnim viljuškarem u dvorište (skladište). Akcija finalizacije paleta dovodi do akumulacije gotovih proizvoda na skladištu. Normativi n7 i n8 određuju koliko je letvi, odnosno eksera, potrebno za formiranje jedne palete. Akcija odliva nastaje usled prodaje paleta kupcima, što je ujedno poslednja akcija. Kada dođe do datuma isporuke, strugara prevozi gotove proizvode do kupaca i izdaje otpremnicu i fakturu. Otpremnica je dokument koji prati isporuku robe kojim se evidentira izlaz robe, odnosno gotovih proizvoda, iz skladišta i služi za zaduženje kupca. Nastaje u trenutku izvršenja naloga [3]. Otpremnica sadrži podatke poput osnovnih informacija o prodavcu, odnosno kupcu, datumu, nazivu robe i količini i druge. Faktura takođe služi za zaduženje kupaca i sadrži objedinjene podatke iz otpremnica, ukoliko ih je bilo više i tada se vrši plaćanje. S obzirom da je preduzeće „Rakić“ paušalnog tipa i nije u sistemu PDV-a, ono je u obavezi da vodi Poslovnu knjigu o ostvarenom prometu. Prema zakonu, potrebno je voditi evidenciju samo o prihodima koje preduzeće ostvari, bez obaveze pravdanja nastalih troškova. Za vođenje ove knjige potrebno je evidentirati sve fakture koje su dokaz prometa [4].

U nastavku rada biće dat dijagram robno-materijalnih tokova. Dijagram robno-materijalnih tokova daje sistematizovan pregled redosleda operacija, njihove povezanosti i pratećih normativa. Dijagram prati robno-materijalni tok od trenutka nabavke sirovina, transformacije ulaza (sirovina) u izlaze (gotove proizvode) kroz niz akcija i akumulacija sa pratećim normativima.



Slika 2. Dijagram robno-materijalnih tokova

III. ANALIZA I PREDLOZI UNAPREĐENJA SA ASPEKTA PRAĆENJA INFORMACIJA O ROBNO-MATERIJALNIM TOKOVIMA

Naručivanje paleta i naručivanje repromaterijala od dobavljača obavljaju se usmeno. Kupci, odnosno preduzeće, putem telefona dogovaraju željeni proizvod, količinu i ostale uslove kupovine. Ne postoji tačno definisan plan nabavke, niti se nabavka planira na osnovu tražnje, već se kupuje proizvoljna količina kada je to potrebno. Preduzeće ne koristi narudžbenice, niti drugi vid dokumenta. Kada je reč o

planiranju sirovina, njihove nabavke, planiranja proizvodnje, praćenja stanja i količina proizvedenih poluproizvoda i proizvoda, kao i prodatih proizvoda, preduzeće ne poseduje informacioni sistem za te svrhe. Nepostojanje ovakvog sistema otežava upravljanje poslovanjem i postoji verovatnoća da se javi neki vid propusta.

A. *Predlog sistema šifriranja*

U slučaju velikog broja proizvoda i materijala javlja se potreba za klasifikacijom, odnosno razvrstavanjem proizvoda [5]. Osim slučaja mnogobrojnosti, da bi se upravljalo subjektima tokova, potrebno je iste prepoznati i učiniti vidljivim, što se može obezbediti uvođenjem sistema šifriranja. Šifra predstavlja sređeni skup znakova (brojeva, slova i drugih znakova) koji na skraćen i sistemski način odražavaju određeni podatak ili skup podataka objekta [5], [6]. Šifra nedvosmisleno ukazuje na sve bitne karakteristike materijala, poluproizvoda, proizvoda i ostataka. U nastavku će biti dat predlog sistema klasičnog, alfanumeričkog šifriranja (Slika 3). Uvođenje ovakvog sistema šifriranja je preduslov za lakše planiranje i upravljanje subjektima tokova, kao i za praćenje njihovih podataka o količini i vrednosti. Šifre sadrže deo koji se odnosi na lokaciju objekta, grupu, podgrupu i deo koji se odnosi na identifikacioni broj (ID) [7].

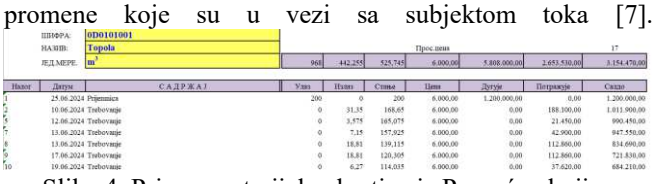
ŠIFRIRANJE	
Naziv	**Šifra**
Topola	D0101001
Lipa	D0101002
Bor	D0101003
Ekseri	M0102001
Ulje za podmazivanje	M0102002
Poluproizvod 1	F0201001
Poluproizvod 2	F0202001
Poluproizvod 3	F0203001
Poluproizvod 4.1.	F0204001
Poluproizvod 4.2	F0204002
Poluproizvod 4.3	F0204003
Poluproizvod 5.1	F0205001
Poluproizvod 5.2	F0205002
Poluproizvod 5.3	F0205003
Obična paleta	S0301001
Euro paleta	S0302001
Krupni drvni otpad	S0302002
Sitni drvni opad	S0302003
Piljevina	S0302004

Mesto	Grupa	Podgrupa	ID	Šifra
D	01	01	001	D0101001
D	01	01	002	D0101002
D	01	01	003	D0101003
M	01	02	001	M0102001
M	01	02	002	M0102002
F1	02	01	001	F0201001
F2	02	02	001	F0202001
F3	02	03	001	F0203001
F4	02	04	001	F0204001
F4	02	04	002	F0204002
F4	02	04	003	F0204003
F5	02	05	001	F0205001
F5	02	05	002	F0205002
F5	02	05	003	F0205003
S	03	01	001	S0301001
S	03	02	001	S0302001
M	04	01	01	S0302002
M	04	02	01	S0302003
M	04	02	02	S0302004

Slika 3. Predlog šifriranja

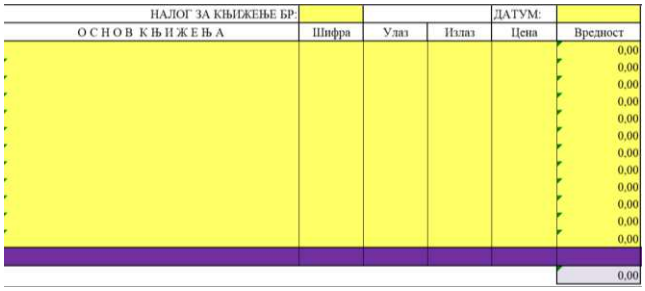
B. *Predlog za uvođenje materijalne kartice i sistema praćenja robno-materijalnih tokova*

Nakon uvođenja sistema šifriranja, mogu se uvesti robno-materijalne kartice. Subjekat toka je materijalne prirode i prolazi kroz niz faza koje čine naizmenične akcije i akumulacije. Za svaki subjekat toka (materijal, poluproizvod, proizvod i ostatke) neophodna je materijalna kartica koja predstavlja zasebnu bazu podataka. Materijalna kartica sadrži podatke o šifri, nazivu, jedinici mere, nalogu, datumu, sadržaju, ulazu, izlazu, stanju, polja duguje i potražuje, saldo, kao i cenu. Materijalnom karticom se prati protok količina, protok vrednosti i na taj način se evidentiraju sve poslovne promene koje su u vezi sa subjektom toka [7].



Slika 4. Primer materijalne kartice iz Pomoćne knjige

Formiranjem materijalnih kartica i sistema za njihovo praćenje u Excel-u, kao i uvođenjem šifriranja, preduzeće bi moglo u svakom trenutku da raspolaže informacijama o stanju svih subjekata tokova od trenutka njihove nabavke, pa sve do prodaje finalizovanih proizvoda. Nakon otvaranja kartice, kada se desi određena poslovna promena, potrebno je promenu proknjižiti. Za knjiženje se koristi forma naloga za materijalno knjiženje, gde se unosi broj naloga, datum, osnov knjiženja, šifra materijalne kartice, ulaz, izlaz, cena i vrednost (Slika 5). Pokretanjem makroa, promena se, na osnovu šifre, knjiži na adekvatnu materijalnu karticu.



Slika 5. Izgled naloga za knjiženje

C. *Predlog uvođenja potrebne dokumentacije*

Uvođenje materijalnih kartica i praćenje promena, zahtevalo bi i uvođenje propratne dokumentacije, odnosno validne knjigovodstvene isprave. Dokument je isprava kojom se verifikuje poslovna promena i služi kao osnov za knjiženje. Knjigovodstveni dokumenti koje bi trebalo uvesti su: narudžbenica materijala, trebovanje materijala, skladišnica za međufazna skladišta i skladište gotovih proizvoda i narudžbenica kupca. Pre knjiženja svaki od dokumenata mora biti likvidiran, odnosno, mora se sprovesti kontrola ispravnosti od strane ovlašćenih lica [7]. Narudžbenica je dokument koji kupac šalje dobavljaču i kojom se iskazuje količina i vrsta robe koju želi da kupi [8]. Trebovanje materijala je interni dokument koji proizvodnja koristi i predstavlja osnov za izdavanje materijala iz magacina. Važnost ovog dokumenta se ogleda u upravljanju zalihama subjekata tokova. Poput narudžbenice, dokument trebovanje služi i za praćenje i upravljanje subjektima materijalnih tokova, kao i za praćenje zaliha i kontrolu troškova. Skladišnica je dokument kojim bi bilo omogućeno praćenje promena i stanja zaliha repromaterijala, poluproizvoda i gotovih proizvoda [9], što je posebno važno zbog postojanja velikog broja međufaznih skladišta.

D. *Predlog za uvođenje upravljačkog modela za planiranje nabavke i planiranje proizvodnje*

S obzirom da preduzeće nema uvid u stanja subjekata tokova u strugari, nema adekvatan sistem planiranja nabavke i proizvodnje, praćenja ostataka i škarta, postoji potreba za uvođenjem adekvatnog sistema, odnosno modela, koji sve to obuhvata. Uvođenjem upravljačkog modela, preduzeću bi bilo omogućeno da prati i planira ne samo potrebnu količinu sirovina, već i troškove, vreme rada, vreme isporuke i drugo. Model bi za unetu traženu količinu, na osnovu normativa, normi i produktivnosti, računao koliko je sirovina potrebno za tu količinu proizvoda, koliko bi trebalo raditi da bi se proizvela tražena količina, kada će tražena količina biti završena, koliko se ostataka može očekivati i drugo.

Imajući u vidu veličinu preduzeća i obim poslovanja, Excel predstavlja najefikasnije rešenje. Pomoću Excel-a korisnici mogu obavljati širok raspon zadataka, od svakodnevnih transakcija do finansijskih izveštaja. Nabavka drugih računovodstvenih informacionih sistema može biti poprilično skupa, dok se u Excel-u može dizajnirati uz veoma niske troškove, kao i prilagoditi industriji u kojoj posluje dato preduzeće [10].

IV. Razvoj spredšit aplikacije

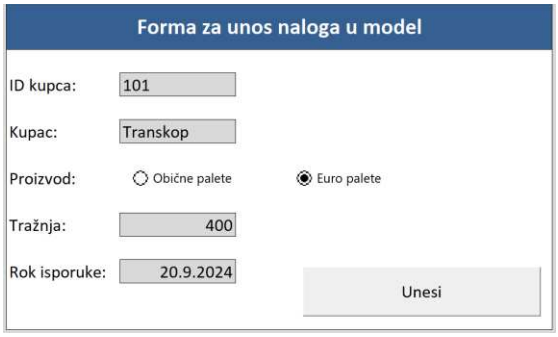
Razvijena spredšit aplikacija namenjena je upravljanju materijalnim tokovima i planiranju proizvodnje. Aplikacija se sastoji iz dve celine: upravljačkog modela i materijalnog knjigovodstva. Upravljački model je napravljen tako da preduzeću olakša planiranje i upravljanje proizvodnjom. Ideja modela je da pruži preduzeću bolji uvid u proizvodnju obuhvatanjem i transformacijom podataka. Unosom osnovnih podataka, model računa podatke poput vrednosti narudžbine, potrebne količine drveta za narudžbinu, proverava da li ima dovoljno drveta na stanju, određuje trajanje proizvodnje, kao i početak i kraj proizvodnje. Model takođe sadrži makroe pomoću kojih se mogu automatski kreirati narudžbenice, trebovanja, skladišnice i fakture. Kako preduzeće nema svoje evidencije, kreirani su i radni listovi „Narudžbine“, „Evidencija prodaje“ i „Poslate narudžbenice“. Radni list „Narudžbine“ služi za evidenciju narudžbenica kupaca koje se kreiraju pomoću modela. Na ovaj način se preduzeću pruža sistematizovan i pregledan uvid u buduću proizvodnju. S obzirom da je preduzeće paušalnog tipa, postoji zakonska obaveza da se vodi Poslovna knjiga o ostvarenom prometu. Za vođenje ove knjige potrebno je evidentirati sve fakture koje su dokaz prometa. Za evidenciju prometa, koristi se radni list „Evidencija prodaje“. Na kraju, radni list „Poslate narudžbenice“ služi za evidenciju narudžbenica preduzeća upućenim dobavljačima.

A. *Upravljački model – struktura, funkcionalnosti i primena*

Nakon pokretanja aplikacije, aktivira se početni ekran aplikacije. Klikom na dugme „Ulaz“ pristupa se glavnom meniju. U glavnom meniju se nalazi dugme „Materijalno knjigovodstvo“, čijim se pokretanjem pristupa meniju. U meniju se nalaze prečice, odnosno dugmad, čijim se pokretanjem direktno otvaraju radni listovi aplikacije. Kroz meni se može pristupiti sledećim radnim listovima: „Forma za nalog“; „Upravljački model“, „Narudžbine kupaca“, „Poslate narudžbine“, „Nalog za knjiženje“, „Otvaranje kartica“, „Pomoćna knjiga“ i „Evidencija prodaje“.

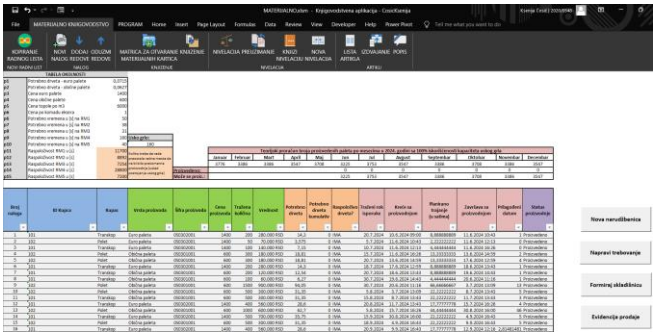
Slika 6. Meni aplikacije

Glavni radni list „Model“ sastoji se iz četiri celine: tabele modela, tabele okolnosti, proširene tabele Teorijski proračun broja proizvedenih paleta po mesecima u 2024. godini sa 100% iskorišćenosti kapaciteta uskog grla i dela sa dugmadima. Da bi se podaci uneli u „Model“ potrebno je uneti ih putem forme (Slika 7), kojoj se može pristupiti iz menija ili klikom na radni list „Forma“.



Slika 7. Forma za unos naloga

Pokretanjem unosa klikom na dugme Unesi, podaci se prenose u model koji se nalazi u radnom listu „Model“. Nakon unosa podataka iz forme, model na osnovu naziva proizvoda unosi šifru proizvoda i cenu. Na osnovu tražene količine računa vrednost te porudžbine i potrebnu količinu drveta. Podatke o ceni palete, na osnovu vrste, preuzima iz tabele okolnosti koja se nalazi iznad tabele modela. Potrebnu količinu drveta računa na osnovu normativa za datu vrstu, koji se takođe nalazi u tabeli okolnosti.



Slika 8. Izgled radnog lista upravljačkog modela

Model potom proverava da li ima dovoljno drveta na stanju i da li se željena količina može proizvesti do traženog roka isporuke. Vrednosti u koloni „Potrebno drveta“ se računaju kao količnik tražene količine i normativ željene palete, preuzet iz tabele okolnosti. Dakle, ukoliko je status proizvodnje „proizvedeno“ ili „u toku“, trebovanje drveta je proknjiženo i već skinuto sa stanja sa materijalne kartice. Ukoliko je status proizvodnje „čeka“, kumulativ dobija vrednost potrebnog drveta. Vreme potrebno za proizvodnju računa se na osnovu dnevne produktivnosti radionice koja je definisana uskim grlom. Za proračun završetka proizvodnje potrebno je uzeti u obzir samo radne dane, jednu osmočasovnu radnu smenu i svakodnevna čišćenja koja je potrebno obaviti. „Planirano trajanje (u satima)“ je pomoćna vrednost i ona se dobija kada se pomnoži tražena količina vremenom rada na uskom grlu za jednu paletu. Vrednost koja se dobije predstavlja ukupno vreme proizvodnje. Na kraju, status proizvodnje se menja automatski u zavisnosti od toga da li proizvodnja čeka, da li je u toku i da li se završila analogno

odnosu planiranog početka, odnosno završetka proizvodnje i sadašnjeg trenutka.

Nakon unosa tražene količine, u tabeli „Teorijski proračun broja proizvedenih paleta po mesecima u 2024. godini (sa 100% iskorišćenosti kapaciteta uskog grla)“ se računa koliko bi paleta bilo proizvedeno u datom mesecu i da li ta količina premašuje teorijski maksimalan broj paleta. Osim toga, u poslednjem redu tabele se prikazuje koliko se još paleta može proizvesti u datom mesecu. Kada su uneti svi podaci, kada je utvrđeno da ima dovoljno drveta i da se proizvodnja može realizovati do traženog roka, moguće je formirati narudžbenicu. Narudžbenica se formira klikom na dugme, koje se nalazi pored tabele modela, sa nazivom „Nova narudžbenica“. Pokretanjem ovog dugmeta, pomoću makroa, preuzimaju se potrebni podaci za popunjavanje narudžbenice i beleže se u radni list „Narudžbine“ (Slika 9). Radni list „Narudžbine“ je jedan od tri raspoloživa lista koja služe preduzeću za evidenciju.

NARUDŽBENICE

Broj naloga	ID kupca	Kupac	Šifra proizvoda	Vrsta proizvoda	Tražena količina	Vrednost	Kreće sa proizvodnjom	Završava sa proizvodnjom
1	101	Transkop	0S0301001	Obična paleta	500	30.000 RSD	10.6.2024	12.6.2024
2	102	Polet	Euro paleta	0S0302001	600	840.000 RSD	12.6.2024	13.6.2024
3	101	Transkop	Euro paleta	0S0302001	100	140.000 RSD	13.6.2024	13.6.2024
4	102	Polet	Obična pale	0S0301001	300	180.000 RSD	13.6.2024	17.6.2024
5	102	Polet	Obična pale	0S0301001	300	180.000 RSD	17.6.2024	19.6.2024
6	101	Transkop	Euro paleta	0S0302001	200	280.000 RSD	19.6.2024	20.6.2024
7	101	Transkop	Euro paleta	0S0302001	200	280.000 RSD	18.6.2024	19.6.2024
8	101	Transkop	Obična pale	0S0301001	100	60.000 RSD	19.6.2024	20.6.2024
9	102	Polet	Obična pale	0S0301001	1500	900.000 RSD	20.6.2024	2.7.2024
10	102	Polet	Obična pale	0S0301001	700	420.000 RSD	2.7.2024	8.7.2024
11	101	Transkop	Obična pale	0S0301001	500	300.000 RSD	8.7.2024	11.7.2024
12	101	Transkop	Euro paleta	0S0302001	400	560.000 RSD	11.7.2024	15.7.2024
13	102	Polet	Obična pale	0S0301001	1000	600.000 RSD	15.7.2024	23.7.2024
14	101	Transkop	Euro paleta	0S0302001	500	700.000 RSD	30.8.2024	4.9.2024
15	102	Polet	Obična pale	0S0301001	500	300.000 RSD	4.9.2024	9.9.2024
16	101	Transkop	Euro paleta	0S0302001	400	560.000 RSD	9.9.2024	12.9.2024

Slika 9. Radni list „Narudžbine”

Nakon evidentirane narudžbenice, na dan početka proizvodnje treba napraviti trebovanje potrebnog drveta za proizvodnju željene količine. Klikom na dugme „Napravi trebovanje“, koje se nalazi pored tabele modela, preuzimaju se potrebni podaci i popunjava nalog za knjiženje u radnom listu „Nalog“. Zatvaranjem prozora sa obaveštenjem se otvara radni list „Nalog“. Nakon pokretanja opcije za knjiženje, trebovanje će biti proknjiženo i evidentirano na kartici sa odgovarajućom šifrom. O tome će biti reči u narednom poglavlju.

Kada se status proizvodnje promeni i dobije status „Proizvedeno“, proizvedena količina ulazi u skladište i trebalo bi je evidentirati pomoću skladišnice. Klikom na dugme „Formiraj skladišnicu“ koje se nalazi pored tabele forme, pokreće se makro, pomoću kog se uzimaju potrebni podaci i popunjava nalog za knjiženje u radnom listu „Nalog“. Nakon pokretanja, prikazuje se upit za unos željenog broja naloga. Nakon unosa i provere, prikazuje se potvrda o uspešnom popunjavanju naloga.

Kada je određena narudžbina proizvedena i evidentirana u skladištu, onda sledi prodaja korisniku. Pokretanjem dugmeta „Evidencija prodaje“ koje se nalazi pored tabele modela, prvo se beleži prodaja u prethodno navedenom radnom listu, a potom se popunjava nalog sa fakturom usled odliva. Pokretanjem ovog dugmeta, prikazuje se upit za broj naloga iz tabele modela koji treba prodati. Prilikom popunjavanja evidencije prenose se podaci iz modela u radni list „Evidencija prodaje“ (Slika 10). Zatim se prikazuje upit o datumu kada se obavlja prodaja.

EVIDENTIRANI PROMET SA KUPCIMA

Broj naloga	ID kupca	Kupac	Šifra proizvoda	Vrsta proizvoda	Tražena količina	Vrednost	Datum prodaje
1	101	Transkop	Obična paleta	0S0301001	500	300.000 RSD	13.6.2024
2	102	Polet	Euro paleta	0S0302001	50	70.000 RSD	13.6.2024
3	101	Transkop	Euro paleta	0S0302001	100	140.000 RSD	14.6.2024
4	102	Polet	Obična paleta	0S0301001	300	180.000 RSD	18.6.2024
5	102	Polet	Obična paleta	0S0301001	300	180.000 RSD	25.6.2024
6	101	Transkop	Euro paleta	0S0302001	200	280.000 RSD	22.6.2024
7	101	Transkop	Euro paleta	0S0302001	200	280.000 RSD	17.6.2024
8	101	Transkop	Obična paleta	0S0301001	100	60.000 RSD	7.7.2024
9	102	Polet	Obična paleta	0S0301001	1500	900.000 RSD	8.7.2024
12	101	Transkop	Euro paleta	0S0302001	400	560.000 RSD	17.7.2024
13	102	Polet	Obična paleta	0S0301001	1000	600.000 RSD	24.7.2024
15	102	Polet	Obična paleta	0S0301001	500	300.000 RSD	9.9.2024
						3.850.000 RSD	

Slika 10. Radni list „Evidencija prodaje"

Na kraju, postoji i treći list za evidenciju – „Poslate narudžbenice“. U ovom listu se kreiraju, evidentiraju i potom knjiže narudžbenice upućene dobavljačima (Slika 11).

NARUDŽBINE ZA DRVO

Broj naloga	Dobavljač	Šifra drveta	Naziv drveta	Tražena količina	Vrednost	Očekivani datum isporuke
1	Popovac Doo	0D0101001	Topola	200	1200000	25.6.2024
2	Popovac DOO	0D0101001	Topola	268	1608000	2.7.2024
3	Popovac DOO	0D0101001	Topola	500	3000000	5.8.2024

Evidentiraj narudžbenicu

Proknjiži prijem

Slika 11. Radni list „Poslate narudžbenice"

Klikom na dugme Evidentiraj narudžbenicu, pokreće se makro za unošenje potrebnih podataka. Da bi se popunila narudžbenica za dobavljača, prikazuju se prozori za unos podataka. Prvo se prikazuje prozor za unos naziva dobavljača, zatim sledeći za vrstu drveta koja se naručuje i poslednji prozor za unos poručene količine drveta. Ostali podaci iz tabele se popunjavaju automatski. Kada porudžbina pristigne, potrebno je da se unese datum, kako bi se obezbediti svi podaci za knjiženje prijemnice. Nakon unosa datuma, pokretanjem dugmeta Proknjiži prijem, aktivira se makro koji služi za popunjavanje naloga za knjiženje prijemnice. Potrebno je uneti broj naloga narudžbenice preko prozora za unos podataka kako bi se preuzeli potrebni podaci. Podaci su potrebni za popunjavanje naloga za knjiženje. Potom se obavlja prenos potrebnih podataka u nalog za knjiženje prijemnice.

B. Materijalno knjigovodstvo – struktura, funkcionalnosti i primena

Materijalno knjigovodstvo je deo upravljačkog računovodstva. Upravljačko računovodstvo podrazumeva obradu i interpretaciju računovodstvenih informacija koje služe kao pomoć preduzeću u procesu donošenja poslovnih odluka [11]. Materijalno knjigovodstvo služi za evidenciju i praćenje robno-materijalnih tokova. Da bi to bilo moguće, neophodno je otvoriti robno-materijalnu karticu za svaki subjekat toka (Slika 4). Svaka robno-materijalna kartica predstavlja zapravo zasebnu bazu podataka [7]. Knjiženje ulaza gotovih proizvoda u skladište može da se vrši po ceni koja je definisana. Gotovi proizvodi u skladištu se mogu voditi po stvarnoj ceni koštanja, planskoj ceni koštanja ili prodajnoj ceni [7]. U datom primeru se koristi prodajna cena. Otvaranje robno-materijalnih kartica se realizuje u radnom listu „Matrica“. Potrebno je uneti šifru, naziv i jedinicu mere

subjekta toka. Klikom na dugme „Otvaranje“, podaci iz „Matrice“ se prenose u pomoćnu knjigu, odnosno u radni list „Poknjiga“, gde će se dalje voditi evidencija o svim promenama u vezi sa datim subjektom toka. Nakon otvaranja robno-materijalnih kartica, kreira se baza podataka svakog subjekta toka. Kada se desi neka poslovna promena, potrebno je evidentirati je knjiženjem. Verifikacija poslovne promene se vrši pomoću dokumenata koji su osnov knjiženja [7]. Neophodno je popuniti nalog za knjiženje, koji se nalazi u radnom listu „Nalog“, odgovarajućim dokumentom i podacima koji opisuju nastalu promenu (Slika 5). Nakon popunjavanja naloga potrebnim podacima, neophodno je pokrenuti opciju „Knjiženje“. Proknjižene promene se evidentiraju na robno-materijalnim karticama odgovarajućih subjekata tokova u pomoćnoj knjizi, tj. u radnom listu „Poknjiga“.

V. ZAKLJUČAK

U radu je predstavljen primer razvoja i implementacije spredšit aplikacije u vidu praktičnog primera spredšit informacionog sistema, kao rezultata primene znanja iz oblasti spredšit inženjerstva.

Uvođenje spredšit aplikacije treba da omogući preduzeću uvid u svakodnevno poslovanje, koje ranije nije postojalo. Pomoću aplikacije svi podaci bi bili objedinjeni na jednom mestu. Automatizovani su procesi formiranja narudžbenica, trebovanja, skladišnica, evidencije prodaje gotovih proizvoda, kao i evidencije prijema nabavljenog repromaterijala. Uvođenje novih dokumenata i automatizacija procesa njihovog formiranja učiniće poslovne promene transparentnim preduzeću i pojednostaviti planiranje. Pomoćni radni listovi evidencija „Narudžbine“, „Evidencija prodaje“ i „Poslate narudžbenice“ treba da omoguće sistematizovan pregled prometa koji bi inače bio evidentiran u finansijskom knjigovodstvu. Na ovaj način će preduzeće imati uvid i u ostvareni promet.

Implementacija predloženih poboljšanja i spredšit aplikacije dovela bi do značajnog unapređenja celokupnog poslovanja preduzeća „Rakić“. Poboljšana produktivnost i efikasnije upravljanje rezultiraće većim koeficijentom obrta novčanih sredstava, većom profitabilnošću i kraćim vremenima isporuke.

LITERATURA

[1] Hadian, S. M., Farughi, H., & Rasay, H. (2021). Joint planning of maintenance, buffer stock and quality control for unreliable, imperfect manufacturing systems. Computers & Industrial Engineering, 157, 107304. https://doi.org/10.1016/j.cie.2021.107304

[2] Paragraf Lex. (n.d.). Pravilnik o obliku i sadržini šumskog žiga, obrascu propratnice, odnosno otpremnice, uslovima i načinu žigosanja posečenog drveta, načinu vođenja evidencije i načinu žigosanja, odnosno obeležavanja četinarskih stabala namenjenih sa novogodišnje I druge praznike., [http://demo.paragraf.rs/demo/combined/Old/t/t2016_11/t11_0253.htm#:~:text=Propratnica%2C%20odnosno%20otpremnica%2C%20izdaju%20se,smatraju%20%C5%A1 umom%20izdaje%20se%20propratnica., datum pristupa 15.02.2025.]

[3] Reclamare, N. (n.d.). Pojam otpremnice, obeležja i pravno uređenje - Knjigovodstvena agencija Tio Trade doo. © Tio Trade. [https://www.knjigovodstveneagencije.com/racunovodstvena-agencija/pojam otpremnice-obelezja-i-pravno-uredjenje , datum pristupa 15.02.2025.]

[4] Selkić, S. (2022, May 11). KPO knjiga. Unija Smart Accounting. [https://unija.com/sr/kpo-knjiga/, датум приступа 18.02.2025.] J. Surutka, “Osnovi elektrotehnike,” 1. izdanje, Naučna knjiga Beograd, 1988, str. 131.

[5] Mrvica Mađarac, S. (2023). Nabavno poslovanje. Veleučilište “Lavoslav Ružička” u Vukovaru. https://www.vevu.hr/wp-content/uploads/2023/03/Nabavno poslovanje_prirucnik.pdf

[6] KNEGO, N., (2013), "Informacijska tehnologija u nabavi, sustavi šifriranja i e-nabava, Ekonomski fakultet-Zagreb, Katedra za trgovinu. http://web.efzg.hr/dok/TRG/bknezevic/nabava bor2013/informacijskatehnologijaunabavi.pdf. [datum pristupa 18.02.2025.].

[7] Kostić, K., Antić, S., & Đorđević Milutinović, L. (2014). Informacioni sistemi preduzeća u Excel-u. Fakultet organizacionih nauka.

[8] Tjuluku, G. R., Windawati, D. P., & Christiani, T. A. (2023). Legal consequences of breach of promise for buyers in the sale and purchase of Purchase order (PO) system. International Journal of Multidisciplinary Research and Analysis, 06(06). https://doi.org/10.47191/ijmra/v6-i6-46

[9] Miranda, M. J., Mulangu, F. M., & Kemeze, F. H. (2019). Warehouse receipt financing for smallholders in developing countries: Challenges and limitations. Agricultural Economics, 50(5), 629–641. https://doi.org/10.1111/agec.12514

[10] Miranda, M. J., Mulangu, F. M., & Kemeze, F. H. (2019). Warehouse receipt financing for smallholders in developing countries: Challenges and limitations. Agricultural Economics, 50(5), 629–641. https://doi.org/10.1111/agec.12514

[11] Mihailović, I. (n.d.). Upravljačko računovodstvo. Autorizovana Predavanja. https://vpsle.edu.rs/wp-content/uploads/2016/04/Upravljacko-racunovodstvo-Ivan Mihailovic.pdf

Management subsystem for monitoring, improving, and controlling material flows in the wood processing industry

Ksenija Ćosić, Slobodan Antić, Nemanja Tulimirović

ABSTRACT

This paper analyzes the production process of a company operating in the wood industry, specializing in the manufacturing of wooden pallets for the transportation and storage of goods. The analysis of the production process was conducted with the aim of improving the existing production process to increase productivity and efficiency, simplify management and planning processes, and enhance monitoring and decision-making based on information collected from production. During the research, based on the collected information about the company and its production process, a management information subsystem was created in the form of a spreadsheet application. This subsystem represents a management system for monitoring and controlling material flows, consisting of two main modules: a management model for production planning, procurement, and sales of wooden products, and a module for tracking information on material flows. The paper presents a successful example of the development and implementation of a spreadsheet application as a practical case of a spreadsheet-based information system, demonstrating the application of knowledge in the field of spreadsheet engineering.

Simulacija rada 3D modela disk kočnice

Dragan R. Stojadinović
Tehnički opitni centar
Beograd Srbija
dragan.stojadinovic90@gmail.com, dragan.stojadinovic@toc.vs
ORCID: https://orcid.org/0009-0004-6534-8443

***Apstrakt* - Podrška računara procesu konstruisanja korišćenjem odgovarajućih softverskih alata sve je više zastupljena u savremenom inženjerstvu. Osim konstruisanja, ona obuhvata direktnu kontrolu mašinskih sistema, proizvodnih linija, kao i logističkih problema u tokovima materijala i tokovima podataka neophodnih u procesu proizvodnje. Različite modele u fazi konstruisanja trebalo bi međusobno formirati i transformisati softverski na što jednostavniji način, pri čemu se dolazi do daleko više informacija i podataka o krajnjem proizvodu, nego na osnovu same konstrukcione dokumentacije. Virtuelne tehnike uvode se u sve faze konstruisanja u vidu modeliranja, kao i simulacije kinematike i strukturne optimizacije modela. U ovom radu može se sagledati kako savremeni softverski alati pomažu u razvoju i projektovanju mašinskih sistema i sklopova koji su u ovom slučaju disk kočnice.**

Ključne reči – projektovanje, model, kočnice, simulacija, softver.

I. UVOD

Savremeno tržište zahteva iznalaženje novih konstrukcionih rešenja, što iziskuje znanje, inovativnost, ekonomičnost i sklonost za usavršavanjem i praćenjem razvoja. Istraživanje i razvoj bitno utiču na osnovne karakteristike proizvoda, kao što su pouzdanost, izdržljivost, kvalitet i adaptibilnost. Izlazak proizvoda na tržište direktno je proporcionalno navedenim osnovnim karakteristikama tog proizvoda. Zato je bitno formirati geometrijski model, koji se dalje može lako analizirati i proučavati, ali i brzo transformisati i dorađivati. [1]

Geometrijski modeli se upotrebljavaju kao virtuelni, računarski modeli, sa ciljem da se prvenstveno omogući vizuelizacija proizvoda. Vizuelizacija se generiše polazeći od neophodnih ortogonalnih projekcija dela. Savremeno projektovanje mašinskih sistema podržano računarom podrazumeva i numeričku simulaciju dinamičkog ponašanja. Težnja je da virtuelno okruženje što više odgovara realnim uslovima eksploatacije.

Računarski model treba biti parametarski definisan geometrijski 3D model. To podrazumeva razvoj proizvoda u virtuelnom okruženju, nakog čega se realizuje fizički model. [2]

Izrada prototipova nekog proizvoda kako bi se na njemu obavila određena probna ispitivanja, zahteva i velika novčana ulaganja. Savremeni softverski alati omogućavaju projektantima provere, analize i testiranja u cilju automatizacije projektovanja i smanjenja troškova. Takvi alati se primenjuju prilikom rešavanja različitih inženjerskih, istaživačko-razvojnih i industrijskih problema.

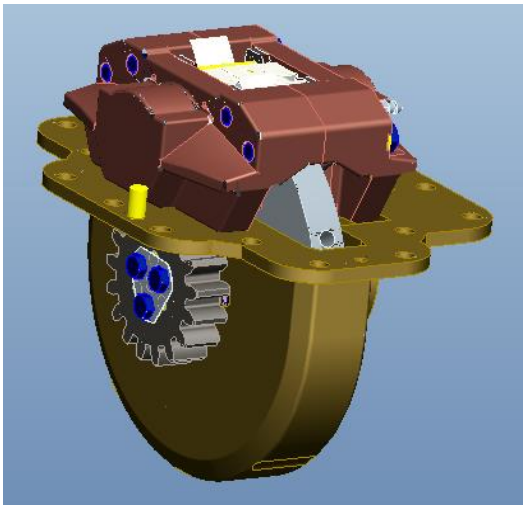
Cilj je da se broj probnih uzoraka smanji ili da oni u potpunosti ne budu potrebni, radi razmatranja određenih problema u razvoju. Umesto na gotovom proizvodu, ti problemi bi se analizirali i rešavali na 3D modelu brzo i lako.

II. FORMIRANJE GEOMETRIJSKOG MODELA SAMOVENTILIRAJUĆE DISK KOČNICE

Kočnice predstavljaju izvršne elemente, pomoću kojih se realizuju osnovni zahtevi sistema za kočenje kod vozila. Princip rada kočnica zasnovan je na jedinstvenoj fizičkoj pojavi – trenju na kontaktnim površinama pokretnih elemenata, vezanih za točkove i nepokretnih elemenata, vezanih za noseću strukturu vozila. Pod uticajem sile trenja guši se kinetička energija vozila i ono usporava, odnosno koči. Prema tome, kočnice su frikcioni mehanizmi.

U ovom radu je kao primer uzeta kočnica pomoću koje se realizuje pravolinijsko kretanje, odnosno izvođenje zaokreta na oklopnom guseničnom vozilu. Njena specifičnost se ogleda u tome što u ovom slučaju ona ima funkciju u sistemu za upravljanje vozila. U pitanju je izvedba kočnice sa samoventilirajućim diskom. Uključeno stanje kočnice ima za posledicu pravolinijsko kretanje, dok isključena uz dopunske sklopove, realizuje zaokret kod datog guseničnog vozila.

Na slici 1 je prikazan 3D model sklopa predmetne disk kočnice, sa klještima i kućištem, formiran u softveru „*CREO*“.



Slika 1. 3D model kočnice pravolinijskog kretanja

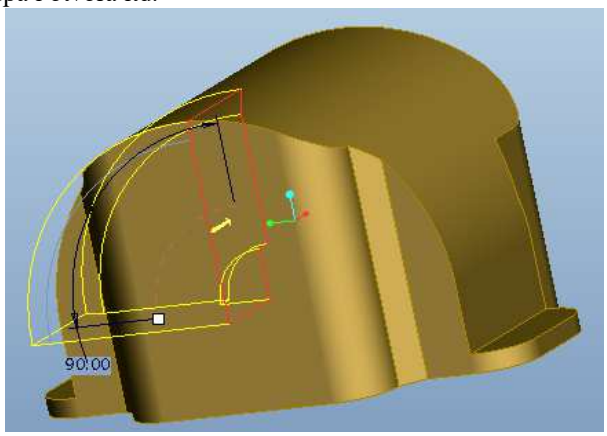
A. *Softver za formiranje geometrijskog modela*

Danas je na raspolaganju velika ponuda softverskih alata za projektovanje i razvoj proizvodnje u virtuelnom, 3D okruženju. To su alati koji kao osnovu imaju baze podataka geometrijskih profila, parametarski definisanih i izrađenih prema važećim standardima.

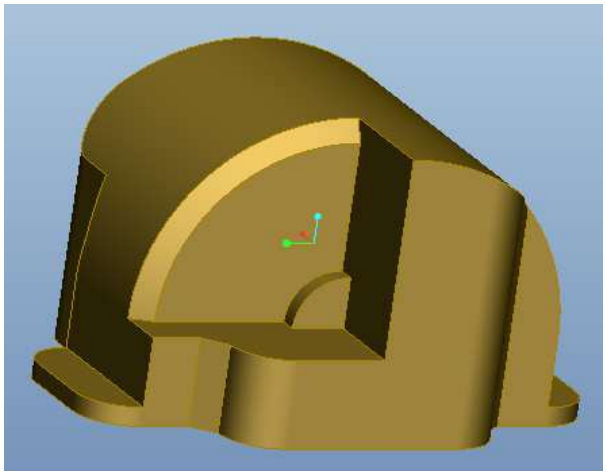
CREO je primer softverskog okruženja za izradu i simulaciju 3D modela, koji postoji na tržištu već dugi niz godina. Pripada familiji Computer-aided design (CAD) softverskih alata, razvijen od strane Parametric Technology Corporation (PTC) i omogućuje da se ubrza inovacija 3D modela, kako bi se došlo do najoptimalnijeg i najsavršenijeg proizvoda.

CREO se koristi kod kompleksnih i odgovornih konstrukcija, u automobilskoj i avionskoj industriji, što potvrđuje njegovu pouzdanost i kvalitet. Omogućava modeliranje, uključujući montažu modela i izradu konstrukcione dokumentacije. Ovaj proces se realizuje na bazi asocijativnosti između delova, sklopova i crteža.

Na slikama 2 i 3 su prikazane neke faze 3D modeliranja u *CREO* okruženju, na primeru kućišta kočnice pravolinijskog kretanja kod guseničnog vozila. Modeliranje je realizovano na osnovu postojeće konstrukcione dokumentacije. Prvo je formiran model sa svojim gabaritnim dimenzijama, a onda se pristupilo izradi detalja, na principu odsecanja viškova, dodavanja materijala, izvlačenja različitih profila, bušenja rupa i otvora itd.



Slika 2. Faza definisanja dela koji se „odseca“



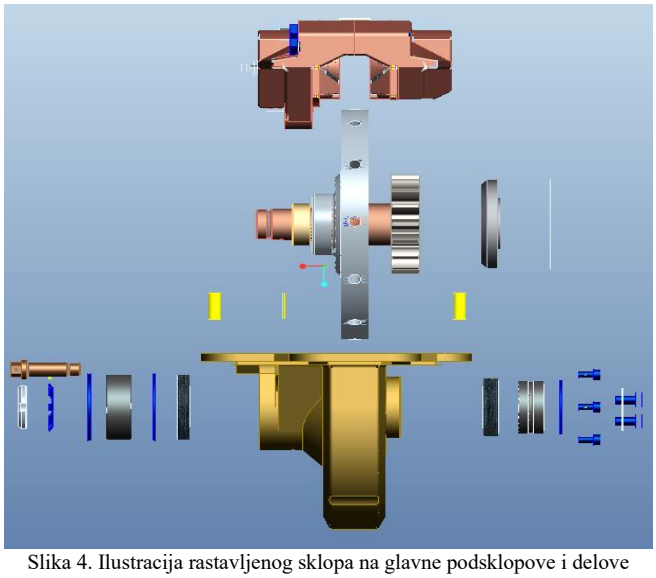
Slika 3. Izgled kućišta nakon „odsečenog“ dela

B. *Formiranje sklopa modela*

Kreiranje sklopa u *CREO* okruženju predstavlja upravljanje sastavnim delovima sklopa po tačno određenom redosledu. Pošto je poznato da geometrijski odnos između bilo koja dva dela ima šest stepeni slobode, da bi se definisao sklop, potrebno je navesti sva ograničenja njegovih delova u odnosu na referentne delove. Najčešće je potrebno upotrebiti više vrsta ograničenja kako bismo potpuno definisali deo u prostoru.

Osim vizuelizacije modela *CREO* omogućava integrisano kreiranje konstrukcione dokumentacije, koja predstavlja osnovu za izradu tehnološke dokumentacije i proces proizvodnje. Konstrukciona dokumentacija izrađena prema zahtevima sistemskog inženjerstva podrazumeva direktnu povezanost sa modelom, uz korišćenje karakteristika asocijativnosti, odnosno mogućnosti automatskog ažuriranja dokumentacije u skladu sa eventualnim izmenama modela. Sve projekcije u okviru jednog tehničkog crteža međusobno su povezane, pa se promena na jednoj od njih odražava na celokupni crtež (dvosmerna asocijativnost). *CREO* takođe podržava uvoz 2D dokumentacije iz drugih CAD programa (npr. preko DXF i IGES formata).

Pored toga što se koristi za prezentaciju geometrije i što ima mogućnost da se unutar modela postave negeometrijski elementi (površinska obrada, dodatne napomene, uputstva), 3D model može da posluži i za simulaciju montaže sklopa, koja je korisna kod održavanja, remonta i u procesu proizvodnje. Montiranje sklopa odgovarajućim redosledom može se predstaviti formiranjem faza montaže delova, koje se povezuju u jedinstven ciklus, a izradom video prikaza (animacije) pojednostavljuje se obuka za radioničko održavanje.



Slika 4. Ilustracija rastavljenog sklopa na glavne podsklopove i delove

III. SIMULACIJA RADA SAMOVENTILIRAJUĆE DISK KOČNICE

Budući da je za simulaciju rada kočnice neophodno raspolagati matematičkim modelima fizičkog procesa trenja

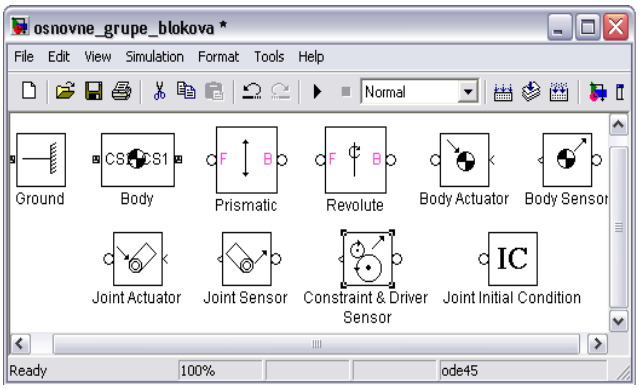
na frikcionim površinama, jasno je da navedeno neće da se realizuje unutar softverskih paketa za 3D modeliranje. Za simulaciju rada razvijeni su posebni softveri, bilo kao moduli vezani za CAD softvere, bilo kao posebni programi. Bez obzira na to o kakvim modelima se radi, neophodno je da se raspolaže osnovnim fizičkim osobinama elemenata sklopa kao što su: masa, težište (centar mase), momenti inercije i sl. Navedeno se sa vrlo velikom preciznošću može dobiti iz računarskog modela, tako da se može konstatovati da je kvalitetna simulacija rada mehaničkih struktura u savremenim okolnostima gotovo nemoguća bez postojanja računarskog modela navedene strukture.

U narednom tekstu će biti opisana simulacija rada kočnice u modulu *Simulink* programskog sistema *MATLAB*, koji ima razvijene modele za simulaciju mehaničkih, hidrauličkih, električnih i drugih sistema, kao i njihovu integraciju, tako da su vrlo zanimljivi za simulaciju rada komponenata motornih vozila.

A. *Osnovni elementi i definisanje procesa simulacije*

Razvoj simulacije rada modela omogućava podmodul *SimMechanics*. Biblioteka u okviru podmodula *SimMechanics* nudi različite tipove blokova za formiranje i precizno definisanje osobina mehaničkih sistema (Slika 5). Dodavanjem adekvatnih blokova i definisanjem međusobnih veza i ograničenja kreira se željeni model.

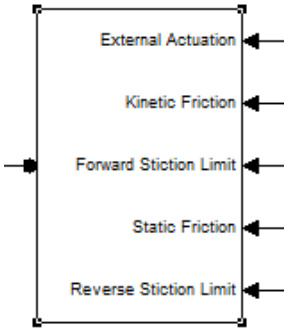
Za predstavljanje modela kočnice od posebne važnosti su *Sensor* i *Actuator* blokovi. *Sensor* blokovi služe za registrovanje željene veličine na ulazu. *Aktuator* blokovi daju fizički smisao ulaznim veličinama, kako bi se došlo do željenog rezultata na izlazu.



Slika 5. Osnovne grupe blokova u SimMechanics biblioteci

Veza između dva rotaciona tela se u podmodulu *SimMechanics* modelira uz pomoć za to posebno razvijenog bloka. Takav blok je prikazan na Slici 6 i nosi naziv *Joint Stiction Actuator*. On služi za definisanje sile trenja kod translatornih sistema, odnosno momenta trenja u kontaktu dva rotaciona tela.

Model trenja ostvaruje vezu rotacionih tela u nekoliko režima: uključenom, prelaznom i isključenom. Promenu režima na modelu realizuje pobuđivač-*Actuator*, na osnovu detekcije parametara na ulazu u sistem.



Slika 6. Blok pobuđivač *Joint Stiction Actuator*

Moment trenja u kontaktu dva rotaciona tela izračunava se u funkciji sledećih veličina:

- spoljašnji moment (M_{ext}) – moment koji deluje na vezu i nezavistan je od momenta trenja,
- trenje kotrljalja/klizanja (M_k) – moment trenja koji deluje na rotaciona tela u prelaznom režimu,
- granica statičkog momenta trenja (M_s) – opseg momenta trenja u kome dolazi do uključivanja veze i taj režim traje dok god moment na vezi dva tela ne izađe iz ovog opsega,
- prag brzine uključivanja (ω_z) – relativna ugaona brzina tela u rotaciji ispod koje prestaje proklizavanje, što dovodi do ostvarivanja čvrste veza tela i njihovog zajedničkog kretanja. [6]

U uključenom stanju, relativna ugaona brzina elemenata u vezi jednaka je nuli ($\omega=0$). Ovakvo stanje traje dok je moment trenja na vezi u granici između donje i gornje granice statičkog trenja $M_{sd} < M_n < M_{sg}$. [7]

U trenutku kada moment trenja na vezi M_n izađe iz granica statičkog momenta trenja, na vezi je ispunjen prvi uslov za isključivanje, simulacija ulazi u prelazno stanje i počinje kretanje u jednom ($\omega > 0$) ili drugom smeru ($\omega < 0$). [7]

U isključenom stanju kretanje tela u vezi se ostvaruje uz proklizavanje pod dejstvom spoljašnjeg momenta (M_{ext}) i kinetičkog momenta trenja (M_k). Prelazak iz isključenog u uključeno stanje na vezi se dešava kada *Simulink* detektuje da je ugaona brzina dostigla vrednost nula. [7]

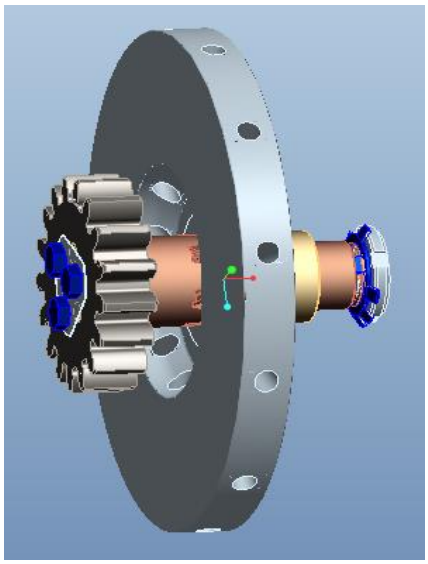
B. *Razvoj simulacionog modela disk kočnice*

Disk kočnice su elementi sistema za kočenje koji imaju karakteristična radna opterećenja u procesu uključivanja, tako da je upravo taj proces najbitniji za projektovanje i za kompjutersku simulaciju rada kočnice. Model je koncipiran tako da predstavlja simulaciju aktiviranja kočnice pri nekom karakterističnom broju obrtaja vratila koje se koči.

Mehanički model samoventilirajuće disk kočnice aproksimira se sklopom rotacionih masa kočnice (Slika 7).

Modeliranjem ovog sklopa u *CREO* softverskom alatu, dolazi se do osnovnih podataka potrebnih za simulaciju rada u *MATLAB*-u, a to su:

- centar gravitacije,
- tenzor inercije,
- masa sklopa koji predstavlja rotacionu masu kočnice.



Slika 7. Mehanički model rotacionih masa disk kočnice

U softveru *CREO* se prvo određuje referentni koordinatni sistem vezan za rotacionu masu, na osnovu čega se dobijaju sledeći podaci:

- Centar gravitacije (pomeren za 45 mm u negativnom smeru Z-ose) i iznosi:

$$Z = -45\ mm,$$

- Tenzor inercije:

$$I = \begin{bmatrix} 21512{,}92 & -0{,}03 & 0{,}94 \\ -0{,}03 & 21512{,}92 & 0{,}37 \\ 0{,}94 & 0{,}37 & 36738{,}84 \end{bmatrix} \cdot 10^6\ kg\ m^2$$

- Masa rotacionog sklopa kočnice kao suma masa svih delova sklopa (koje je softver sam izračunao na osnovu dimenzija delova i definisanih standardnih materijala od kojih su delovi izrađeni) i iznosi:

$$m = 7{,}2\ kg.$$

Na disk deluje normalna sila F (aktivaciona sila). Ova sila deluje na srednjem poluprečniku trenja. Proces uključivanja kočnice opisuje diferencijalna jednačina:

$$M_k = M_u \pm I \cdot \dot{\omega} \qquad (2)$$

Gde su:

M_u - ulazni moment koji se zadaje vratilu,

M_k - moment kočenja,

I - moment inercije rotirajućih delova kočnice i

ω - ugaona brzina okretanja diska.

Moment kočenja je u opštem slučaju funkcija karakteristike trenja, veličine površine trenja, broja frikcionih površina i normalne sile pritiska. [6] Izračunava se prema izrazu:

$$M_k = F \cdot \mu \cdot r_{sr} \cdot z \qquad (3)$$

Gde su:

F - normalna sila pritiska,

μ - koeficijent trenja,

r_{sr} - srednji poluprečnik trenja i

z - broj površina trenja.

Srednji poluprečnik trenja izračunava se prema izrazu:

$$r_{sr} = \frac{2}{3} \cdot \frac{R^3 - r^3}{R^2 - r^2} \qquad (4)$$

Gde su:

R - spoljni poluprečnik diska (rastojanje od ose rotacije do spoljne ivice kočne pločice) i iznosi

$R = 108{,}73\ mm$, a

r - unutrašnji poluprečnik diska (rastojanje od ose rotacije do unutrašnje ivice kočne pločice) i iznosi

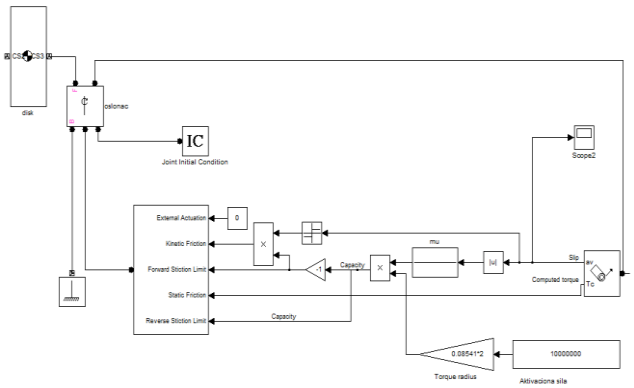
$r = 56{,}6\ mm.$

Na osnovu toga dobija se:

$$r_{sr} = 85{,}41\ mm.$$

Mehanički model kočnice u modulu *SimMechanics* simulira se blokovima i vezama opisanim u odeljku III A. Potrebno je modelirati rotacionu masu, koja prenosi obrtni moment putem sile trenja.

Prvo se vratilu zadaje ugaona brzina, a onda se u željenom trenutku aktivira kočnica zadatom aktivacionom silom na kočne pločice, pri čemu usled trenja između pločica i diska dolazi do smanjenja ugaone brzine diska, a samim tim i vratila i njegovih gonjenih elemenata. Tokom ovog procesa prati se i promena momenta kočenja. Ostali granični uslovi potrebni za ovu simulaciju mogu se definisati parametrima dobijenim eksperimentalnim metodama ispitivanja ili iskustveno. Parametri se na jednostavan način mogu u *MATLAB*-u prevesti u *Simulink* signale i kao takvi mogu predstavljati ulaz u formirani model.



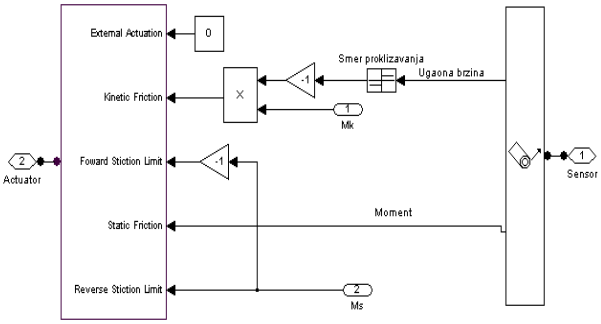
Slika 8. Grafički prikaz simulacionog modela disk kočnice

Grafički prikaz simulacionog modela kočnice razvijen u podmodulu *SimMechanics* prikazan je na Slici 8. Simulacioni model je aproksimiran mehaničkim modelom rotacionih masa disk kočnice. Izabrana je jedna osa rotacije, što odgovara stvarnosti, gde se svi rotirajući elementi kočnice nalaze u jednoj osi. Rotaciona masa je vezana *Revolute* vezom za blok oslonac, čime je definisano njeno oslanjanje.

Prethodno opisani blok *Joint Stiction Actuator* će se iskoristiti za modeliranje procesa uključivanja kočnice. Početna ugaona brzina (broj obrtaja) vratila se unosi preko bloka koji ima oznaku IC (Initial conditions) i omogućava određivanje početne pozicije ili početne brzine elemenata u vezi (Slika 8). Za potrebe ove simulacije bilo je neophodno uneti i početni broj obrtaja rotacionih masa. Blok, kojim se definiše rotaciona masa, kao glavne parametre ima masu (7,2 kg) i moment inercije definisan tenzorom inercije. Za blok su vezani davači, koji mere ugaonu brzinu kočnice. Izmerene veličine se sabiraju u jedan signal i zajedno sa izmerenim momentom kočenja predstavljaju izlaz sistema (blok *Scope 2*).

Za uključivanje kočnice, osim definisane ugaone brzine i rotacionih masa, potrebna je i vrednost normalne-aktivacione sile (*F*). Za definisanje aktivacione sile korišćen je blok *Aktivaciona sila* (Slika 8) iz podmodula *SimMechanics*.

Zajedno sa normalnom silom, kao ulazni parametar zadaje se i koeficijent trenja u podsistem koji se vidi na Slici 8, gde se množenjem sa srednjim poluprečnikom trenja na kome normalna sila deluje i brojem frikcionih površina dobija vrednost momenta kočenja.



Slika 9. Model trenja kočnice

Za rad disk kočnice karakteristični su dinamički i statički koeficijenti trenja (d, s), od kojih je prvi promenljiv u funkciji vremena proklizavanja, a drugi konstantan i deluje na rotaciona tela po prestanku procesa klizanja. Na osnovu toga, rezultat definisanog podsistema su kinetički moment trenja M_k i statički moment trenja M_s. Ove dve veličine predstavljaju ulaz u podsistem *Model trenja* (Slika 9), koji je centralni element razmatranog modela kočnice. Podsistem *Model trenja* je formiran od blok pobuđivača *Joint Stiction Actuator*, a definisani su mu ulazni parametri koji su od važnosti za model kočnice. [6]

C. Rezultati simulacije rada disk kočnice

Za kreiran blok-sistem u podmodulu *SimMechanics*, softverskog alata *MATLAB*, pristupilo se formiranju relevantnih parametara koji karakterišu rotacionu masu:

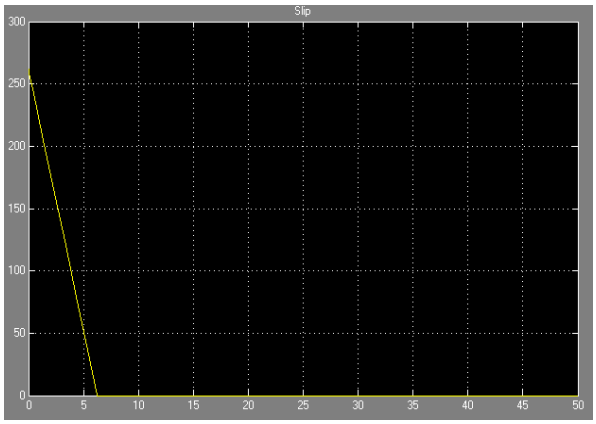
z [-] - broj frikcionih površina kočnice,

R [m]- spoljni poluprečnik diska,

r [m] - unutrašnji poluprečnik diska,

I [kgm^2] - moment inercije rotacione mase.

Prethodno nabrojani parametri su veličine koje ne zavise od vremena aktiviranja uključivanja kočnice i predstavljaju promenljive koje su karakteristika različitih vrsta kočnica za koje se izvodi simulacija. Zadavanjem svih potrebnih parametara i u zavisnosti od aktivacione sile formiran je model na osnovu koga je moguće izvesti veliki broj simulacija, pri kojima se može zaključiti da najviše uticaja na vreme klizanja imaju aktivaciona sila, koeficijent trenja i moment inercije rotacione mase kočnice. Prikazan je jedan od mogućih izgleda krive zavisnosti vremena zaustavljanja od brzine klizanja što predstavlja reprezentativan primer i osnovni cilj izvedenih simulacija (Slika 10).



Slika 10. Dijagram zavisnosti vremena zaustavljanja od brzine klizanja

Rezultati simulacije pokazuju da model trenja koji je formiran u *MATLAB*-u, može da se primeni na konkretan primer pri uključivanju bilo koje disk kočnice. Formirani simulacioni model disk kočnice ima dvojaku upotrebnu vrednost:

- s obzirom da je formiran tako da omogućava jednostavnu izmenu osnovnih parametara kočnice (Slika 8), model se može koristiti za simulaciju rada bilo koje disk kočnice za koju su poznati osnovni navedeni parametri;

- pri simulaciji rada određene kočnice mogu se uzimati u obzir različiti uticajni parametri, kao što su različite promene normalne sile, koeficijenta trenja u toku vremena aktivacije, kao i momenata inercije rotacione mase kočnice.

IV. ZAKLJUČAK

Računarsko modeliranje mašinskih sistema sve više postaje neizostavan deo razvoja. Fleksibilni računarski alati omogućavaju da se relativno jednostavno i brzo razviju potrebni modeli u virtuelnom okruženju, da se izvrši geometrijska prezentacija modela, kao i različite analize i testiranja. 3D model, pored toga što se koristi za prezentaciju geometrije, ima mogućnost definisanja negeometrijskih elemenata (površinska obrada, dodatne napomene, uputstva), kao i simulacije montaže, odnosno demontaže (za potrebe remonta, proizvodnje).

Računarski model samoventilirajuće disk kočnice razvijen u okviru ovog rada, realno predstavlja geometriju sklopa i njihove međusobne veze, pa je s toga primenljiv za analizu geometrijske regularnosti sklopa, simulaciju montaže, kao i simulaciju rada sklopa kočnice. Na osnovu modela moguće je jednostavno i lako formirati i elektronsku konstrukcionu dokumentaciju, koja je integrisana sa modelom, čime se rešava problem ažuriranja dokumentacije.

CREO programski paket na izlazu daje podatke o masama, momentima inercije, centru gravitacije, koji su potrebni kao ulazni podaci za simulacije u drugim softverima. Simulacijom rada disk kočnice dobijaju se parametri potrebni za analizu sklopa, njegovu optimizaciju i modifikaciju u cilju efikasnijeg i ekonomičnijeg razvoja.

Izrada 3D modela softverskim alatima, može izazvati nove ideje i nove načine razmišljanja o dizajniranju delova i celokupnog sklopa, njegovih elemenata i izboru međusobnih veza tih elemenata. Sledeći koraci se mogu ogledati u više pravaca, kao što su analiza inženjeringa metodom konačnih elemenata, modeliranje proizvodnje i analize, dizajn kalupa itd.

LITERATURA

[1] M. Galjak, G. Devedžić, S. Ćuković, "Primena principa integrisanog razvoja proizvoda", academia.edu, str.1-6.

[2] S. Muždeka, M. Pantić, M. Vesić, "Računarski model planetarnog prenosnika tipa RAVIGNEAUX", 1. Naučni skup odbrambene tehnologije u funkciji mira, Beograd, 2005. godina, str. II-14-19.

[3] Č. Duboka, Ž. Arsenić, "Sistemsko inženjerstvo u razvoju, proizvodnji i korišćenju mašinskih sistema", 25. savetovanje proizvodnog mašinstva Jugoslavije, Beograd, 1994. godina.

[4] Č. Duboka, Ž. Arsenić, "Kvalitet koji zadovoljava zahteve korisnika", 21. JUPITER konferencija, Beograd, 1995. godina.

[5] D. Stamenković, M. Milošević, "Projektovanje računarom mašinskih sistema uzimajući u obzir trenje", Niš, 2003. godina.

[6] M. Krsmanović, S. Muždeka, Ž. Arsenić, "Modeliranje procesa uključivanja glavne frikcione spojnice motornog vozila", 1. Naučni skup odbrambene tehnologije u funkciji mira, Beograd, 2005. godina, str. II-37-41.

[7] M. Krsmanović, "Primena programskog paketa Matlab pri simulaciji rada sistema za prenos snage", Vojnotehnički glasnik, Beograd, 2008. godina, str. 50-68.

[8] A. Grkić, Č. Duboka, S. Muždeka, "Simulacioni model višelamelastih frikcionih sklopova", Vojnotehnički glasnik, Beograd, 2009. godina, str. 65-80.

[9] MATLAB Using Simulink and Stateflow™ in Automotive Aplication, 1999.

Simulation of disc brake 3D model operation

Dragan R. Stojadinović

ABSTRACT

Computer support of the design process using appropriate software tools is increasingly present in modern engineering. In addition to design, it includes direct control of machine systems, production lines, as well as logistical problems in material flows and data flows necessary in the production process. Different models in the design phase should be mutually formed and transformed programmatically in the simplest possible way, thereby obtaining far more information and data about the final product than on the basis of the same design documentation. Virtual techniques are introduced into all types and phases of design in the form of modeling, as well as kinematic simulations and structural optimization of models. This paper can be seen how modern software tools help in the development and design of machine systems and assemblies, which in this case are disc brakes.

Keywords – design, model, brakes, simulation, software.

ChatGPT vs. DeepSeek: Prevođenje prirodnog jezika u SQL kod

Ivan Jovanović
Fakultet organizacionih nauka, Univerzitet u Beogradu
Beograd, Republika Srbija
ij20233020@student.fon.bg.ac.rs

Milica Škembarević
Fakultet organizacionih nauka, Univerzitet u Beogradu
Beograd, Republika Srbija
milica.skembarevic@fon.bg.ac.rs
0000-0003-0649-3005

Olga Jejić
Fakultet organizacionih nauka, Univerzitet u Beogradu
Beograd, Republika Srbija
olga.jejic@fon.bg.ac.rs
0000-0002-6594-6388

Marija Đukić
Fakultet organizacionih nauka, Univerzitet u Beogradu
Beograd, Republika Srbija
marija.djukic@fon.bg.ac.rs
0000-0002-1136-4278

***Apstrakt* - Veliki jezički modeli su privukli značajnu pažnju zbog sposobnosti generisanja smislenog i preciznog teksta. Prevođenje prirodnog jezika u SQL kod, kao i primena velikih jezičkih modela u ovoj oblasti, postaje sve popularnija tema među istraživačima. U radu se analiziraju performanse dva velika jezička modela, GPT-4o i DeepSeek R1, u generisanju SQL koda iz prirodnog jezika koristeći BIRD skup podataka. Fokus istraživanja je na proceni tačnosti generisanja SQL koda u zavisnosti od složenosti upita, s posebnim akcentom na složenost šema baza podataka. Rezultati pokazuju umerene performanse oba modela, pri čemu GPT-4o postiže tačnost od 60.17%, a DeepSeek R1 tačnost od 60.37%. Iako su ove vrednosti gotovo identične, detaljna analiza otkriva značajne razlike u performansama, pri čemu DeepSeek R1 pokazuje bolje rezultate u rešavanju izazovnijih upita, dok oba modela ostvaruju sličnu tačnost kod jednostavnijih upita. Istraživanje ukazuje na potrebu za detaljnijim testiranjima modela na upitima različite složenosti, kako bi se stekla preciznija slika o njihovim stvarnim sposobnostima.**

Ključne reči – NL-to-SQL, ChatGPT, DeepSeek, veliki jezički modeli, obrada prirodnog jezika

I. Uvod

Organizacije koriste relacione baze podataka za upravljanje i analizu velikih količina strukturiranih podataka, čineći ih ključnim delom poslovnih sistema [1]. Kako se obim podataka povećava, raste i potreba za efikasnim preuzimanjem i analizom podataka. Međutim, rad sa bazama podataka zahteva poznavanje odgovarajućeg jezika, kao što je struktuirani upitni jezik (SQL). Iako više od polovine profesionalnih programera (51%) koristi SQL jezik, svega četvrtina (23.6%) poseduje formalno obrazovanje u ovoj oblasti [2]. Zbog svoje tehničke prirode, SQL često predstavlja prepreku korisnicima bez odgovarajuće obuke, čime se stvara jaz između potrebe za podacima i znanja neophodnog za njihovo preuzimanje.

Prevođenje prirodnog jezika u SQL (NL-to-SQL) ima potencijal da u velikoj meri promeni pristup podacima omogućavajući korisnicima da pretražuju baze podataka koristeći prirodni jezik, bez potrebe za poznavanjem SQL-a ili šeme baze podataka [1]. Na taj način korisnici koji nisu tehnički obučeni mogu jednostavno pristupati i analizirati podatke, što može biti posebno korisno u oblastima poput poslovne inteligencije, korisničke podrške, ali i naučnog istraživanja.

Kako baze podataka postaju sve složenije, formulisanje odgovarajućih upita postaje sve izazovnije. Nedavno, veliki jezički modeli (LLM) su privukli značajnu pažnju zbog sposobnosti generisanja smislenog i preciznog teksta. Njihova primena može unaprediti obradu prirodnog jezika, jer se mogu koristiti za generisanje SQL koda na osnovu korisničkog pitanja i odgovarajuće šeme baze podataka. Zahvaljujući obuci na velikim skupovima podataka, LLM modeli bolje razumeju složene odnose između prirodnog jezika i šeme baze podataka u poređenju sa ranijim modelima [3], što im omogućava da generišu preciznije i sveobuhvatnije odgovore.

U radu su analizirane performanse velikih jezičkih modela za generisanju SQL koda na osnovu prirodnog jezika, uz analizu njihove sposobnost da odgovore na upite različitih nivoa složenosti. Za potrebe analize korišćen je skup podataka BIRD (BIg Bench for LaRge-scale Database Grounded Text-to-SQL Evaluation), koji omogućava verodostojniju procenu zahvaljujući složenijim šemama baza podataka koje vernije odražavaju kompleksnost stvarnih sistema.

Rad je organizovan na sledeći način: poglavlje 2 predstavlja pregled literature iz oblasti, dok su u poglavlju 3 predstavljeni korišćeni veliki jezički modeli. Metodologija istraživanja i opis testiranja, odnosno studije slučaja, prikazani su u poglavljima 4 i 5 respektivno. Diskusija rezultata, kao i identifikovana ograničenja, opisani su u poglavlju 6, a zaključak je dat u poglavlju 7.

II. Pregled literature

Problem prevođenja prirodnog jezika u SQL upite, kao i sve šira primena velikih jezičkih modela u toj oblasti, postaje predmet sve intenzivnijeg interesovanja istraživačke zajednice. Iako je broj publikacija na ovu temu još uvek relativno ograničen, istraživači ispituju različite pristupe korišćenju LLM-a kako bi unapredili proces prevođenja prirodnog jezika u SQL kod.

Prirodni jezik može biti nejasan jer korisnički upiti nisu dovoljno precizni. [4] su razvili sistem zasnovan na GPT-3 modelu kako bi poboljšali razumevanje upita na prirodnom jeziku i razrešili potencijalne nejasnoće. [5] su primenili

LLM modele kako bi pravilno "poravnali" SQL koda sa šemom baze podataka, a predloženi sistem su testirali na skupovima podataka Spider i BIRD. [6] su istraživali kako prilagođavanje šeme baze podataka može poboljšati performanse NL-to-SQL alata uvodeći opise vrednosti baze podataka koji prilagođeni LLM modelima. [7] navodi da performanse postojećih NL-to-SQL modela opadaju zbog nedostatka domensko-specifičnih podataka za domen za treniranje. Autori predlažu rešenje koje pojednostavljuje kreiranje skupova podataka za treniranje NL-to-SQL modela.

Autori [8] predlažu *Chain-of-Programs* pristup za NL-to-SQL, koji koristi Pandas za dekompoziciju složenih SQL upita na jednostavne operacije, čime omogućava open-source LLM modelima da postignu performanse slične GPT-4, uz niže troškove. U radu [9] su zbog domensko-specifičnih podataka razvijene posebne instrukcije za primenu GPT modela za NL-to-SQL zadatke. DBCopilot [10] je sistem namenjen generisanju SQL koda iz prirodnog jezika u okviru velikih i složenih baza podataka. Korišćenjem jednostavnog modela za upravljanje šemom i LLM modela za generaciju upita, omogućeno je skalabilno i precizno kreiranje SQL izraza u dinamičnim okruženjima.

U svetlu dosadašnjih nalaza, ovaj rad istražuje sposobnost velikih jezičkih modela da razumeju prirodni jezik u kontekstu upita različitih nivoa složenosti, sa ciljem generisanja tačnih SQL izraza za baze podataka sa kompleksnim šemama. Na ovaj način simulira se složenost stvarnih informacionih sistema, čime se obezbeđuje verodostojnija i praktično relevantna evaluacija performansi analiziranih modela.

III. VELIKI JEZIČKI MODELI

Zbog sposobnosti generisanja smislenog i preciznog teksta, veliki jezički modeli izazvali su značajan interes u istraživačkim krugovima i među korisnicima, posebno nakon pojave ChatGPT-a. Jedan od modela analiziranih u ovom radu je GPT-4o, koji je razvijen od strane OpenAI fondacije i postao dostupan javnosti u maju 2024. godine [11]. Funkcionalnosti vezane za napredno generisanje koda uvedene su ranije, sa verzijom GPT-4 u martu 2023. godine [12]. Drugi model koji je obuhvaćen analizom je DeepSeek R1, razvijen od strane kompanije DeepSeekAI, koji je postao dostupan u januaru 2025. godine [13]. Poboljšanja u generisanju složenog koda uvedena su sa verzijom DeepSeekCoder u januaru 2024. godine [14]. GPT-4o je odabran zbog svoje široke primene i velike rasprostranjenosti među korisnicima, dok je DeepSeek privukao pažnju istraživačke zajednice kao novo i perspektivno open-source rešenje.

IV. METODOLOGIJA

Performanse dva navedena modela analizirane su na skupu podataka BIRD (BIg Bench for LaRge-scale Database Grounded Text-to-SQL Evaluation). BIRD [15] predstavlja jedan od često korišćenih skupova podataka u istraživanjima na temu NL-to-SQL, a odlikuje ga prisustvo složenijih šeme baza podataka, čime bolje odražava realne izazove u transformaciji prirodnog jezika u SQL kod. Drugi popularni skupovi, kao što su Spider i WikiSQL, sastoje se uglavnom od jednostavnih baza podataka koje su kreirane namenski za istraživačke svrhe i ne oslikavaju u potpunosti složenost stvarnih baza podataka [16]. Dodatno, WikiSQL skup podataka je orijentisan na upite koji se odnose na pojedinačne tabele, bez mogućnosti testiranja sposobnosti modela u kontekstu relacija unutar složenijih baza podataka.

U okviru istraživanja analizirane su performanse dva velika jezička modela u generisanju SQL koda na osnovu prirodnog jezika. Evaluacija modela sprovedena je korišćenjem unapred pripremljenih pitanja iz BIRD skupa podataka, koji obuhvata upite različitih nivoa složenosti – od jednostavnih do složenih. Svaka instanca u skupu podataka sastoji se od pitanja formulisanog na prirodnom jeziku, konteksta određene baze podataka, kao i odgovarajućeg SQL upita, takozvani „zlatni“ kod. Za ocenu tačnosti generisanog koda korišćena je metrika Exact Match, a po uzoru na rad [17], primenjena je njena varijanta – Exact-Set-Match Accuracy. Ova metrika se uglavnom koristi u zadacima u kojima je očekivani izlaz skup vrednosti, a u radu je primenjena za određivanje stepena podudarnosti između generisanog SQL upita i „zlatnog“ koda. Metrika ima binarnu vrednost: rezultat 1 dodeljuje se kada generisani upit u potpunosti odgovara referentnom, dok rezultat 0 označava svako odstupanje. Zbog mogućnosti da za jedno pitanje postoji više ispravnih SQL upita, ova metrika omogućava precizniju procenu performansi modela.

A. *Istraživačka pitanja*

Cilj ovog rada je analiza performansi velikih jezičkih modela u generisanju SQL koda na osnovu prirodnog jezika (NL-to-SQL). Istraživanje obuhvata analizu dva modela: GPT-4o, koji je razvijen od strane kompanije OpenAI, i DeepSeek R1, koji je proizvod kompanije DeepSeekAI.

Rad daje odgovor na sledeća istraživačka pitanja:

1. Koji od dva analizirana jezička modela pokazuje bolje performanse u generisanju SQL koda na osnovu prirodnog jezika?
2. Kakve su performanse LLM modela za generisanje SQL upita različitih nivoa složenosti?

V. STUDIJA SLUČAJA

Performanse dva navedena modela ispitane su na skupu podataka BIRD. Po uzoru na [16], a radi smanjenja troškova i očuvanja reprezentativnosti, korišćen je uzorak koji obuhvata 10% podataka iz svake baze u razvojnom delu BIRD skupa podataka. Uzorak se sastoji od ukupno 113 upita, od kojih je 68 jednostavnih, 36 upita umerene složenosti i 9 izazovnih upita. Za interakciju sa LLM modelima korišćen je API pristup, pri čemu je Python skripta korišćena za automatizaciju procesa postavljanja upita. Svaki prompt upućen modelima sastojao se od instrukcija za model, korisničkog pitanja formulisanog na prirodnom jeziku i konteksta baze podataka na osnovu kojeg se generiše SQL kod. Kao odgovor, modeli su vraćali generisane SQL upite za svaki postavljeni zahtev.

VI. REZULTATI

Rezultati istraživanja pružaju koristan uvid u sposobnosti velikih jezičkih modela za generisanje SQL koda. Oba modela pokazuju zadovoljavajuće ukupne performanse, sa tačnošću od 60.17% za GPT-4o i 60.37% za DeepSeek R1. Iako su ove vrednosti na prvi pogled gotovo identične, detaljnija analiza pokazuje značajne razlike u načinu na koji svaki model odgovara na upite različitih nivoa složenosti.

Za jednostavne upite, oba modela pokazuju podjednaku uspešnost, uz blagu prednost na strani DeepSeek R1 modela (68.25% u odnosu na 67.64%). Ovo ukazuje na to da oba modela podjednako dobro rešavaju osnovne NL-to-SQL zadataka, poput direktnog izbora kolona i jednostavnih uslova filtriranja. Kod umerenih upita, koji obično uključuju složenije logičke operacije ili višestruke uslove u okviru WHERE klauzule, primetan je pad performansi kod oba modela. GPT-4o beleži nešto bolje rezultate (47.22%) u poređenju sa DeepSeek R1 modelom (44.44%). Pad performansi je donekle očekivan, obzirom da upiti zahtevaju dublje razumevanje semantike i konteksta, što može predstavljati izazov za velike jezičke modele. Kod izazovnih upita, DeepSeek R1 je značajno nadmašio GPT-4o, sa tačnošću od 71.42% naspram 55.55%. Ovakav rezultat može ukazivati da DeepSeek R1 uspešnije prepoznaje šablone i bolje se prilagođava jezičkim obrascima prisutnim u prirodnom jeziku. Nasuprot tome, lošije performanse GPT-4o modela mogu ukazivati na probleme generalizacije sa porastom složenosti zahteva.

Tabela 1. Rezultati testiranja

Složenost upita / LLM	Jednostavan	Umeren	Izazovan	Ukupne performanse
GPT-4o	67.64%	47.22%	55.55%	**60.17%**
DeepSeek R1	68.25%	44.44%	71.42%	**60.37%**

Neke od najčešćih grešaka koje su modeli pravili kod jednostavnih upita uključuju izostavljanje DISTINCT klauzule ili JOIN operacija. Kod umerenih upita, greške su se uglavnom odnosile na netačne uslove u WHERE klauzuli, spajanje pogrešnih tabela ili izbor neodgovarajućih kolona prilikom JOIN operacija. Iako su oba modela pokazala relativno visok nivo efikasnosti u rešavanju NL-to-SQL zadataka, njihova uspešnost varira u zavisnosti od složenosti upita. Samim tim, rezultati ističu značaj testiranja modela na upitima različite složenosti, kako bi se dobila jasniji uvid u stvarne mogućnosti I ograničenja.

A. *Ograničenja*

Važno je naglasiti potencijalna ograničenja u istraživanju koja mogu uticati na tumačenje rezultata. Dizajn prompta može uticati na rezultate i oblikovati odgovore modela. Takođe, obim uzorka koji je korišćen u evaluaciji može umanjiti mogućnost generalizacije zaključaka. Dodatno, sama prirode velikih jezičkih modela koji "rade" na principu „crne kutije“ ograničava mogućnost tumačenja njihovog ponašanja, naročito kod vlasničkih modela poput GPT-4o. Odsustvo uvida u mehanizme rezonovanja otežava identifikaciju razloga za uspeh ili neuspeh kod pojedinačnih upita, čime se onemogućava dublje sagledavanje rada modela.

VII. Zaključak

U ovom istraživanju su upoređene performanse dva velika jezička modela GPT-4o i DeepSeek R1 u kontekstu njihove sposobnosti prevođenja prirodnog jezika u SQL kod. Za testiranje je korišćen BIRD skup podataka zbog složenosti šema baza podataka koje su dostupne u ovom skupu podataka. Iako su oba modela ostvarila slične ukupne rezultate, sa tačnostima od 60.17% za GPT-4o i 60.37% za DeepSeek R1, detaljna analiza je ukazala na značajne razlike u njihovoj sposobnosti da rešavaju upite različite složenosti. DeepSeek R1 je pokazao bolje performanse pri rešavanju izazovnih upita, dok su oba modela bila podjednako uspešna u rešavanju jednostavnijih zadataka. Dobijeni rezultati ukazuju da oba modela imaju zadovoljavajući potencijal za generisanje SQL koda, ali njihova efikasnost varira u zavisnosti od složenosti zadatka.

Iako su oba modela ostvarila zadovoljavajuće rezultate, istraživanje ukazuje na prostor za unapređenje daljim testiranjem na većem skupu podataka. Takođe dublje razumevanje mehanizama na kojima ovi modeli funkcionišu doprinelo bi interpretaciji rezultata, uzimajući u obzir njihovu nedostatak transparentnosti i ograničenu mogućnost direktnog uvida u operativne principe procesiranja. Dodatno, istraživanje generisanja SQL koda na srpskom jeziku predstavlja važnu temu za buduća istraživanja, što bi doprinosilo primeni ovih tehnologija u lokalnom kontekstu. Dugoročno, istraživački napori trebali bi biti usmereni ka unapređenju sposobnosti modela za rešavanje kompleksnih upita i poboljšanju razumevanja jezičkih i semantičkih specifičnosti u različitim domenima primene.

A. *Budući rad*

Buduća istraživanja uključiće evaluaciju performansi velikih jezičkih modela u generisanju SQL koda za baze podataka stvarnih poslovnih sistema, s ciljem procene njihove sposobnosti u razumevanju složenih i realističnih šema baza podataka. Takođe, jedan od mogućih pravaca daljeg istraživanja obuhvata analizu primene LLM modela u kontekstu NoSQL baza podataka, uzimajući u obzir ograničenu dostupnost odgovarajućih skupova podataka u ovoj oblasti. Dodatno, buduća istraživanja mogu uključiti ispitivanje performansi modela u generisanju upita na osnovu prirodnog jezika na srpskom jeziku, čime bi se unapredila lokalizacija i omogućila šira primena ovih tehnologija u domaćem okruženju.

Zahvalnica

Ovo istraživanje je finansirano od strane Fakulteta organizacionih nauka, Univerziteta u Beogradu.

Literatura

[1] A. B. Kanburoğlu and F. B. Tek, “Text-to-SQL: A methodical review of challenges and models,” Turkish J. Electr. Eng. Comput. Sci., vol. 32, no. 3, str. 403–419, 2024.

[2] Stack Overflow, “Stack Overflow Developer Survey 2024,” Stack Exchange Inc., https://survey.stackoverflow.co/2024, 2024.

[3] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, et al., “Recent advances in natural language processing via large pre-trained language models: A survey,” ACM Comput. Surv., vol. 56, no. 2, str. 1–40, 2023.

[4] A. Attawar, S. Vora, P. Narechania, V. Sawant, and H. Vora, “NLSQL: generating and executing SQL queries via natural language using large language models,” in Proc. Int. Conf. Adv. Comput. Technol. Appl. (ICACTA), pp. 1–6, okt. 2023.

[5] Y. Shen, X. Lin, J. Liu, Z. Huang, S. Wang, and Q. Liu, “SA-SQL: A Schema-Aligned Framework for Text-to-SQL through Large Language Models,” in Proc. Int. Conf. Comput. Linguist. Nat. Lang. Process. (CLNLP), pp. 71–77, jul. 2024.

[6] E. R. Nascimento, G. García, Y. T. Izquierdo, L. Feijó, G. M. Coelho, A. R. de Oliveira, et al., “LLM-Based Text-to-SQL for Real-World Databases,” SN Comput. Sci., vol. 6, no. 2, str. 130, 2025.

[7] Y. Tian, D. Lee, F. Wu, T. Mai, K. Qian, S. Sahai, et al., "Text-to-SQL Domain Adaptation via Human-LLM Collaborative Data Annotation," in Proc. 30th Int. Conf. Intell. User Interfaces, pp. 1398–1425, mart 2025.

[8] B. Xu, S. Li, Y. Wu, S. Wei, M. Du, H. Wang, and H. Song, "Chain-of-Program Prompting with Open-Source Large Language Models for Text-to-SQL," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), pp. 1–8, jun. 2024.

[9] G. Sun, R. Shen, L. Jin, Y. Wang, S. Xu, J. Chen, and W. Jiang, "Instruction tuning text-to-SQL with large language models in the power grid domain," in Proc. 4th Int. Conf. Control, Robot. Intell. Syst., pp. 59–63, avg. 2023.

[10] W. Zhang, Y. Shen, W. Lu, and Y. Zhuang, "Data-Copilot: Bridging Billions of Data and Humans with Autonomous Workflow," in ICLR Workshop on Large Language Model (LLM) Agents, 2024.

[11] OpenAI, "GPT-4o System Card," OpenAI, https://openai.com/index/gpt-4o-system-card/, 8. avg. 2024.

[12] OpenAI, "Chat Completions API," OpenAI, https://platform.openai.com/docs/guides/gpt/chat-completions-api, pristupljeno 11. apr. 2025.

[13] DeepSeek, "DeepSeek R1: Reasoning-focused language model," DeepSeek, https://github.com/deepseek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf, pristupljeno 11. apr. 2025.

[14] D. Guo, Q. Zhu, D. Yang, Z. Xie, K. Dong, W. Zhang, et al., "DeepSeek-Coder: When the Large Language Model Meets Programming--The Rise of Code Intelligence," arXiv e-prints, arXiv:2401, 2024.

[15] J. Li, B. Hui, G. Qu, J. Yang, B. Li, B. Li, et al., "Can LLM already serve as a database interface? A big bench for large-scale database grounded text-to-SQLs," in Proc. 37th Int. Conf. Neural Inf. Process. Syst., pp. 42330–42357, dec. 2023.

[16] S. Talaei, M. Pourreza, Y. C. Chang, A. Mirhoseini, and A. Saberi, "CHESS: Contextual Harnessing for Efficient SQL Synthesis," arXiv e-prints, arXiv:2405, 2024.

[17] R. Zhong, T. Yu, and D. Klein, "Semantic evaluation for text-to-SQL with distilled test suites," in Proc. 2020 Conf. Empir. Methods Nat. Lang. Process. (EMNLP), pp. 396–411, nov. 2020.

ChatGPT vs. DeepSeek: Translating natural language into SQL code

Ivan Jovanović, Milica Škembarević, Olga Jejić, Marija Đukić

ABSTRACT

Large language models have attracted considerable attention for their ability to generate meaningful and precise text. The translation of natural language into SQL code, as well as the application of large language models in this area, is becoming an increasingly popular topic among researchers. This paper analyzes the performance of two large language models, GPT-4o and DeepSeek R1, in generating SQL code from natural language using the BIRD dataset. The focus of the research is on evaluating the accuracy of SQL query predictions based on query complexity, with a particular emphasis on the complexity of database schemas. The results show moderate performance for both models, with GPT-4o achieving an accuracy of 60.17% and DeepSeek R1 achieving an accuracy of 60.37%. Although these values are nearly identical, a detailed analysis reveals significant performance differences, with DeepSeek R1 demonstrating better results on more challenging queries, while both models perform similarly on simpler queries. The research highlights the need for further testing of the models on queries of varying complexity to obtain a more accurate picture of their true capabilities.

YU #6: Sesija 6

Vojne simulacije i primene

Realizacija autonomnog kretanja na besposadnoj platformi

Rade Pavlović
Vojnotehnički institut
Beograd, Republika Srbija
rade.pavlovic@mod.gov.rs
0000-0002-4921-3950

Nina Mitričević
Vojnotehnički institut
Beograd, Republika Srbija
nmitricevic@gmail.com

***Apstrakt* – Autonomna vozila i robotski sistemi sve su više zastupljeni u mnogim istraživanjima. Najčešće se vrši prevođenje postojećih besposadnih ili vozila sa posadom u autonomna. U ovom radu izvršena je implementacija postojećih algoritama na besposadnoj zemaljskoj platformi, kao i validacija dobijenih rezultata u laboratorijskim uslovima. Robotska platforma koja je korišćenja u ispitivanju je Rosbot 2 Pro, koja u sebi ima implementiran ROS 2 operativni sistem, kao i podršku u simulacionim okruženjima. Korišćen je najpre simulacioni model kako bi se pokazala opravdanost algoritama, a zatim implementacija na besposadnu platformu. Rezultati koji su dobijeni validacijom pokazali su da robotska platforma može uspešno da izvrši zadatke kao što je odlazak do željene lokacije, povratak na početnu poziciju, kao i praćenje operatera. Pri tome korišćene su i dinamičke prepreke koje je robotska platforma uspešno uspela da savlada i dođe do željenog cilja.**

Ključne reči autonomno vozilo, besposadna platforma, robotski sistem.

I. UVOD

Robotski sistemi i vojni roboti imaju sposobnost da izvršavaju zadatke u autonomnom ili bilo kom drugom režimu prema potrebi. Ovakvi sistemi imaju mogućnost da izvršavaju operacije za čoveka. Naoružanje i sistemi naoružanja u kojima se nalaze robotski sistemi nazivaju se hibridnim čovek-mašina sistemi [1].

U poslednjim godinama, može se videti ogroman razvoj u oblasti robotike u vojnoj i civilnoj sferi. U vojnoj sferi to uglavnom uključuje njegovo angažovanje u vojnim operacijama, gde se izvršava raznovrstan spektar zadataka. Ovi sistemi i njihovi podsistemi razvijaju se intenzivno i testiraju u vojnim primenama. Napor lidera u oblasti vojne robotike i autonomnih sistema ima za cilj da se dobije prednost na bojnom polju. Tehnologije koje se primenjuju u robotskim sistemima u stanju su da značajno promene odnos sila koristeći karakteristike kao što su kao velika brzina obrade informacija i donošenja odluka, preciznost, kao i niska cena u poređenju sa upotrebom tradicionalnih sistema i ljudskih resursa. Na kraju krajeva, ovakvi sistemi predstavljaju opciju koja omogućuje da vojnici imaju veće šanse da prežive na bojnom polju.

Robotski sistemi poboljšavaju C4I sposobnosti, održivost i mobilnost u vojnim operacijama [2]. Dakle, oni favorizuju stranu sukoba koja poseduje ovakve sisteme. S druge strane, potrebno je obezbediti otpornost celog sistema u kome robotski sistem radi, jer njegove greške mogu prouzrokovati destrukciju čitavog sistema. Obično bi to uključivalo sajber napade ili rad u slučaju elektromagnetnog ometanja. Ovo pitanje je veoma dinamično i svake godine možemo da posmatramo značajne tehničke promene, inovacije kao i rezultirajuće promene u mogućnostima primene.

II. KATEGORIZACIJA ROBOTSKIH SISTEMA

Robotski sistemi se mogu podeliti (kategorizovati) prema većem broju različitih aspekata [3]. U zavisnosti od sredine u kojoj se nalazi robot (daljinski upravljani/bespilotni) sistemi se koriste, možemo ih podeliti na sledeći način:

- Vazdušni domen – Bespilotni vazduhoplovni sistemi, UAS (engl. unmanned aerial systems);
- Zemljište – Bespilotni zemaljski sistemi, UGS (engl. unmanned ground systems);
- Morsko područje – Bespilotni pomorski sistemi;
- Svemirski domen – Bespilotni svemirski sistemi;
- Domen sajberprostora – samoaktivirajući softveri.

Prema načinu upravljanja sistemi se mogu podeliti na sledeći način:

- Direktno upravljani sistemi – upravljanje se vrši na osnovu direktne kontrolu od strane operatera, bez sopstvene sposobnosti donošenja odluka osim ljudskog uticaja. Nije potrebna nikakva interakcija sa okolinom jer je upravljanje u potpunosti osigurano od strane ljudskog operatera;
- Kontrolisani sistemi – upravljaju se na osnovu trenutnog uputstva koja daje operater, imaju jednostavnu logičku sposobnost donošenja odluka poput mašine konačnog stanja;
- Regulisani sistemi – u osnovi, to su kontrolisani sistemi koji ostvaruju ciljeve na unapred određen način, tj. postiže se cilj pod različitim uslovima (vožnja automobila koji je opremljen automatskim menjačem, ABS, ASR, ESP funkcijama itd.);
- Daljinski upravljani sistemi – rade na osnovu instrukcija koje daje operater koji se nalazi odvojeno od sistema. Operacije koje izvršava takav sistem zavise od prenosa operacija operater-mašina. U svim gore navedenim slučajevima ljudski faktor je deo kontrolne petlje sistema;
- Poluautonomni sistemi - koja se takođe nazivaju i sistemi sa nekim autonomnim funkcijama, postižu cilj na način koji sam sistem izabere. Međutim, sistem ne može predvideti iznenadne prepreke. U slučaju iznenadnih prepreka sistem zahteva angažovanje

ljudskog faktora za dalje odlučivanje. Ljudski faktor proverava postizanje ciljeva ili ispravlja svoje ciljeve;

- Potpuno autonomni sistemi biraju sopstvene ciljeve i načine za njihovo postizanje (na osnovu algoritama). U toku izvođenja operacija ovi sistemi ne zahtevaju ljudske resurse. Takvi sistemi su u stvari robotski sistemi sa veštačkom inteligencijom.

Robotski autonomni sistem, u varijanti autonomnog vozila potrebno je da imaju sledeće mogućnosti i karakteristike [4]:

- prikupljanje informacija o neposrednom okruženju;
- detekcija objekata od interesa kao što su ljudi i vozila;
- kretanje između tačaka puta bez pomoći čoveka - navigacija;
- rad bez ljudske intervencije duži vremenski period;
- izbegavanje situacije koje su štetne po ljude, imovinu ili sebe;
- traženje ili uklanjanje eksploziva;
- popravljanje bez spoljne pomoći;
- robot može i samostalno da uči. Autonomno učenje podrazumeva sposobnost učenja ili sticanja novih funkcionalnosti bez spoljne pomoći;
- obavljanjanje zadatka u zavisnosti od okruženja sredine, prilagođavati se okruženju bez eksterne pomoći;
- razvijanje etičkog osećaja za postizanje misije.

III. REALIZACIJA AUTONOMNOG KRETANJA NA LABORATORIJSKOJ ROBOTSKOJ PLATFORMI

Celokupan sistem je implementiran na robotskoj platformi Rosbot 2 Pro (Slika 1), proizvođača Husarion. Ovo je mobilna robotska platforma sa četiri točka i dolazi opremljena sa više senzora od kojih su nama najbitniji enkoderi na motorima, pomoću kojih se vrši odometrijsko merenje položaja i LIDAR, tj. laserski senzor koji skenira prostor u jednoj ravni oko robota. Platforma poseduje set običnih točkova i set omnidirekcionih „mecanum“ točkova, što omogućava implementaciju dva različita tipa pogona. Prvi je tzv. „skid steer“ pogon karakterističan za vozila sa gusenicama, dok je drugi pogon sa omnidirekcionim kretanjem koji omogućava translaciju u svim pravcima i pretvara platformu u ''holonomni'' mehanički sistem. Pogon „skid steer“ se suštinski isto modeluje kao i diferencijalni pogon ali su performanse odometrije koje se dobijaju lošije zbog proklizavanja točkova.

Slika 1. Мобилна роботска платформа ROSbot 2 Pro.

Mobilna platforma sadrži integrisani računar SBC UP board sa sledećim performansama: procesor Intel Atom x5-Z8350 1.44/1.92 GHz, 4 GB RAM, Intel® HD 400 GPU, 32GB eMMC. Ova razvojna ploča podržava veliki broj operativnih sistema, uključujući Ubuntu koji je potreban za instalaciju ROS paketa biblioteka.

Za razvoj matematičkih modela i simulaciju rada senzora i različitih algoritama za lokalizaciju, mapiranje terena, lokalno i globalno planiranje trajektorije na raspolaganju je veliki broj različitih softverskih paketa i okruženja koji podržavaju fizičku simulaciju mobilnih robotskih sistema [6]. Neki od dostupnih simulatora fizike su Gazebo, Webots, PyBullet, Mujoco, itd. U krajnjoj realizaciji koristili smo Gazebo simulator sa sa integrisanim modulom za simulaciju fizike (engl. physics engine). Iako većina ovih simulatora može da radi samostalno, koristili smo izabrani simulator u kombinaciji sa robotskim operativnim sistemom (ROS), koji olakšava kasniju implementaciju razvijenih algoritama na stvarnoj hardverskoj platformi i integraciju samog robota sa različitim senzorima koji predstavljaju deo sistema. U kasnijoj fazi rada, nakon implementacije na stvarnom hardveru, koristili smo biblioteku RviZ za vizualizaciju kretanja robota na mapi.

Razmatran je veći broj biblioteka za mapiranje, lokalizaciju i planiranje kretanja mobilnih robota u nestrukturiranom dinamičkom okruženju. Planer treba da se sastoji od globalnog i lokalnog planera i da ima integrisan algoritam za izbegavanje dinamičkih prepreka. Za te potrebe izabran je programski paket Navigation 2 u ROS 2 okruženju, koji podržava sve potrebne funkcionalnosti [7,8]. Za mapiranje terena i lokalizaciju robota dostupne su biblioteke poput RTAB Map i SLAM Toolbox. Za implementaciju je izabrana biblioteka SLAM Toolbox. Za softversku bazu izabran je robotski operativni sistem – ROS. Ovo nije pravi operativni sistem u tom smislu, već se radi o kolekciji softverskih paketa za olakšanu implementaciju robotskih sistema i njihovu integraciju sa drugim hardverom.

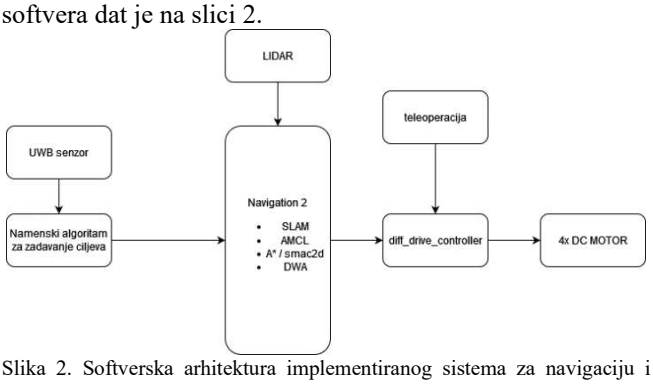
Prvo testiranje platforme zahteva pokretanje robota u manulenom režimu, primenom teleoperacije. Na robotu je potrebno pokrenuti *launch* skriptu proizvođača koja uspostavlja komunikaciju sa motor drajverima, učitava fizičke parametre robota i pokreće *diff_drive_controller* paket koji na osnovu enkodera računa poziciju robota, a na osnovu komandovane brzine platforme preračunava brzine točkova i šalje tu informaciju drajverima motora. Nakon pokretanja pojavljuje se ROS tema */cmd_vel* na koju šaljemo refrentnu željenu brzinu. Na ličnom računaru pokreće se paket *teleop_twist_keyboard* koji na osnovu unosa sa tastature omogućava intuitivno upravljanje robotom. On objavljuje zadatu referentnu brzinu na temu */cmd_vel*, što zbog distribuiranog pristupa preko bežične mreže i rutera stiže robotu i on se pokreće.

Usled odometrijskih grešaka, pogotovo zbog proklizavanja točkova tokom skretanja, u rad smo uključili i lidar senzor pomoću paketa *rplidar*. On omogućava skeniranje prostora u jednoj ravni i to 360 stepeni oko robota. Integracija celog sistema je olakšana popularnim i robusnim paketom *navigation2*. Unutar njega pomoću paketa *slam_toolbox* pokrenuli smo algoritam za mapiranje prostora. Ovaj paket

pomoću odometrijskih podataka sa enkodera i skeniranja lidarom postepeno gradi mapu prostora. Robota je potrebno ručno pokretati teleoperacijom kako bi istražio prostor i izgradio mapu. Celokupan proces se prati na ličnom računaru alatom za vizuelizaciju *RViz*. U okviru njega vidimo položaj robota, mapu, koordinatne sistema, kao i očitavanja senzora od interesa. Nakon što smo istražili prostor robotom, sačuvali smo generisanu mapu na disk.

Moguće je stalno istraživati prostor i koristiti algoritam za generisanje mape, ali ovo opterećuje procesor. S obzirom da smo dobili statičku mapu prostora, dovoljno je samo lokalizovati se na njoj (dinamičke prepreke će i dalje biti detektovane). Lokalizacija je izvršena paketom *nav2_amcl* koji implementira adaptivnu Monte-Karlo lokalizaciju. Ovaj paket učitava statičku mapu sa diska i pomoću odometrije i lidara lokalizuje robota na mapi.

Robot je lokalizovan, ali za autonomnu vožnju potrebno je učitati planere. Globalni planer nalazi putanju na mapi do ciljne tačke i isprobali smo *A** i *smac2d* algoritme. Lokalni planer koristi *dynamic window approach* algoritam i omogućava dinamičko izbegavanje prepreka. Blok dijagram softvera dat je na slici 2.



Slika 2. Softverska arhitektura implementiranog sistema za navigaciju i praćenje pokretne mete.

Pomoću softvera *Rviz* ili *nav2_simple_commander* paketa moguće je zadati cilj robotu. On se do cilja kreće autonomno i izbegava prepreke. Na blok dijagramu se vidi i UWB senzor, odabran je NooploopLinktrack AOA koji pruža dobar odnos cene i kvaliteta i poseduje zadovoljavajuće performanse. Napisan je ROS čvor koji komunicira sa senzorom preko USB-a i poziciju taga (objekta praćenja), koji operater na terenu nosi sa sobom, objavljuje kao koordinatni sistem pomoću paketa *tf* za sisteme koordinatnih transformacija. Pored ovog koordinatnog sistema, objavljuje se još jedan koji ima istu orijentaciju ali je bliži robotu za predefinisanu udaljenost. Ova udaljenost predstavlja željenu distancu (engl. offset) od mete i u pitanju je kontrolni parametar koji se može podešavati.

Algoritam za praćenje čoveka poseduje dva osnovna režima [9]. Prvi zadaje cilj robotu koji ga vodi direktno ka tagu, najkraćim putem, uz periodično ažuriranje pozicije cilja. Tačnije, cilj nije sam tag, već ofsetovani koordinatni sistem kako robot ne bi previše prišao osobi koja nosi tag. Drugi režim prati putanju (trag) po kojoj se kretao operater sa tagom. Ovo se postiže periodičnim memorijskim skladištenjem očitanih položaja taga u listu, nakon čega se memorisane pozicije sekvencijalno prosleđuju robotu kao ciljevi.

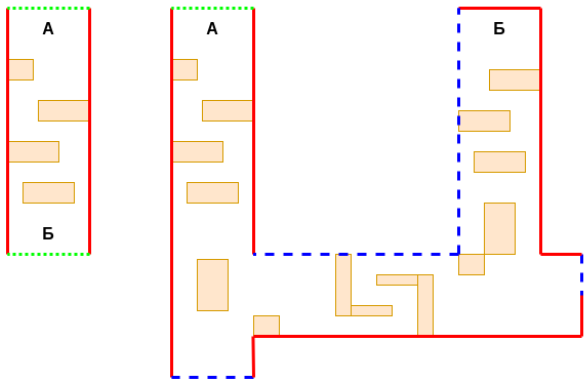
IV. АНАЛИЗА РЕЗУЛТАТА

Najpre je bilo potrebno definisati i modelovati adekvatan poligon sa *''dummy''* preprekama u virtuelnom simulacionom okruženju i niz scenarija simulacije u kojima će se generisati sve željene situacije potrebne za adekvatno testiranje algoritama. Poligon treba da bude odgovarajuće kompleksnosti i sa dovoljnim brojem objekata, čiji položaj, oblik i dimenzije mogu varirati, kako bi se testirao *SLAM* algoritam za mapiranje prostora i lokalizaciju robota. Potrebno je obezbediti dovoljan broj situacija u kojima robot nailazi na statičke ili dinamičke prepreke, kako bi se testirao lokalni planer i sistem za izbegavanje prepreka. Definisan je niz metrika za evaluaciju performansi razvijenih algoritama. Izvršeno je poredjenje iterativno generisanu mapu iz *SLAM*-a sa unapred poznatim *CAD* modelom mape iz simulatora, zatim meriti akumuliranu grešku pozicije i orijentacije robota, nakon pređenih N krugova po nekoj zatvorenoj putanji na mapi, meriti tačnost i ponovljivost nakon višestrukog slanja robota iz jedne fiksirane početne pozicije u jednu te istu ciljnu tačku, itd.

Validacija sistema putem simulacije – na osnovu definisane metodologije validacije i postavke eksperimenta, simulirane su sve predviđene scenarije i logovani potrebni podaci za dalju analizu. Time se proverava i integracija izabranih senzorskih sistema na postojeću mobilnu robotsku platformu sa omnidirekcionim pogonom i upravljanje kretanjem robotske platforme i prenošenje svih funkcionalnosti iz simulacionog okruženja na stvarni sistem. Ovaj korak je znatno olakašan zbog primene ROS-a.

Validacija sistema u laboratorijskim uslovima – Na osnovu definisane metodologije validacije i postavke eksperimenta, realizovani su predviđeni scenariji na laboratorijskom poligonu sa stvarnom hardverskom platformom i logovani potrebni podaci za dalju analizu.

Za potrebe validacije tri različite funkcionalnosti robota osmišljena su tri odgovarajuća testa, tj. eksperimentalna protokola. Prvi test je predstavljao izbegavanje prepreka i navigacija. Ovim testom se radi validacija funkcionalnosti senzorskog sistema robota, implementiranog *SLAM* algoritam i algoritma za izbegavanje prepreka. Testira se mogućnost kretanja robota u nesturkturiranoj okolini, čija mapa nije u potpunosti poznata (sadrži nepoznate statičke prepreke) i koja može da sadrži dinamičke prepreke (Slika 3).



Slika 3. Ilustracija poligona koji će se koristiti za validaciju sistema za izbegavanje prepreka.

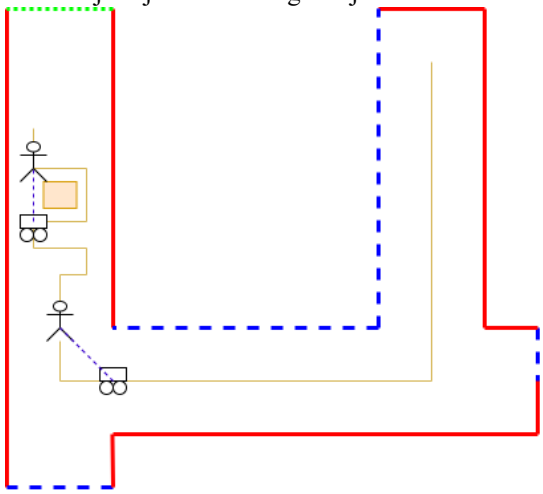
U okviru hodnika na koji je u obliku ćiliričnog slova P, robot se kretao od tačke A do tačke B i pritom da izbegavao

prepreke sastavljene od kartonskih kutija, nameštaja, zidova i ljudi. Za potrebe testa, tačke A i B su zadavane komandama iz ROS okruženja.

Prvi deo testa predstavlja kretanje gde se tačke A i B vide u okviru vazdušne linije, pri čemu je između njih postavljen set prepreka. U drugom delu se ispitivalo kretanje robota između tačaka A i B koje se ne vide vazdušnom linijom, definisanjem početne pozicije i cilja na krajevima hodnika u obliku ćiliričnog slova P. Na ovaj način sami zidovi predstavljaju prepreku algoritmu za planiranje kretanja robota, što je realističan scenario u stvarnoj upotrebi. Prikaz poligona je dat na je na slici 3.

Funkcionalnosti *SLAM* algoritma koji vrši simulatanu lokalizaciju i mapiranje terena, kao i algoritma za globalno i lokalno planiranje putanje testirane su na istom poliogonu ali uz dodatno uvođenje dinamičkih prepreka, koje će se simulirati kretanjem ljudi u sceni ili pomeranjem kutija. Na ovaj način su validirane su funkcionalnosti robota za odlaženje u predefinisanu ciljnu tačku (engl. go to target) i povratak u početnu poziciju (engl. return home).

Drugi test je predstavljao praćenje mete. Ovim testom se validira funkcionalnost praćenja UWB tag-a koji operater na terenu nosi sa sobom. Na podu je izlepljena traka kao predefinisana putanja kojom će se kretati čovek sa UWB tagom. Zadatak robota jeste da prati putanju po kojoj je prošao čovek, tj. operater na terenu. Pri tome se posmatrano je da li će robot ispratiti svaku tačku kroz koju je čovek prošao i time i on pratiti izlepljenu traku ili će praviti prečice po vazdušnoj liniji. Prikaz ovog test je ilustrovan na slici 4.

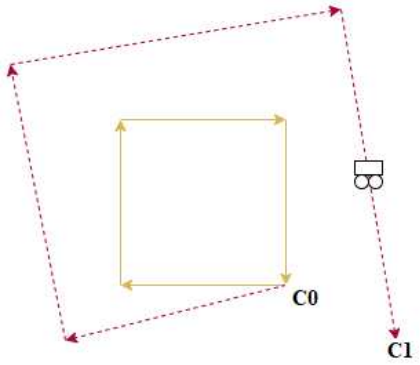


Slika 4. Ilustracija poligona koji će se koristiti za validaciju sistema za praćenje pokretne mete na osnovu UWB senzorske tehnologije.

U drugoj fazi ovog testiranja analizirana je sposobnost robota da se vrati u početni položaj nakon završene misije (engl. Return to Home), pri čemu se u ovom slučaju invertuje cilj, pa trenutna pozicija taga predstavlja početnu tačku A, a kuća će biti tačka B. Ovim testom se validira funkcionalnost *SLAM* algoritma zasnovanog na senszorskom sistemu koji čine lidar i enkoderska odometrija.

Robot se kreće po isprogramiranoj zatvorenoj putanji, gde se početna i krajnja tačka C poklapaju. U testiranjima ovog tipa kao standardna putanja koriste se kvadrat ili putanja u obliku broja osam. Početna tačka C predstavlja inicijalnu tačku C0. Nakon jednog pređenog ciklusa po putanji robot će se nalaziti na tački C1, pa posle još jednog na tački C2, itd. Nakon predefinisanog broja ciklusa, u našem slučaju 10, akumulirana greška je ogledana u tome da se tačka C10 neće poklapati sa početnom tačkom C0, kao što je ilustrovano na slici 5. Takođe, osim odstupanja u poziciji tačke C posmatrano je i odstupanje u krajnjoj orijentaciji robota u odnosu na početnu.

Kao referentni sistem merenja korišćen je sistem za snimanje kretanja VICON koji se sastoji iz skupa infracrvenih kamera sa visokom učestalošću osvežavanja slike (f>600Hz) koji omogućava tačnost merenja pozicije ispod 1mm. Ovaj sistem je zasnovan na primeni pasivnih refleksivnih markera koji se postavljaju na objekat čije se kretanje snima. Postavljanjem najmanje tri markera koji formiraju koordinatni sistem moguće je odrediti i poziciju i orijentaciju objekta u prostoru.



Slika 5. Ilustracija poligona koji će se koristiti za validaciju tačnosti i ponovljivosti sistema za navigaciju5.

V. ZAKLJUČAK

Autonomna besposadna vozila imaju sve veću primenu kako u civilnoj tako i u vojnoj industriji. U civilnoj industriji teži se uvođenju totalne autonomije u vozilima u cilju smanjenja ljudske greške, koja je direktan uzrok saobraćajnih nezgoda. Takođe, u industriji se koristi veliki broj autonomnih vozila koja obavljaju uspešno zadatke.

U vojnoj industriji autonomna besposadna vozila imaju kao osnovni zadatak da smanje ljudske gubitke na bojištima. Pored toga autonomna vozila mogu obavljati zadatke koje nije moguće izvršiti postojećim sistemima (odlazak na zadate lokacije prilikom gubitka komunikacije sa vozilom). U okviru istraživanja u ovom radu dati su rezultati dobijeni korišćenjem standardnih algoritama za autonomno kretanje implementiranih na realnoj robotskoj platformi. Algoritmi su se pokazali kao pouzdani za korišćenje u laboratorijskim uslovima sa statičkim i dinamičkim preprekama. Autonomno vozilo je uspešno uspelo da prepozna i savlada sve prepreke koje je imalo na putu do zadate lokacije. Pored toga uspešno je izvršeno i praćenje operatera po putanji kojom se kretao.

Svi navedeni rezultati otvaraju mogućnost da se isti algoritmi implementiraju i na robotskim platformama koje su namenjene za rad u spoljnim uslovima i izvrši njihovo testiranje.

ZAHVALNICA

Ovaj rad je podržan od strane Ministarstva nauke, tehnološkog razvoja i inovacija Republike Srbije, Ugovor broj 451-03-47/2023-01/202325.

LITERATURA

[1] Corke, P. (2017). Navigation. In: Robotics, Vision and Control. Springer Tracts in Advanced Robotics, vol 118. Springer, Cham. https://doi.org/10.1007/978-3-319-54413-7_5

[2] Li, Z.; Gong, J.; Lu, C.; Xi, J. Importance Weighted Gaussian Process Regression for Transferable Driver Behaviour Learning in the Lane Change Scenario. IEEE Trans. Veh. Technol. 2020, 69, 12497–12509.

[3] Rosique, F.; Lorente, P.N.; Fernandez, C.; Padilla, A. A Systematic Review of Perception System and Simulators for Autonomous Vehicles Research. Sensors 2019, 19, 648.

[4] Chen, T.; Wang, R.; Dai, B.; Liu, D.; Song, J. Likelihood-Field-Model-Based Dynamic Vehicle Detection and Tracking for Self-Driving. IEEE Trans. Intell. Transp. Syst. 2016, 11, 3142–3158.

[5] Patole, S.M.; Torlak, M.; Wang, D.; Ali, M. Automotive radars: A review of signal processing techniques. IEEE Signal Process. Mag. 2017, 34, 22–35.

[6] Pech, H.; Nauth, P.M.; Michalik, R. A new Approach for Pedestrian Detection in Vehicles by Ultrasonic Signal Analysis. In Proceedings of the IEEE EUROCON 2019-18th International Conference on Smart Technologies, Novi Sad, Serbia, 1–4 July 2019; pp. 1–5.

[7] Koenig, N.; Howard, A. Design and use paradigms for gazebo, an open-source multi-robot simulator. In Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No. 04CH37566), Sendai, Japan, 8 September–2 October 2004; IEEE: Piscataway, NJ, USA, 2004; Volume 3, pp. 2149–2154.

[8] Kato, S.; Takeuchi, E.; Ishiguro, Y.; Ninomiya, Y.; Takeda, K.; Hamada, T. An open approach to autonomous vehicles. IEEE Micro 2015, 35, 60–68.

[9] Oh Seong Park, Jae Hoon Lee and Shingo Okamoto, "Position Estimation Method Using Multiple UWB Radio Communication Modules and Its Application to Mobile Robot"

Realization of autonomous movement on an unmanned platform

Rade Pavlović, Nina Mitričević

ABSTRACT

Autonomous vehicles and robotic systems are increasingly represented in many researches. The conversion of existing unmanned or manned vehicles into autonomous ones is most often carried out. In this paper, the implementation of the existing algorithms on an unmanned ground platform, as well as the validation of the obtained results in laboratory conditions, was presented. The robotic platform used in the research is Rosbot 2 Pro, which has the ROS 2 operating system implemented, as well as support in simulation environments.. First, a simulation model was used to demonstrate the justification of the algorithms, and then the implementation on an unmanned platform. The results obtained by validation showed that the robotic platform can successfully perform tasks such as going to the desired location, returning to the initial position, as well as following the operator. In doing so, dynamic obstacles were used that the robotic platform successfully managed to overcome and reach the desired goal.

Merenje intenziteta svetlosti, dometa svetlosnog snopa i provera zaptivenosti ručne taktičke lampe

Milena Jovanović
Tehnički opitni centar GŠ VS
Beograd Srbija
milenakv1996@gmail.com

Marina Tripkovic
Tehnički opitni centar GŠ VS
Beograd Srbija
marinamakitripkovic9@gmail.com

***Apstrakt* - Predmet ovog ispitivanja je lampa taktička ručna namenjena za osvetljavanje prostora prilikom pregleda motornih vozila, prostorija i objekata, pretresa terena u noćnim uslovima i uslovima smanjene vidljivosti (magla, sneg, kiša itd.). Namenjena je kao svetlosna signalizacija i upozoravanje u slučaju uočavanja potencijalno opasnih predmeta. Lampa ima mogućnost podešavanja jačine svetlosnog fluksa u četiri režima rada, za različite upotrebe. Mora da zadrži odabrani intenzitet svetlosti u svim uslovima upotrebe, u temperaturnom opsegu od -30°C do +50°C sa specijalnim izvorom napajanja za niske temperature i opsegu od -10°C do +50°C sa komercijalnim izvorom napajanja. U pogledu otpornosti mora ispunjavati zahteve funkcionalnosti prilikom pada i prevrtanja, dugotrajne vlage i slobodanog pada, a konstrukcija mora da sprečava njeno kotrljanje. Svetlosni snop lampe se sastoji od centralne tačke (snop većeg osvetljenja) i kruga (snop manjeg osvetljenja). Ispitivanje merenja intenziteta svetlosti, dometa i zaptivenosti se vrši u cilju provere deklarisanih vrednosti lampe u svim režimima upotrebe koja se odnosi na jačinu snopa svetlosti, dometa i vodootpornosti.**

Ključne reči – taktička lampa; intenzitet svetlosti; jačina snopa svetlosti; domet svetlosnog snopa; osvetljenost; luks; režimi rada; kandela; izvor napajanja; digitalni luksmetar

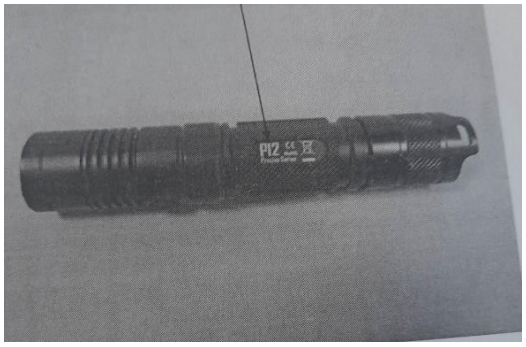
I. UVOD

Ručna taktička lampa koja se koristi u opitovanju je izrađena od aluminijumske legure sa zaštitom od korozije u crnoj mat boji. Lampa koristi punjivi izvor napajanja koji ima autonomiju rada sa jednim punjenjem baterija na maksimalnoj snazi od jednog časa na temperaturi od 20°C. Lampa ima zaštitu od napajanja obrnutim polaritetom. Baterije koriste punjač za izvor za napajanje za mrežni napon 220V i za 12/24V za vozila. Svetlosni izvor je LED dioda koja daje svetlost bele boje [1].

Lampa ima mogućnost uključenja preko osnovnog tastera na zadnjem delu i preko daljinske komande kada se koristi sa postavljenim kablom sa tasterom za daljinsku komandu. Osnovni taster je konstrukcijom kućišta zaštićen od slučajnog uključenja. Na prednjem delu lampe je taster sa oznakom „MODE", koji se koristi za promenu jačine svetlosti i rad sa specijalnim režimima rada. Jača svetlosti se podešava u četiri nivoa jačine kraćim pritiskom na taster. Dužim pritiskom na taster ulazi se u specijalni režim rada po sledećem rasporedu Strobe/ Location Beacon/ SOS. Kraćim pritiskom na taster izlazi se iz specijalnog režima rada.

Lampa ima mogućnost pamćenja nivoa osvetljenja. Odnosno pri ponovnom uključenju lampa radi u nivou osvetljenja u kom je radila pre isključenja [2]. Osnovne tehničke karakteristike:

1. masa..... 91g
2. dimenzije.....140mm x 25,4 mm x 25,4mm
3. intenzitet svetlosti u centru snopa... 25700cd
4. domet svetlosnog snopa..... 320m
5. fluks..... 1000lm



Slika 1. Ručna taktička lampa

II. MERNA I ISPITNA OPREMA

Merenje svetlosnog osvetljaja ručne taktičke lampe je izvršeno digitalnim luksmetrom koji se sastoji od mernog instrumenta i detektorske sonde. [3] Slika 1.

Tehnički podaci za DIGILUX 9500 A30 ser. br. 0049.14.502:

1. merni opsezi…….. 200mlx sa rezolucijom 0,01mlx; 2lx sa rezolucijom 0,1mlx; 20lx, 200lx, 2 klx, 20klx i 200klx (merni opsezi se mogu birati ručno ili automatski);
2. tačnost merenja… … greška je manja od 4%.

Digitalni luksmetar se sastoji od detektorske sonde sa fotoelementom i mernog instrumenta sa digitalnim displejem na kome se osvetljenost čita direktno u luksima (lx)[4]. Instrument se baždari u razmaku od jedne godine.

Slika 2. DIGILUX 9500 A30

Tačnost merenja je na svim opsezima merenja manja od ±4% što je znato bolje od tačnosti zahtevane u standardu SRPS U.C9.100, tač.6.(±10%).

III. REZULTATI MERENJA I TUMAČENJE

Opitovanje je otpočelo proverom funkcionalnosti, uključenjem i isključenjem uređaja i promenom svetlosnih modova. Izvršeno je takođe testiranje funkcionalnosti punjača, potpunim punjenjem baterija koja su korišćene u toku merenja. Utvrđeno je da su lampa, punjač i baterije funkcionalni.

A. Merenje intenziteta svetlosti

Intenzitet ili jačina svetlosti je jedna od sedam osnovnih fizičkih veličina koja opisuje snagu elektromagnetnog zračenja u području frekvencija vidljive svetlosti. Jednaka je svetlosnom fluksu Φs koji se izrači po jediničnom prostornom uglu ω [5][6][7]. Jedinica mere za jačinu svetlosti je kandela (cd) [4].

Pre početka merenja izvršili smo pripremu za opitovanje:

- instrument mrežnim kablom priključili smo na mrežu 22V/51 Hz i uključili napajanje prekidačem sa oznakom „NETZ" koji se nalazi na levoj strani prednje ploče,
- priključili detektorsku sondu na dvopolni priključak sa oznakom „EINGANG" na desnoj strani prednje ploče,
- detektorsku sondu postaviti tako da ne bude izložena svetlosti, odnosno ne skidati zaštitni poklopac,
- potenciometar označen sa „1-0.1" postaviti na vrednost 10.00,
- ostaviti instrument da radi 15 min,
- preklopnik za izbor opsega postaviti na položaj 2lx i potenciometrom „NULL" podesiti pokazivanje displeja na vrednost 0.000, odnosno 0.0000,
- proveriti sve ostale opsege (pokazivanje displeja treba da bude kao na opsegu 2lx).

Posle izvršenih provera i podešavanja instrument je spreman za rad [3]. Ispitna lampa postavljena je na optičkoj klupi i fiksirana je sa svih strana da ne bi došlo do njenog pomeranja. Lampa je u odnosu od sonde instrumenta udaljena 0,4m. Meren je intenzitet osvetljenosti za četiri različita režima lampe. Snop svetlosti od lampe je bio usmeren na centar sonde instrumenta.

Izvršena je provera intenziteta svetlosti ručne taktičke lampe, a zahtev je bio da jačina svetlosti ne sme biti manja od 25700 cd. Meren je svetlosni osvetljaj Ev, za četiri različita nivoa osvetljenja u luksima [lx][1]. Svetlosni osvetljaj predstavlja količnik svetlosnog fluksa i površine sa koje taj svetlosni fluks odlazi [5][6][7]. Merenje je sprovedeno u potpuno zamračenoj optoelektronskoj laboratoriji.

Intenzitet svetlosti (I_v) se izračunava po formuli:

$$E_v \circ R^2 = I_v \qquad (1)$$

R – rastojanje lampe od sonde instrumenta u metrima (m).

R. br. merenja	svetlosni osvetljaj Ev [lx]	rastojanje lampe od sonde instrumenta R [m]	kvadrat rastojanja R^2 [m^2]	intenzitet svetlosti Iv [cd]	funkcionalnost lampe
1.	161590	0,4	0,16	25854,4	funkcioniše
2.	161650			25864	funkcioniše
3.	161870			25900	funkcioniše
4.	162100			25936	funkcioniše

Izmereni intenzitet svetlosti taktičkog svetla u centralnom delu svetlosnog snopa je u opsegu od 25854,4cd do 25936cd. Zahtev za intenzitet svetlosti, po kome je intenzitet svetlosti u centru svetlosnog snopa ne sme biti manji od 25700cd, je ispunjen.

B. Merenje dometa svetlosnog snopa

Provera dometa svetlosnog snopa taktičke lampe izvršena je osmatranjem mete dimenzija 1m x 1m, noću, pri horizontalnoj osvetljenosti, pod otvorenim nebom, pri skoro punom Mesecu od oko 20mlx. Na udaljenosti od 340m od test mete, svetlosni snop taktičke lampe, jasno osvetljava celu površinu mete. Opitno-eksplatacionim ispitivanjem je utvrđeno, da lampa ima jasan domet snopa od minimalno 100m i da, takođe ima mogućnost osvetljenja predmeta i prostora na daljinama do 100m.

C. Provera zaptivenosti lampe

Usled korišćenja ručne taktičke lampe u uslovima smanjene vidljivosti (magla, sneg, kiša itd.), ispitana je po pitanju vodootpornosti. Lampa je ispitana prema standardu SNO 5706/84 na veštačku kišu i potapanje [8].

Vreme izlaganja na veštačku kišu je 1h. Vreme

potapanja lampe je 30 minuta na dubini 50cm. Temperatura vode u kojoj se potapa lampa iznosi 35̊C, a temperatura lampe je 10̊C iznad temperature vode. Nakon izlaganja na uticaj veštačke kiše i potapanja, lampa se obriše i osuši, nakon čega je utvrđeno da u lampu nije prodrla voda i njena funkcija je bila pravilna.

IV. ZAKLJUČAK

Ručna taktička lampa je na osnovu primenjenog tehničko-tehnološkog rešenja u skladu sa savremenim tendencijama razvoja i primenjenih tehničko-tehnoloških rešenja u svetu za proizvode iste ili slične namene. Na osnovu rezultata sprovedenih merenja lampa ima mogućnost rada sa različitim nivoima jačine svetlosti, mogućnost pamćenja nivoa osvetljenja, odgovarajući intenzitet svetlosti, domet svetlosnog snopa i mogućnost eksploatacije u svim klimatskim uslovima, tako da se može zaključiti da je ručna taktička lampa u skladu sa svojom namenom.

LITERATURA

[1] Sysmax IndustryTrading Co.,Ltd, Declaration of Conformity LED Flashlight

[2] .Sysmax IndustryTrading Co.,Ltd, P12GT User Manual

[3] M. Marković, Uputstvo za rukovanje digitalnim luksmetrom DIGILUX 9500 A30, Beograd, Srbija, 2022., Strana 2.

[4] Medjunarodni Sistem jedinica (SI sistem), 1960.

[5] P. Matavulj, Optoelektronika, Beograd, Srbija, 2007, strana 6. i 11.

[6] „Službeni glasnik RS" br. 132/2021, Uredba o zakonskim mernim jedinicama i načinu njihove upotrebe

[7] Građevinski fakultet, Odsek za Geodeziju I geoinformatiku, Tehnička fizika 2, predavanja Fotometrija, Beograd, 2020/2021.

[8] Biro za SiM u JNA, Standard Narodne Odbrane SNO 5706/84, Ispitivanje uticaja okoline na elektronske i elektromehaničke uređaje i pribor za potrebe KoV- a,1969.

Termovizijsko praćenje toplotnih efekata borbene opreme pri različitim fizičkim aktivnostima korisnika

Marina Tripkovic
Tehnički opitni centar GŠ VS
Beograd Srbija
marinamakitripkovic9@gmail.com
ORCID broj

Milena Jovanović
Tehnički opitni centar GŠ VS
Beograd Srbija
milenakv1996@gmail.com

ORCID broj

***Apstrakt* - Objekat ispitivanja je odećna oprema koja treba da zauzme bitnu ulogu u izgradnji operativne sposobnosti i izvršenju namenskih zadataka korisnika. U prvom delu namenjena je da pruži što bolju zaštitu od hladnoće, suvišne toplote, padavina i raznih mogućih povreda. Namenjena je, takođe da zaštiti telo od negativnih atmosferskih uticaja, a istovremeno mu obezbedi slobodu pokreta i brzinu reagovanja. Izrađena je od materijala koji ima mogućnost brzog sušenja i omogućava dobro disanje tela. Oprema mora da omogući potrebnu zaštitu korisnika u bliskom infracrvenom i termalnom delu elektromagnetnog spektra u svim uslovima danju i noću. U okviru ispitivanja izvršena je procena promene efektivne temperature na istom objektu ispitivanja pri različitim fizičkim aktivnostima i različitim ambijentalnim temperaturama. Odećna oprema ima za cilj da što više potisne signaturu termalnog odraza tela korisnika i smanji intenzitet reflektovanog snopa zraka, pa je zato potrebno da se izvrši laboratorijsko opitovanje toplotnih efekata termovizijskim praćenjem.**

Ključne reči – IC zračenje; stepen potiskivanja signature; termalni odraz; termovizijska kamera; termografski snimak; temperaturne merne tačke; temperaturna razlika;

I. UVOD

Odelo koje se koristi u termovizijskom opitovanju je od izolacionog materijala, brend sintetičkih vlakana koji se koristi za toplotnu izolaciju odeće, male debljine ali obezbeđuje potrebna izolaciona svojstva, tanka – izolacija. Materijal ima smanjenu veličinu i povećanu gustinu vlakana. Praznine između vlakana smanjuju protok toplote i omogućavaju odvod vlage. Izolacioni materijal predstavlja netkani materijal koji se izgrađuje od mikrovlakana, finih vlakana koja čine izolaciju na način da zadržavaju molekule vazduha između tela i spoljašnje sredine. Što je više vazduha u određenom prostoru raspoređeno u materijalu, to je bolja izolacija od hladnog spoljašnjeg vazduha. Vlakna u izolaciji su finija od vlakana koja se koriste u većini drugih sintetičkih ili prirodnih izolacija, ona zadržavaju više vazduha u manje prostora, što izolaciju čini efikasnijom.

Odeća sa ugrađenim izolacionim materijalom ne ograničava kretanje i pruža potpunu udobnost u veoma hladnim vrmenskim uslovima. Tanki sloj izolacije od mikrovlakana ima dobra svojstva toplotne izolaije u kombinaciji sa izdržljivošću i lakoćom održavanja. Životni vek materijala treba da odgovara životnom veku spoljašnje tkanine, od koje je izrađen odevni predmet. Klimatski uslovi u kojima će se odeća koristiti određuje vrstu izolacije, koja se proizvodi u različitim debljinama i različitim modifikacijama za različite klimatske uslove.

Odeća ima sposobnost u zaštiti od hladnoće, ali i u uslovima visoke vlažnosti. Izolacija na odelu dobro zadržava toplotu. Izolacijska vlakna upijaju malo vlage - manje od 1% svoje težine, tako da se izolacija postiže i u vlažnom okruženju (mokar materijal se brzo suši). Očuvanje toplote kroz materijal obezbeđuje vazduh koji se nalazi između vlakana. Što je više vazduha zadržano, izolacija materijala je bolja.

Sirovinski sastav materijala koji se ispituje je 100% poliester.

Svojstva materijala koji se ispituje:

- zadržava toplotu i omogućava jednostavno isparavanje viška vlage;
- elastičan je odnosto prima bilo koji oblik bez gubitaka osnovnih svojstava;
- ne gori, već se samo topi, ne upija vlagu, ima dobru propustljivost vazduha, čime se sprečava preterano znojenje, otporan je na habanje;
- nakon pranja ne gubi svoja svojstva.

Osnovna tkanina od koje se izrađuje odeća[1] namenjena je za povećanje maskirne zaštite [2], čime se obezbeđuje zaštita od osmatranja termovizijskim uređajima za osmatranje [3]. Tkanina se izađuje iz prediva bojenog u zeleno-drap boju, a zatim se vrši njeno štampanje u maskirnom dezenu [2].

Opitivanje se sprovodi u cilju provere toplotno-izolacionih osobina materijala na temperaturi od -10°C do +20°C. Radi poređenja karakteristika dva ispitna kompleta, uporedno su se pratili parametri pri nošenju dve različite vrste zimskih kompleta. U cilju pribavljanja što više informacija o ponašanju ugrađenjih materijala, vršeno je termovizijsko praćenje toplotnih efekata [3] pri fizičkoj aktivnosti ispitanika.

Ispitivanje je spovedno na istom ispitaniku koje je trajalo tri dana.

II. MERNA I ISPITNA OPREMA

Pri snimanju IC scene i merenju zračenja [4] korišćena je merna termovizijska kamera FLIR SC 7200 (slika 1) čije su deklarisane karakteristike date u Tabeli 1 [5]:

TABELA 1
Karakteristike termovizijske kamere FLIR SC 620

Spektralni opseg osetljivosti	(7, 5 – 13) µm
Rezolucija detektora	640 x 480 piksela
Osetljivost kamere (NETD)	< 40 mK
Temperaturni merni opseg	- 40˚C - + 120˚C
Ugao vidnog polja objektiva	24˚x 18˚
Žižna daljina korišćenog objektiva	38 mm

Dugotalasno IC zračenje (8-14) µm ima primenu u termoviziji, jer se prostire kroz atmosferu bez većeg slabljenja [2]

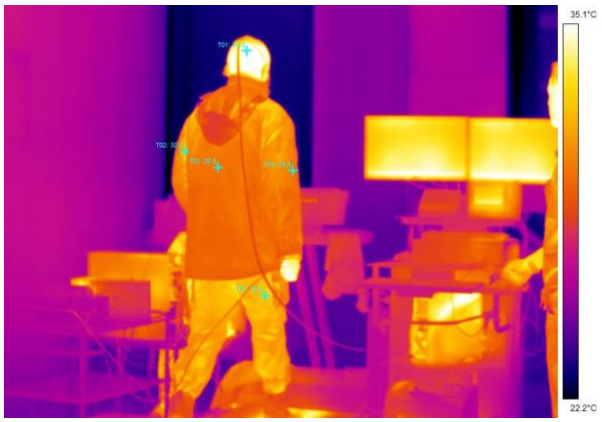
Slika 1. Merna termovizijska kamera FLIR SC 620

III. REZULTATI MERENJA I TUMAČENJE

Prvi deo ispitivanja vršen je na zimskom odelu komleta koji se satoji od:

- pantalona (poliesterski filamenti),
- majce dugih rukava (sintetička pletenina),
- bluze (sintetička pletenina) i
- jakne sa uloškom (poliesterska vlakna).

Ispitanik se nalazio na pokretnoj traci, čija se brzina povećavala u toku vremena, ako i nagib trake. Ispitivanje na pokretnoj traci možemo smatrati improvizacijom marširanja vojnika u realnim uslovima. Objekat ispitivanja je bio jasno uočljiv, a takođe je bio izažen utisak refleksije [6] od uređaja korišćenih pri opitu. Ambijentalna temperatura vazduha za ovaj deo ispitivanja je bila oko 24°C.



Slika 2. Dat je prikaz ispitanika sa opremom bez opterećenja

Slika 3. Dat je prikaz ispitanika sa opremom kada je brzina trake 4 km/h, bez nagiba i vremenom trajanja nakon tri minuta

Slika 4. Dat je prikaz ispitanika sa opremom kada je brzina trake 5 km/h, sa nagibom od 4% i vremenom trajanja nakon šest minuta

Slika 5. Dat je prikaz ispitanika sa opremom kada je brzina trake 6 km/h, sa nagibom od 7% i vremenom trajanja nakon tri minuta

Izabrano je po pet mernih tračaka na prvom kompletu i na osnovu dobijenih rezulatata temperatura termovizijskom kamerom uočeno je smanjenje temperature u predelu jakne sa uloškom u odnosu na glavu. Prilikom fizičke akrivnosti ispitanika došlo je do malog povećanja termalnog odraza u odnosu na početak ispitivanja. Termalni odraz ispitanika u oblasti jakne je prigušeniji u odnosu na termalni odraz pantalona istog kompleta.

Drugi deo ispitivanja vršen je na zimskom odelu komleta koji se satoji od:

- pantalona (poliesterski filamenti),
- majce kratnih rukava (sintetička pletenina),
- rolke (vuneno-sitetičko predivo),
- bluze (sintetička pletenina),

– kape (sitetička pletenina „polar”) i
– jakne sa uloškom (poliakrilat i Ag sa dodatim primesama Cu i Ni).

Ispitivanje je sprovedeno u komori [7], čija je temperatura -10°C. Ispitanik deo vremena provedenog u komori okreće pedale bicikla, što dovodi do postepenog povećanja opterećenja u toku opita. Objekat ispitivanja je bio jasno uočljiv i nije postojao uticaj refleksije drugih objekata na objekat ispitivanja.

U toku ovog dela opita napravljene su sledeće fotografije.

Slika 6. Dat je prikaz ispitanika sa opremom u komori bez opterećenja

Slika 7. Dat je prikaz ispitanika u komori nakon 3 minuta, kada bicika vrši opterećenje na ispitanika 50W

Slika 8. Dat je prikaz ispitanika u komori nakon 3 minuta, kada bicikla vrši opterećenje na ispitanika 70W

Slika 9. Dat je prikaz ispitanika nakon 30sec, kada bicikla vrši opterećenje na ispitanika 90W

Poređenjem karakteristika dva ispitna kompleta može se uočiti da je slabije prigušenje toplote u predelu glave i kada ispitanik nosi kapu i u predelu pantalona ispitanika kod oba kompleta. Najefikasnije prigušenje toplote je u predelu jakne sa uloškom kod oba ispitna kompleta.

Treći deo ispitivanja vršeno je na prvom kompletu koji se sastojao od:

– pantalona (poliesterski filamenti),
– majce dugih rukava (sintetička pletenina),
– bluze (sintetička pletenina) i
– jakne sa uloškom (poliesterska vlakna).

Ispitivanje sa ovim kompletom sprovedeno je u komori [7] čija je tempetarura -10°C. Ispitanik u komori okreće pedale bicikla, što dovodi do postepenog povećanja opterećenja u toku opita. Objekat ispitivanja je bio jasno uočljiv i nije postojao uticaj refleksije drugih objekata na objekat ispitivanja.

U toku ovog dela opita napravljene su sledeće fotografije.

Slika 10. Dat je prikaz ispitanika nakon 10 min u komori, kada se bicikla nalazi u stanju mirovanja

Slika 11. Dat je prikaz ispitanika u komori nakon 3 minuta, kada bicika vrši opterećenje na ispitanika 50W

Slika 12. Dat je prikaz ispitanika u komori nakon 3 minuta, kada bicikla vrši opterećenje na ispitanika 70W

Slika 13. Dat je prikaz ispitanika nakon 30sec, kada bicikla vrši opterećenje na ispitanika 90W

Termalni odraz sa ispitanika sa drugim kompletom je prigušeniji u odnosu na termalni odraz istog ispitanika sa prvim kompletom. Može se uočiti da drugi komplet smanjuje temperaturu IC odraza ispitanika u oblastima gde se koristi.

IV. Zaključak

Na osnovu snimaka napravljenih termovizijskom kamerom, gde je napravljena procena prigušenja toplote materijala ispitanika u infracrvenom području elektromagnetnog spektra od 8μm do 12μm može se zaključiti da materijal korišćen za drugi komplet, daje vidljivo prigušenje toplote u odnosu na materijal upotrebljen u sastavu prvog kompleta. Drugi komplet može zbog svojih karakteristika da obezbedi da konture ispitanika budu manje uočljive nego konture ispitanika sa prvim kompletom. Termalni odraz ispitanika će biti slabiji u koliko je u sastavu njegovog kompleta i kapa, a takođe i ako materijal ima obostrani nanos poliakrilata sa dodatim primesama srebra, bakra i nikla. Iz opita se vidi i važnost temeprature bitnog kontrolnog faktora radi što efikasnije zaštite ljudstva u vojci.

Literatura

Reference otkucati tekstvom veličine 8*pt*. Brojevi referenci treba da se automatski pojavljuju ispred naziva reference, kao u primeru dole i u uglastim zagradama numerisanja [1]. Takođe, tekst same reference treba da bude na istom nivou kao i u šablonu. U svim referencama samo se prezime daje puno, a ime se skraćuje na inicijal i stavlja pre prezimena. Molimo navedite imena svih autora; ne koristite "et al", osim ukoliko broj koautora rada nije veći od 8. Ne kombinujte reference: pod jednim brojem može biti samo jedna referenca. Ukoliko postoji DOI broj ili URL adresa elektronskog izvora, možete ih uneti na kraju reference. Uvek pišite pune naslove. Skratite imena časopisa prema standardima. Primeri različitih tipova referenci (članci u časopisima, knjige, poglavlja u knjigama, patenti itd.) dati su u primeru.

[1] Vojnotehnički institut, Izveštaj o ispitivanju spektralne refleksije materijala u spektralnom opsegu od 650-1000 nm – uzorak: tkanina osnovna Beograd, Srbija, 2023.
[2] Maskirna zaštita – Opšti propisi za proveru maskirnih karakteristika, SNO 8655, II – 2002.
[3] Infrared training center Serbia, Damiba trade, Kurs infracrvene termografije, nivo I, tehnički priručnik ITC.
[4] Petar Matavulj, Optoelektronika, Beograd, Srbija, 2007, strana 3,197. i 198.
[5] Tehničko uputstvo za kameru FLIR SC620
[6] Maskirna zaštita – Refleksija maskirnih materijala u UV, V i BIC području EM spectra, SNO 7511, II –
[7] Maskirna zaštita – Laboratorijske metode ispitivanja maskirnih karakteristika, SNO 8670, II – 2002.

Analiza podataka o saobraćajnim nezgodama sa socio-ekonomskog aspekta korišćenjem sistema poslovne inteligencije

Jordan Atanasijević
Centar za primenjenu matematiku i elektroniku, UTI (J-6) GŠ VS
Beograd
jordan.atanasijevic@vs.rs

Dejan Djukic
Fakultet za informacione tehnologije
Alfa BK Univerzitet
Beograd
dejan.djukic@alfa.edu.rs

Ivan Tot
Vojna akademija Univerzitet odbrane
Beograd
ivan.tot@va.mod.gov.rs

***Apstrakt* - Osnovna postavka rada vezana je za sagledavanje posledica saobraćajnih nezgoda na teritoriji Republike Srbije, sa socio-ekonomskog aspekta, korišćenjem sistema poslovne inteligencije. To je veoma značajno za odlučivanje o konkretnom problemu, a u cilju sagledavanja posledica pri donošenju odluka na najvišem nivou. Analiza mora da bude primerena i usklađena sa razvojem sistema poslovne inteligencije u uslovima kada već postoji ranije akumulirano znanje o problemu. Dobijeni rezultati treba da pokažu da se sistemom poslovne inteligencije i savremenog alata za transformaciju sirovih podataka u strateške informacije, *Power BI*, mogu identifikovati trendovi i posledice saobraćajnih nezgoda sa smrtno stradalim i teško povređenim licima, sa socio-ekonomskog aspekta, a sve sa ciljem preventivnog delovanja i smanjenja negativnih efekata.**

Ključne reči – poslovna inteligencija, saobraćajne nezgode, Power BI, analiza podataka.

I. UVOD

Skladištenje podataka i poslovna inteligencija su tehnike koje obezbeđuju poslovnim ljudima informacije i alate koji su im potrebni za donošenje odluka o operativnom i strateškom poslovanju.

Korisnici su uglavnom poslovni ljudi u kompanijama. Ali ne nose svi poslovni ljudi isti značaj. Potrebno je posebno se pobrinuti za one koji donose strateške poslovne odluke. Jedna dobra poslovna odluka može doneti milione dolara kompanijama. Zato su glavni korisnici rukovodioci, menadžeri i analitičari u kompanijama. Skladište podataka i poslovna inteligencija (*DW/BI*) jesu sistemi visokog uticaja. Termin strateški, takođe označava važnost. To su odluke koje mogu da donesu dobitak, ali i gubitak preduzeću. Prema tome, *DW/BI* sistem je sistem visokog poslovnog rizika. Kada se donese strateški važna odluka, neko dobija, a neko često i gubi. [1]

DW/BI sistem takođe podržava i donošenje operativnih odluka, naročito kada donosilac istih mora da vidi arhivu podataka ili integrisane podatke iz više izvora. Mnoge analitičke aplikacije imaju takav operativni fokus. [2]

Bilo da je odlučivanje strateško ili operativno, *DW/BI* sistem mora da pruži informaciju koja je potrebna za donošenje tih odluka.

Odluke najčešće zahtevaju jedinstveni podskup informacija, koji u osnovi nije predodređen. Potrebno je da se izgradi informaciona infrastruktura koja integriše podatke iz cele organizacije i potencijalno izvan organizacije, a zatim čisti, ispravlja i restruktuira podatke da budu fleksibilni i upotrebljivi što je više moguće. Dok većina modula transakcionog sistema radi sa jednim tipom informacija, kao što su fakture, nalozi ili potraživanja, *DW/BI* sistem mora da ih integriše sve zajedno. *DW/BI* sistem zahteva tehnički sofisticirano upravljanje i prikupljanje podataka.

Konačno, neophodno je donosioca odluke obezbediti i podržati alatima kojima će iskorišćavati podatke. U ovom slučaju, termin „alati“ je mnogo više od samog softvera. On označava u ovom slučaju sve što je korisnicima potrebno da mogu razumeti koje su im informacije dostupne, zatim da mogu pronaći podskup informacija koji im je potreban i da na kraju mogu struktuirati podatke kako bi uvideli poslovnu dinamiku. „Alati“ obuhvataju obuku, dokumentaciju i podršku, zajedno sa *ad hoc* upitima, izveštajima i analitičkim aplikacijama.

II. CILJ RADA

Osnovni cilj rada može se definisati kao istraživanje mogućnosti identifikacije trendova i kretanja troškova prouzrokovanih saobraćajnim nezgodama, pre svega primenom programskog paketa *Power Query* u okviru *Power BI* koji omogućava automatsko izdvajanje, transformaciju i učitavanja podataka, tj. razvoj *ETL* sistema i otkrivanja zakonitosti u podacima za potrebe brojnih analiza i upotrebe rezultata u cilju donošenja pravovremenih i ispravnih odluka.

Poseban cilj rada je praktična implementacija dobijenog novog i objedinjenog pristupa u cilju korišćenja istog u svrhu predikcije budućih stanja i postizanja optimalnih parametara rada u različitim situacijama kod donošenja odluka.

III. POSLOVNA INTELIGENCIJA

Izraz poslovna inteligencija prvi put je upotrebljen 1996. godine kako bi označio kategoriju sredstava analize podatka, postavljanja upita, izveštavanja, koji korisniku u poslovnim procesima mogu pomoći da iz ogromne količine podataka sintetizuju vredne informacije na kojima će zasnivati svoje poslovne odluke. Fenomen poslovne inteligencije može se posmatrati sa dva aspekta - makro i mikro aspekta. Posmatrana sa makro aspekta, poslovna inteligencija označava složenu agregiranu kategoriju, koja se stvara sistematskim, ali ne ciljanim prikupljanjem podataka o makro ekonomskim kretanjima u određenoj geopolitičkoj sredini. Ona, takođe, podrazumeva njihovo organizovanje i strukturirano beleženje, kao i logičko-računsku obradu radi otkrivanja trendova. Danas posebnu pažnju inženjera sve više

pobuđuje fenomen poslovne inteligencije posmatran sa mikro aspekta. U ovom slučaju se radi o otkrivanju prikrivenih znanja iz poslovnih podataka, koje neka organizacija prikuplja rutinski, obavljajući svoje svakodnevne poslovne transakcije.

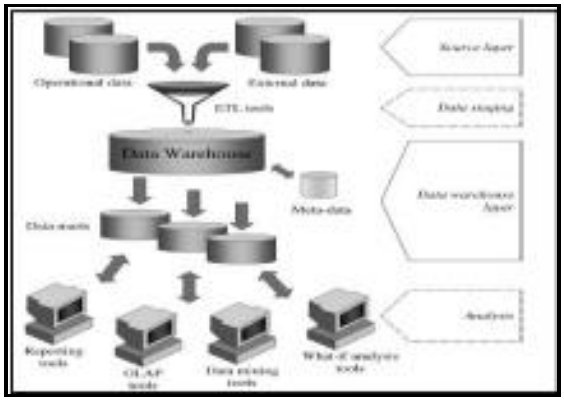
Poslovna inteligencija je relativno mlad pojam za koji postoji više definicija, ali svaka se svodi na to da je poslovna inteligencija proces prikupljanja podataka, pretvaranje tih podataka u informacije koje su korisne za donošenje važnih odluka. Poslovna inteligencija kao termin se najčešće koristi za označavanje računarske podrške odlučivanju. Sistem poslovne inteligencije je deo informacionog sistema organizacije namenski razvijen da podrži upravljanje. Upravljanje zahteva sveobuhvatan i blagovremen uvid u pokazatelje funkcionisanja organizacije kako bi donošene odluke bile pouzdane i precizne. Prema savremenim teorijama ovaj uvid treba da omogući što većem broju zaposlenih od kojih se, kada ga dobiju, može očekivati i veća efikasnost i odgovornost za ostvarene rezultate. Tehnike poslovne inteligencije (*data warehousing, reporting, OLAP, data mining, dashboards* i dr.) ekstrahuju podatke iz postojećeg informacionog sistema i transformišu ih u oblik pogodan za odlučivanje. Implementacija *BI* tehnika povećava upotrebnu vrednost postojećeg informacionog sistema organizacije, usled čega je interesovanje za sisteme poslovne inteligencije veliko i još uvek raste. [7]

Pošto ne postoji univerzalna definicija pojma poslovna inteligencija, različiti autori ga definišu na različite načine. Jedna od najčešće korišćenih i opštijih definicija je sledeća: "Poslovna inteligencija je takvo korišćenje podataka koje vodi ka donošenju boljih poslovnih odluka. Ono se odnosi na pristup, analizu i otkrivanje novih mogućnosti" [8]

IV. KORIŠĆENI MODEL ARHITEKTURE SKLADIŠTA PODATAKA

Zahtev za razvdvajanje igra ključnu ulogu u definisanju tipične arhitekture u istemu skladišta podataka, kao što je prikazano na narednoj slici (slika.1.). [1] Iako se obično naziva dvoslojnom arhitekturom, da naglasi razdvajanje između fizički raspoloživih izvora i skladišta podataka, zapravo se sastoji od sledeće četiri faze kroz koje podaci prolaze [4]:

1. **Izvorni sloj**: Sistem skladišta podataka koristi heterogene izvore podataka. Ti podaci se obično čuvaju u korporativnim relacionim bazama podataka a mogu doći iz informacionih sistema koji su izvan korporativnih zidova.
2. **Prenos podataka**: Memorisane podatke u izvorima treba izvući, očistiti od nedoslednosti, popuniti praznine i integrisati da se spoje heterogeni izvori u zajedničku šemu. Takozvani *ETL (Extraction, Transformation and Loading tools*) može da spojiti heterogenu šemu, ekstrahuje, transformiše, očisti, potvrdi, filtrira i učita podatke iz nekog izvora u skladište podataka. Tehnički rečeno, ova faza se bavi problemima koji su tipični za distribuirane informacione sisteme, kao što je i nekonzistentno upravljanje podacima nekompatibilne strukture podataka.
3. **Sloj skladišta podataka**: Informacije se čuvaju u jednom logičnom centralizovanom skladištu, skladištu podataka. Skladištu podataka se može direktno pristupiti, ali se takođe može koristiti kao izvor za stvaranje „*data marts*"- ova podataka, koji delimično ponavljaju sadržaj skladišta podataka i dizajnirani su za određena odeljenja preduzeća.
4. **Analiza**: U ovom sloju, integrisani podaci su na efikasan i fleksibilan način dostupni za izdavanje izveštaja, dinamičku analizu informacije i simulaciju hipotetičkih poslovih scenarija. Tehnički rečeno, trebali bi imati sledeća svojstva: navigaciju nad agregiranim podacima, optimizere kompleksnih upita i korisnički orijentisan *GUI*. Arhitekturna razlika između skladišta podataka i data-marts svakako treba biti jasno izražena. Komponenta označena kao skladište podataka na narednoj slici (slika.1.), se takođe često naziva primarnim skladištem podataka ili korporativnim skladištem podataka.



Slika 1. Ekstrakcija, transformacija i učitavanje podatke

Svaka analiza stvarnih podataka uključuje manipulaciju podacima, vizualizacija i modeliranje. Vizualizacija i modeliranje su komplementarni. Vizualizacije će nam pomoći da poboljšamo nejasna pitanja i oslanja se na ljudsku interpretaciju. Modeli su mnogo bolji i omogućavaju složenije računanje, ali su ograničeni svojim pretpostavkama. [5]

Krajnji proizvod analize nije model, već je to retorika. Analiza je besmislena ukoliko ne ubedi nekoga da preduzme akciju. U poslovanju, to obično znači ubediti više rukovodioce koji imaju malo statističkog znanja, pri donošenju neke strateške odluke. [6]

V. KORIŠĆENI MODEL ARHITEKTURE SKLADIŠTA PODATAKA

Trend razvoja Republike Srbije prati poređenje i praćenje najboljih i najnaprednijih zemalja u svim oblastima, pa tako i u saobraćaju. Iz tih razloga, potrebno je strateški delovati kako bi se aktivirao sistem bezbednosti saobraćaja u Republici Srbiji. Republika Srbija je počela 2019.godine da otvara svoje podatke, a nadležna institucija koja je zadužena za uvođenje otvorenih podataka (*engl. open data*) je Kancelarija za informacione tehnologije i elektronsku upravu u okviru Direkcije za elektronsku upravu. Otvoreni podaci su digitalni podaci, dostupni javnosti. Imaju takve tehničke i pravne karakteristike da svako, u svakom trenutku i svuda može da ih koristi, ponovo koristi i preraspodeljuje. Otvoreni podaci mogu da pomognu vladama, građanima i organizacijama da postignu bolje rezultate na polju javnih usluga.

Skupovi podataka na potralu otvorenih podataka dati su u *.CSV* formatu, koji je izuzetno pogodan za transformaciju, sažimanje i analizu, a takođe je dostupan za konverziju u mnoge druge formate, pogodne za analizu podataka. U objavljenim dokumentima o broju saobraćajnih nezgoda, koje su se desile na teritoriji Republike Srbije za peiod od u 2018. do 2023. godine, nalazi se sedam promenljivih (varijabli), odnosno kolona. Nazivi kolona definisani su sledećom tabelom. [3]

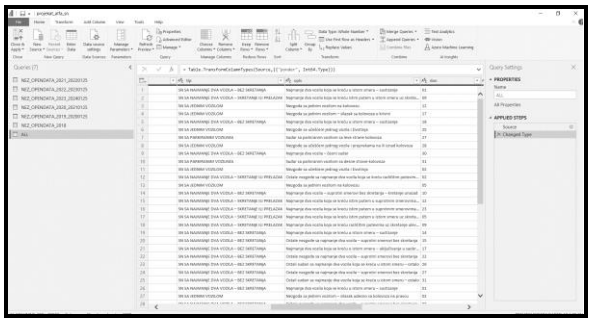
Tabela 1. Nazivi i objašnjenje promenljivih koje se nalaze u skupu otvorenih podataka korišćenog u istraživanju

Redni broj	***Naziv promenljive***	***Objašnjenje***
1.	NEZG_ID	Identifikacioni broj saobraćajne
2.	VREME_NEZ	Datum i vreme saobraćajne nezgode
3.	VRSTA_NEZ	Vrsta saobraćajne nezgode
4.	NAZIV_TIP	Naziv i tip saobraćajne nezgode
5.	NAZIV_DET	Detaljniji opis saobraćajne nezgode
6.	WGS_X	X koordinata WGS (engl. World Geodetic System) ili Geografska dužina
7.	WGS_Y	koordinata WGS (engl. World Geodetic System) ili Geografska širina

NEZG_ID - Identifikacioni broj saobraćajne nezgode, koji je predstavljen kao broj od sedam cifara. VREME_NEZ - Datum i vreme saobraćajne nezgode je dato u formatu dd.mm.yyyy, hh:mm (dan.mesec.godina, čas:minut). Iz datih podataka, može se zaključiti da su u bazi dostupni kompletni podaci za svih dvanaest meseci za 2018, 2019, 2020, 2021, 2022 i 2023 godinu. VRSTA_NEZ - Vrsta saobraćajne nezgode se deli na: 1) saobraćajne nezgode sa materijalnom štetom; 2) saobraćajne nezgode sa povređenim licima; 3) Saobraćajne nezgode sa poginulim licima. NAZIV_TIP - Naziv i tip saobraćajne nezgode se deli na: 1) saobraćajne nezgode sa jednim vozilom; 2) saobraćajne nezgode sa najmanje dva vozila (bez skretanja); 3) saobraćajne nezgode sa najmanje dva vozila (skretanje ili prelaz); 4) saobraćajne nezgode sa parkiranim vozilima; 5) saobraćajne nezgode sa pešacima. NAZIV_DET - Detaljniji opis saobraćajne nezgode sadrži 68 događaja: 1) saobraćajne nezgode sa jednim vozilom (11 događaja); 2) saobraćajne nezgode sa najmanje dva vozila (bez skretanja) (9 događaja); 3) saobraćajne nezgode sa najmanje dva vozila (skretanje ili prelaz) (18 događaja); 4) saobraćajne nezgode sa parkiranim vozilima (5 događaja); 5) saobraćajne nezgode sa pešacima (25 događaja).

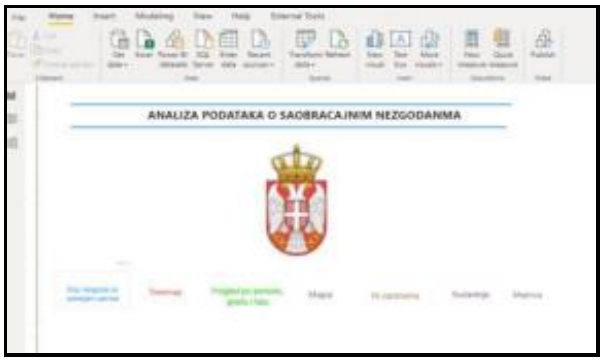
VI. REZULTATI ISTRAŽIVANJA

Za analizu podataka korišćen je *Microsoft Power BI* koji omogućava automatizovani process izdvajanja, transformacije i učitavanja podataka, tj. razvoj *ETL* sistema, korišćenjem *Power Query-a*. Kao što je predhodno rečeno, kao izvori podataka korišćeni su *excel* fajlovi koji su dostupni na portalu otvorenih podataka. Na slici broj 2 prikazan je proces učitavanja podataka sa predhodno navedenog portala, koji će se kasnije koristiti za analizu. Podaci su parsirani po redovima.



Slika 2. Učitavanje podataka u *Power Query*

Izgled glavnog menija aplikacije prikazan je na slici broj 3.



Slika 3. Glavni meni aplikacije

U nedostatku zvanične nacionalne metodologije, za proračun troškova saobraćajnih nezgoda u Srbiji u 2019. godini, korišćena je metodologija Evropske komisije za procenu ukupnih društveno-ekonomskih posledica saobraćajnih nezgoda.

Tako su ukupni društveno-ekonomski troškovi saobraćajnih nezgoda u Srbiji u 2019. godini iznosili oko 4,1 milijardu evra. [9]

Vrednosti su dobijene na osnovu prethodno sprovedenih istraživanja u Velikoj Britaniji. Istraživanja su objavljena u izveštaju Svetske asocijacije za puteve PIARC [9]. Troškovi se proračunavaju prema tome koliko neka osoba doprinosi Bruto domaćem proizvodu, svojim radom, kupovinom, transportom, kreditnim zaduživanjima itd. S obzirom da Republika Srbija nije sprovodila istraživanja utvrđivanja koeficijenata ili pondera koji se dobijaju na osnovu ukupnih društvenih troškova za pojedine vrste saobraćajnih nezgoda i posledica, rezultati su preuzeti iz britanskih istraživanja. Dobijene vrednosti su primenjive u velikom broju evropskih država, pa tako i u Srbiji.

Uzimajući u obzir da je ukupni bruto domaći proizvod (BDP) Srbije za 2019. godinu iznosio 46,1 milijardi evra, dolazi se do zaključka da je udeo troškova saobraćajnih nezgoda u ukupnom BDP-u Srbije 8,8 odsto. [9]

Prema ovoj metodologiji, trošak jedne saobraćajne nezgode sa poginulom osobom iznosi 3.300.100 eura, trošak jedne saobraćajne nezgode sa teško povređenom osobom 498.591 eura, dok trošak jedne saobraćajne nezgode sa lakše povređenom osobom iznosi 38.514 eura. [9]

Vrednosti koeficijenata 1, 10 i 85, koje se pridružuju saobraćajnim nezgodama preuzete su od Ministarstva

transporta Velike Britanije. Vrednosti su dobijene na osnovu prethodno sprovedenih istraživanja u Velikoj Britaniji. S obzirom da Republika Srbija nije sprovodila istraživanja utvrđivanja koeficijenata ili pondera koji se dobijaju na osnovu ukupnih društvenih troškova za pojedine vrste saobraćajnih nezgoda i posledica, rezultati su preuzeti iz britanskih istraživanja. Dobijene vrednosti su primenjive u velikom broju evropskih država, pa tako i u Srbiji. Pomenute vrednosti pondera su najčešće korišćene u istraživanjima i analizama rizika stradanja u saobraćaju po opštinama i policijskim upravama Republike Srbije. [10]

Tabela 2. Ponderisane vrednosti prema vrsti saobraćajne nezgode

Redni broj	*Naziv promenljive*	*Vrednost*
1.	Saobraćajna nezgoda sa smrtno stradalim	85
2.	Saobraćajna nezgoda sa povređenima	10
3.	Saobraćajna nezgoda sa materijalnom štetom	1

Ukoliko uzmemo bruto domaći proizvod za period od 2018-2023 godine, za koje posedujemo zvanične podatke i za koje postoje podaci o saobraćajnim nezgodama, onda možemo proračunati trošak jedne saobraćajne nezgode sa stradalima. [11]

Tabela 3. Proračun troškova jedne saobraćajne nezgode sa stradalima, na osnovu podataka o BDP-u

Godina	*Iznos BDP- a izražen u milijardama eura*	*Trošak jedne saobraćajne nezgode sa stradalima izražen u milionima eura*
1.	42.9	3.077.474
2.	46	3.300.100
3.	46.8	3.357.245
4.	53.3	3.823.544
5.	60.4	4.332.855
6.	69	4.949.784

Navedene vrednosti, obzirom da se radi o saobraćajnoj nezgodi sa stradalima, mogu se konvertovati u ponder 85. Troškovi saobraćajne nezgode sa povređenim licima, čiji je ponder 10, dobijaju se kada navedenu vrednost podelimo sa 8.5, a troškovi saobraćajne nezgode sa materijalnom štetom, dobijaju se deljenjem prikazanih vrednosti sa 85, respektivno, za svaki zapis u tabeli, jer se vrednosti menjaju po godinama.

Na osnovu prethodno navedenog, nakon učitavanja, transformacije i vizuelizacije podataka, dobijaju se sledeći izveštaji.

Na slici 4 prikazan je pregled proračunatih ponderisanih vrednosti i ukupnih troškova po godinama.

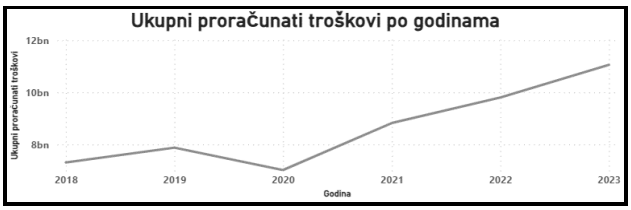
Year	Suma ponderisanih vrednosti	Ukupni proračunati troskovi
2018	201869	7,308,783,516.54
2019	202912	7,877,760,000.00
2020	177633	7,015,970,601.00
2021	196133	8,822,740,729.44
2022	192261	9,800,459,237.12
2023	189928	11,060,030,300.61
Total	**1160736**	**51,885,744,384.70**

Slika 4. Uporedni prikaz proračunatih ponderisanih vrednosti i proračunatih troškova

Sa slike se može videti da postoji pozitivan trend jer godinama suma ponderisanih vrednosti, koje predstavljaju zbir prema težini nezgode, opada, što je prikazano plavom bojom koja je sa opadanjem vrednosti tamnija, što je pozitivno. Sa druge strane, ukupni troškovi rastu. To je prikazano crvenom bojom, koja je godinama sve tamnija, što je negativno.

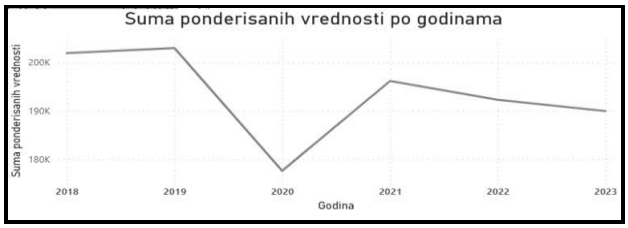
Možemo uočiti izuzetak u 2020 godini, koji je izazvan pandemijom korona virusa, kada je bilo uvedeno vanredno stanje i kada u određenom periodu nije bilo saobraćaja.

Na slici 5 prikazan je negativan trend rasta troškova.



Slika 5. Negativan trend rasta ukupnih troškova po godinama

Na slici broj 6 prikazan je pozitivan trend pada proračunatih ponderisanih vrednosti po godinama.



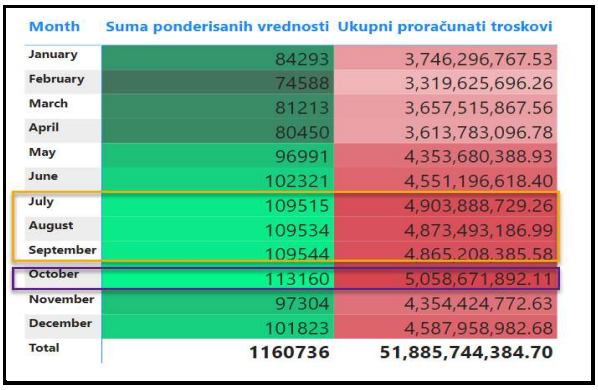
Slika 6. Pozitivan trend pada ukupnih proračunatih ponderisanih vrednosti

Microsoft Power BI daje nam mogućnost pravljenja dinamičkih upita, po bilo kom kriterijumu. Ukoliko kao kriterijum uzmemo vremensku dimenziju, konkretno tromesečje, dolazimo do zaključka da se najviše saobraćajnih nezgoda događa u 3 kvartalu (jul, avgust, septembar), pa su i troškovi koji to prouzrokuje za taj period najviši. Izveštaj se može videti na slici 7.

Quarter	Suma ponderisanih vrednosti	Ukupni proračunati troskovi
Qtr 1	240094	10,723,438,331.35
Qtr 2	279762	12,518,660,104.10
Qtr 3	328593	14,642,590,301.82
Qtr 4	312287	14,001,055,647.42
Total	**1160736**	**51,885,744,384.69**

Slika 7. Uporedni prikaz proračunatih ponderisanih vrednosti i proračunatih troškova po kvartalima

Ukoliko analizu izvršimo po mesecima, pored predhodno zaključenog da se najviše saobraćajnih nezgoda događa u 3 kvartalu (jul, avgust, septembar), dolazimo i do nove informacije, a to je da se najviše saobraćajnih nezgoda događa u mesecu oktobru. Izveštaj se može videti na slici 8.

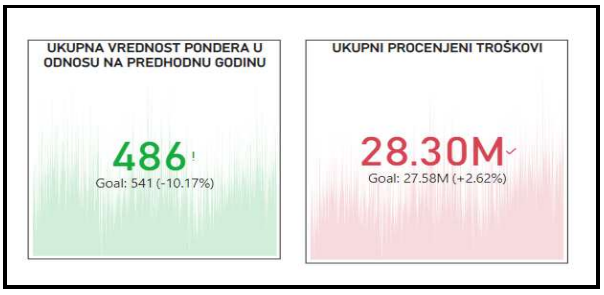


Month	Suma ponderisanih vrednosti	Ukupni proračunati troskovi
January	84293	3,746,296,767.53
February	74588	3,319,625,696.26
March	81213	3,657,515,867.56
April	80450	3,613,783,096.78
May	96991	4,353,680,388.93
June	102321	4,551,196,618.40
July	109515	4,903,888,729.26
August	109534	4,873,493,186.99
September	109544	4,865,208,385.58
October	113160	5,058,671,892.11
November	97304	4,354,424,772.63
December	101823	4,587,958,982.68
Total	1160736	51,885,744,384.70

Slika 8. Uporedni prikaz proračunatih ponderisanih vrednosti i proračunatih troškova po mesecima

Microsoft Power BI je alat koji nudi velike mogućnosti, jer u svojoj strukturi podržava *DAX*. *DAX (Data Analysis Expressions*) je jezik koji se koristi za kreiranje izraza (formula) radi pravljenja izveštajnih dimenzija koje se koriste u *PowerPivot* tabelama. Isti izrazi mogu da se koriste i za tabularni model u okviru rešenja *MS SQL Server Analysis Services*. Veliki broj *DAX* funkcija ima istu sintaksu kao *Excel* funkcije, dok neke druge mogu da rade sa relacionim podacima i vrše dinamičku agregaciju podataka, njihovo filtriranje itd.

Time Intelligence funkcije se često koriste za sagledavanje stanja u odnosu na isto vreme prethodne godine. U radu je korišćena jedna od njih, funkcija *SAMEPERIODLASTYEAR*. Korišćenjem ove funkcije, omogućeno je da izborom godine u bilo kom trenutku sagledamo trenutno stanje u poređenju sa prošlogodišnjim. Izgled izveštaja prikazan je na slici broj 9.



Slika 9. Uporedni prikaz proračunatih ponderisanih vrednosti i proračunatih troškova korišćenjem *DAX* funkcija, za isto vreme prethodne godine

VII. Zaključak

U radu je analizirano trenutno stanje u području istraživanja, korišćenjem objavljenih radova u prethodnom periodu koji su fokusirani u potpunosti ili delom na elemente korišćenja programskog paketa *Microsoft Power BI*.

Pri izradi i razvoju sistema poslovne inteligencije analizira se da li on olakšava rad krajnjim korisnicima, u procesu projektovanja, izgradnje, korišćenja i održavanja sistema poslovne inteligencije, one još ne daju pozitivne rezultate. Neophodno je pronaći načine i metode na koji način promeniti ustaljeni način rada, i unaprediti trenutne procese analize, korišćenjem savremenih alata.

Da bi se ova problematika rešila, neophodni su pozitivni stavovi ljudi iz upravljačkih struktura sistema, zatim prihvatanje novih sistema od strane krajnjih korisnika i profesionalnost, tj. poznavanje te tehnologije i mogućnost njene primene. U radu je predložena primena programskog paketa *Microsoft Power BI* u cilju unapređenja sistema poslovne inteligencije, kada su podaci predhodno učitani, transformisani i skladišteni u nekom od *DBMS*.

Literatura

[1] Matteo Golfarelli and Stefano Rizzi: "Data Warehouse Design: Modern principles andtehnologies", 2009

[2] Kimball, R., Ross, M., Thornthwaite, W., Mundy, J.,& Becker, B. (2008). The data warehouse lifecycle toolkit. John Wiley & Sons.

[3] Portal otvorenih podataka Republike Srbije [Internet]. [citirano 04.11.2020] Dostupno na: https://data.gov.rs/sr/datasets/

[4] Joseph Rickert, "Big Data Analysis with Revolution R Enterprise", 2011

[5] Wang, J., & Gu, L. (2016). Challenges of teaching data science in a business school. Issues in Information Systems, 17(3).

[6] Fotache, M., & Strimbei, C. (2015). SQL and data analysis. Some implications for data analysits and higher education. Procedia Economics and Finance, 20, 243-251.

[7] Wickham, H. (2019). Data science: how is it different to statistics?. IMS Bulletin, 48.

[8] Kimball, R., & Ross, M. (2016). Dimension Table Core Concepts. The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence, 327-384

[9] Godthelp, H., & Ksentini, A. (2024). Specific road safety issues in low- and middle income countries (LMICs): an overview and some illustrative examples. Traffic Safety Research, 8, e000068-e000068.

[10] Kukic, D., Milinkovic, B., Petrovic, D., Nojkovic, D. (2016). Uticajni faktori u procesu nastanka saobracajne nezgode – prva iskustva prikupljanja podataka u Republici Srbiji. XIII International Symposium „Road Accidents Prevention 2016". Novi Sad. Srbija.

[11] https://www.stat.gov.rs/sr-latn/oblasti/nacionalni-racuni/godisnji-nacionalni-racuni, pristupljeno dana 13.12.2024.godine

Analysis of traffic accident data from a socio-economic aspect using a business intelligence system

Atanasijević Jordan

ABSTRACT

The basic setting of the work is related to the assessment of the consequences of traffic accidents on the territory of the Republic of Serbia, from the socio-economic aspect, using the business intelligence system. This is very important for deciding on a specific problem, and in order to see the consequences when making decisions at the highest level. The analysis must be appropriate and coordinated with the development of the business intelligence system in conditions where previously accumulated knowledge about the problem already exists. The obtained results should show that the business intelligence system and the modern tool for transforming raw data into strategic information, Power BI, can identify trends and consequences of traffic accidents with fatal and seriously injured persons, from a socio-economic aspect, all with the aim of preventive action and reduction of negative effects..

Primena CMMN notacije u modelovanju dinamičnih poslovnih procesa i upravljanju slučajevima

Jordan Atanasijević
Centar za primenjenu matematiku i elektroniku, UTI (J-6) GŠ VS
Beograd
jordan.atanasijevic@vs.rs

Dejan Viduka
Fakultet za informacione tehnologije
Alfa BK Univerzitet
Beograd
dejan.viduka@alfa.edu.rs

Ivan Tot
Vojna akademija Univerzitet odbrane
Beograd
ivan.tot@va.mod.gov.rs

***Apstrakt* - Osnovna postavka rada i problematika koja se obrađuje značajna je iz više razloga, pa u budućnosti može pomoći u mnogim poslovnim slučajevima. Jedan od glavnih razloga je taj što se korišćenjem predloženog modela postiže standardizacija i razumljivost. Sama notacija pruža standarde kojih se moramo pridržavati, pa se pomoću tih standarda svaki proces može analizirati i razumeti bez nesporazuma s drugim osobama. Korišćenjem ove notacije izbegavaju se nepotrebni koraci u procesima, pa organizacijama omogućuje optimizaciju svih procesa uz smanjenje troškova i povećanje efikasnosti. Notacija omogućava organizacijama lako prilagođavanje poslovnim okruženjima i brže prilagođavanje novim prilikama i izazovima. Još jedan razlog zašto je ova tema značajna je mogućnost lakše analitike zbog toga što su procesi i odluke jasno definisani. Bolje izveštavanje i analitika omogućuju nam lakši izbor poslovne strategije.**

Ključne reči – notacija, poslovni proces, modelovanje.

I. UVOD

Razvoj softverskih sistema na današnjem nivou, mogućnosti računara i očekivanja korisnika, zahteva sveobuhvatni rad vezan za realizaciju bilo kog informacionog sistema. Pošto je ručno razvijanje softvera od najnižeg nivoa skupo i dugotrajno i sa ne uvek predvidivim rezultatima, postoji potreba da se razvoj softvera olakša, zbog čega je, pre više od dvadeset godina, nastalo softversko inženjerstvo kao disciplina.

Automatizacija softverskog inženjeringa na računaru se izvodi posebnim alatom, čiji je naziv CASE (*Computer Aided System Engineering*).

CASE sistem predstavlja alat koji služi kao pomoć projektantu informacionih sistema. Od efikasnosti ovog alata može da zavisi kvalitet gotovog proizvoda (informacionog sistema), tako da je projektantu veoma važno da odabere pravi alat koji će ga zamenjivati u većini manuelnih poslova vezanih za projektovanje. CASE alati treba da omoguće da [1]:

- Dalji rad na projektu IS ne zavisi od prethodnog izvodjača.
- Nezavisnost od budućeg sistema za upravljanje bazom podataka (SUBP) koji se definiše tek sa izvođačkim projektom iz čega se zaključuje da donošenje odluke o vrsti buduće baze podataka nije prioritet.
- Treća prednost je što za izradu projekata nije potrebno kupovati hardver jer je u pitanju prototipski način rada.

Modeli u svetu softvera su planovi za sistem. Planovi pomažu da se isplanira izgradnja pre nego što se stvarno krene na izradu aplikacija. Rezlutat procesa modelovanja je mogućnost da se prate poslovni zahtevi.

Vizuelno modelovanje je proces uzimanja informacija sa modela i njihovo grafičko prikazivanje korišćenjem niza standardnih grafičkih elemenata. Standardi su neophodni kako bi se izvukla bitna korist od modelovanja: komunikacija. [2]

Komunikacija između korisnika, programera, analitičara, menadžera i svih onih koji su uključeni u projekat, je osnovna svrha vizuelnog modelovanja. Komunikaciju možete ostvariti korišćenjem nevizuelnih (tekstualnih) informacija, ali ipak ljudi mogu vizuelno mnogo lakše prihvatati i uočavati određene stvari.

Jedno od važnih pitanja u vizuelnom modelovanju je grafička notacija i koristi se za opis različitih aspekata sistema. Ta notacija mora da bude poznata svim zainteresovanim grupama, u suprotnom model neće biti koristan.

Na početku treba definisati granice do kojih će se ići u modelovanju poslovnih procesa. Da li se modeluje čitava organizacija ili samo jedno odeljenje? Koji tokovi poslovnih procesa su bitni za projekat ? i sl.

Kad se definiše obim projekta, veoma je važno okupiti pravi tim. Potrebni su pojedinci koji poznaju poslovne procese, kao i pojedinci koji poznaju modelovanje poslovnih procesa. Uopšteno, članovi tima ne moraju da budu informatičari, čak je bolje da ne budu. Informatičari vrlo brzo kreću ka rešenju, dizajnu sistema, ne analizirajući dovoljno poslovne procese.

Obavezni članovi tima su [3]:

- Vođa tima - On treba da poseduje dovoljno znanja i o poslovnim procesima i o modelovanju poslovnih procesa. On ili ona će biti zadužen za koordinaciju napora članova tima i za održavanje pravca rada.
- Predstavnik(ci) poslovnih procesa-Oni su predstavnici različitih delova organizacije koje modelujemo. Oni treba da su dosta upoznati sa tokovima poslovnih procesa, uključujući i probleme koji se u njima javljaju i dobit od tih tokova poslovnih procesa.
- Predstavnik(ci) menadžmenta kompanije - Neko ko ima autoritet da odluči koji će delovi ili poslovni procesi biti modelovani. Oni mogu da pomognu

timu da razume tokove poslovnih procesa iz perspektive menadžmenta.

II. CILJ RADA

Osnovni cilj rada predstavlja mogućnost da se istražiti i analizira primena CMMN (*Case Management Model and Notation*) notacije u modelovanju i optimizaciji poslovnih procesa koji se temelje na slučajevima i donošenju odluka. Rad ima za cilj:

- **Razumevanje teorijskog okvira CMMN notacije** – pružiti detaljan pregled ključnih elemenata CMMN-a, kao što su slučajevi, aktivnosti, događaji i planovi, te način na koji ova notacija omogućava fleksibilnost u upravljanju poslovnim procesima.
- **Istraživanje primene CMMN u različitim industrijama** – analizirati konkretne slučajeve upotrebe CMMN notacije u različitim poslovnim domenima, kao što su pravni, medicinski, i uslužni sektori, i identifikovati prednosti koje donosi u odnosu na tradicionalne pristupe upravljanju procesima.
- **Procena uticaja na donošenje odluka** – istražiti kako primena CMMN notacije omogućava bolje, brže i fleksibilnije donošenje odluka u složenim i dinamičnim poslovnim okruženjima.
- **Razvijanje metodologije za implementaciju CMMN u organizacijama** – pružiti smernice i preporuke za uspešnu implementaciju CMMN notacije u organizacijama, uključujući izazove i strategije prevazilaženja istih.
- **Predložiti moguće pravce za dalja istraživanja i unapređenja CMMN notacije** – identifikovati oblasti u kojima se CMMN može dalje razvijati i unapređivati

III. OSNOVNI KONCEPTI CMMN

Na samom početku, pre objašnjenja same notacije CMMN, biće objašnjen jedan od osnovnih pojmova, a to je POSLOVNI PROCES. Poslovni proces je zapravo skup određenih zadataka ili aktivnosti koje su povezane i strukturirane, pa predstavljaju niz koraka, čijim se praćenjm dolazi do određene usluge ili proizvoda.[4] Poslovni proces je povezani skup aktivnosti i odluka, koji se realizuje radi dostizanja određenog cilja organizacije, traje određeno vreme i troši neke ulazne resurse pretvarajući ih u specifične proizvode ili usluge od značaja za kupca ili korisnika. [5] Postoji mnoštvo definicija poslovnog procesa i svaka vodi do nekog drugog zaključka. Kako bi se izbeglo da se svaka osoba drži svoje definicije i proizvodi usluge i proizvode na svoj način, uvedena je notacija BPMN (*Bussines Process Modeling and Notation*).

BPMN se bavi dizajniranjem, upravljanjem i izvršavanjem poslovnog procesa, a njegova snaga leži u integrisanju i proširenju postojećih procesno orijentisanih tehnika i tehnologija. [6]

BPMN je skup je konvencija za modeliranje poslovnog procesa, sastavljen od grafičkih elemenata i formaliziranih zapisa, koji ima status profesionalne norme. BPMN sadrži specifikacije potrebne za generiranje aplikacije za izvršavanje poslovnog procesa. BPMN je međunarodni standard koga se svi pridržavaju, a nastao je kako bi sva poduzeća mogla da vizuelizuju poslovne procese i kako bi sam tok rada bio efikasniji i efektivniji. [7]

BPMN je doneo brojne prednosti, kao što su smanjenje vremena poslovnog procesa, rutinska automatizacija aktivnosti, poslovna integracija i vidljivost rezultata do kraja. Međutim, pored strukturiranih procesa, važni su i takozvani nestrukturirani procesi koji nisu unapred definisani ili ponovljivi, ali zavise od razvoja određenih okolnosti i odluka donetih u određenim situacijama (slučajevima). Koristi se CMMN da se odredi ova grupa procesa. [8]

CMMN je standard koji je razvio OMG (*Object Management Group*) i koristi se za modelovanje upravljanja slučajevima (*case management*). Ova notacija korisna je u situacijama gde procesi nisu strogo definisani i gde je potrebna fleksibilnost u rukovanju različitim scenarijima i izuzetnim situacijama. CMMN proširuje granice onoga što se može modelovati BPMN-om, uključujući manje strukturirane radne aktivnosti i one koje pokreću stručnjaci. U kombinaciji BPMN i CMMN omogućuju korisnicima pokrivanje puno šireg spektra radnih metoda. [9]

CMMN se pojavio kao grafički prikaz procesa zasnovan na slučajevima koji su nestrukturirani i nepredvidljivi. Činilo se da je to bila sjajna ideja, obzirom da su sve ostale notacije i standardi bile usmerene na strukturirane slučajeve. Ključna tačka za uspon CMMN-a bila je kada je postao standard koji je implementirao OMG.

Prva verzija CMMN-a nastala je 2014. godine, a ispravljena verzija je nastala 2016. godine. Tri godine nakon izdavanja ispravljene verzije, CMMN nije postigao očekivanu popularnost, ali su ga određeni alati podržavali. Trenutno stanje CMMN-a je takvo da se određeni ljudi udaljavaju od njega, a čak više ni ne žele da razvijaju podršku za CMMN. Mnoštvo ljudi se slaže da je CMMN složen za učenje i korišćenje, a još jedan razlog za udaljavanje od CMMN-a je što su ljudi već naučili BPMN. Mnogi imaju otpor ka učenju nove notacije. [10]

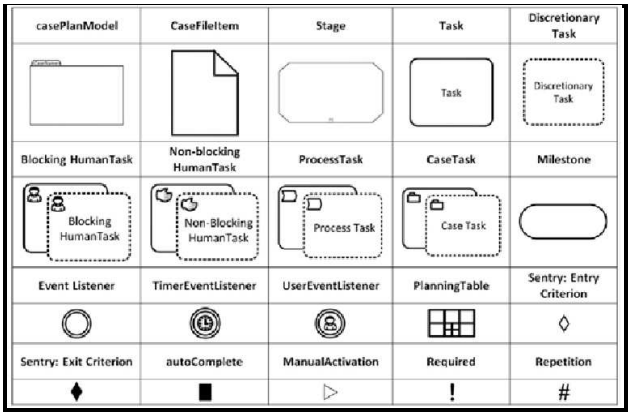
Bez obzira na određene komentare i manju podršku ljudi CMMN je bitan za korišćenje i svakako nam jako pomaže u kombinaciji s BPMN-om. Istraživanje i praksa poslovnih informacijskih sistema fokusiraju se na dobro strukturirane poslovne procese, ali mnoge poslovne procese je teško modelovati. Ovo se najviše odnosi na siteme koji podrazumevaju široko znanje u navedenoj oblasti, kao što je upravljanje incidentima, savetovanje ili prodaja. Mnoge aktivnosti vrše se na *ad-hoc* način, umesto da se unapred isplaniraju. To se događa kod aktivnosti koje zahtevaju intenzivno znanje ili se temelje na projektima. Takve aktivnosti inače predstavljaju ključne aktivnosti organizacije.

Ad-hoc zapravo predstavlja nešto što se odnosi na konkretan slučaj ili situaciju. U tom delu svi koristili BPMN notaciju. Kada su poznate informacije, odluke i proizvodi lako je napraviti određeni poslovni proces. Sa druge strane CMMN nam omogućuje da stvaramo modele u kojima je situacija nepredvidiva i gde se stvarne aktivnosti i njihov redosled razlikuju od slučaja do slučaja. Uz CMMN možemo stvoriti procese koji nisu u potpunosti definisani na početku jer zahtevaju informacije koje postaju dostupne tek tokom projekta. [11]

Sam tok procesa ne može biti strukturiran i novi zadaci pojavljuju se u toku izrade projekta. Aktivnosti su delimično poznate unapred i ljudi imaju visok stepen slobode u toku izrade samog projekta.

IV. SINTAKSA I SEMANTIKA CMMN-A

Sintaksa CMMN-a obuhvata grafičke elemente i njihove veze koji se koriste za modelovanje upravljanja slučajevima. Slučaj (*Case*) osnovna je jedinica CMMN-a i predstavlja situaciju koja zahteva upravljanje. Slučaj se može sastojati od različitih aktivnosti i događaja. Planirani elementi (*Planning elements*) su aktivnosti, događaji i zadaci unutar slučaja koji se mogu izvršavati različitim redosledima ili u isto vreme tj. paralelno. Odluke (*Decisions*) definišu uslove ili kriterijume na osnovu kojih se donose određene akcije unutar slučaja. Dijagram slučaja (*Case diagram*) prikazuje sve planirane elemente, odluke, događaje i njihove odnose, pa omogućavaju jasno razumevanje toka upravljanja slučajevima. Semantika CMMN-a će biti objašnjena u nastavku. Pregled elemenata sa nazivom i simbolom, a koji se koriste u imenovanoj notaciji prikazan je na slici 1.



Slika 1. Prikaz elemenata koje koristimo u CMNN notaciji [12]

Nakon prikaza bitnih elemente koji se koriste u izradi CMMN modela, sledi objašnjenje značenja svakog od tih.

Case Plan Model (Model Plana Slučaja) predstavlja glavni dijagram slučaja u CMMN-u. To je centralna struktura koja sadrži sve ostale elemente i definiše kako će se slučaj izvršavati.

Case File Item (Stavka Datoteke Slučaja) predstavlja podatke ili informacije koje se koriste unutar slučaja. Ove stavke mogu biti dokumenti, obrasci ili bilo koji drugi izvori podataka koji su potrebni za obradu slučaja.

Stage (Faza) je element koji grupiše srodne aktivnosti ili događaje unutar slučaja. Koristi se za organizovanje i strukturiranje različitih faza ili faza koje su u tokuizvršavanja slučaja.

Task (Zadatak) je osnovna jedinica aktivnosti unutar slučaja koju mora izvršiti neka osoba ili organiyaciona jedinica. Postoje različite vrste kao što su *human tasks* (zadaci koje realiyuju ljudi) ili *automated tasks* (automatizirani zadaci).

Discretionary Task (Proizvoljni zadatak) je zadatak koji nije obvezan za izvršavanje tokom trajanja slučaja, pa se može izvršiti prema potrebi ili odluci.

Blocking Human Task (Blokirajući ljudski zadatak) je zadatak koji realizuje određena osoba i može zaustaviti dalji tok slučaja dok se ne izvrši u potpunosti.

Non-blocking Human Task (Ne-blokirajući ljudski zadatak) je zadatak koji se izvršava paralelno s drugim aktivnostima, ukoliko je potrebno, i ne zahteva zaustavljanje ostalih aktivnosti.

Process Task (Procesni zadatak) je zadatak koji predstavlja integraciju sa spoljnim poslovnim procesima ili uslugama koje se izvršavaju unutar slučaja.

Case Task (Zadatak slučaja) je zadatak koji se koristi unutar slučaja za upravljanje specifičnim aspektima slučaja ili njegovim tokom.

Milestone (Prekretnica) označava dostignuće ili važan događaj unutar slučaja koji označava određenu tačku napretka u toku izvršavanja.

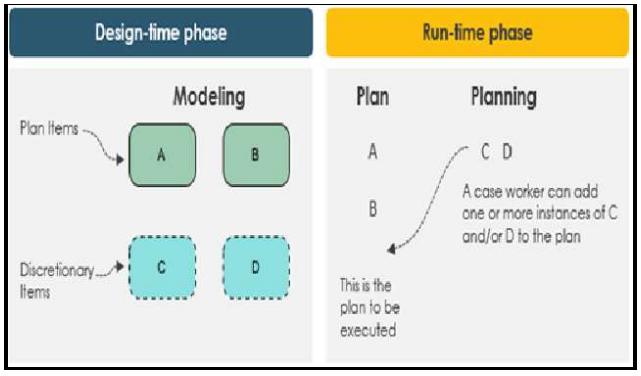
Event Listener (Slušatelj događaja) je element koji reaguje na određene događaje i prati ih.

Timer Event Listener (Osluškivač rokova) je specifičan tip osluškivača koji alarmira na određene rokove unutar slučaja ili tajmere.

User Event Listener (Slušatelj korisničkog događaja) je osluškivač koji reaguje na specifične događaje koje generišu korisnici.

Ono što se mora naglasiti je da u CMMN-u nema modela sekvencijskog toka. Sva izvršenja zadataka zavise od događaja i uuslova koji se nazivaju straže (*sentries*). Kada straža zabeleži pojavu određenog događaja ili ispunjenje uslova unutar slučaj, tek onda se kreće sa izvršavanjem zadataka. Straže se koriste kao ulazni i izlazni kriterijumi. Crni i beli dijamanti koje možete videti na slici broj 1, predstavljaju oznaku za uslove. Slučaj CMMN-a ima dve faze, a to su faza dizajna i faza izvođenja. Za vreme faze dizajna, poslovni analitičari se bave modelovanjem, a to uključuje planiranje stavki tj. definisanje zadataka koji su uvek deo unapred definianih segmenata u modelu slučaja, kao i proizvoljnih zadataka koji su dostupni radniku na slučaju i koji se mogu primeniti opciono. [13]

U fazi izvođenja radnici koji rade na slučaju izvršavaju plan tako što obavljaju zadatke prema planu, a po potrebi dodaju proizvoljne zadatke u instancu plana slučaja tokom izvođenja. Na slici 2 prikazane su faza dizajna i faza izvođenja.



Slika 2. Faza dizajna i faza izvođenja [14]

V. PODRUČJA PRIMENE CMMN-A

CMMN može pružiti kvalitetna rešenja u širokom opsegu poslovnih situacija. Često se koristi u državnim institucijama, zahtevima za kredite, zahtevima za osiguranje, reklamacija kupaca i u zdravstvu. Takođe se koristi za upravljanje pravnih postupaka, istraživanje sumnjivih aktivnosti, administraciju socijalnih programa i za rad u nekim humanitarnim organizacijama. CMMN se može koristiti na različite načine pa zato zadovoljava razne organizacijske potrebe.

CMMN standard nije podržan od mnogih distributera alata jer ostali više podržavaju BPMN. Finansijske usluge su jedno od najpogodnijih područja gde CMMN ima veliku ulogu. Potražnja za svim vrstama finansijskih usluga je velika, pa se zato reinženjering tih usluga vrši učestalo što je pogodno za CMMN. [15]

U zdravstvu se u današnje vreme koriste tekstualni opisi, dijagrami i dijagrami toka podataka za dokumentovanje kliničkih procesa. To je pogodno za CMMN jer se pomoću njega mogu stvoriti kvalitetni i sigurni automatizovani grafički prikazi zdravstvenih slučajeva i kliničkih stanja. Takvi CMMN dijagrami su vizualni, jednoznačni i intuitivni, pa su lako razumljivi pružaocima usluga i IT stručnjacima. Ovakve vrste podataka su superiornije u odnosu na tekstualne dokumente.

Osim u bankama i zdravstvu, CMMN se koristi i u različitim sektorima i industrijama kao što su osiguranje, obrazovanje, telekomunikacija, maloprodaja i IT.

U osiguranju se CMMN koristi kako bi se obradili različiti zahtevi, proverile prijave i kako bi se upravljalo pritužbama osiguranika.

U obrazovanju se CMMN koristi kako bi se upravljalo studentskim slučajevima, akademskim postupcima, prijavama i žalbama.

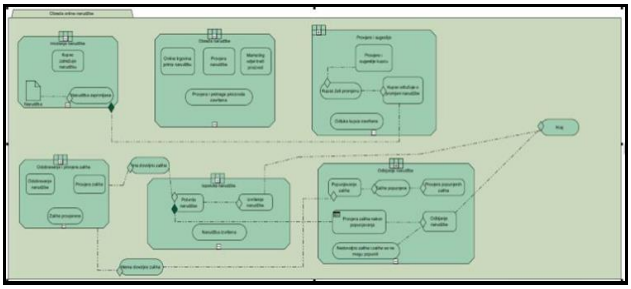
U telekomunikaciji se CMMN koristi kako bi se rešili problemi korisničke podrške, pratili kvarovi na mreži i kako bi se upravljalo tužbama korisnika.

U maloprodaji CMMN koristimo za rešavanje pritužbi kupaca, upravljanje reklamacijama i zamenu proizvoda.

U IT-u CMMN koristimo za upravljanje IT incidentima, praćenje promjena, rješavanje tehničkih problema i upravljanje projektima

VI. PRIMER UPOTREBE

U nastavku će biti objašnjeno zašto je CMMN notacija važna i kako funkcionišu dijagrami pravljeni CMMN standardom.



Slika 3. Primer CMMN dijagrama u procesu kupnje proizvoda preko Web Shopa

U ovom poglavlju će biti objašnjen model CMMN dijagrama kroz primer procesa kupovine proizvoda putem web trgovine. Za razliku od BPMN-a koji je idealan za jasno definisane događaje, CMMN nam pruža fleksibilnost za procese koji nisu jasno definisani i za koje ne znamo kada će se dogoditi. Istraži ćemo kako CMMN može biti korišćen za modelovanje procesa koji uključuju razne korake i odluke pokrenute određenim prekretnicama. Na primeru procesa online porudžbine videćemo koliko je zapravo lakše CMMN-om modelovati fleksibilan tok aktivnosti. U samom dijagramu biće uključene razne interakcije poput interakcije kupca, online tprodavnice, dela za marketing i skladišta. Razmatraćemo različite odluke, odobrenja, provere i eventualne promene porudžbine zbog dostupnosti proizvoda.

U nastavku će biti objašnjene prednosti korišćenja CMMN-a u odnosu na tradicionalne BPMN dijagrame i opisani sami CMMN dijagram. Posebna pažnja biće posvećena tom delu zbog konteksta upravljanja poslovnim procesima koji zahtevaju prilagodljivost i reakcije na nepredviđene događaje.

CMMN dijagram „Obrada online porudžbine“ pokazuje da ovaj proces započinje kada kupac inicira porudžbinu. Svi procesi izvršavaju se unutar velikog prostora koji se zapravo zove „*Case Plan Model*“. Unutar njega postavljaju se sve ostale sekcije procesa.

Unutar *Case Plan Modela* nalaze se faze koje se zovu „*Stage*“. Svaka faza je zapravo skup zadataka, uslova i kriterijuma. Faza 1. „Iniciranje porudžbine“ objašnjava kako kupac inicira porudžbinu kroz zadatak „Kupac zatražuje porudžbinu“. U tom delu potrebni su nam Case File Item „Narudžba“ koji označava dokument koji se šalje Online trgovini. Case File Item „Narudžba“ je input prekretnici „Narudžba zaprimljena“. Prekretnica je zapravo događaj nakon kojeg se stvaraju nove odluke i procesi. Faza 2. je „Obrada porudžbine“, a on nam zapravo pokazuje procese koji se ostvaruju nakon što je narudžba zaprimljena. Sljedeći procesi nakon dobijanja same porudžbine su primanje porudžbine, provjera porudžbine i traženje proizvoda. Na kraju imamo prekretnicu koja označava novi događaj gde su provjera i pretraga proizvoda završeni. Faza 3. „Provjere i sugestije“ označava procese vezane uz provjere i sugestije. U tom delu odvijaju se procesi i odluke vezani za slučaj kada kupac želi promeniti određene elemente porudžbine. Ako kupac nešto želi promeniti onda se vraćamo na fazu 1. gde se ponavljaju procesi od dobijanja porudžbine pa nadalje. Nakon toga imamo uslov koji pokazuje da je odluka kupca završena. Faza 4. „Odobravanje i projera zaliha“ izvršava se nakon završne odluke kupca, a u tom delu imamo događaje kao što su „Odobravanje porudžbine“ i „Provera zaliha“. Nakon toga imamo slov „Zalihe proverene“ koja pokazuje da je provera napravljena. Sledeći uslov „Ima dovoljno zaliha“ izvršava se ako su svi elementi porudžbine kupca dostupni. Taj uslov je izvan svih faza i input je sledećem zadatku „Potvrda porudžbine“ koji se nalazi u fazi 5. „Isporuka porudžbine“. Zadatak „Potvrda porudžbine“ input je zadatku „Izvršenje porudžbine“ koji se izvršava nakon potvrde porudžbine.

Nakon toga je zadatak „Izvršenje porudžbine“ input uslovu „Kraj“ koji označava kraj procesa. Ako se nakon provere zaliha aktivira uslov „Nema dovoljno zaliha“ onda se izvršavaju zadaci u fazi 6. „Odbijanje porudžbine“. Prvi zadatak je „Popunjavanje zaliha“, a nakon njega sledi uslov „Zalihe popunjene“ koji je input događaju „Provera popunjenih zaliha“. To označava da su zalihe dostupne i da se nakon toga može izvršavati proces potvrde i isporuke porudžbine. Zadatak „Provera zaliha nakon popunjavanja“.je drugačiji od drugih jer je to zapravo zadatak odluke, pa je on input za dva zadataka. Ako su zalihe koje su tražene u elementu porudžbine dostupne onda je taj zadatak input zadatku „Potvrda porudžbine“ u fazi 5., a ako zaliha nema onda je taj zadatak input zadatku „Odbijanje porudžbine“ koji se nalazi u fazi 6. Samim tim ako zaliha nema onda se aktivira uslov „Nedovoljno zaliha i zalihe se ne mogu popuniti“, pa zadatak „Odbijanje porudžbine“ postaje input uslovu „Kraj“.

Ovim su objašnjeni svi mogući slučajevi u ovom dijagramu, pa možemo zaključiti da je ovaj dijagram izgledom dosta drugačiji od BPMN dijagrama, a u nastavku je objašnjeno zašto.

CMMN dijagrami rade se na ovakav način zato što su namenjeni za nepredvidive procese koji su često temeljeni na

ljudskim odlukama. Ovakvi dijagrami se koriste za modelovanje slučajeva za koje nije moguće unapred tačno definisati tok aktivnosti. Akcenat je na fazama (*Stage*), zadacima (*Task*), uslovima (*Milestone*) i kriterijumima (*Sentries*) koji su zapravo inputi i outputi određenih zadataka. Što se tiče vizualnih razlika, BPMN dijagrami su linearni i prikazuju jasan sekvencijalni tok, a aktivnosti i događaji su povezani strelicama koje označavaju tok procesa. CMMN dijagrami su više modularni i sadrže faze koje imaju svoje vlastite interne tokove. Umesto povezivanja strelicama koriste se kriterijumi (*sentries*) za definisanje uslova kada se zadaci ili faze mogu aktivirati. Obe notacije imaju svoje specifične primene i mogu se koristiti zajedno za kompleksne poslovne procese. [16]

VII. PREDNOSTI I NEDOSTACI CMMN-A

CMMN je standard koji je uveliko zadovoljio potrebe I zahteve mnoštva korisnika i stručnjaka u području modelovanja. Jedna od velikih prednosti CMMN-a je modelovanje *ad-hoc* i nepredvidivih poslovnih procesa. Ovaj standard je idealan za scenarije u kojima procesi nisu jasno definisani. CMMN takođe omogućuje uvođenje promena u proces u realnom vremenu, a to je jako korisno za prilagođavanje u promjenjivim okolnostima. Ono što je jako bitno napomenuti, to je da ovaj standard ne može zameniti ljude i da su ljudi i dalje najbitniji faktor. CMMN standard nam omogućuje baš to. On uključe korisnika u sam proces donošenja odluka. Time se omogućavaju i biraju koraci na temelju realne situacije. CMMN se može kombinovati I sa drugim standardima, pa u kombinaciji sa njima mogu se dobiti kvalitetniji i celishodniji moeli. Jedna od mana CMMN-a je složenost učenja. Ovaj standard ipak zahteva određeni nivo znanja, razumevanja i stručnosti. Mnoge industrije bazirane su samo na BPMN-u i na strogo definisanim procesima. Zato je jedna od mana CMMN-a to što je manje prikladan većini industrija. CMMN tako ne možemo koristiti u proizvodnim sektorima jer se u njima često koriste samo predefinisani procesi. Za sam kraj treba dodati da je CMMN moćan alat i nudi značajnu fleksibilnost i podršku za prilagodljive procese. Bitno je pokušati pronaći ravnotežu između fleksibilnosti i kontrole kako bi se ostvarila maksimalna korist od primene samog CMMN-a. [17]

VIII. ZAKLJUČAK

CMMN je koristan i idealan u situacijama kada se kompanija mora prilagoditi određenim situacijama. Takođe je bitno naglasiti da je CMMN notacija najbolja za rad sa procesima koji su nepredvidivi i koji se konstantno menjaju. Može se lako integrisati sa BPMN-om i tako se poslovne potrebe mogu u celosti pokriti. [18]

U radu je prikazana mogućnost primene CMMN notacije u modelovanju poslovnih procesa koji se temelje na slučajevima i donošenju odluka, s posebnim fokusom na fleksibilnost i dinamičnost koju ova notacija pruža u odnosu na tradicionalne pristupe. Kroz analizu ključnih komponenti CMMN-a, kao što su slučajevi, aktivnosti, događaji i planovi, utvrđeno je da CMMN omogućava efikasnije upravljanje poslovnim procesima u složenim i nepredvidivim okruženjima, gde se odluke donose na osnovu situacija koje se razvijaju tokom vremena.

Primena CMMN notacije pokazuje jasne prednosti u industrijama kao što su pravo i medicina, uslužni sector i sl., jer omogućava veću fleksibilnost, bolju kontrolu nad dinamičkim procesima i brže donošenje odluka. Obzirom na njenu sposobnost da integriše različite izvore podataka i događaje, CMMN doprinosi optimizaciji resursa i smanjenju rizika u složenim poslovnim slučajevima.

Međutim, implementacija CMMN notacije nije bez izazova. Potrebna je visoka tehnička ekspertiza i prilagođavanje organizacijskih procesa kako bi se u potpunosti iskoristile prednosti ove notacije. Dalja istraživanja u ovoj oblasti mogu doprineti poboljšanju metodologije implementacije i integraciji CMMN-a sa drugim poslovnim standardima i tehnologijama, čime bi se omogućilo još efikasnije upravljanje poslovnim procesima i odlučivanjem u dinamičnim poslovnim okruženjima.

Na kraju, CMMN predstavlja značajan korak ka transformaciji načina na koji organizacije upravljaju složenim slučajevima i procesima, te pruža temelje za buduća istraživanja i unapređenja u oblasti poslovnog procesnog menadžmenta.

LITERATURA

[1] Freund, J., & Rücker, B. (2012). Real-life BPMN. Camunda.

[2] Marin, M. A. (2016). Introduction to the case management model and notation (CMMN). arXiv preprint arXiv:1608.05011.

[3] Kurz, M., Schmidt, W., Fleischmann, A., & Lederer, M. (2015, April). Leveraging CMMN for ACM: examining the applicability of a new OMG standard for adaptive case management. In Proceedings of the 7th international conference on subject-oriented business process management (pp. 1-9).

[4] Routis, I., Nikolaidou, M., & Anagnostopoulos, D. (2018). Using CMMN to model social processes. In Business Process Management Workshops: BPM 2017 International Workshops, Barcelona, Spain, September 10-11, 2017, Revised Papers 15 (pp. 335-347). Springer International Publishing.

[5] Routis, I., Bardaki, C., Dede, G., Nikolaidou, M., Kamalakis, T., & Anagnostopoulos, D. (2021). CMMN evaluation: the modelers' perceptions of the main notation elements. Software and Systems Modeling, 20(6), 2089-2109.

[6] Herzberg, N., Kirchner, K., & Weske, M. (2015). Modeling and monitoring variability in hospital treatments: a scenario using CMMN. In Business Process Management Workshops: BPM 2014 International Workshops, Eindhoven, The Netherlands, September 7-8, 2014, Revised Papers 12 (pp. 3-15). Springer International Publishing.

[7] Zensen, A., & Küster, J. (2018, October). A comparison of flexible BPMN and CMMN in practice: a case study on component release processes. In 2018 IEEE 22nd international enterprise distributed object computing conference (EDOC) (pp. 105-114). IEEE.

[8] Routis, I., Bardaki, C., Nikolaidou, M., Dede, G., & Anagnostopoulos, D. (2023). Exploring CMMN applicability to knowledge-intensive process modeling: An empirical evaluation by modelers. Knowledge and Process Management, 30(1), 33-54.

[9] Wiemuth, M., Junger, D., Leitritz, M. A., Neumann, J., Neumuth, T., & Burgert, O. (2017). Application fields for the new object management group (OMG) standards case management model and notation (CMMN) and decision management notation (DMN) in the perioperative field. International journal of computer assisted radiology and surgery, 12, 1439-1449.

[10] Routis, I., Nikolaidou, M., & Anagnostopoulos, D. (2018). Modeling collaborative processes with CMMN: success or failure? An experience report. In Enterprise, Business-Process and Information Systems Modeling: 19th International Conference, BPMDS 2018, 23rd International Conference, EMMSAD 2018, Held at CAiSE 2018, Tallinn, Estonia, June 11-12, 2018, Proceedings 19 (pp. 199-210). Springer International Publishing.

[11] Routis, I., Nikolaidou, M., & Anagnostopoulos, D. (2020). Empirical evaluation of CMMN models: a collaborative process case study. Software and Systems Modeling, 19, 1395-1413.

[12] Bule, M. K., Polančič, G., Huber, J., & Jošt, G. (2019). Semiotic clarity of case management model and notation (CMMN). Computer Standards & Interfaces, 66, 103354.

[13] Suchenia, A., Kluza, K., Wiśniewski, P., Jobczyk, K., & Ligęza, A. R. (2018). Towards knowledge interoperability between the UML, DMN, BPMN and CMMN models. In MATEC Web of Conferences (Vol. 252).

[14] Niemz, S., Gehrke, S., & Ruhland, J. (2021). On process organization in crisis situations with bpmn, cmmn and dmn. In 37th ibima conference.

[15] https://www.researchgate.net/publication/306258104_Introduction_to_the_Case_Management_Model_and_Notation_CMMN/ , pristupljeno dana 25.01.2025.godine

[16] http://knut.hinkelmann.ch/lectures/bpm2013-14/06_CMMN.pdf, pristupljeno dana 25.01.2025.godine

[17] https://docs.camunda.org/manual/7.8/reference/cmmn11/, pristupljeno dana 25.01.2025.godine

[18] https://www.omg.org/spec/CMMN/, pristupljeno dana 25.01.2025.godine

Application of CMMN notation in dynamic business process modeling and case management

Atanasijević Jordan

ABSTRACT

The basic setting of the work and the issues that are dealt with are important for several reasons, and in the future it can help in many business cases. One of the biggest reasons is that standardization and comprehensibility are achieved by using the proposed model. The notation itself provides the standards that we must adhere to, and with the help of those standards, any process can be analyzed and understood without misunderstandings with other people. Using this notation avoids unnecessary steps in the processes, thus enabling organizations to optimize all processes while reducing costs and increasing efficiency. Notation enables organizations to easily adapt to business environments and to adapt more quickly to new opportunities and challenges. Another reason why this topic is important is the possibility of easier analysis because the processes and decisions are clearly defined. Better reporting and analytics allow us to choose a business strategy more easily.

Интерполација кубним сплајном коришћењем програмског пакета *Wolfram Mathematica*

Јордан Атанасијевић
Центар за примењену математику и електронику, УТИ (Ј-6) ГШ ВС
Београд
jordan.atanasijevic@vs.rs

Дејан Ђукић
Факултет за информационе технологије Алфа БК Универзитет
Београд
dejan.djukic@alfa.edu.rs

Иван Тот
Војна академија Универзитет одбране odbrane
Београд
ivan.tot@va.mod.gov.rs

***Апстракт* - Развој науке и технике, посебно рачунарске технике након другог света рата, условио је бржи и систематичнији развој нумеричке математике, која омогућава решавање веома сложених проблема помоћу рачунара. Наиме, способност рачунара да у реалном времену обави велики број рачунарских операција уз аутоматизовани процес рачунања, пружа неслућене могућности нумеричкој математици. На тај начин низ математичких проблема који се класичним математичким методама не могу увек тачно решити или би њихово решавање било нецелисходно, ефикасно се решавају коришћењем апарата нумеричке математике. Програмски реализовани нумерички методи омогућавају корисницима брзо решавање проблема са произвољном тачношћу, а да при томе не морају бити експерти у области нумеричке математике. Ова околност има позитивно повратно дејство на развој нових технологија и на развој науке уопште.**

Кључне речи – математичка метода, рачунарска техника, рачунарска наука.

I. УВОД

Нумеричко решавање система једначина је област нумеричке математике која се дуго и успешно развија. За већину система једначина немогуће је одредити тачно решење, па се овакве једначине решавају само приближно, применом неког нумеричког поступка.

Многи итеративни поступци за нумеричко решавање система једначина с једном непознатом откривани су више пута, [1]. Сваки пут је дат и доказ конвергенције, па за поједине поступке постоји више различитих доказа конвергенције. То је често и оправдано, пошто су претпоставке под којима се доказује конвергенција различите. Многи од поступака, као што је Њутнов, појављују се и као специјални случајеви неких поступака. У таквим случајевима се и доказ њихове конвергенције добија из доказа конвергенције целе фамилије.

II. ЦИЉ РАДА

Проблем решавања једначина је један од најстаријих проблема у математици. Познато је да у општем случају није могуће изразити корене полинома, степена већег од 4, преко својих коефицијената и операција сабирања, множења и кореновања. Циљ овог рада је да покаже значај програмског пакета *Wolfram Mathematica* у нумеричкој математици. Поред тога, рад треба да прикаже постојање директне везе између математике као теоријске науке и примене рачунара у решавању сложених проблема, брзо, ефикасно и тачно.

III. ИНТЕРПОЛАЦИЈА КУБНИМ СПЛАЈНОМ

Полиномна интерполација високог степена може имати врло лоша својства, па се уместо тога често користи полиномна интерполација по деловима, тј. на сваком подинтервалу важи:

$$g\,|_{x_k,\,x_{k+1}} = P_k,\ k = 0,\ 1,\ \ldots,\ n - 1$$

где су P_k полиноми ниског (али фиксног) степена. За разлику од полиномне интерполације функцијских вредности, где је било довољно да су чворови интерполације међусобно различити, овде подразумевамо да су границе подинтервала интерполације узлазно нумерисане, тј. да важи $a = x_0 < x_1 < \cdots < x_n = b$. То још не осигурава да је g функција јер је могућа двозначност у додирним тачкама подинтервала.

Прецизније, предпостављамо да на сваком подинтервалу $[x_k, x_{k+1}]$ користимо полином P_k степена m који је одређен $s(m + 1)$ – им коефицијентом. Укупно морамо одредити коефицијенте полинома P_k у n подинтервала, тј. укупно $(m + 1)\cdot n$ коефицијената. Интерполацијски услови су

$$g(x_k) = y_k,\ k = 0,\ \ldots,\ n$$

што за сваки полином даје по два услова:

$$P_k(x_k) = y_k,\ P_k(x_{k+1}) = y_{k+1},\ \ k = 0,\ \ldots,\ n - 1,$$

а укупно даје $2n$ услова интерполације. Уочимо да смо постављањем предходних услова интерполације осигурали непрекидност функције g , јер је

$$P_{k-1}(x_k) = y_k,\ P_k(x_k)\ k = 1,\ \ldots,\ n - 1,$$

Приметимо да услова интерполације има $2n$, а морамо наћи $(m + 1)\cdot n$ коефицијената. Без додатних услова то је могуће направити само за $m = 1$, тј. за по деловима линеарну интерполацију.

За $m > 1$ морају се додати услови на глаткоћу интерполације функције g у чворовима интерполације.

A. Кубна интерполација по деловима

Нека су дати подаци $(x_0, y_0),\ (x_1, y_1),\ \ldots,\ (x_n, y_n)$. Поставља се проблем налажења функције g дефинисане на интервалу $[x_0, x_n]$ тако да је $g/[x_k, x_{k+1}] = P_k$, $k = 0,\ \ldots,\ n - 1$, при чему су $P_k,\ k = 0,\ \ldots.,\ n - 1$ полиноми трећег степена. Сада су услови:

$$P_k(x_k) = y_k,\ P_k(x_{k+1}) = y_{k+1},$$
$$P_k'(x_k) = s_k,\ P_k'(x_{k+1}) = s_{k+1}.$$

Приметимо да тако дефинисана функција g има непрекидну прву деривацију за произвољан избор "нагиба" s_k, $k = 0,\ 1,\ \ldots,\ n$ тј. мора бити:

$$P'_{k-1}(x_k) =, P'_k(x_k),\ k = 1,\ \ldots,\ n - 1$$

Из Хермитеовог облика интерполационог полинома, лако се види:

$$P_k(x) = y_k + (x - x_k)f[x_k, x_k] + (x - x_k)^2 f[x_k, x_k, x_{k+1}] + (x - x_k)^2(x - x_{k+1})f[x_k, x_k, x_{k+1}, x_{k+1}]$$

Б. Кубни сплајн

Бројеве $s_0, \ldots, s_n$ можемо одредити на још један начин. Уместо да су реални бројеви s_k нека апроксимација деривације функције f у чворовима, можемо захтевати да се s_k бирају тако да функција g буде још глађа – тј. да јој је и друга деривација непрекидна.

Нагибе $s_1, \ldots, s_{n-1}$, одређујемо из услова непрекидности друге деривације у унутрашњим чворовима $x_1, \ldots, x_{n-1}$. Таква интерполација се зове (кубна) сплајн интерполација. Приметимо да сада имамо $2n + (n - 1) + (n - 1) = 4n - 2$ једначина, при чему имамо n – полинома трећег степена, тј. $4n$ непознатих коефицијената. Приметимо осим тога да у чворовима x_0, x_n немамо захтева да вредност прве и друге деривације полинома P_0 и P_{n-1}. Управо то ће нам касније омогућити да формирамо два задата захтева да полиноме P_0 и P_{n-1}. који ће нам дати две недостајуће једначине.

Захтев на услове лепљења друге деривације у унутрашњим чворовима сада је:

$$P''_{k-1}(x_k) = P''_k(x_k),\ k = 1,\ \ldots,\ n - 1.$$

Полиноме можемо написати и у следећој форми

$$P''_{k-1}(x) = 2c_{3,k-1} + 6c_{4,k-1}(x - x_{k-1}),$$
$$P''_k(x) = 2c_{3,k-1} + 6c_{4,k-1}(x - x_k),$$

па је

$$P''_{k-1}(x_k) = 2c_{3,k-1} + 6c_{4,k-1}(x - x_{k-1}),$$
$$P''_k(x) = 2c_{3,k}.$$

Другим речима, ако предходне једначине поделимо са 2, услов лепљења гласи

$$c_{3,k-1} + 3c_{4,k-1}(x_k - x_{k-1}) = c_{3,k}.$$

Користећи релације 1 и 2 добијамо

$$\frac{f[x_{k-1},x_k] - s_{k-1}}{\Delta x_{k-1}} + 2\frac{s_{k-1} + s_k - 2f[x_{k-1},x_k]}{\Delta x_{k-1}} = \frac{f[x_k,x_{k+1}] - s_k}{\Delta x_k} + \frac{s_k + s_{k+1} - 2f[x_k,x_{k+1}]}{\Delta x_{k-1}},$$

где је $\Delta x_{k-1} = x_k - x_{k-1}$. Сређивањем добијамо

$$\frac{-3f[x_{k-1},x_k] + s_{k-1} + 2s_k}{\Delta x_{k-1}} = \frac{3f[x_k,x_{k+1}] - 2s_k - s_{k+1}}{\Delta x_k}.$$

Уколико предходну једначину помножимо са $\Delta x_{k-1}\Delta x_k$ и ако пребацимо све s_k на леву страну, а чланове који немају s_k на десну страну, за $k = 1, 2, \ldots, n - 1$, добијамо

$$\Delta x_k s_{k-1} + 2(\Delta x_{k-1} + \Delta x_k)s_k + \Delta x_{k-1}s_{k+1} = 3(\Delta x_k f[x_{k-1},x_k] + \Delta x_{k-1}f[x_k,x_{k+1}]).$$

Ово је линеарни систем са $(n+1)$ - непознатих и $(n-1)$ - ом једначином.

У наставку ће бити приказана три начина добијања још додатне две једначине.

1. *Задани нагиби у рубовима, $s_0 = f'(x_0)$, $s_n = f'(x_n)$.*

Како су сада задати, имамо линеарни систем једначина са непознатих. Матрица тако добијеног линеарног система је тродијагонална

$$\begin{vmatrix} 2(\Delta x_0 + \Delta x_1) & \Delta x_0 & & & \\ \Delta x_2 & 2(\Delta x_1 + \Delta x_2) & \Delta x_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \Delta x_{n-2} & 2(\Delta x_{n-3} + \Delta x_{n-2}) & \Delta x_{n-3} \\ & & & \Delta x_{n-1} & 2(\Delta x_{n-2} + \Delta x_{n-1}) \end{vmatrix}$$

и стога дијагонално доминантна по редовима, јер за свако k вреди

$$2(\Delta x_{k-1} + \Delta x_k) > \Delta x_{k-1} + \Delta x_k,$$

па је регуларна. Према томе, овај линеарни састав сигурно има јединствено решење $s_1, \ldots, s_{n-1}$. Овакав састав је сигурно веома погодан за различите итеративне методе решавања система линеарних једначина.

2. *Задани угиби (друга деривација) у рубовима $f''(x_0)$, $s_n = f''(x_n)$.*

Из захтева $f''(x_0) = P''_0(x_0)$ и облика полинома P_0, лако се добија једначина:

$$2s_0 + s_1 = 3f[x_0, x_1] - \frac{1}{2}f''(x_0)\Delta x_0.$$

Аналогно из захтева $f''(x_n) = P''_{n-1}(x_n)$ и облика полинома P_{n-1} лако се добија једначина:

$$s_{n-1} + 2s_n = 3f[x_{n-1}, x_n] - \frac{1}{2}f''(x_n)\Delta x_{n-1}.$$

Матрица тако добијеног линеарног система је

$$\begin{vmatrix} 2 & 1 & & & \\ \Delta x_1 & 2(\Delta x_0 + \Delta x_1) & \Delta x_0 & & \\ & \ddots & \ddots & \ddots & \\ & & \Delta x_{n-2} & 2(\Delta x_{n-3} + \Delta x_{n-2}) & \Delta x_{n-3} \\ & & & \Delta x_{n-1} & 2(\Delta x_{n-2} + \Delta x_{n-1}) \\ & & & 1 & 2 \end{vmatrix}$$

Треба напоменути да је систем формата $(n + 1)\ x\ (n + 1)$, а такође је тродијагоналан и дијагонално доминантан.

3. Захтеви цикличности $f'(x_0) = f'(x_n)$, $f''(x_0) = f''(x_n)$.

Из захтева $f'(x_0) = f'(x_n)$, очито се добије $s_0 = s_n$. Из захтева $f''(x_0) = f''(x_n)$, се добије $P''_0(x_0) = P''_{n-1}(x_0)$ што коришћењем облика полинома P_0 и P_{n-1} даје једначину

$$\Delta x_{n-1}s_1 + \Delta x_0 s_{n-1} + 2(\Delta x_0 + \Delta x_{n-1})s_n = 3(\Delta x_{n-1}f[x_0, x_1] + \Delta x_0 f[x_{n-1}, x_n]),$$

па је матрица облика

$$\begin{vmatrix} 2(\Delta x_0 + \Delta x_1) & \Delta x_0 & & & \Delta x_1 \\ \Delta x_2 & 2(\Delta x_1 + \Delta x_2) & \Delta x_1 & & \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ & \Delta x_{n-2} & 2(\Delta x_{n-3} + \Delta x_{n-2}) & \Delta x_{n-3} & \\ & \Delta x_{n-1} & 2(\Delta x_{n-2} + \Delta x_{n-1}) & \Delta x_{n-2} & \\ \Delta x_{n-1} & & \Delta x_0 & 2(\Delta x_0 + \Delta x_{n-1}) & \end{vmatrix}$$

Уопшено, ако саставе у којима је матрица састава тријагонална напишемо у облику $AT = D$, где је:

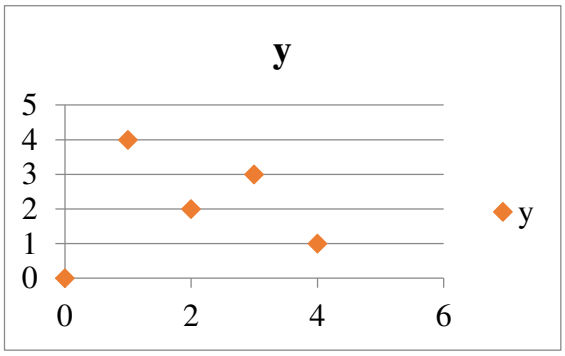
$$A = \begin{vmatrix} B_1 & C_1 & & & \\ A_2 & B_2 & C_2 & & \\ \ddots & \ddots & \ddots & \ddots & \\ & A_{n-1} & B_{n-1} & C_{n-1} & \\ & & A_n & B_n & \end{vmatrix}, \quad T = \begin{vmatrix} T_1 \\ T_2 \\ \vdots \\ T_{n-1} \\ T_n \end{vmatrix}, \quad D = \begin{vmatrix} D_1 \\ D_2 \\ \vdots \\ D_{n-1} \\ D_n \end{vmatrix}$$

а рекурзивна метода решавање је:

$$b_1 = \frac{C_1}{B_1},\ b_i = \frac{C_I}{B_i - A_i b_{i-1}},\ i = 2, 3, \ldots, n - 1$$
$$q_1 = \frac{D_1}{B_1},\ b_i = \frac{D_i - A_i q_{i-1}}{B_i - A_i b_{i-1}},\ i = 2, 3, \ldots, n$$
$$T_n = q_n,\ T_i = q_i - b_i T_{i+1},\ i = n - 1, \ldots, 1.$$

IV. Пример употребе коришћењем програмског пакета Wolfram Mathematica

Израчунавање кубног интерполационог сплајна за функцију дату графиконом:



Графикон 1. Параметри функције

Како пет датих чворова одређују три сегмента, јасно је да ће се резултујући сплајн састојати од четири полинома трећег степена.

Услов да је резултујући сплајн природни, намеће услове на други извод у рубним тачкама посматраних интервала.

Дакле, одредићемо шеснаест непознатих коефицијената за четири полинома трећег степена:

$$S_0 = a_0 + b_0(x-0) + c_0(x-0)^2 + d_0(x-0)^3$$
$$S_1 = a_1 + b_1(x-1) + c_1(x-1)^2 + d_1(x-1)^3$$
$$S_2 = a_2 + b_2(x-2) + c_2(x-2)^2 + d_2(x-2)^3$$
$$S_3 = a_3 + b_3(x-3) + c_3(x-3)^2 + d_3(x-3)^3$$

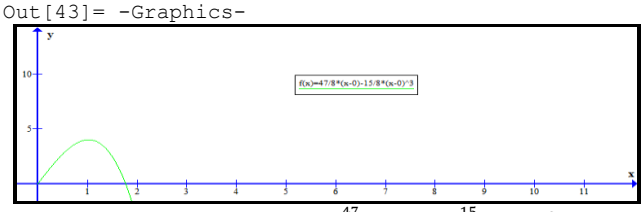
Решење занаведени задатак у програмском језику Wolfram Mathematica, дато је у наставку:

```
In[1]:= data = {{0, 0}, {1, 4}, {2, 2}, {3, 3},
{4, 1}}

Out[1]= {{0, 0}, {1, 4}, {2, 2}, {3, 3}, {4, 1}}

In[2]:= S0[x_] := a0 + b0*(x - 0) + c0*(x - 0)^2 +
d0*(x - 0)^3

In[3]:= S1[x_] := a1 + b1*(x - 1) + c1*(x - 1)^2 +
d1*(x - 1)^3

In[4]:= S2[x_] := a2 + b2*(x - 2) + c2*(x - 2)^2 +
d2*(x - 2)^3

In[5]:= S3[x_] := a3 + b3*(x - 3) + c3*(x - 3)^2 +
d3*(x - 3)^3

In[6]:= data[[1]][[2]] == S0[x] /. x -> 0 (* 1 *)

Out[6]= 0 == a0

In[7]:= data[[2]][[2]] == S0[x] /. x -> 1 (* 2 *)

Out[7]= 4 == a0 + b0 + c0 + d0

In[8]:= data[[2]][[2]] == S1[x] /. x -> 1 (* 3 *)

Out[8]= 4 == a1

In[9]:= data[[3]][[2]] == S1[x] /. x -> 2 (* 4 *)

Out[9]= 2 == a1 + b1 + c1 + d1

In[10]:= data[[3]][[2]] == S2[x] /. x -> 2 (* 5 *)

Out[10]= 2 == a2

In[11]:= data[[4]][[2]] == S2[x] /. x -> 3 (* 6 *)

Out[11]= 3 == a2 + b2 + c2 + d2

In[12]:= data[[4]][[2]] == S3[x] /. x -> 3 (* 7 *)

Out[12]= 3 == a3

In[13]:= data[[5]][[2]] == S3[x] /. x -> 4 (* 8 *)

Out[13]= 1 == a3 + b3 + c3 + d3

In[14]:= PrviIzvodS0[x_] := D[S0[x], {x, 1}]

In[15]:= PrviIzvodS1[x_] := D[S1[x], {x, 1}]

In[16]:= (PrviIzvodS0[x] /. x -> 1) ==
(PrviIzvodS1[x] /. x -> 1)  (* 9 *)

Out[16]= b0 + 2 c0 + 3 d0 == b1

In[17]:= DrugiIzvodS0[x_] := D[S0[x], {x, 2}]

In[18]:= DrugiIzvodS1[x_] := D[S1[x], {x, 2}]

In[19]:= (DrugiIzvodS0[x] /. x -> 1) ==
(DrugiIzvodS1[x] /. x -> 1)  (* 10 *)

Out[19]= 2 c0 + 6 d0 == 2 c1

In[20]:= PrviIzvodS2[x_] := D[S2[x], {x, 1}]

In[21]:= (PrviIzvodS1[x] /. x -> 2) ==
(PrviIzvodS2[x] /. x -> 2)  (* 11 *)

Out[21]= b1 + 2 c1 + 3 d1 == b2

In[22]:= DrugiIzvodS2[x_] := D[S2[x], {x, 2}]

In[23]:= (DrugiIzvodS1[x] /. x -> 2) ==
(DrugiIzvodS2[x] /. x -> 2)  (* 12 *)

Out[23]= 2 c1 + 6 d1 == 2 c2

In[24]:= PrviIzvodS3[x_] := D[S3[x], {x, 1}]

In[25]:= (PrviIzvodS2[x] /. x -> 3) ==
(PrviIzvodS3[x] /. x -> 3)  (* 13 *)

Out[25]= b2 + 2 c2 + 3 d2 == b3

In[26]:= DrugiIzvodS3[x_] := D[S3[x], {x, 2}]

In[27]:= (DrugiIzvodS2[x] /. x -> 3) ==
(DrugiIzvodS3[x] /. x -> 3)  (* 14 *)

Out[27]= 2 c2 + 6 d2 == 2 c3

In[28]:= (DrugiIzvodS0[x] /. x -> 0) == 0 (* 15 *)

Out[28]= 2 c0 == 0

In[29]:= (DrugiIzvodS3[x] /. x -> 4) == 0 (* 16 *)

Out[29]= 2 c3 + 6 d3 == 0

In[30]:= systemEqu = {data[[1]][[2]] == S0[x] /. x
-> 0,
          data[[2]][[2]] == S0[x] /. x -> 1 ,
          data[[2]][[2]] == S1[x] /. x -> 1 ,
          data[[3]][[2]] == S1[x] /. x -> 2 ,
          data[[3]][[2]] == S2[x] /. x -> 2 ,
          data[[4]][[2]] == S2[x] /. x -> 3 ,
```

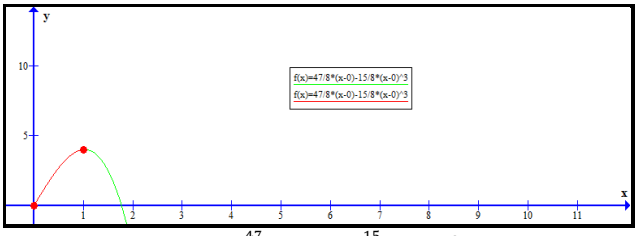
```
          data[[4]][[2]] == S3[x] /. x -> 3 ,
          data[[5]][[2]] == S3[x] /.
           x -> 4 , (PrviIzvodS0[x] /. x -> 1) ==
(PrviIzvodS1[x] /.
             x -> 1), (DrugiIzvodS0[x] /. x -> 1)
== (DrugiIzvodS1[x] /.
             x -> 1), (PrviIzvodS1[x] /. x -> 2)
== (PrviIzvodS2[x] /.
             x -> 2), (DrugiIzvodS1[x] /. x -> 2)
== (DrugiIzvodS2[x] /.
             x -> 2), (PrviIzvodS2[x] /. x -> 3)
== (PrviIzvodS3[x] /.
             x -> 3), (DrugiIzvodS2[x] /. x -> 3)
== (DrugiIzvodS3[x] /.
             x -> 3), (DrugiIzvodS0[x] /. x -> 0)
==
           0, (DrugiIzvodS3[x] /. x -> 4)}

Out[30]= {0 == a0, 4 == a0 + b0 + c0 + d0, 4 ==
a1, 2 == a1 + b1 + c1 + d1, 2 == a2, 3 == a2 + b2
+ c2 + d2, 3 == a3,

>    1 == a3 + b3 + c3 + d3, b0 + 2 c0 + 3 d0 ==
b1, 2 c0 + 6 d0 == 2 c1, b1 + 2 c1 + 3 d1 == b2, 2
c1 + 6 d1 == 2 c2,

>    b2 + 2 c2 + 3 d2 == b3, 2 c2 + 6 d2 == 2 c3,
2 c0 == 0, 2 c3 + 6 d3}

In[31]:= SplineSolutions =
          NSolve[systemEqu, {a0, b0, c0, d0, a1,
b1, c1, d1, a2, b2, c2, d2,
            a3, b3, c3, d3}]

Out[31]= {{a0 -> 0., b0 -> 5.875, c0 -> 0., d0 ->
-1.875, a1 -> 4., b1 -> 0.25, c1 -> -5.625, d1 ->
3.375, a2 -> 2.,

>     b2 -> -0.875, c2 -> 4.5, d2 -> -2.625, a3 ->
3., b3 -> 0.25, c3 -> -3.375, d3 -> 1.125}}

In[32]:= SplineSolutions // Transpose // TableForm

Out[32]//TableForm= a0 -> 0.

                    b0 -> 5.875

                    c0 -> 0.

                    d0 -> -1.875

                    a1 -> 4.

                    b1 -> 0.25

                    c1 -> -5.625

                    d1 -> 3.375

                    a2 -> 2.

                    b2 -> -0.875

                    c2 -> 4.5

                    d2 -> -2.625

                    a3 -> 3.

                    b3 -> 0.25

                    c3 -> -3.375

                    d3 -> 1.125

In[33]:= SplineSolutions // Transpose // TableForm
// Rationalize

Out[33]//TableForm= a0 -> 0

                          47
                    b0 -> --
                          8

                    c0 -> 0

                            15
                    d0 -> -(--)
                            8

                    a1 -> 4

                          1
                    b1 -> -
                          4

                            45
                    c1 -> -(--)
                            8

                          27
                    d1 -> --
                          8

                    a2 -> 2

                            7
                    b2 -> -(-)
                            8

                          9
                    c2 -> -
                          2

                            21
                    d2 -> -(--)
                            8

                    a3 -> 3

                          1
                    b3 -> -
                          4

                            27
                    c3 -> -(--)
                            8

                          9
                    d3 -> -
                          8

In[34]:= {a0, b0, c0, d0, a1, b1, c1, d1, a2, b2,
c2, d2, a3, b3, c3, d3} /.
            SplineSolutions // Rationalize //
Transpose // TableForm

Out[34]//TableForm= 0

                    47
                    --
                    8

                    0

                      15
                    -(--)
                      8

                    4

                    1
                    -
                    4
```

```
                    45
                  -(--)
                    8

                  27
                  --
                  8

                  2

                    7
                  -(-)
                    8

                  9
                  -
                  2

                    21
                  -(--)
                    8

                  3

                  1
                  -
                  4

                    27
                  -(--)
                    8

                  9
                  -
                  8
```

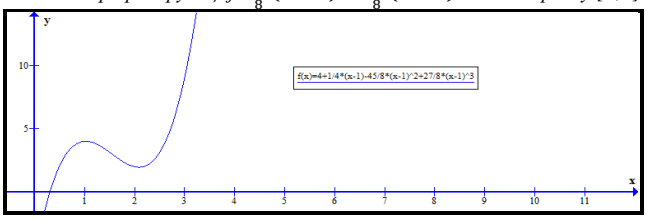
```
In[35]:= S0[x] /. SplineSolutions // Rationalize

                      3
           47 x    15 x
Out[35]= {---- - -----}
           8        8

In[36]:= S1[x] /. SplineSolutions // Rationalize

                                  2              3
              -1 + x    45 (-1 + x)     27 (-1 + x)
Out[36]= {4 + ------ - ------------ + ------------
}
                4            8               8

In[37]:= S2[x] /. SplineSolutions // Rationalize

                                       2
3
              7 (-2 + x)    9 (-2 + x)     21 (-2 +
x)
Out[37]= {2 - ---------- + ----------- - ---------
---}
                 8             2              8

In[38]:= S3[x] /. SplineSolutions // Rationalize

                                  2              3
              -3 + x    27 (-3 + x)     9 (-3 + x)
Out[38]= {3 + ------ - ------------ + -----------}
                4            8               8

In[39]:= SL1 = Plot[
           Piecewise[{{S0[x] /. SplineSolutions,
             0 <= x <= 1}, {S1[x] /.
SplineSolutions,
             1 <= x <= 2}, {S2[x] /.
SplineSolutions,
             2 <= x <= 3}, {S3[x] /.
SplineSolutions, 3 <= x <= 4}}], {x, 0,
           4}, PlotStyle -> {Black, Thick}]

Out[39]= -Graphics-

In[40]:= << Splines`

In[41]:= cub = SplineFit[data, Cubic]

Out[41]= SplineFunction[Cubic, {0., 4.}, <>]

In[42]:= SL2 = ParametricPlot[cub[t], {t, 0, 4},
Epilog -> Point[data],
          AspectRatio -> 1/GoldenRatio]

Out[42]= -Graphics-

In[43]:= Show[SL1, SL2]

Out[43]= -Graphics-
```



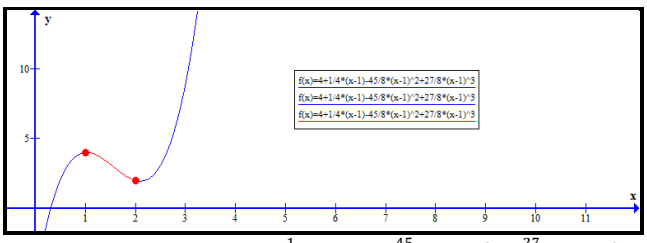
Слика 1. График функције $\frac{47}{8}(x-0)-\frac{15}{8}(x-0)^3$



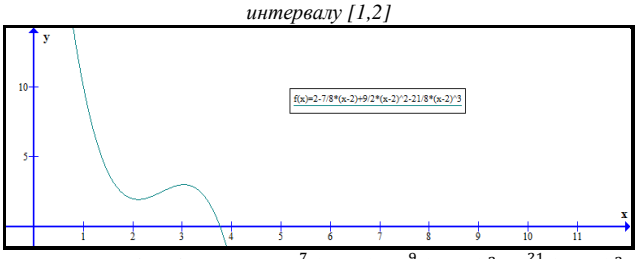
Слика 2. График функције $\frac{47}{8}(x-0)-\frac{15}{8}(x-0)^3$ *на интервалу [0,1]*



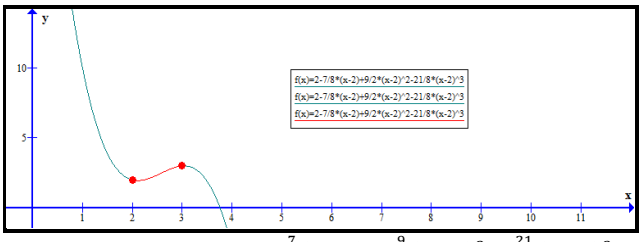
Слика 3. График функције $4+\frac{1}{4}(x-1)-\frac{45}{8}(x-1)^2+\frac{27}{8}(x-1)^3$



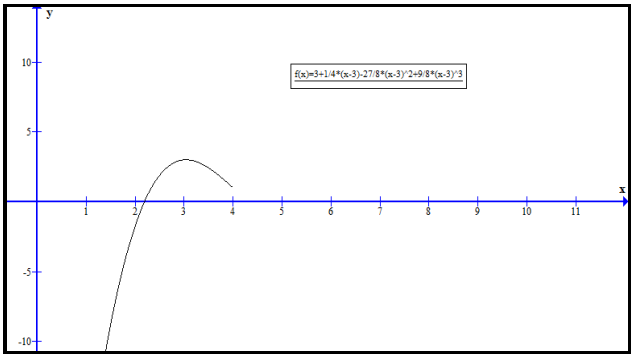
Слика 4. График функције $4+\frac{1}{4}(x-1)-\frac{45}{8}(x-1)^2+\frac{27}{8}(x-1)^3$ *на интервалу [1,2]*



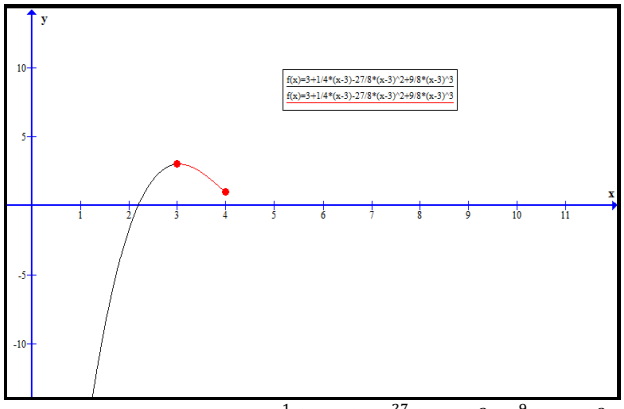
Слика 5. График функције $2-\frac{7}{8}(x-2)+\frac{9}{2}(x-2)^2-\frac{21}{8}(x-2)^3$



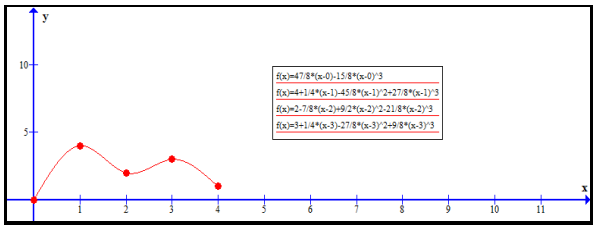
Слика 6. График функције $2-\frac{7}{8}(x-2)+\frac{9}{2}(x-2)^2-\frac{21}{8}(x-2)^3$ *на интервалу [2,3]*



Слика 7. График функције $3+\frac{1}{4}(x-3)-\frac{27}{8}(x-3)^2+\frac{9}{8}(x-3)^3$



Слика 8. График функције $3+\frac{1}{4}(x-3)-\frac{27}{8}(x-3)^2+\frac{9}{8}(x-3)^3$*на интервалу [3,4]*



Слика 9. Приказ комплетног решење за дати опсег у програмском пакету *Wolfram Mathematica* за пример 4

V. ЗАКЉУЧАК

Кроз овај рад смо видели велики значај симболичких и нумеричких могућности програмског пакета *Wolfram Mathematica* у нумеричкој математици.

Практично је показано да се коришћењем наведеног софтвера лако и ефикасно може израчунати вредност сложених параметара, ефикасно, поуздано, брзо и тачно. Како су нелинеарне једначине саставни део сваке озбиљне науке, онда је овај допринос још већи.

ЛИТЕРАТУРА

[1] Dennis, Jr, J. E., & Moré, J. J. (1977). Quasi-Newton methods, motivation and theory. SIAM review, 19(1), 46-89.

[2] Миловановић, М. Ковачевић, М. Спалевић, Нумеричка математика, Збирка решених проблема, Универзитет у Крагујевцу, 2002.

[3] Ortega, J. M., & Rheinboldt, W. C. (1970). Iterative solution of nonlinear equations in several variables (Vol. 30). Siam

[4] Vuković, N., Bulajić M.,Osnove statistike, Beograd: FON, 2014.

Cubic spline interpolation using the Wolfram Mathematica software package

Атанасијевић Јордан

ABSTRACT

The development of science and technology, especially computer technology after the Second World War, has led to a faster and more systematic development of numerical mathematics, which allows solving very complex problems using computers. Namely, the ability of computers to perform a large number of computer operations in real time with an automated calculation process provides unimaginable opportunities for numerical mathematics. In this way, a number of mathematical problems that cannot always be solved accurately by classical mathematical methods or their solution would be impractical are effectively solved using the apparatus of numerical mathematics. Programmatically implemented numerical methods allow users to quickly solve problems with arbitrary accuracy, without having to be experts in the field of numerical mathematics. This circumstance has a positive feedback effect on the development of new technologies and on the development of science in general.

Indeks svih autora konferencije "*YU INFO 2025*":

Autor	Sesija	Broj rada u sesiji	Broj strana u zborniku
Aleksić, Danijela	YU-2	2.07	71 - 74
Andrić Gušavac, Bisera	YU-5	5.02	155 - 159
Aničić, Nenad	YU-5	5.02	155 - 159
Antić, Slobodan	YU-5	5.04	167 - 172
Atanasijevic, Jordan	YU-6	6.07	196 - 200
Atanasijevic, Jordan	YU-6	6.08	201 - 206
Atanasijevic, Jordan	YU-6	6.09	207 - 212
Bogdanović, Natalija	YU-3	3.06	99 - 103
Bulatović, Ana	YU-3	3.06	99 - 103
Cincović, Jelica	YU-3	3.05	93 - 98
Cincović, Jelica	YU-4	4.02	109 - 113
Ćosić, Ksenija	YU-5	5.04	167 - 172
Damnjanovic, Jasmina	YU-3	3.03	84 - 88
Delibašić, Boris	YU-2	2.03	53 - 55
Delić, Vlado	YU-3	3.04	89 - 92
Djurović, Sanja	YU-4	4.01	105 - 108
Đokić, Nataša	YU-5	5.01	151 - 154
Đošić, Danijel	YU-4	4.01	105 - 108
Drašković, Dražen	YU-2	2.05	62 - 65
Drašković, Dražen	YU-3	3.05	93 - 98
Drašković, Dražen	YU-3	3.06	99 - 103
Drašković, Dražen	YU-4	4.03	114 - 119
Đukic, Dejan	YU-6	6.07	196 - 200
Đukic, Dejan	YU-6	6.09	207 - 212
Đukić, Marija	YU-5	5.02	155 - 159
Đukić, Marija	YU-5	5.09	179 - 182
Đurić, Zoran	YU-2	2.04	56 - 61
Gačić, Miodrag	YU-5	5.01	151 - 154
Hrvačević, Luka	YU-4	4.02	109 - 113
Ilić, Aleksa	YU-4	4.03	114 - 119
Ivankovic, Milana	YU-3	3.03	84 - 88
Ivankovic, Zdravko	YU-3	3.03	84 - 88
Jakovljević, Nikša	YU-3	3.01	76 - 79
Janković, Filip	YU-4	4.03	114 - 119
Jaško, Ondrej	YU-1	1.05	26 - 30
Jejić, Olga	YU-5	5.09	179 - 182
Jelović, Marko	YU-3	3.06	99 - 103
Jocović, Vladimir	YU-3	3.05	93 - 98

Jocović, Vladimir	YU-3	3.06	99 - 103
Joksimović, Aleksandar	YU-5	5.03	160 - 166
Jolović, Miloš	YU-5	5.03	160 - 166
Jovanović, Ivan	YU-5	5.09	179 - 182
Jovanović, Milan	YU-1	1.05	26 - 30
Jovanović, Milena	YU-6	6.05	189 - 191
Jovanović, Milena	YU-6	6.06	192 - 195
Kostić, Dušan	YU-5	5.03	160 - 166
Kovačević, Vladimir	YU-1	1.01	15 - 19
Krstanović, Lidija	YU-3	3.01	76 - 79
Lazović, Luka	YU-3	3.06	99 - 103
Lečić-Cvetković, Danica	YU-5	5.02	155 - 159
Lukovac, Petar	YU-5	5.03	160 - 166
Majstorović, Milosav	YU-2	2.06	66 - 70
Majstorović, Vidosav	YU-1	1.06	31 - 34
Mandić, Ana	YU-3	3.06	99 - 103
Marič, Miha	YU-1	1.05	26 - 30
Matvejev, Valerijan	YU-2	2.05	62 - 65
Mićić, Aleksije	YU-2	2.04	56 - 61
Mićović, Marko	YU-3	3.05	93 - 98
Mihajlov, Anja	YU-3	3.06	99 - 103
Mijatović, Hana	YU-3	3.06	99 - 103
Milaković, Adrian	YU-3	3.05	93 - 98
Milaković, Adrian	YU-3	3.06	99 - 103
Milić, Dejan	YU-4	4.01	105 - 108
Milojević, Milan	YU-5	5.01	151 - 154
Mišić, Marko	YU-1	1.01	15 - 19
Mišić, Marko	YU-4	4.04	120 - 125
Mitričević, Nina	YU-6	6.01	184 - 188
Mutavdžić, Uroš	YU-4	4.02	109 - 113
Nastić, Miloš	YU-4	4.04	120 - 125
Naumović, Tamara	YU-5	5.03	160 - 166
Negoicić, Rastko	YU-1	1.06	31 - 34
Nemec, Dejan	YU-4	4.07	132 - 137
Nemec, Dejan	YU-4	4.09	144 - 149
Nikolić, Filip	YU-3	3.06	99 - 103
Nosek, Tijana	YU-3	3.01	76 - 79
Nosek, Tijana	YU-3	3.04	89 - 92
Obradović, Miloš	YU-4	4.05	126 - 131
Ogrizović, Mihajlo	YU-4	4.03	114 - 119
Pantović, Vladan	YU-2	2.06	66 - 70
Pavlović, Dejan	YU-2	2.02	48 - 52
Pavlovic, Rade	YU-6	6.01	184 - 188

Pecev, Predrag	YU-3	3.03	84 - 88
Pekar, Darko	YU-3	3.04	89 - 92
Popović, Isidora	YU-1	1.02	20 - 25
Popović, Lana	YU-3	3.06	99 - 103
Popović, Milena	YU-5	5.02	155 - 159
Potkonjak, Iva	YU-4	4.02	109 - 113
Protić, Jelica	YU-4	4.04	120 - 125
Punt, Marija	YU-4	4.02	109 - 113
Radenković, Uroš	YU-3	3.05	93 - 98
Radojičić, Dragana	YU-3	3.02	80 - 83
Radovanović, Sandro	YU-2	2.03	53 - 55
Ristić, Jadranka	YU-1	1.08	35 - 37
Savić, Gordana	YU-5	5.02	155 - 159
Sečujski, Milan	YU-3	3.01	76 - 79
Simeunović, Vladimir	YU-1	1.06	31 - 34
Simić, Dejan	YU-2	2.01	39 - 47
Simić, Nikola	YU-3	3.04	89 - 92
Škembarević, Milica	YU-5	5.02	155 - 159
Škembarević, Milica	YU-5	5.09	179 - 182
Smilić, Marko	YU-4	4.01	105 - 108
Smiljković, Lazar	YU-1	1.01	15 - 19
Stamenković, Mladen	YU-3	3.02	80 - 83
Stanić, Ana	YU-3	3.06	99 - 103
Stanimirović, Petar	YU-1	1.05	26 - 30
Stanojev, Vuk	YU-3	3.01	76 - 79
Stefanović, Časlav	YU-4	4.01	105 - 108
Stefanović, Katarina	YU-4	4.08	138 - 143
Stojadinović, Dragan	YU-5	5.07	173 - 178
Stošić, Dragan	YU-1	1.06	31 - 34
Sudar, Sasa	YU-3	3.03	84 - 88
Suknović, Milija	YU-2	2.03	53 - 55
Suzić, Siniša	YU-3	3.01	76 - 79
Suzić, Siniša	YU-3	3.04	89 - 92
Terzić, Dušan	YU-2	2.06	66 - 70
Terzić, Rajko	YU-2	2.06	66 - 70
Todorović, Filip	YU-1	1.06	31 - 34
Todorović, Ivan	YU-1	1.05	26 - 30
Tot, Ivan	YU-6	6.07	196 - 200
Tot, Ivan	YU-6	6.08	201 - 206
Tot, Ivan	YU-6	6.09	207 - 212
Tripković, Marina	YU-6	6.05	189 - 191
Tripković, Marina	YU-6	6.06	192 - 195
Tufegdžić, Janko	YU-4	4.02	109 - 113

Tulimirović, Nemanja	YU-5	5.04	167 - 172
Viduka, Dejan	YU-6	6.08	201 - 206
Vučićević, Nikola	YU-3	3.06	99 - 103
Vuletić, Pavle	YU-4	4.04	120 - 125
Vuletić, Pavle	YU-4	4.05	126 - 131
Zečević, Anđelka	YU-1	1.02	20 - 25
Živadinović, Miloš	YU-2	2.01	39 - 47