

UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET



Nikola Stojanović

KLASIFIKACIJA TIPOVA KORONAVIRUSA NA OSNOVU KARAKTERISTIKA PROTEINA PREDSTAVLJENIH STRUKTURNIM ALFABETOM

master rad

Beograd, 2026.

Mentor:

dr Mirjana MALJKOVIĆ RUŽIČIĆ, docent
Univerzitet u Beogradu, Matematički fakultet

Članovi komisije:

prof. dr Jovana KOVAČEVIĆ, vanredni profesor
Univerzitet u Beogradu, Matematički fakultet

dr Saša MALKOV, vanredni profesor
Univerzitet u Beogradu, Matematički fakultet

Datum odbrane: _____

Naslov master rada: Klasifikacija tipova koronavirusa na osnovu karakteristika proteina predstavljenih strukturnim alfabetom

Rezime: Klasifikacija predstavlja važan zadatak u oblasti istraživanja podataka i ima značajnu primenu u bioinformatici, naročito u analizi bioloških sekvenci i proteinskih struktura. Centralna tema ovog rada je klasifikacija tipova koronavirusa na osnovu strukturnih karakteristika spajk-proteina. U radu su proteini predstavljeni pomoću strukturnog alfabeta, odnosno proteinskih blokova, čime je lokalna prostorna organizacija proteinske okosnice prevedena u oblik pogodan za primenu metoda istraživanja podataka.

U radu su najpre izložene biološke i strukturne osnove potrebne za razumevanje problema, kao i osnovni koncepti strukturnih alfabeta sa akcentom na strukturni alfabet Proteinski blokovi. Nakon toga opisan je skup podataka, postupak pripreme proteinskih struktura i izdvajanja karakteristika iz sekvenci proteinskih blokova. Prikazani su korišćeni algoritmi klasifikacije i mere za ocenjivanje njihovih performansi. Na kraju su predstavljeni i analizirani dobijeni rezultati, uz poređenje različitih skupova karakteristika, algoritama i uticaja proširivanja skupa podataka predviđenim proteinskim strukturama.

Ključne reči: koronavirus, klasifikacija, spajk-protein, strukturni alfabet, proteinski blokovi, n-grami

Sadržaj

1	Uvod	1
2	Biološke i strukturne osnove	4
2.1	Proteini i proteinske strukture	4
2.2	Strukturni alfabeti	6
2.3	Proteinski blokovi	7
2.4	Virusi i koronavirusi	10
2.5	Spajk-protein koronavirusa	11
3	Opis skupa podataka i metoda klasifikacije	13
3.1	Skup podataka	13
3.2	Priprema podataka	14
3.3	Problem nebalansiranosti klasa	17
3.4	Algoritmi klasifikacije	18
3.5	Ocenjivanje performansi klasifikatora	31
4	Rezultati i analiza istraživanja	34
4.1	Postavka eksperimenata	34
4.2	Analiza rezultata modela	36
4.3	Analiza rezultata po klasama	45
5	Zaključak	51
6	Dodatne tabele rezultata po klasama	53
	Literatura	72

Glava 1

Uvod

Koronavirusi (eng. *coronaviruses*) predstavljaju veliku grupu virusa. Njihov genetski materijal zapisan je u obliku jednolančane ribonukleinske kiseline (*RNK*). Ime su dobili po karakterističnom izgledu virusne čestice, čiju površinu prekrivaju proteinski izdanci nalik kruni. Različiti tipovi koronavirusa mogu inficirati različite domaćine, uključujući ljude, druge sisare i ptice, a njihova biološka raznovrsnost ogleda se u postojanju više rodova i tipova sa različitim osobinama. Posebnu pažnju naučne javnosti koronavirusi su privukli usled epidemija izazvanih virusima SARS-CoV i MERS-CoV, kao i pandemije bolesti COVID-19 izazvane virusom SARS-CoV-2 [11].

Razlikovanje i klasifikacija tipova koronavirusa imaju značajnu ulogu u istraživanju podataka, bioinformatički, virusologiji i javnom zdravlju. Precizna klasifikacija koronavirusa omogućava bolje razumevanje sličnosti i razlika između pojedinačnih tipova virusa, njihovih potencijalnih domaćina, evolutivnih odnosa i mehanizama infekcije. Sa računarskog stanovišta, ovaj zadatak se može posmatrati kao problem višeklasne klasifikacije, u kome se na osnovu izdvojenih karakteristika bioloških podataka određuje kom tipu koronavirusa pripada posmatrana instanca.

U dosadašnjim istraživanjima klasifikacija virusa često se zasnivala na informacijama dobijenim iz nukleotidnih ili aminokiselinskih sekvenci. Takav pristup je prirodan, jer sekvence nose osnovne informacije o građi i funkciji bioloških molekula. Na primer, na osnovu aminokiselinske sekvence proteina moguće je analizirati obrasce koji se ponavljaju, sličnosti između proteina, kao i razlike koje mogu biti karakteristične za određene tipove virusa.

Međutim, biološka funkcija proteina nije određena samo njegovom primarnom sekvencom, već i načinom na koji se proteinski lanac savija u prostoru. Zbog

toga prostorna struktura proteina može sadržati dodatne informacije koje nisu neposredno vidljive iz same aminokiselinske sekvence.

Posebno važan protein u građi koronavirusa jeste površinski šiljkasti protein (eng. *surface spike protein*) ili kraće spajk-protein. On se nalazi na površini virusne čestice i ima ključnu ulogu u procesu vezivanja virusa za receptor ćelije domaćina i omogućavanju ulaska virusa u ćeliju. Upravo zbog te uloge, spajk-protein je jedan od najčešće proučavanih proteina koronavirusa, kako u eksperimentalnim, tako i u računarskim analizama [20, 16]. Razlike u njegovoj sekvenci i strukturi mogu doprineti razlikama u infektivnosti, tropizmu¹ i evolutivnim osobinama pojedinačnih tipova koronavirusa. Zbog toga je spajk-protein pogodan izbor za ispitivanje mogućnosti klasifikacije tipova koronavirusa na osnovu proteinskih karakteristika.

U ovom radu razmatra se pristup koji se ne zasniva neposredno na nukleotidnim ili aminokiselinskim sekvencama, već na karakteristikama izvedenim iz strukturne reprezentacije spajk-proteina. Da bi se trodimenziona struktura proteina predstavila u obliku pogodnom za računarsku obradu, koristi se koncept *strukturnog alfabeta*. Strukturni alfabet predstavlja konačan skup reprezentativnih lokalnih strukturnih oblika, odnosno prototipova, kojima se mogu aproksimirati delovi proteinske okosnice. Na taj način se složena trodimenziona struktura proteina prevodi u diskretnu sekvencu simbola, slično načinu na koji se primarna struktura proteina predstavlja nizom aminokiselina [21].

Jedan od najpoznatijih strukturnih alfabeta je *proteinski blokovi*. Proteinski blokovi predstavljaju skup lokalnih strukturnih prototipova koji opisuju konformacije kratkih fragmenata, preciznije 5 uzastopnih aminokiselina proteinske okosnice. Svakom fragmentu strukture može se dodeliti odgovarajući proteinski blok, čime se početna proteinska struktura prevodi u sekvencu oznaka proteinskih blokova [12, 18]. Ovakva reprezentacija omogućava da se metode razvijene za obradu sekvenci primene i na proteinske strukture, uz očuvanje informacija o lokalnoj prostornoj organizaciji proteina.

Cilj ovog rada je da se ispita da li karakteristike spajk-proteina predstavljenih pomoću strukturnog alfabeta mogu biti korisne za klasifikaciju tipova koronavirusa. U tu svrhu, proteinske strukture spajk-proteina transformisane su u sekvence proteinskih blokova, a zatim su iz tih sekvenci izdvojene različite grupe karakteristika. Razmatrane su učestalosti pojedinačnih proteinskih blokova, n-grami proteinskih blokova, kao i njihove kombinacije. Nad tako formiranim skupovima

¹Sklonost virusa da inficira specifičnog domaćina, vrstu, organ, tkivo ili tip ćelije.

podataka primenjeni su različiti algoritmi mašinskog učenja, a njihovi rezultati upoređeni su pomoću standardnih mera evaluacije za klasifikacione modele.

Dodatna motivacija za ovaj rad potiče iz potrebe da se uporede različiti načini predstavljanja bioloških podataka u zadatku klasifikacije koronavirusa. Dok se pristupi zasnovani na nukleotidnim i aminokiselinskim sekvencama oslanjaju na primarnu strukturu bioloških molekula, pristup korišćen u ovom radu polazi od lokalne prostorne strukture proteina. Time se ispituje da li strukturna reprezentacija spajk-proteina sadrži dovoljno diskriminativnih informacija za razlikovanje tipova koronavirusa i u kojoj meri se rezultati dobijeni na ovaj način mogu porediti sa rezultatima zasnovanim na sekvencijalnim karakteristikama.

U glavi 2 opisane su biološke i strukturne osnove neophodne za razumevanje problema, sa posebnim osvrtom na koronaviruse, spajk-protein, proteinske strukture, strukturne alfabete i proteinske blokove. U glavi 3 opisani su skup podataka, postupak pripreme podataka, problem nebalansiranosti klasa, korišćeni algoritmi klasifikacije i mere za evaluaciju modela. U glavi 4 prikazani su i analizirani rezultati eksperimenata. U glavi 5 izneti su zaključci rada i mogući pravci daljeg istraživanja, dok su u dodatku prikazane dodatne tabele rezultata po klasama.

Glava 2

Biološke i strukturne osnove

2.1 Proteini i proteinske strukture

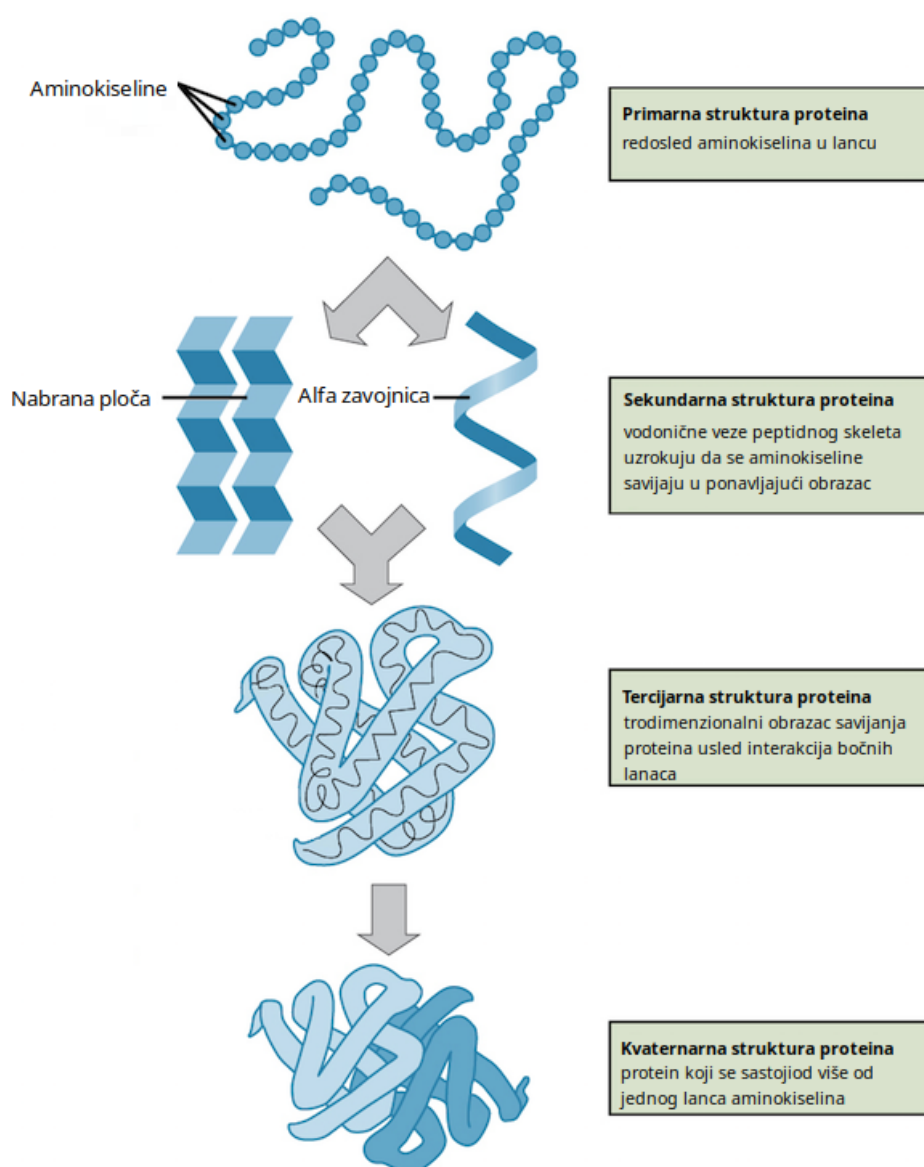
Proteini su biološki makromolekuli koji učestvuju u velikom broju procesa neophodnih za funkcionisanje živih organizama. Neki proteini imaju strukturnu ulogu, neki deluju kao enzimi i ubrzavaju biohemijske reakcije, dok drugi učestvuju u prenosu signala, transportu molekula, imunološkoj odbrani ili regulaciji ćelijskih procesa. Bez obzira na razlike u funkciji, svi proteini građeni su od aminokiselina povezanih peptidnim vezama u polipeptidni lanac [6, 8].

Struktura proteina najčešće se opisuje kroz nekoliko nivoa organizacije. *Primarna struktura proteina* predstavlja redosled aminokiselina u polipeptidnom lancu. Iako je ovaj redosled osnovni zapis proteina, on nije dovoljan da se u potpunosti opiše njegovo ponašanje, jer biološka uloga proteina zavisi i od načina na koji se proteinski lanac organizuje u prostoru.

Pored primarne strukture, razlikuju se i viši nivoi proteinske organizacije. *Sekundarna struktura* opisuje lokalne pravilne oblike u proteinskom lancu, među kojima su najpoznatiji alfa-heliksi (eng. α -*helix*) i beta-nabrani listovi (eng. β -*sheets*). *Tercijarna struktura* odnosi se na ukupan trodimenzioni oblik, kojim je opisan položaj svakog atoma jednog polipeptidnog lanca, dok se o *kvaternarnoj strukturi* govori kada se funkcionalni protein sastoji od više međusobno povezanih proteinskih lanaca [6, 8]. Navedeni nivoi organizacije proteinske strukture prikazani su na slici 2.1.

Za ovaj rad posebno je značajna tercijarna struktura, jer ona opisuje ukupan trodimenzioni oblik jednog polipeptidnog lanca. Tercijarna struktura nastaje kao posledica interakcija između bočnih grupa aminokiselina, među kojima se mogu

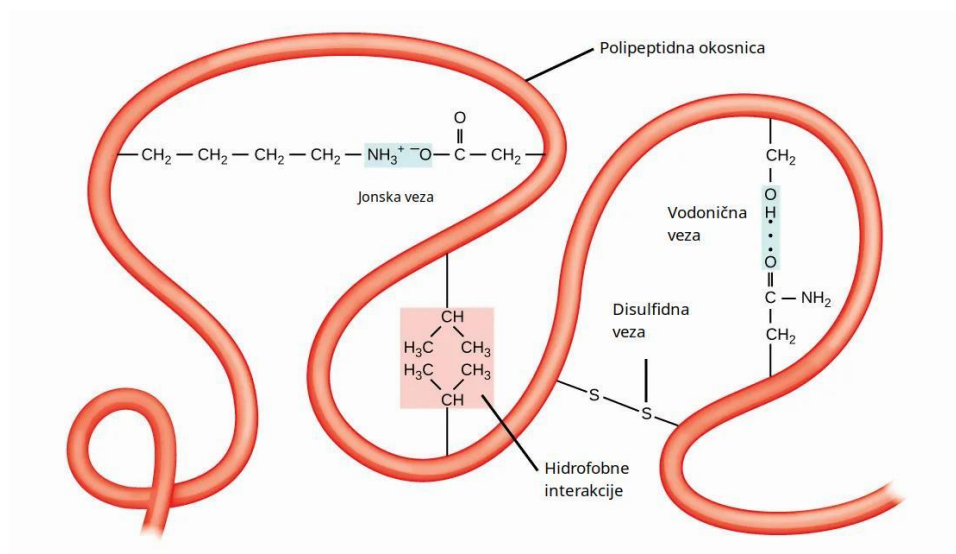
javiti hidrofobne interakcije, vodonične veze, jonske veze i disulfidni mostovi, što je prikazano na slici 2.2. Upravo prostorni raspored delova proteinskog lanca određuje koje će površine proteina biti dostupne za interakciju sa drugim molekulima, pa je zato trodimenziona struktura usko povezana sa funkcijom proteina [6, 8]. U



Slika 2.1: Nivoi organizacije proteinske strukture: primarna, sekundarna, terciarna i kvaternarna struktura. (Preuzeto iz [25])

kontekstu računarske analize proteina, posebno je važno pronaći način da se složena trodimenziona struktura predstavi jednostavnije, ali bez potpunog gubitka

informacija o njenoj prostornoj organizaciji. Jedna mogućnost je da se protein ne posmatra samo kao celina, već i kroz lokalne fragmente njegove okosnice. Takvi fragmenti mogu imati slične oblike u različitim proteinima ili različitim delovima istog proteina. Na toj ideji zasnivaju se strukturni alfabeti, koji omogućavaju da se lokalne konformacije proteinske strukture predstave konačnim skupom simbola [18, 21, 24].



Slika 2.2: Interakcije koje učestvuju u formiranju tercijarne strukture proteina. (Preuzeto iz [8])

2.2 Strukturni alfabeti

Trodimenziona struktura proteina sadrži veliki broj informacija o njegovoj građi i mogućoj funkciji, ali je takav oblik podataka složen za direktnu računarsku obradu. Za razliku od aminokiselinske sekvence, koja je prirodno predstavljena kao niz simbola, prostorna struktura proteina opisana je koordinatama atoma i geometrijskim odnosima između delova proteinskog lanca. Zbog toga se u strukturnoj bioinformatici koriste pojednostavljene reprezentacije, koje nastoje da očuvaju važne informacije o lokalnoj strukturi, ali u obliku pogodnijem za analizu. Jedan od takvih pristupa predstavljaju strukturni alfabeti.

Strukturni alfabet može se posmatrati kao konačan skup reprezentativnih lokalnih strukturnih oblika, odnosno prototipova, kojima se opisuju kratki fragmenti proteinske okosnice. Na taj način se složena trodimenziona struktura proteina

može prevesti u sekvencu diskretnih oznaka. Takva sekvenca je jednostavnija za obradu, jer se nad njom mogu primeniti metode slične onima koje se koriste za analizu običnih bioloških sekvenci, ali pritom ona i dalje nosi informacije o lokalnoj prostornoj organizaciji proteina [18, 21, 24].

Osnovna ideja strukturnog alfabeta jeste da se veliki broj mogućih lokalnih konformacija proteinske okosnice aproksimira manjim brojem tipičnih oblika. Svakom posmatranom fragmentu proteina dodeljuje se oznaka onog prototipa koji mu je geometrijski najbliži. Dobijeni niz oznaka ne predstavlja više samo i neposredno aminokiselinski sastav proteina, već redosled lokalnih strukturnih stanja kroz koja prolazi proteinska okosnica. Zbog toga se ovakva reprezentacija može posmatrati kao veza između punog trodimenzionog opisa proteina i jednostavnijeg sekvencijalnog zapisa.

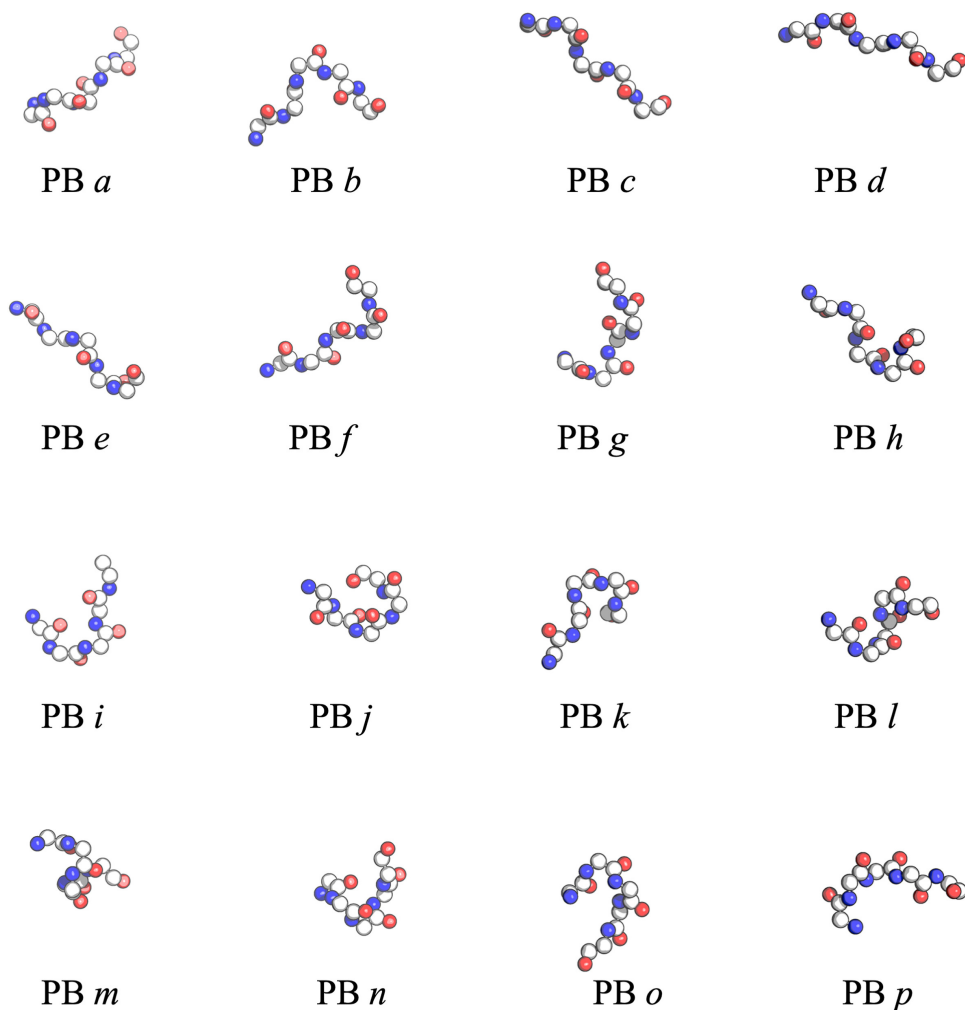
Prednost ovakvog pristupa je u tome što omogućava da se proteinske strukture analiziraju na sistematičan i računarski pogodniji način. Umesto rada sa velikim brojem atomskih koordinata, analiza se može sprovesti nad sekvencom simbola strukturnog alfabeta. Time se olakšava poređenje lokalnih strukturnih obrazaca, izdvajanje karakteristika i primena algoritama mašinskog učenja. U kontekstu ovog rada takav pristup je značajan zato što omogućava da se strukture spajk-proteina koronavirusa predstave skupom numeričkih karakteristika dobijenih iz sekvenci strukturnog alfabeta.

Važno je naglasiti da strukturni alfabet ne zamenjuje potpunu trodimenzionu strukturu proteina, već predstavlja njenu pojednostavljenu reprezentaciju. Pri tom pojednostavljenju deo detalja se gubi, ali se dobija oblik podataka koji je pogodniji za poređenje većeg broja proteina i za primenu metoda klasifikacije. Upravo ta ravnoteža između očuvanja strukturne informacije i jednostavnije reprezentacije čini strukturne alfabete korisnim u bioinformatičkim analizama proteinskih struktura.

2.3 Proteinski blokovi

Jedan od najpoznatijih strukturnih alfabeta jeste alfabet „proteinski blokovi”. Proteinski blokovi predstavljaju skup lokalnih strukturnih prototipova kojima se opisuju fragmenti proteinske okosnice dužine 5 aminokiselina. Ovaj alfabet sastoji se od 16 proteinskih blokova, koji se najčešće označavaju slovima od a do p . Svaki blok opisuje jedan karakterističan lokalni oblik fragmenta proteina, pa se dodeljivanjem proteinskih blokova duž proteinske strukture dobija sekvenca simbola koja

predstavlja lokalnu strukturnu organizaciju proteina [12, 18, 24].



Slika 2.3: Šesnaest proteinskih blokova koji čine PB strukturni alfabet. (Preuzeto iz [24])

Proteinski blokovi definisani su na osnovu lokalnih fragmenata proteinske okosnice. U originalnom pristupu, lokalni strukturni alfabet dobijen je analizom poznatih proteinskih struktura i izdvojen je skup od 16 obrazaca savijanja za *pentapeptidne* fragmente proteinske okosnice, odnosno fragmente od pet uzastopnih aminokiselina [12, 24]. Drugim rečima, umesto da se svaki deo proteina opisuje velikim brojem koordinata i uglova, lokalna konformacija se aproksimira jednim od unapred definisanih proteinskih blokova. Na taj način se trodimenziona struktura proteina prevodi u diskretnu sekvencu oznaka.

Dodeljivanje proteinskih blokova može se razumeti kao postupak lokalne aprok-

simacije strukture. Njihovo računanje zasniva se isključivo na trodimenzionalnom rasporedu atoma proteinske okosnice, odnosno na uglovima koji opisuju konformaciju posmatranog fragmenta. Za svaki posmatrani pentapeptid proteinske okosnice određuje se koji mu od 16 prototipova najbolje odgovara. Oznaka tog prototipa zatim postaje deo PB sekvence. Tako dobijena sekvenca ne opisuje redosled aminokiselina, već redosled lokalnih strukturnih stanja. Zbog toga različite aminokiselinske sekvence mogu deliti slične lokalne strukturne obrasce. Analogno tome, ista aminokiselinska sekvenca može biti povezana sa različitim lokalnim strukturnim stanjima ukoliko je posmatrana u različitim prostornim konformacijama.

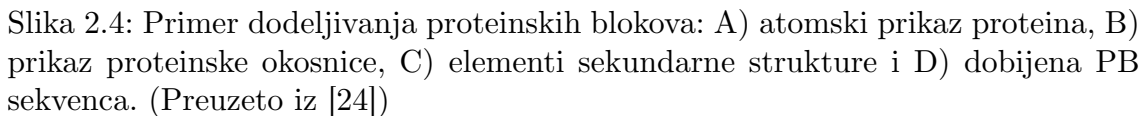
Ovakva reprezentacija je pogodna za računarsku obradu jer strukturu proteina prevodi u oblik sličan tekstualnoj ili biološkoj sekvenci. Iz sekvenci moguće je izdvajati različite numeričke karakteristike, kao što su učestalosti pojedinačnih blokova, učestalosti susednih parova ili trojki blokova, kao i dodatne statističke osobine sekvence. Na taj način proteinski blokovi predstavljaju osnovu za formiranje skupova karakteristika koji se mogu koristiti u algoritmima istraživanja podataka [21, 24].

U ovom radu proteinski blokovi imaju centralnu ulogu, jer se na njima zasniva reprezentacija spajk-proteina koronavirusa. Polazne proteinske strukture najpre se prevode u PB sekvence, a zatim se iz tih sekvenci izdvajaju karakteristike korišćene za klasifikaciju tipova koronavirusa. Time se ne analizira samo aminokiselinski sastav proteina, već i lokalna prostorna organizacija njegove okosnice, predstavljena kroz konačan skup strukturnih prototipova.

Na slici 2.4 prikazano je kako se proteinska struktura može prevesti u sekvencu proteinskih blokova. Polazna struktura proteina se može posmatrati kroz atomski prikaz, zatim kroz proteinsku okosnicu i elemente sekundarne strukture. Nakon toga se svakom lokalnom fragmentu proteinske okosnice dodeljuje odgovarajući proteinski blok, čime se trodimenziona strukturna informacija prevodi u jednodimenzioni zapis, odnosno PB sekvencu.

Oznaka Z^1 predstavlja pomoćnu oznaku u PB sekvenci koja se koristi za pozicije za koje nije moguće dodeliti regularan proteinski blok. Pošto se dodela proteinskog bloka zasniva na pentapeptidnom fragmentu i centrirana je oko trećeg (srednjeg) aminokiselinskog ostatka u pentapeptidnom fragmentu, za prve dve i poslednje dve pozicije u sekvenci ne postoji dovoljan broj susednih ostataka za formiranje pot-

¹Nemogućnost dodele može biti posledica nedostajućih ili neodgovarajućih lokalnih strukturnih informacija.



in i protein ovojnice. Ovi proteini učestvuju u formiranju virusne čestice, zaštiti genetskog materijala, kontaktu sa ćelijom domaćina i procesu infekcije [11, 13].

Na osnovu genetskih i evolutivnih sličnosti, koronavirusi se dele na alfa-, beta-, gama-i deltakoronaviruse [11]. Neki od njih inficiraju ljude, dok su drugi vezani za različite životinjske domaćine, uključujući sisare i ptice. Posebnu pažnju privukli su visokopatogeni koronavirusi SARS-CoV, MERS-CoV i SARS-CoV-2, jer su pokazali da razlike između pojedinačnih tipova koronavirusa mogu imati veliki medicinski i epidemiološki značaj. Zbog toga je njihovo razlikovanje važno ne samo u virusologiji, već i u bioinformatičkim istraživanjima koja pokušavaju da te razlike opišu i računarski iskoriste.

2.5 Spajk-protein koronavirusa

Jedan od najvažnijih strukturnih proteina koronavirusa jeste spajk-protein, često označen i kao S protein. Ovaj protein nalazi se na površini virusne čestice i formira karakteristične izdanke po kojima su koronavirusi i dobili ime. Njegova osnovna uloga ogleda se u prepoznavanju receptora na površini ćelije domaćina i omogućavanju ulaska virusa u ćeliju. Zbog toga spajk-protein ima centralno mesto u proučavanju mehanizama infekcije, ali i u razvoju vakcina, terapijskih pristupa i bioinformatičkih analiza koronavirusa [20, 16].

Spajk-protein je glikoprotein koji se javlja u obliku trimera, što znači da funkcionalnu strukturu čine tri međusobno povezana proteinska lanca. Svaki od tih lanaca sastoji se od dve funkcionalne podjedinice, označene kao S1 i S2. Podjedinica S1 učestvuje u prepoznavanju i vezivanju za receptor ćelije domaćina, dok je podjedinica S2 važna za spajanje virusa i ćelijske membrane, čime se omogućava ulazak virusnog genetskog materijala u ćeliju [16].

Značaj spajk-proteina nije isti kod svih tipova koronavirusa. Razlike u njegovoj sekvenci i strukturi mogu uticati na to koje ćelije i organizme virus može da inficira, koliko se efikasno vezuje za receptore domaćina i na koji način se prenosi. Upravo zato se spajk-protein često koristi kao polazna tačka u istraživanjima koja se bave poređenjem, klasifikacijom i evolucijom koronavirusa. U ovom radu pažnja je usmerena na strukturne karakteristike spajk-proteina, jer one mogu pružiti drugačiji pogled na sličnosti i razlike između tipova koronavirusa u odnosu na pristupe koji koriste samo primarnu sekvencu.

Za razliku od analiza koje polaze isključivo od aminokiselinske ili nukleotidne

sekvence, ovde se posmatra način na koji je spajk-protein organizovan u prostoru. Takav pristup je značajan zato što biološka funkcija proteina zavisi ne samo od redosleda aminokiselina, već i od njegove trodimenzionalne strukture.

Glava 3

Opis skupa podataka i metoda klasifikacije

3.1 Skup podataka

Polazni skup podataka formiran je od eksperimentalno određenih struktura spajk-proteina dostupnih u PDB formatu (eng. *Protein Data Bank format*). Eksperimentalne strukture preuzete su iz baze *RCSB Protein Data Bank* [2]. Usled velikog broja istraživanja sprovedenih tokom i nakon pandemije bolesti COVID-19, eksperimentalne strukture proteina virusa SARS-CoV-2 su dostupne u velikom broju, dok je broj struktura ostalih razmatranih tipova koronavirusa izrazito mali ili čak nepostojeći. Zbog toga je osnovni eksperimentalni skup proširen generisanim skupom podataka pomoću visoko preciznog modela neuronske mreže i alata AlphaFold2, odnosno njegove implementacije ColabFold¹, koji vrši predviđanje 3D proteinske strukture na osnovu sekvenci aminokiselina [19, 22, 26]. Za potrebe generisanja predviđenih struktura korišćene su aminokiselinske FASTA sekvence preuzete iz baze NCBI Protein [1].

Dopunjavanje skupa podataka predviđenim strukturama primenjeno je za klase sa manjim brojem dostupnih eksperimentalnih instanci. Za tipove HCoV-HKU1, HCoV-OC43, MERS-CoV, HCoV-229E i Porcine-EDV generisano je po 50, za tip BatCoV-HKU5 generisano je 27, a za tip SARS-CoV 21 dodatnih struktura, u skladu sa brojem dostupnih sekvenci, dok za tip SARS-CoV-2 nisu generisane, jer je za ovu klasu već postojao veliki broj eksperimentalno određenih struktura. Na broj generisanih struktura uticala su i računarska ograničenja, budući da je

¹Dostupna implementacija kroz Google Colab okruženje.

predikcija proteinskih struktura pomoću AlphaFold2/ColabFold alata vremenski zahtevna. Odabrane sekvence za generisanje predviđenih struktura su provereno bez međusobnih duplikata na nivou aminokiselinskih sekvenci, kao i na nivou dobijenih PB sekvenci. Takođe, nema ni duplikata, odnosno preklapanja sa instancama PB sekvenci eksperimentalnog skupa.

Na ovaj način dobijena su dva skupa za eksperimentalnu analizu: osnovni skup, koji sadrži samo eksperimentalno određene strukture, i prošireni skup, koji pored eksperimentalnih sadrži i predviđene strukture za manje zastupljene klase. Time je omogućeno da se ispita koliko dobro klasifikacioni modeli mogu razlikovati tipove virusa i da li dodati podaci poboljšavaju njihove performanse.

Raspodela instanci po klasama nije ujednačena, pri čemu je klasa SARS-CoV-2 znatno zastupljenija od ostalih klasa. Zbog toga je pri obučavanju i ocenjivanju modela posebna pažnja posvećena problemu nebalansiranosti klasa.

Tabela 3.1: Početni broj instanci po klasama u eksperimentalnom i proširenom skupu podataka

Tip koronavirusa	Eksperimentalni skup	Predviđeni skup	Prošireni skup
SARS-CoV-2	1159	0	1159
HCoV-HKU1	30	50	80
MERS-CoV	20	50	70
HCoV-OC43	15	50	65
HCoV-229E	7	50	57
Porcine-EDV	6	50	56
SARS-CoV	26	21	47
BatCoV-HKU5	6	27	33
Ukupno	1269	298	1567

Tabela 3.1 prikazuje početni skup podataka. Dostupnih eksperimentalnih struktura je 1269, dodato je 298 predviđenih struktura, čime prošireni skup sadrži ukupno 1567 struktura.

3.2 Priprema podataka

Nakon formiranja skupa podataka, izvršena je priprema proteinskih struktura za izdvajanje karakteristika i primenu algoritama mašinskog učenja. Polazni podaci bili su PDB fajlovi koji sadrže atomske koordinate trodimenzionalnih struktura spajk-proteina. Kod eksperimentalno određenih struktura najpre su izdvojeni lanci

koji odgovaraju spajk-proteinu, jer su pojedini PDB zapisi, pored spajk-proteina, sadržali i dodatne molekule ili proteinske komponente.

Pošto očitane strukture sadrže više lanaca spajk-proteina, kako bi se izbeglo dupliranje podataka odabran je samo jedan reprezentativni lanac, jer su lanci međusobno gotovo identični (dodatni lanci mogu posedovati male razlike u očitanim koordinatama, stoga ne bi donosili bitno novu informaciju). Kao reprezentativni lanac biran je najduži dostupni lanac, budući da se na taj način zadržava najpotpunija dostupna strukturna informacija.

Uz pomoć alata PBxplore [4] izdvojene su sekvence proteinskih blokova na osnovu podataka iz PDB datoteka. Ovim postupkom lokalna trodimenziona struktura proteinske okosnice prevodi se u niz simbola nad alfabetom od 16 proteinskih blokova, što je objašnjeno u prethodnoj glavi.

Za svaku obrađenu strukturu sačuvane su osnovne informacije kao što su *identifikator strukture*, *tip koronavirusa*, *oznaka izabranog lanca*, *dužina lanca*, *dužina PB sekvence* i *broj simbola Z*. Ove informacije korišćene su za kontrolu kvaliteta i filtriranje podataka.

U početnom skupu podataka, koji je obuhvatao 1567 struktura, uočene su instance sa identičnim PB sekvencama. Duplikati se javljaju samo u eksperimentalnom delu skupa. Pronađeno je 10 grupa identičnih PB sekvenci, koje su obuhvatale 24 instance od kojih je 14 redundantnih sekvenci uklonjeno. Nakon uklanjanja PB duplikata izvršeno je filtriranje po dužini PB sekvence. Pošto su pojedine eksperimentalne strukture predstavljale kraće fragmente spajk-proteina, kako bi se izbeglo uključivanje izrazito kratkih i nepotpunih struktura u nastavku analize zadržane su samo PB sekvence dužine najmanje 800 simbola. Nakon uklanjanja PB duplikata i filtriranja po dužini, konačan eksperimentalni skup sadržao je 1027 instanci, dok je konačan prošireni skup sadržao 1325 instanci.

Tabela 3.2 prikazuje konačnu raspodelu instanci po klasama, kao i osnovne statistike PB sekvenci nakon uklanjanja tačnih PB duplikata i filtriranja po dužini. Udeo simbola Z izračunat je kao odnos ukupnog broja simbola Z i ukupne dužine svih PB sekvenci u odgovarajućem skupu.

Nakon dobijanja PB sekvenci, iz njih su izdvojene numeričke karakteristike korišćene za obučavanje klasifikacionih modela. Prvu grupu karakteristika čine *učestalosti pojedinačnih proteinskih blokova*. Kako PB alfabet sadrži 16 simbola, ova reprezentacija daje 16 karakteristika po instanci i opisuje zastupljenost pojedinačnih lokalnih strukturnih oblika u sekvenci.

Tabela 3.2: Konačan skup podataka nakon uklanjanja PB duplikata i filtriranja po dužini PB sekvence

Tip koronavirusa	Eksperimentalni skup	Predviđeni skup	Prošireni skup
SARS-CoV-2	939	0	939
HCoV-HKU1	23	50	73
HCoV-OC43	14	50	64
MERS-CoV	11	50	61
HCoV-229E	7	50	57
Porcine-EDV	6	50	56
SARS-CoV	22	21	43
BatCoV-HKU5	5	27	32
Statistike			
Ukupno instanci	1027	298	1325
Prosečna PB dužina	1045.15	1312.48	1105.27
Prosečan broj simbola Z	30.18	5.41	24.61
Udeo simbola Z (%)	2.89	0.41	2.23

Pored učestalosti pojedinačnih blokova, izdvojene su i n-gramske karakteristike. *Bigram* predstavlja uređeni par uzastopnih proteinskih blokova, pa je ukupan broj mogućih bigrama jednak $16^2 = 256$. *Trigram* predstavlja uređenu trojku uzastopnih proteinskih blokova, zbog čega je ukupan broj mogućih trigrama jednak $16^3 = 4096$. Ove karakteristike omogućavaju da se, pored same zastupljenosti pojedinačnih blokova, u obzir uzme i lokalni redosled strukturnih simbola u PB sekvenci.

U radu nisu razmatrani n-grami duži od trigrama, jer bi se povećanjem dužine n-grama broj karakteristika eksponencijalno povećavao. Već bigramske karakteristike uvode znatno bogatiji opis lokalnog redosleda proteinskih blokova, dok trigrama omogućavaju uključivanje nešto dužih lokalnih obrazaca. Zbog toga su u eksperimentima razmatrane učestalosti pojedinačnih blokova, bigrama, trigrama i njihove kombinacije.

Tabela 3.3: Korišćeni skupovi karakteristika

Skup karakteristika	Broj karakteristika
Učestalosti proteinskih blokova	16
Bigrama proteinskih blokova	256
Trigrama proteinskih blokova	4096
Učestalosti i bigrama	272
Učestalosti, bigrama i trigrama	4368

Ovako formirani skupovi karakteristika korišćeni su kao ulaz za klasifikacione

modele. Time je omogućeno da se ispita koliko različiti nivoi opisa PB sekvenci doprinose razlikovanju tipova koronavirusa.

3.3 Problem nebalansiranosti klasa

Kao što je prikazano u tabeli 3.1, klasa SARS-CoV-2 je znatno zastupljenija od ostalih razmatranih tipova koronavirusa. Ovakva raspodela može uticati na obučavanje klasifikacionih modela, jer algoritmi često teže da bolje prepoznaju klase sa većim brojem instanci, dok manje zastupljene klase mogu biti pogrešno klasifikovane. Problem nebalansiranosti klasa posebno je značajan kada su upravo ređe klase važne za analizu i kada ukupna tačnost ne daje potpunu sliku o kvalitetu modela [17].

Postoji više pristupa za ublažavanje problema nebalansiranosti, kao što su smanjenje broja instanci većinske klase ili povećavanje broja instanci manjinskih klasa, dodavanje novih ili veštački generisanih instanci i dodela težina klasama. Smanjivanje većinske klase može dovesti do gubitka korisnih informacija, naročito kada ta klasa sadrži veliki broj raznovrsnih primera. Sa druge strane, jednostavno dupliranje instanci manjinskih klasa može povećati rizik od preprilagođavanja (eng. *overfitting*), jer model više puta vidi iste ili gotovo iste primere.

U ovom radu nisu primenjeni uklanjanje instanci iz većinske klase niti prosto dupliranje instanci manjinskih klasa. Umesto toga, korišćena su dva pristupa za ublažavanje posledica nebalansiranosti.

Prvi pristup predstavlja dopunjavanje osnovnog skupa predviđenim strukturama dobijenim pomoću AlphaFold2/ColabFold alata. Na taj način u radu se dodatno ispituje i važno istraživačko pitanje: da li se strukture spajk-proteina predviđene pomoću AlphaFold2 alata mogu koristiti za dopunjavanje manje zastupljenih klasa i da li takvo proširivanje skupa doprinosi boljoj klasifikaciji tipova koronavirusa. Ovo proširenje ne uklanja u potpunosti nebalansiranost, ali povećava broj instanci u slabije zastupljenim klasama i omogućava analizu uticaja predviđenih struktura na rezultate klasifikacije.

Drugi pristup zasniva se na težinama klasa (eng. *class weights*). Kod algoritama koji podržavaju ovaj mehanizam, manje zastupljenim klasama dodeljuje se veća težina tokom obučavanja, dok se uticaj najzastupljenije klase relativno smanjuje. Na taj način model pri učenju više kažnjava greške nad ređim klasama, bez potrebe da se postojeći podaci uklanjaju ili da se manjinske instance veštački dupliraju.

3.4 Algoritmi klasifikacije

Nakon pripreme podataka, sledeći korak je izbor algoritama klasifikacije. Klasifikacioni algoritmi pripadaju metodama nadgledanog učenja i koriste označene podatke kako bi naučili vezu između ulaznih karakteristika i odgovarajućih klasa. U ovom radu ulazne karakteristike dobijene su iz PB sekvenci spajk-proteina, dok klase predstavljaju razmatrane tipove koronavirusa.

U cilju poređenja različitih pristupa, korišćeno je više algoritama mašinskog učenja. Izabrani algoritmi zasnivaju se na različitim principima: pravilima odlučivanja, prostornoj blizini instanci, probabilističkom modelovanju, neuronskim mrežama i ansambl metodama. Na taj način omogućeno je da se ispita kako se različite vrste klasifikatora ponašaju nad karakteristikama izvedenim iz strukturnog alfa-beta.

Drvo odlučivanja

Drvo odlučivanja (eng. *decision tree*) predstavlja jedan od najintuitivnijih algoritama klasifikacije. Model ima strukturu stabla i odluku donosi nizom uzastopnih provera nad vrednostima ulaznih karakteristika [27, 40]. Osnovni elementi stabla su:

- **koren stabla**, koji predstavlja početni čvor od koga počinje klasifikacija;
- **unutrašnji čvorovi**, u kojima se proverava uslov nad nekom karakteristikom;
- **grane**, koje predstavljaju ishode tih uslova;
- **listovi**, koji predstavljaju konačne odluke, odnosno klase koje se dodeljuju instancama.

Klasifikacija nove instance vrši se prolaskom od korena stabla ka jednom od listova. U svakom unutrašnjem čvoru bira se odgovarajuća grana na osnovu vrednosti posmatrane karakteristike. Kada instanca dođe do lista, dodeljuje joj se klasa povezana sa tim listom.

Tokom obučavanja, stablo se gradi rekursivnim deljenjem skupa podataka. U svakom čvoru bira se karakteristika i granična vrednost podele tako da se dobijeni podskupovi što bolje razdvoje po klasama. Cilj je da čvorovi nastali podelom budu

što homogeniji, odnosno da sadrže instance koje u najvećoj meri pripadaju istoj klasi. Za merenje kvaliteta podele mogu se koristiti različiti kriterijumi nečistoće, kao što su Džinijeva nečistoća (eng. *Gini impurity*), entropija (eng. *entropy*) i logaritamski gubitak (eng. *log loss*) [28].

Važan problem kod stabala odlučivanja jeste mogućnost preprilagođavanja trening podacima. Ako se stablu dozvoli da raste bez ograničenja, ono može formirati veoma specifična pravila koja dobro opisuju trening skup, ali dovode do slabije generalizacije na nove podatke. Zbog toga se u praksi koriste hiperparametri kojima se ograničava složenost stabla. Među najvažnijim hiperparametrima su:

- `max_depth`, koji određuje najveću dozvoljenu dubinu stabla;
- `min_samples_split`, koji određuje najmanji broj instanci potreban da bi se čvor dalje delio;
- `min_samples_leaf`, koji određuje najmanji broj instanci koje moraju ostati u listu;
- `max_leaf_nodes`, kojim se može ograničiti ukupan broj listova;
- `ccp_alpha`, koji se koristi za naknadno potkresivanje stabla;
- `class_weight`, kojim se klasama mogu dodeliti različite težine tokom obučavanja.

Parametar `class_weight` posebno je značajan u ovom radu zbog nebalansirane raspodele klasa. Njegovim korišćenjem greške nad manje zastupljenim klasama mogu imati veći uticaj tokom obučavanja, čime se umanjuje pristrasnost modela ka najzastupljenijoj klasi.

Prednost drveta odlučivanja je njegova interpretabilnost, jer se odluke modela mogu posmatrati kao niz pravila. Sa druge strane, pojedinačno stablo može biti nestabilno, jer male promene u trening podacima mogu dovesti do drugačije strukture stabla. Zbog toga se stabla odlučivanja često koriste i kao osnova za ansambl metode, kao što su nasumične šume i metode pojačavanja.

Metod k najbližih suseda

Metod k najbližih suseda (eng. *k-nearest neighbors*, *KNN*) predstavlja algoritam nadgledanog učenja koji se može koristiti za klasifikaciju i regresiju. U problemu klasifikacije, osnovna pretpostavka ovog metoda jeste da instance koje su

međusobno bliske u prostoru karakteristika često pripadaju istoj klasi. Za razliku od algoritama koji tokom obučavanja eksplicitno grade model, metod najbližih suseda čuva podatke za obučavanje, a klasifikaciju nove instance vrši tek u trenutku predviđanja [10, 35, 40].

Klasifikacija nove instance odvija se kroz nekoliko koraka:

- izračunava se rastojanje između nove instance i instanci iz skupa za obučavanje;
- bira se k najbližih instanci, odnosno k najbližih suseda;
- posmatraju se klase izabranih suseda;
- novoj instanci se najčešće dodeljuje klasa koja je najzastupljenija među izabranim susedima.

Vrednost parametra k ima značajan uticaj na ponašanje klasifikatora. Ako je k malo, model može biti osetljiv na šum i pojedinačne instance, jer se odluka donosi na osnovu malog broja suseda. Sa druge strane, veće vrednosti parametra k daju stabilnije odluke, ali mogu dovesti do grubljih granica odlučivanja i slabijeg razlikovanja lokalnih obrazaca. Zbog toga se vrednost ovog parametra najčešće određuje eksperimentalno, na osnovu performansi modela na validacionom skupu podataka.

Pored broja suseda, važni su i parametri koji određuju način glasanja, meru rastojanja i način pretrage najbližih suseda [30]. Među najvažnijim parametrima su:

- `n_neighbors`, koji određuje broj suseda k korišćenih pri klasifikaciji;
- `weights`, koji određuje da li svi susedi imaju jednak uticaj ili bliži susedi imaju veću težinu;
- `metric`, koji određuje metriku rastojanja između instanci, na primer rastojanje Minkovskog;
- `p`, koji određuje oblik rastojanja;

Parametar `weights` može imati vrednost `uniform`, kada svi izabrani susedi imaju isti uticaj, ili `distance`, kada bliži susedi imaju veći uticaj na konačnu

odluku. Ova druga mogućnost može biti korisna kada se očekuje da veoma bliske instance nose pouzdaniju informaciju o klasi nove instance od udaljenijih suseda.

Izbor mere rastojanja takođe je važan, jer KNN direktno zavisi od pojma bliskosti između instanci. Uobičajeno se koristi rastojanje Minkovskog, pri čemu vrednost $p=1$ odgovara Menhetn rastojanju (eng. *Manhattan distance*), dok vrednost $p=2$ odgovara Euklidskom rastojanju (eng. *Euclidean distance*). U ovom radu instance su predstavljene numeričkim karakteristikama dobijenim iz PB sekvenci, pa izbor mere rastojanja utiče na način na koji se porede učestalosti proteinskih blokova i njihovih n -grama.

Prednost metoda k najbližih suseda je njegova jednostavnost i činjenica da ne uvodi stroge pretpostavke o obliku granica između klasa. Međutim, ovaj algoritam može biti osetljiv na izbor mere rastojanja, vrednost parametra k , skaliranje karakteristika i dimenzionalnost prostora. Kod reprezentacija sa velikim brojem karakteristika, kao što su trigramske reprezentacije PB sekvenci, rastojanja između instanci mogu postati manje informativna, što može uticati na kvalitet klasifikacije. Zbog toga su u eksperimentalnom delu rada vrednost parametra k , kao i ostale važne postavke modela, određivane tokom podešavanja hiperparametara, na osnovu rezultata dobijenih na validacionim delovima unakrsne validacije.

Naivni Bajesov klasifikator

Naivni Bajesov klasifikator (eng. *Naive Bayes classifier*) predstavlja probablistički algoritam nadgledanog učenja koji se zasniva na Bajesovoj teoremi (3.1).

Bajesova teorema može se zapisati u obliku:

$$P(y | X) = \frac{P(y)P(X | y)}{P(X)}, \quad (3.1)$$

gde je y oznaka klase kojoj pripada instanca, a X skup karakteristika instance koja se klasifikuje.

U problemu klasifikacije, za svaku klasu procenjuje se verovatnoća da posmatrana instanca pripada toj klasi na osnovu vrednosti njenih karakteristika. Instanci se zatim dodeljuje klasa za koju je procenjena verovatnoća najveća [34].

Naziv „naivni” potiče od pretpostavke da su karakteristike međusobno uslovno nezavisne kada je poznata klasa. Drugim rečima, pretpostavlja se da vrednost jedne karakteristike ne zavisi od vrednosti drugih karakteristika unutar iste klase. Ova pretpostavka u praksi često nije u potpunosti ispunjena, naročito kod n -gramskih

karakteristika, jer se pojedinačni proteinski blokovi, bigrami i trigrami prirodno mogu pojavljivati u međusobno povezanim obrascima. Ipak, Naivni Bajesov klasifikator se često koristi kao jednostavan i efikasan model, posebno kada su podaci predstavljeni učestalostima ili brojevima pojavljivanja karakteristika.

U zavisnosti od pretpostavki o raspodeli karakteristika, postoje različite varijante Naivnog Bajesovog klasifikatora. Neke od najpoznatijih su Gausova, multinomijalna i Bernulijeva varijanta ovog algoritma. Gausov Naivni Bajes (eng. *Gaussian Naive Bayes*) koristi se kada se karakteristike modeluju normalnom raspodelom, Bernulijev Naivni Bajes (eng. *Bernoulli Naive Bayes*) kada su karakteristike binarne, dok je multinomijalni Naivni Bajes (eng. *Multinomial Naive Bayes*) pogodan za nenegativne karakteristike koje predstavljaju brojanja ili učestalosti [33, 34, 40].

U ovom radu korišćena je multinomijalna varijanta Naivnog Bajesovog klasifikatora, jer su instance predstavljene nenegativnim numeričkim karakteristikama izvedenim iz PB sekvenci. Te karakteristike predstavljaju učestalosti pojedinačnih proteinskih blokova, bigrama i trigrama, što je po obliku slično reprezentacijama koje se koriste u zadacima klasifikacije tekstualnih podataka.

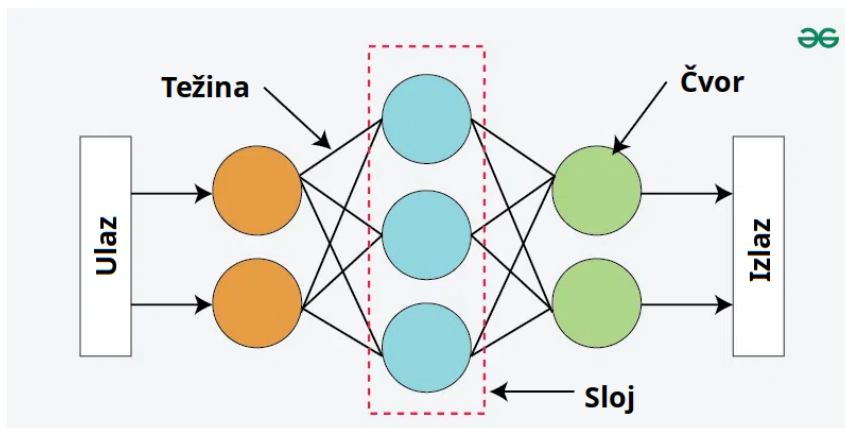
Kod implementacije `MultinomialNB` jedan od najvažnijih parametara je `alpha`, kojim se kontroliše aditivno glačanje (eng. *additive smoothing*). Aditivno glačanje podrazumeva da se pri proceni uslovnih verovatnoća karakteristikama dodaje mala pozitivna konstanta α . Na taj način se sprečava da karakteristika koja se ne pojavljuje u instancama neke klase dobije verovatnoću jednaku nuli, što bi moglo da dovede do toga da cela verovatnoća pripadnosti toj klasi postane nula.

Prednost Naivnog Bajesovog klasifikatora je njegova jednostavnost, brzina obučavanja i dobra primenljivost na podatke sa velikim brojem karakteristika. To ga čini pogodnim za ispitivanje n -gramskih reprezentacija PB sekvenci, posebno u slučaju trigrama, gde je broj mogućih karakteristika velik. Sa druge strane, osnovno ograničenje ovog algoritma je pretpostavka uslovne nezavisnosti karakteristika, koja kod strukturnih n -grama ne mora biti realistična. Zbog toga se Naivni Bajesov klasifikator u ovom radu posmatra kao jednostavan probabilistički model sa kojim se mogu porediti složeniji klasifikatori.

Neuronske mreže

Neuronske mreže (eng. *neural networks*) predstavljaju grupu modela mašinskog učenja koji se sastoje od međusobno povezanih računskih jedinica, odnosno

neurona. Neuroni su organizovani u slojeve, a svaka veza između neurona ima pridruženu težinu koja se podešava tokom obučavanja, kao što se može videti na slici 3.1. Na taj način neuronska mreža uči preslikavanje između ulaznih karakteristika i izlaznih vrednosti, odnosno klasa [23, 36].



Slika 3.1: Opšta struktura neuronske mreže sa ulaznim slojem, skrivenim slojem, težinama veza i izlaznim slojem. (Preuzeto iz [14]).

Neuronske mreže mogu imati različite arhitekture, u zavisnosti od vrste podataka i problema koji se rešava. Među najčešće korišćenim arhitekturama su:

- **potpuno povezane neuronske mreže** (eng. *fully connected neural networks*), kod kojih je svaki neuron jednog sloja povezan sa neuronima narednog sloja;
- **konvolucione neuronske mreže** (eng. *convolutional neural networks, CNN*), koje se često koriste za obradu slika i podataka sa prostornom strukturom;
- **rekurentne neuronske mreže** (eng. *recurrent neural networks, RNN*), koje su namenjene obradi sekvencijalnih podataka;
- **duboke neuronske mreže** (eng. *deep neural networks, DNN*), koje sadrže veći broj slojeva i mogu učiti složenije hijerarhijske reprezentacije podataka.

U ovom radu korišćena je potpuno povezana neuronska mreža, odnosno višeslojni perceptron (eng. *multilayer perceptron, MLP*). Ovakva mreža se sastoji od ulaznog sloja, jednog ili više skrivenih slojeva i izlaznog sloja. Ako mreža sadrži više skrivenih slojeva, može se posmatrati i kao primer duboke neuronske mreže. Ulazni sloj prima vrednosti karakteristika instance, skriveni slojevi uče njihove

nelinearne kombinacije, dok izlazni sloj daje procenu klase kojoj instanca pripada [23, 36].

Svaki neuron računa ponderisanu kombinaciju svojih ulaza, nakon čega se na dobijenu vrednost primenjuje funkcija aktivacije. Uvođenjem nelinearnih funkcija aktivacije mreža može da modeluje složenije odnose među karakteristikama nego linearni modeli. Neke od najčešće korišćenih funkcija aktivacije su:

- **sigmoidna funkcija** (eng. *sigmoid function*) (3.2), koja preslikava vrednosti u interval od 0 do 1. Često se koristi kada je potrebno dobiti izlaz koji se može tumačiti kao verovatnoća, naročito kod binarne klasifikacije;

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \quad (3.2)$$

- **ReLU funkcija** (eng. *rectified linear unit*) (3.3), koja negativne vrednosti preslikava u nulu, dok pozitivne vrednosti ostavlja nepromenjenim. Zbog jednostavnosti i efikasnosti često se koristi u skrivenim slojevima neuronskih mreža;

$$\text{ReLU}(x) = \max(0, x), \quad (3.3)$$

- **softmax funkcija** (eng. *softmax function*) (3.4), koja se najčešće koristi u izlaznom sloju kod višeklasne klasifikacije. Ona daje raspodelu verovatnoća po klasama, pri čemu je zbir izlaznih vrednosti jednak 1.

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}, \quad (3.4)$$

Obučavanje neuronske mreže zasniva se na postepenom podešavanju težina između neurona. Tokom obučavanja mreža najpre izračunava predikciju za date ulazne karakteristike, zatim se izračunava greška između predikcije i tačne klase, a potom se težine ažuriraju tako da se greška smanji. Ovaj postupak se ponavlja kroz više iteracija, sve dok se ne postigne zadovoljavajuće ponašanje modela ili dok se ne ispuni zadati kriterijum za zaustavljanje obučavanja.

Prednost neuronskih mreža je sposobnost modelovanja složenih nelinearnih odnosa među karakteristikama. Sa druge strane, one zahtevaju pažljivo podešavanje hiperparametara i manje su interpretabilne od jednostavnijih modela, kao što su drvo odlučivanja ili logistička regresija. U ovom radu neuronska mreža korišćena je kao jedan od modela za poređenje, kako bi se ispitalo da li složeniji nelinearni pristup doprinosi boljoj klasifikaciji tipova koronavirusa na osnovu PB karakteristika.

Ansambl metode

Ansambl metode (eng. *ensemble methods*) predstavljaju pristupe u kojima se konačna odluka ne donosi na osnovu jednog modela, već kombinovanjem predikcija više pojedinačnih modela. Osnovna ideja ovih metoda jeste da se povezivanjem većeg broja slabijih ili jednostavnijih modela dobije stabilniji i tačniji klasifikator. Pojedinačni modeli mogu praviti različite greške, pa njihovo kombinovanje može smanjiti uticaj grešaka pojedinačnih modela i poboljšati sposobnost generalizacije [29, 40].

Ansambl metode se mogu zasnivati na različitim principima. Dva često korišćena pristupa su:

- **prosta agregacija** (eng. *bootstrap aggregating, bagging*), kod koga se više modela obučava relativno nezavisno, najčešće na različitim podskupovima podataka, a njihove predikcije se zatim kombinuju;
- **pojačavanje** (eng. *boosting*), kod koga se modeli obučavaju redom, tako da svaki naredni model pokušava da ispravi greške prethodnih modela.

Nasumične šume

Nasumične šume (eng. *random forests*) predstavljaju ansambl metodu zasnovanu na većem broju stabala odlučivanja. Umesto jednog stabla, koje može biti osetljivo na promene u podacima i sklono preprilagođavanju, nasumična šuma kombinuje predikcije više stabala i na taj način dobija stabilniji model [5, 37].

Svako stablo u nasumičnoj šumi obučava se na nasumično izabranom podskupu instanci iz skupa za obučavanje. Ovaj izbor se najčešće vrši sa ponavljanjem, što znači da se ista instanca može pojaviti više puta u podskupu za jedno stablo, dok neke instance mogu biti izostavljene. Pored toga, pri izboru najbolje podele u čvorovima stabla ne razmatraju se uvek sve karakteristike, već samo njihov nasumično izabran podskup. Na taj način se povećava raznovrsnost pojedinačnih stabala i smanjuje korelacija između njihovih grešaka.

Klasifikacija nove instance vrši se tako što svako stablo u šumi daje svoju predikciju, a konačna klasa se najčešće određuje većinskim glasanjem. Drugim rečima, instanci se dodeljuje ona klasa za koju je glasao najveći broj stabala. Kod višeklasne klasifikacije, kao u ovom radu, ovaj princip se prirodno proširuje tako što svako stablo bira jednu od mogućih klasa koronavirusa, a ansambl kao celina donosi konačnu odluku.

Najvažniji parametri koji utiču na ponašanje nasumične šume su:

- `n_estimators`, koji određuje broj stabala u šumi;
- `max_depth`, koji ograničava najveću dubinu pojedinačnih stabala;
- `min_samples_leaf`, koji određuje najmanji broj instanci u listu;
- `class_weight`, kojim se mogu zadati težine klasa u slučaju nebalansiranih podataka.

Prednost nasumičnih šuma je u tome što smanjuju varijansu i rizik od preprilagođavanja u odnosu na pojedinačno stablo odlučivanja. Takođe, mogu dobro raditi i kada postoji veliki broj karakteristika, što je značajno u ovom radu zbog n-gramskih reprezentacija PB sekvenci. Sa druge strane, u odnosu na jedno stablo, nasumična šuma je manje interpretabilna, jer se konačna odluka dobija kombinovanjem većeg broja pojedinačnih stabala.

XGBoost

XGBoost (eng. *Extreme Gradient Boosting*) predstavlja ansambl metodu zasnovanu na gradijentnom pojačavanju stabala odlučivanja (eng. *gradient tree boosting*). Za razliku od nasumičnih šuma, gde se stabla grade relativno nezavisno, kod XGBoost algoritma stabla se grade sekvencijalno. Svako novo stablo dodaje se ansamblu sa ciljem da popravi greške koje su napravila prethodna stabla [7, 41].

Osnovna ideja boosting pristupa jeste da se više slabijih modela kombinuje u jedan jači model. U početku se formira jednostavan model, a zatim se kroz naredne korake dodaju nova stabla koja pokušavaju da smanje preostalu grešku. Na taj način model postepeno poboljšava svoje predikcije. Kod XGBoost algoritma ovaj postupak je dodatno optimizovan i uključuje regularizaciju, što pomaže u kontroli složenosti modela i smanjenju rizika od preprilagođavanja.

XGBoost može raditi sa različitim funkcijama cilja, u zavisnosti od vrste problema. Kod višeklasne klasifikacije cilj je da model nauči da razlikuje više mogućih klasa. U ovom radu te klase odgovaraju tipovima koronavirusa, dok ulaz čine karakteristike izdvojene iz PB sekvenci. Pošto XGBoost kombinuje više stabala odlučivanja, on može modelovati nelinearne odnose među karakteristikama i koristiti interakcije između različitih PB n-grama.

Među najvažnijim parametrima XGBoost algoritma su:

- `n_estimators`, koji određuje broj stabala koja se dodaju u ansambl;
- `max_depth`, koji ograničava dubinu pojedinačnih stabala;
- `learning_rate`, koji određuje koliko svako novo stablo utiče na konačni model;
- `subsample`, koji određuje udeo instanci korišćenih pri obučavanju pojedinačnih stabala;
- `objective`, koji određuje funkciju cilja koja se optimizuje;
- `eval_metric`, koji određuje meru korišćenu za praćenje performansi modela.

Parametar `learning_rate` ima važnu ulogu u kontroli procesa učenja. Manja vrednost ovog parametra znači da svako novo stablo ima manji pojedinačni uticaj, pa je često potreban veći broj stabala, ali model može učiti stabilnije. Sa druge strane, veća vrednost može ubrzati učenje, ali i povećati rizik od preprilagođavanja. Parametar `subsample` uvodi dodatnu nasumičnost, što može doprineti boljoj generalizaciji. Slično kao kod neuronskih mreža, kod višeklasne klasifikacije, kao funkcija cilja, najčešće se koristi funkcija `multi:softmax` koja je pandan *softmax* funkciji.

Prednost XGBoost algoritma je u tome što često postiže dobre performanse na strukturiranim tabelarnim podacima i može uspešno modelovati složene nelinearne odnose. Zbog toga je pogodan za poređenje sa drugim modelima korišćenim u ovom radu. Međutim, XGBoost ima veći broj hiperparametara i zahteva pažljivo podešavanje. U slučaju nebalansiranih skupova podataka, njegove performanse mogu zavisi od načina podešavanja modela i od mera koje se koriste za evaluaciju.

Logistička regresija

Logistička regresija (eng. *logistic regression*) predstavlja linearni model koji se često koristi za zadatke klasifikacije. Iako naziv sadrži reč regresija, u ovom kontekstu model se koristi za predviđanje klase instance, odnosno za procenu verovatnoće pripadnosti određenoj klasi. Osnovna ideja je da se najpre formira linearna kombinacija ulaznih karakteristika, a zatim se ona transformiše u vrednost koja se može tumačiti kao verovatnoća [15, 31, 32].

U binarnoj klasifikaciji, za instancu opisanu vektorom karakteristika X , model najpre računa linearnu kombinaciju:

$$z = w^T X + b, \quad (3.5)$$

gde je w vektor težina, a b slobodan član. Zatim se ova vrednost transformiše pomoću sigmoidne funkcije:

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-z}}. \quad (3.6)$$

Dobijena vrednost nalazi se u intervalu od 0 do 1 i može se tumačiti kao procenjena verovatnoća pripadnosti pozitivnoj klasi.

Kod višeklasne klasifikacije, logistička regresija se proširuje tako da može razlikovati više klasa. To se može postići obučavanjem više binarnih klasifikatora ili direktnim modelovanjem verovatnoća za sve klase.

Važna osobina logističke regresije je interpretabilnost. Pošto model uči težine uz svaku karakteristiku, moguće je posmatrati na koji način pojedine karakteristike utiču na odluku modela. Sa druge strane, logistička regresija je linearni model, pa ne može direktno modelovati složene nelinearne odnose među karakteristikama, osim ako su takvi odnosi već predstavljeni kroz ulazne karakteristike. U ovom radu to je delimično omogućeno korišćenjem bigrama i trigramata proteinskih blokova, koji uvode informacije o lokalnom redosledu PB simbola.

Među najvažnijim parametrima algoritma su:

- `C`, koji određuje jačinu regularizacije², pri čemu manje vrednosti označavaju jaču regularizaciju;
- `penalty`, koji određuje tip regularizacije;
- `solver`, koji određuje algoritam optimizacije;
- `max_iter`, koji određuje maksimalan broj iteracija optimizacije;
- `class_weight`, kojim se klasama mogu dodeliti različite težine tokom obučavanja.

Parametar `penalty` određuje tip regularizacije koji se primenjuje nad koeficijentima modela. Najčešće se koriste L_1 i L_2 regularizacija. L_2 regularizacija kažnjava

²Regularizacija je tehnika za ograničavanje preprilagođavanja modela podacima koja podrazumeva dodavanje određenih kaznenih članova funkciji greške

velike vrednosti koeficijenata i teži ka njihovom smanjivanju, dok L_1 regularizacija može dovesti do toga da neki koeficijenti postanu jednaki nuli, čime se dobija ređi model. Pored toga, moguće je koristiti i kombinaciju ova dva pristupa, poznatu kao elastična mreža (eng. *elastic net*). Parametar *solver* određuje algoritam optimizacije koji se koristi za pronalaženje koeficijenata modela, međutim različiti algoritmi ne podržavaju uvek iste tipove regularizacije, pa izbor parametara *penalty* i *solver* mora biti međusobno usklađen [32].

Metod potpornih vektora

Metod potpornih vektora (eng. *support vector machine, SVM*) je algoritam koji se koristi za klasifikaciju i regresiju. Osnovna ideja ovog metoda je pronalaženje razdvajajuće hiperravni koja što bolje razdvaja instance različitih klasa. Kada postoji više mogućih hiperravni koje razdvajaju klase, SVM bira onu za koju je margina između klasa najveća [9, 38].

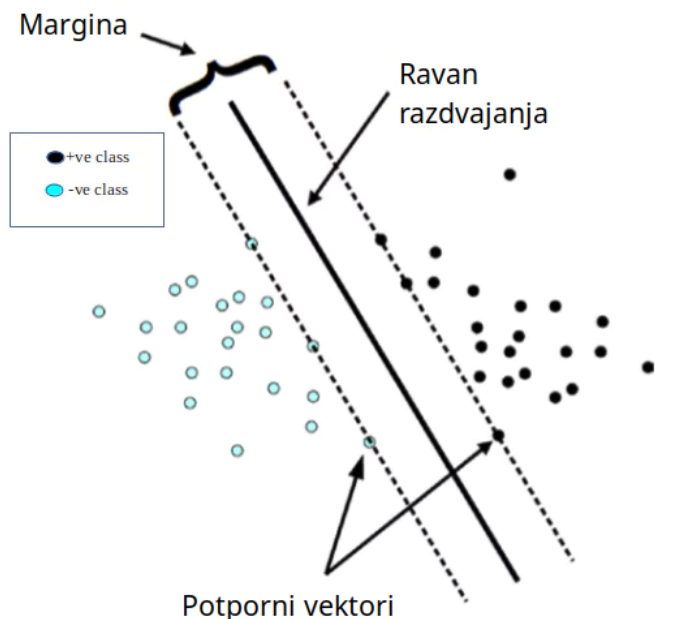
Margina predstavlja rastojanje između razdvajajuće hiperravni i najbližih instanci iz različitih klasa. Instance koje se nalaze najbliže granici odlučivanja nazivaju se potporni vektori (eng. *support vectors*), jer upravo one najviše utiču na položaj razdvajajuće hiperravni. Zbog toga SVM ne zavisi jednako od svih instanci iz skupa za obučavanje, već je granica odlučivanja određena najvažnijim instancama koje leže blizu granice između klasa.

U najjednostavnijem slučaju koristi se linearni SVM, kod koga je granica odlučivanja linearna. Međutim, podaci često nisu linearno razdvojivi u originalnom prostoru karakteristika. U takvim slučajevima koristi se ideja jezgra (eng. *kernel*), kojom se implicitno omogućava razdvajanje podataka u prostoru veće dimenzionalnosti, bez eksplicitnog računanja koordinata u tom prostoru. Na taj način SVM može modelovati i nelinearne granice odlučivanja.

U ovom radu korišćene su dve varijante metoda potpornih vektora: linearni SVM i SVM sa RBF jezgrom (eng. *radial basis function kernel*). Linearni SVM koristi linearnu granicu odlučivanja i pogodan je za ispitivanje da li se klase mogu razdvojiti jednostavnijim linearnim odnosima među karakteristikama. RBF kernel omogućava formiranje nelinearnih granica odlučivanja. Razlika između linearne i nelinearne granice odlučivanja prikazana je na slici 3.3.

Neki od najvažnijih hiperparametara SVM algoritma su:

- C , koji određuje jačinu kažnjavanja pogrešnih klasifikacija;



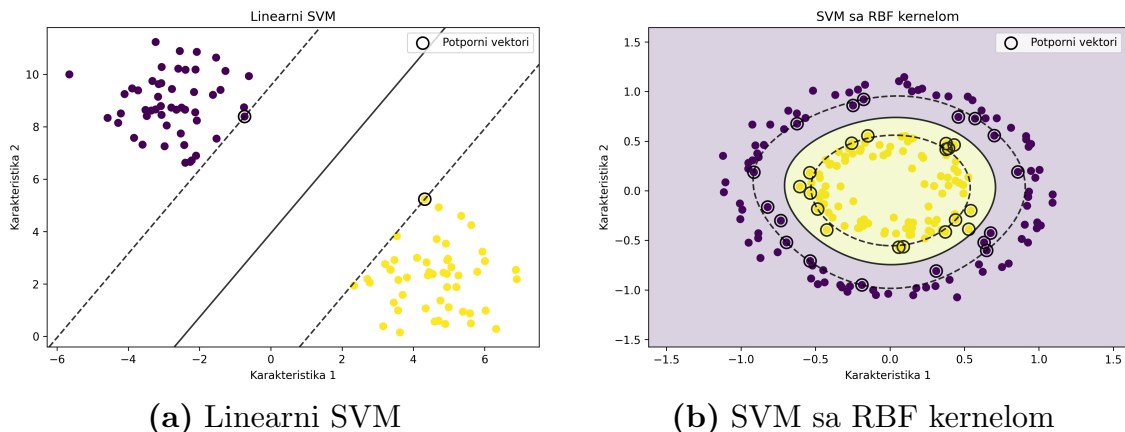
Slika 3.2: Ilustracija razdvajajuće hiperravni, margine i potpornih vektora kod SVM algoritma. (Preuzeto iz [3])

- `kernel`, koji određuje tip jezgra, na primer linearno ili RBF jezgro;
- `gamma`, koji utiče na oblik granice odlučivanja kod nelinearnih kernela;
- `class_weight`, kojim se klasama mogu dodeliti različite težine tokom obučavanja.

Parametar `C` kontroliše odnos između širine margine i grešaka klasifikacije na skupu za obučavanje. Manja vrednost parametra `C` dozvoljava širu marginu i veći broj grešaka na trening podacima, što može doprineti boljoj generalizaciji. Veća vrednost parametra `C` više kažnjava pogrešne klasifikacije i može dovesti do složenije granice odlučivanja, ali i do većeg rizika od preprilagodavanja.

Kod SVM-a sa RBF kernelom značajan je i parametar `gamma`, koji određuje koliko pojedinačne instance utiču na oblik granice odlučivanja. Veće vrednosti parametra `gamma` znače da pojedinačne instance imaju lokalniji uticaj, što može dovesti do složenijih granica odlučivanja. Manje vrednosti daju glade granice, ali mogu dovesti do modela koji nije dovoljno fleksibilan.

Prednost SVM algoritma je u tome što može dobro raditi u prostorima sa velikim brojem karakteristika, što je važno za n-gramske reprezentacije PB sekvenci.



Slika 3.3: Poređenje granica odlučivanja kod linearnog SVM-a i SVM-a sa RBF kernelom na sintetički generisanim podacima.

Linearna varijanta omogućava poređenje sa drugim jednostavnijim modelima, dok RBF varijanta omogućava ispitivanje nelinearnih odnosa među karakteristikama. Sa druge strane, SVM može biti osetljiv na izbor parametara C i γ , pa je njihovo podešavanje važno za dobijanje dobrih rezultata.

3.5 Ocenjivanje performansi klasifikatora

Pre ocenjivanja performansi klasifikacionih modela potrebno je definisati način podele podataka, pripreme karakteristika i izbora hiperparametara.

Podela skupa podataka podrazumeva razdvajanje dostupnih instanci na delove koji imaju različitu ulogu u postupku obučavanja i ocenjivanja modela. *Skup za obučavanje* koristi se za učenje parametara modela, *validacioni skup* za izbor hiperparametara i poređenje različitih podešavanja, dok se *test skup* koristi za konačnu procenu performansi nad podacima koji nisu korišćeni tokom obučavanja [40].

Priprema karakteristika često uključuje skaliranje numeričkih vrednosti. Kod algoritama koji zavise od rastojanja, veličine koeficijenata ili gradijentne optimizacije, različite skale karakteristika mogu uticati na ponašanje modela. Zbog toga se često koristi *standardizacija*, kojom se karakteristike transformišu tako da imaju srednju vrednost nula i standardnu devijaciju jedan. U nekim slučajevima koristi se i preslikavanje vrednosti u zadati opseg, na primer kada algoritam zahteva nenegativne ulazne vrednosti.

Hiperparametri su vrednosti koje se ne uče direktno iz podataka, već se zadaju pre obučavanja modela. Primeri hiperparametara su broj suseda kod metoda k

najbližih suseda, dubina stabla odlučivanja, jačina regularizacije kod logističke regresije i SVM algoritma, kao i broj stabala u ansambl metodama. Jedan od načina njihovog izbora je *mrežna pretraga* (eng. *grid search*), pri kojoj se ispituju unapred zadate kombinacije vrednosti hiperparametara.

Unakrsna validacija (eng. *cross-validation*) koristi se za stabilnije ocenjivanje modela. Kod k -struke unakrsne validacije skup podataka se deli na k delova, pri čemu se model obučava više puta, tako da se svaki deo jednom koristi za validaciju, a preostali delovi za obučavanje. Kod klasifikacije nad nebalansiranim skupovima podataka posebno je korisna *stratifikovana unakrsna validacija* (eng. *stratified cross-validation*), kod koje se u svakoj podeli približno čuva raspodela klasa iz celog skupa podataka.

Najzad, potrebno je oceniti kvalitet predikcija obučenih klasifikacionih modela. Za opis rezultata često se koriste oznake TP, TN, FP i FN. One se definišu u odnosu na jednu posmatranu klasu, pri čemu se ta klasa posmatra kao pozitivna, a sve ostale klase kao negativne:

- **stvarno pozitivni** (eng. *true positive, TP*) – broj instanci koje pripadaju posmatranoj klasi i model ih je ispravno klasifikovao u tu klasu;
- **lažno pozitivni** (eng. *false positive, FP*) – broj instanci koje ne pripadaju posmatranoj klasi, ali ih je model pogrešno klasifikovao u tu klasu;
- **stvarno negativni** (eng. *true negative, TN*) – broj instanci koje ne pripadaju posmatranoj klasi i model ih nije klasifikovao u tu klasu;
- **lažno negativni** (eng. *false negative, FN*) – broj instanci koje pripadaju posmatranoj klasi, ali ih je model pogrešno klasifikovao u neku drugu klasu.

Na osnovu ovih vrednosti mogu se definisati osnovne metrike za ocenjivanje klasifikatora. Tačnost (eng. *accuracy*) predstavlja udeo tačno klasifikovanih instanci u ukupnom broju instanci:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (3.7)$$

Preciznost (eng. *precision*) pokazuje koliki udeo instanci koje je model klasifikovao u posmatranu klasu zaista pripada toj klasi:

$$\text{precision} = \frac{TP}{TP + FP}. \quad (3.8)$$

Odziv (eng. *recall*) pokazuje koliki udeo instanci koje stvarno pripadaju posmatranoj klasi model uspeva da prepozna:

$$\text{recall} = \frac{TP}{TP + FN}. \quad (3.9)$$

F1-mera (eng. *F1-score*) predstavlja harmonijsku sredinu preciznosti i odziva:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \quad (3.10)$$

U višeklasnoj klasifikaciji preciznost, odziv i F1-mera mogu se računati za svaku klasu posebno, a zatim objediniti pomoću različitih proseka. Najčešće korišćeni su:

- *Makro prosek* - računa se kao obična aritmetička sredina vrednosti po klasama;
- *Težinski prosek* - uzima u obzir broj instanci u svakoj klasi.

Glava 4

Rezultati i analiza istraživanja

4.1 Postavka eksperimenata

Eksperimenti su sprovedeni nad dva skupa podataka opisana u prethodnoj glavi. Prvi skup čine eksperimentalno određene strukture spajk-proteina, dok drugi predstavlja njegovu proširenu varijantu, dopunjenu predviđenim strukturama, alatom AlphaFold2/ColabFold za manje zastupljene klase. Iz početnog celokupnog skupa od 1567 PB sekvenci, uklonjeni su duplikati i sekvence kraće od 800 simbola, kako bi se izbeglo uključivanje kratkih, nepotpunih i dupliranih struktura. Nakon toga je eksperimentalni skup sadržao 1027 instanci, dok je prošireni skup sadržao 1325 instanci.

Za svaki od ova dva skupa razmatrano je pet skupova karakteristika: učestalosti pojedinačnih proteinskih blokova, bigrami, trigrami, kombinacija učestalosti i bigrama, kao i kombinacija učestalosti, bigrama i trigrama. Na ovaj način ispitano je da li se bolji rezultati dobijaju na osnovu jednostavnijih karakteristika, koje opisuju samo zastupljenost pojedinačnih blokova, ili na osnovu n-gramskih karakteristika, koje uključuju i lokalni redosled PB simbola.

Za klasifikaciju je korišćeno devet modela. Modeli su implementirani pomoću biblioteke `scikit-learn`, dok je za XGBoost korišćena biblioteka `xgboost`. Korišćene klase i najvažniji parametri prikazani su u tabeli 4.1. Parametri su odabrani na osnovu prethodno sprovedene mrežne pretrage, a u tabeli su prikazane vrednosti koje su korišćene u konačnoj evaluaciji modela.¹

¹Izvorni kod korišćen za pripremu podataka, izdvajanje PB sekvenci, formiranje skupova karakteristika i evaluaciju modela dostupan je na adresi: https://github.com/nikollaa23/master_theses. Pristup se može dobiti od autora na zahtev.

Tabela 4.1: Klasifikacioni modeli i najvažniji parametri korišćeni u eksperimentima

Algoritam	Oznaka	scikit-learn klasa / implementacija	Parametri
Logistička regresija	LR	LogisticRegression	C=10, max_iter=5000, class_weight=balanced
Linearni SVM	LinSVM	LinearSVC	C=0.01, class_weight=balanced, max_iter=20000
SVM sa RBF jezgrom	RBF	SVC	kernel='rbf', C=1, gamma=0.001, class_weight=balanced
Drvo odlučivanja	DTC	DecisionTreeClassifier	max_depth=5, min_samples_leaf=5, class_weight=balanced
Nasumične šume	RF	RandomForestClassifier	n_estimators=200, max_depth=None, min_samples_leaf=5, class_weight=balanced_subsample
K najbližih suseda	KNN	KNeighborsClassifier	n_neighbors=3, weights='distance', metric='manhattan'
Naivni Bajes	NB	MultinomialNB	alpha=0.1
Neuronska mreža	NN	MLPClassifier	hidden_layer_sizes=(128,), activation='relu', alpha=0.0001, learning_rate_init=0.001, max_iter=2000
XGBoost	XGB	XGBClassifier	objective='multi:softprob', eval_metric='mlogloss', learning_rate=0.1, max_depth=3, n_estimators=300, subsample=0.8

Kod modela koji zahtevaju skaliranje karakteristika korišćena je standardizacija podataka. To se odnosi na logističku regresiju, SVM modele, metod k najbližih suseda i neuronsku mrežu. Kod Naivnog Bajesovog klasifikatora nije korišćena standardizacija, jer ona može proizvesti negativne vrednosti, dok multinomijalna varijanta ovog algoritma zahteva nenegativne ulazne vrednosti. Umesto toga, za ovaj model korišćeno je preslikavanje karakteristika u nenegativan opseg. Iako su učestalosti PB simbola i njihovih n -grama već nenegativne po svojoj prirodi, ovaj korak samo obezbeđuje da obrada podataka za Naivni Bajesov klasifikator ostane usklađena sa pretpostavkama korišćene implementacije.

Za ocenjivanje modela korišćena je ponovljena stratifikovana unakrsna validacija sa pet podela i deset ponavljanja. U svakoj podeli jedan deo podataka koristi se za validaciju, dok se preostali delovi koriste za obučavanje modela. Postupak se zatim ponavlja više puta, čime se dobija stabilnija procena performansi modela nego pri jednoj podeli na trening i test skup.

Stratifikacija je korišćena kako bi se u svakoj podeli približno očuvala raspodela klasa iz celog skupa podataka. Ovo je posebno važno u ovom radu zbog nebalansirane raspodele klasa, jer bi obična nasumična podela mogla dovesti do toga da manje zastupljene klase budu slabo predstavljene ili potpuno izostavljene

u pojedinim validacionim skupovima.

4.2 Analiza rezultata modela

Pre detaljne analize rezultata izvršena je dodatna provera mogućeg preprilagođavanja modela. U tu svrhu upoređene su prosečne makro F1-mere ostvarene na trening delovima i na odgovarajućim validacionim delovima ponovljene stratifikovane unakrsne validacije.

U sledećim tabelama i u nastavku rada, radi lakšeg organizovanja se koriste sledeće oznake za skupove karakteristika: PB - učestalosti pojedinačnih proteinskih blokova, PB2 - bigrami proteinskih blokova, PB3 - trigrami proteinskih blokova, PB+PB2 - kombinacija učestalosti i bigrama, PB+PB2+PB3 - kombinacija učestalosti, bigrama i trigrama.

Tabela 4.2: Poređenje makro F1-mere na trening i validacionim delovima za LR

Eksperimentalni skup			Prošireni skup		
Karakteristike	Trening	Validacioni	Karakteristike	Trening	Validacioni
PB	0.7550	0.5233	PB	0.7851	0.6995
PB2	1.0000	0.8461	PB2	1.0000	0.9144
PB3	1.0000	0.8181	PB3	1.0000	0.9410
PB+PB2	1.0000	0.8634	PB+PB2	1.0000	0.9158
PB+PB2+PB3	1.0000	0.8283	PB+PB2+PB3	1.0000	0.9409

Tabela 4.3: Poređenje makro F1-mere na trening i validacionim delovima za LinSVM

Eksperimentalni skup			Prošireni skup		
Karakteristike	Trening	Validacioni	Karakteristike	Trening	Validacioni
PB	0.6384	0.4945	PB	0.7948	0.7608
PB2	0.9817	0.8526	PB2	0.9954	0.9111
PB3	0.9979	0.7085	PB3	1.0000	0.8986
PB+PB2	0.9818	0.8558	PB+PB2	0.9954	0.9116
PB+PB2+PB3	0.9980	0.7124	PB+PB2+PB3	1.0000	0.9020

Na osnovu tabela 4.2–4.10 može se uočiti da se stepen prilagođenosti modela znatno razlikuje u zavisnosti od algoritma i skupa karakteristika. Pošto je u ovom radu makro F1-mera glavna mera poređenja modela, poređenje vrednosti makro F1-mere izračunate na trening delovima i odgovarajućim validacionim delovima ponovljene stratifikovane unakrsne validacije korišćeno je za uočavanje mogućeg

Tabela 4.4: Poređenje makro F1-mere na trening i validacionim delovima za RBF SVM

Eksperimentalni skup			Prošireni skup		
Karakteristike	Trening	Validacioni	Karakteristike	Trening	Validacioni
PB	0.2933	0.2467	PB	0.7317	0.7182
PB2	0.9404	0.7584	PB2	0.9735	0.9183
PB3	0.9978	0.1931	PB3	0.9988	0.2274
PB+PB2	0.9411	0.7633	PB+PB2	0.9738	0.9199
PB+PB2+PB3	0.9978	0.1721	PB+PB2+PB3	0.9988	0.2091

Tabela 4.5: Poređenje makro F1-mere na trening i validacionim delovima za DTC

Eksperimentalni skup			Prošireni skup		
Karakteristike	Trening	Validacioni	Karakteristike	Trening	Validacioni
PB	0.5463	0.2731	PB	0.6575	0.5826
PB2	0.6963	0.4389	PB2	0.7050	0.6493
PB3	0.5909	0.4108	PB3	0.8271	0.7295
PB+PB2	0.6942	0.4393	PB+PB2	0.7119	0.6519
PB+PB2+PB3	0.5890	0.4137	PB+PB2+PB3	0.8310	0.7232

Tabela 4.6: Poređenje makro F1-mere na trening i validacionim delovima za RF

Eksperimentalni skup			Prošireni skup		
Karakteristike	Trening	Validacioni	Karakteristike	Trening	Validacioni
PB	0.9801	0.4049	PB	0.9811	0.8050
PB2	0.9955	0.5466	PB2	0.9990	0.9034
PB3	0.9956	0.6221	PB3	0.9984	0.9167
PB+PB2	0.9957	0.5274	PB+PB2	0.9990	0.9008
PB+PB2+PB3	0.9956	0.5880	PB+PB2+PB3	0.9988	0.9143

preprilagođavanja i potprilagođavanja. Velika razlika između vrednosti na trening i validacionom delu može ukazivati na preprilagođavanje, dok niske vrednosti na oba dela mogu ukazivati na potprilagođavanje.

Najizraženije preprilagođavanje uočava se kod SVM modela sa RBF jezgrom, prikazanog u tabeli 4.4. Na eksperimentalnom skupu, za PB3 karakteristike trening makro F1 iznosi 0.9978, dok makro F1-mera na validacionom delu iznosi samo 0.1931. Slično tome, za skup PB+PB2+PB3 trening makro F1 iznosi 0.9978, a makro F1-mera na validacionom delu 0.1721. Isti obrazac javlja se i na proširenom skupu.

Sličan, ali manje izražen oblik preprilagođavanja može se uočiti kod modela KNN, nasumične šume, neuronske mreže i XGBoost modela, posebno na eksperimentalnom skupu. Na primer, kod KNN modela u tabeli 4.7 trening makro F1 za

Tabela 4.7: Poređenje makro F1-mere na trening i validacionim delovima za KNN

Eksperimentalni skup			Prošireni skup		
Karakteristike	Trening	Validacioni	Karakteristike	Trening	Validacioni
PB	1.0000	0.5737	PB	1.0000	0.8165
PB2	1.0000	0.7240	PB2	1.0000	0.9097
PB3	1.0000	0.6899	PB3	1.0000	0.9034
PB+PB2	1.0000	0.7273	PB+PB2	1.0000	0.9172
PB+PB2+PB3	1.0000	0.7100	PB+PB2+PB3	1.0000	0.9321

Tabela 4.8: Poređenje makro F1-mere na trening i validacionim delovima za NB

Eksperimentalni skup			Prošireni skup		
Karakteristike	Trening	Validacioni	Karakteristike	Trening	Validacioni
PB	0.1194	0.1194	PB	0.1116	0.1081
PB2	0.9425	0.6431	PB2	0.8738	0.8155
PB3	0.9763	0.6637	PB3	0.9611	0.8797
PB+PB2	0.9427	0.6494	PB+PB2	0.8717	0.8162
PB+PB2+PB3	0.9764	0.6798	PB+PB2+PB3	0.9555	0.8782

Tabela 4.9: Poređenje makro F1-mere na trening i validacionim delovima za NN

Eksperimentalni skup			Prošireni skup		
Karakteristike	Trening	Validacioni	Karakteristike	Trening	Validacioni
PB	1.0000	0.5380	PB	1.0000	0.8261
PB2	1.0000	0.6660	PB2	1.0000	0.9032
PB3	1.0000	0.5941	PB3	1.0000	0.9010
PB+PB2	1.0000	0.7379	PB+PB2	1.0000	0.9132
PB+PB2+PB3	1.0000	0.6346	PB+PB2+PB3	1.0000	0.9171

sve skupove karakteristika iznosi 1.0000, dok se validacione vrednosti na eksperimentalnom skupu kreću od 0.5737 za PB do 0.7273 za PB+PB2. Kod nasumične šume, prikazane u tabeli 4.6, trening vrednosti su takođe veoma visoke, od 0.9801 do 0.9957, dok su validacione vrednosti na eksperimentalnom skupu znatno niže, od 0.4049 do 0.6221. Slično se uočava i kod XGBoost modela u tabeli 4.10, gde je trening makro F1 jednak 1.0000 za sve skupove karakteristika, dok je makro F1-mera na validacionom delu eksperimentalnog skupa znatno niža, na primer 0.3819 za PB i 0.6412 za PB+PB2.

Sa druge strane, tabele 4.2 i 4.3 pokazuju da logistička regresija i linearni SVM imaju povoljniji odnos između trening i validacionih vrednosti makro F1-mere, posebno kod bigramskih i kombinovanih karakteristika.

Potprilagođavanje se najjasnije uočava kod modela koji imaju niske vrednosti makro F1-mere i na trening i na validacionim delovima. Primer za to je naivni

Tabela 4.10: Poređenje makro F1-mere na trening i validacionim delovima za XGB

Eksperimentalni skup			Prošireni skup		
Karakteristike	Trening	Validacioni	Karakteristike	Trening	Validacioni
PB	1.0000	0.3819	PB	1.0000	0.8156
PB2	1.0000	0.6401	PB2	1.0000	0.9058
PB3	1.0000	0.7417	PB3	1.0000	0.9237
PB+PB2	1.0000	0.6412	PB+PB2	1.0000	0.9038
PB+PB2+PB3	1.0000	0.7383	PB+PB2+PB3	1.0000	0.9241

Bajesov klasifikator nad skupom PB, prikazan u tabeli 4.8, gde su trening i validaciona makro F1-mera na eksperimentalnom skupu jednake 0.1194, a na proširenom skupu približno jednako niske, 0.1116 na trening delu i 0.1081 na validacionom delu. Sličan oblik potprilagođavanja javlja se i kod RBF SVM modela nad skupom PB na eksperimentalnom skupu, gde trening makro F1 iznosi 0.2933, a makro F1-mera na validacionom delu 0.2467. Kod stabla odlučivanja, prikazanog u tabeli 4.5, vrednosti su takođe relativno niske, naročito na eksperimentalnom skupu, gde za PB karakteristike trening makro F1 iznosi 0.5463, a makro F1-mera na validacionom delu 0.2731. Ovi rezultati ukazuju da pojedini modeli i jednostavnije reprezentacije ne uspevaju dovoljno dobro da nauče obrasce koji razlikuju klase.

U celini, rezultati pokazuju da učestalosti proteinskih blokova ne sadrže dovoljno informacija za uspešno razlikovanje klasa, dok se kod višedimenzionalnih n-gramskih reprezentacija, naročito u kombinaciji sa složenijim modelima, može javiti preprilagođavanje. Najpovoljniji odnos između prilagođenosti trening podacima i generalizacije pokazuju jednostavniji linearni modeli kod n-gramskih reprezentacija.

U tabelama 4.11 i 4.12 prikazani su najbolji modeli za svaki skup karakteristika, posebno za eksperimentalni i prošireni skup podataka. Vrednosti u tabelama predstavljaju srednje vrednosti metrika izračunate na validacionim delovima ponovljene stratifikovane unakrsne validacije. Za svaki model i skup karakteristika prikazana je prosečna makro F1-mera i prosečna težinska F1-mera kroz sve podele validacije.

Tabela 4.11: Najbolji rezultati po skupovima karakteristika na eksperimentalnom skupu

Skup karakteristika	Model	Makro F1	Težinski F1
Učestalosti PB	KNN	0.5737	0.9383
PB bigrami	Linearni SVM	0.8526	0.9637
PB trigrami	Logistička regresija	0.8181	0.9584
Učestalosti PB i bigrami	Logistička regresija	0.8634	0.9678
Učestalosti PB, bigrami i trigrami	Logistička regresija	0.8283	0.9609

Tabela 4.12: Najbolji rezultati po skupovima karakteristika na proširenom skupu

Skup karakteristika	Model	Makro F1	Težinski F1
Učestalosti PB	Neuronska mreža	0.8261	0.9274
PB bigrami	RBF SVM	0.9183	0.9617
PB trigrami	Logistička regresija	0.9410	0.9757
Učestalosti PB i bigrami	RBF SVM	0.9199	0.9623
Učestalosti PB, bigrami i trigrami	Logistička regresija	0.9409	0.9745

Na eksperimentalnom skupu najbolji rezultat postignut je korišćenjem kombinacije učestalosti proteinskih blokova i bigrama, uz logističku regresiju. Ova kombinacija ostvarila je makro F1-meru od 0.8634. Veoma sličan rezultat postignut je korišćenjem samo bigrama, gde je najbolji model bio linearni SVM sa makro F1-merom od 0.8526. Ovi rezultati ukazuju da se na eksperimentalnom skupu najstabilnije ponašaju karakteristike koje opisuju pojedinačne proteinske blokove i lokalne parove uzastopnih PB simbola.

Sa druge strane, trigramske karakteristike na eksperimentalnom skupu nisu dale najbolji rezultat, iako nose detaljnije informacije o lokalnom redosledu proteinskih blokova. Logistička regresija sa trigramima ostvarila je makro F1-meru od 0.8181, što je niže od rezultata dobijenih bigramima i kombinacijom učestalosti i bigrama. Ovo se može povezati sa većim brojem karakteristika kod trigrama, jer trigramska reprezentacija ima 4096 mogućih karakteristika. Kod eksperimentalnog skupa, u kome pojedine klase imaju mali broj instanci, ovako veliki prostor karakteristika može dovesti do manje stabilnog učenja za manjinske klase.

Na proširenom skupu podataka rezultati su se znatno poboljšali. Najbolji rezultat postignut je logističkom regresijom nad trigramskim karakteristikama, sa makro F1-merom od 0.9410. Gotovo identičan rezultat ostvaren je i kombinacijom učestalosti, bigrama i trigrama, takođe uz logističku regresiju, gde makro F1-mera iznosi 0.9409. Za razliku od eksperimentalnog skupa, na proširenom skupu trigramske karakteristike dolaze do izražaja, što ukazuje da veći broj instanci, naročito u

manjinskim klasama, omogućava stabilnije korišćenje bogatijih n-gramskih reprezentacija.

Poređenje bigramskih i trigramskih karakteristika pokazuje različito ponašanje na eksperimentalnom i proširenom skupu. Na eksperimentalnom skupu najbolji rezultat daju bigrami u kombinaciji sa učestalostima PB, dok trigrami ne dovode do poboljšanja. To ukazuje da u osnovnom eksperimentalnom skupu dodatni kontekst dužine tri PB simbola uvodi veliki broj novih karakteristika, ali bez dovoljno podataka za njihovo stabilno iskorišćavanje. Nasuprot tome, na proširenom skupu trigramske i kombinovane n-gramske karakteristike ostvaruju najbolje rezultate. Ovo ukazuje da povećanje broja instanci, posebno u manje zastupljenim klasama, omogućava stabilnije korišćenje bogatijih n-gramskih reprezentacija. Ipak, ovaj zaključak treba tumačiti oprezno, jer deo proširenog skupa čine strukture dobijene računarskim predviđanjem.

Posmatrano po algoritmima, logistička regresija pokazala se kao najuspešniji i najstabilniji model u većini najboljih kombinacija. Na eksperimentalnom skupu logistička regresija daje najbolje rezultate za trigramsku i obe kombinovane reprezentacije, dok je za PB bigrame najbolji linearni SVM. Na proširenom skupu daje najbolje rezultate za trigram i kombinovanu n-gramsku reprezentaciju. Ovakav rezultat ukazuje da su PB n-gramske karakteristike dovoljno informativne da ih i linearni model može uspešno koristiti za razlikovanje tipova koronavirusa, što se vidi u tabelama 4.11, 4.12 i 4.16 kao i u tabelama iz dodatka 6.1–6.5.

U tabelama 4.11 i 4.12 može se videti i da SVM modeli postižu dobre rezultate kod određenih skupova karakteristika. Linearni SVM je najbolji model za PB bigrame na eksperimentalnom skupu, sa makro F1-merom od 0.8526, dok je RBF SVM najbolji model za bigrame i za kombinaciju učestalosti i bigrama na proširenom skupu, sa makro F1-merama oko 0.92. To pokazuje da nelinearna granica odlučivanja može biti korisna kod određenih skupova karakteristika. Ipak, najbolji ukupni rezultat i dalje postiže logistička regresija, što sugerise da povećanje složenosti modela ne dovodi nužno do boljih rezultata. Detaljniji rezultati za linearni SVM i SVM sa RBF jezgrom dati su u dodatku, u tabelama 6.6–6.10 i 6.11–6.15.

Kod najjednostavnijeg skupa karakteristika, odnosno kod učestalosti pojedinačnih proteinskih blokova, najbolji modeli se razlikuju između dva skupa podataka. Na eksperimentalnom skupu najbolji rezultat ostvaruje KNN sa makro F1-merom od 0.5737, dok na proširenom skupu najbolji rezultat ostvaruje neuronska mreža sa makro F1-merom od 0.8261. Iako je rezultat na proširenom

skupu znatno bolji, učestalosti pojedinačnih PB i dalje daju slabije rezultate od n-gramskih reprezentacija. To ukazuje da same učestalosti pojedinačnih blokova ne nose dovoljno informacija o lokalnom redosledu strukturnih obrazaca koji mogu doprineti razlikovanju pojedinačnih tipova koronavirusa, odnosno ciljnih klasa. Različiti tipovi virusa mogu imati sličnu ukupnu zastupljenost pojedinačnih PB simbola, ali se mogu razlikovati po njihovom lokalnom rasporedu u PB sekvenci. Zbog toga bigrami i trigrami pružaju informativniji opis za potrebe klasifikacije, jer pored učestalosti pojedinačnih lokalnih stanja uključuju i njihov redosled. Ovaj zaključak se vidi poređenjem rezultata za različite skupove karakteristika u tabelama 4.11 i 4.12, dok su detaljni rezultati za KNN i NN dati u tabelama 6.16–6.20 i 6.36–6.40.

Na osnovu ovih rezultata može se zaključiti da su n-gramske karakteristike značajno korisnije od samih učestalosti proteinskih blokova za klasifikaciju tipova koronavirusa, odnosno za razlikovanje ciljnih klasa. Na eksperimentalnom skupu najbolji balans postižu bigrami i kombinacija učestalosti sa bigramima, dok na proširenom skupu najbolji rezultat daju trigramske i kombinovane n-gramske karakteristike. Time se pokazuje da povećanje broja instanci, posebno u manjinskim klasama, omogućava uspešnije korišćenje složenijih karakteristika.

Što se ostalih algoritama tiče, najmanje i najveće vrednosti metrika po klasama, za svaki skup karakteristika, za oba skupa podataka i na osnovu zbirnih validacionih predikcija, prikazane su u tabelama 4.13, 4.14 i 4.15, a detaljniji rezultati mogu se videti u tabelama 6.21–6.45.

Kolone *Min* i *Max* označavaju najmanju i najveću vrednost posmatrane metrike među pojedinačnim klasama za dati algoritam i dati skup karakteristika. Redovi *Prosek makro* i *Prosek težinski* prikazuju usrednjene makro, odnosno težinske vrednosti odgovarajuće metrike za dati algoritam kroz sve PB reprezentacije. Na taj način ovi redovi daju zbirni prikaz ponašanja algoritma preko različitih skupova karakteristika.

GLAVA 4. REZULTATI I ANALIZA ISTRAŽIVANJA

Tabela 4.13: Upoređivanje preciznosti (eng. *precision*) algoritama na eksperimentalnom i proširenom skupu

Karakteristike	LR	LinSVM	RBF	DTC	RF	KNN	NB	NN	XGB
	Min--Max	Min--Max	Min--Max	Min--Max	Min--Max	Min--Max	Min--Max	Min--Max	Min--Max
Eksperimentalni skup									
PB	0.07--0.98	0.00--0.97	0.03--0.98	0.03--0.95	0.33--1.00	0.50--1.00	0.00--0.91	0.50--1.00	0.00--1.00
PB2	0.41--1.00	0.33--1.00	0.24--1.00	0.04--0.99	0.00--1.00	0.62--1.00	0.10--1.00	0.50--1.00	0.50--1.00
PB3	0.44--1.00	0.29--1.00	0.00--1.00	0.00--0.88	0.90--1.00	0.00--1.00	0.25--1.00	0.20--0.98	0.75--1.00
PB+PB2	0.37--1.00	0.33--1.00	0.25--1.00	0.04--0.99	0.00--1.00	0.71--1.00	0.18--1.00	0.75--1.00	0.50--1.00
PB+PB2+PB3	0.30--1.00	0.33--1.00	0.00--1.00	0.00--1.00	0.89--1.00	0.62--1.00	0.43--1.00	0.22--0.98	0.50--1.00
Prosek makro	0.80	0.70	0.58	0.40	0.88	0.84	0.71	0.73	0.82
Prosek težinski	0.96	0.95	0.92	0.90	0.96	0.96	0.93	0.95	0.95
Prošireni skup									
PB	0.22--0.98	0.68--0.95	0.55--1.00	0.20--0.94	0.77--0.98	0.78--1.00	0.00--1.00	0.74--0.96	0.74--1.00
PB2	0.72--1.00	0.81--1.00	0.55--1.00	0.13--0.99	0.90--1.00	0.82--1.00	0.80--1.00	0.88--1.00	0.86--1.00
PB3	0.86--1.00	0.83--1.00	0.00--1.00	0.17--0.98	0.95--1.00	0.75--1.00	0.86--1.00	0.88--0.98	0.91--1.00
PB+PB2	0.72--1.00	0.81--1.00	0.58--1.00	0.14--0.98	0.93--1.00	0.80--1.00	0.80--1.00	0.90--1.00	0.85--1.00
PB+PB2+PB3	0.91--1.00	0.85--1.00	0.00--1.00	0.20--0.97	0.93--1.00	0.84--1.00	0.89--1.00	0.88--1.00	0.90--1.00
Prosek makro	0.89	0.90	0.87	0.64	0.96	0.93	0.81	0.92	0.94
Prosek težinski	0.96	0.95	0.88	0.87	0.96	0.96	0.87	0.96	0.96

Tabela 4.14: Upoređivanje odziva (eng. *recall*) algoritama na eksperimentalnom i proširenom skupu

Karakteristike	LR	LinSVM	RBF	DTC	RF	KNN	NB	NN	XGB
	Min--Max	Min--Max	Min--Max	Min--Max	Min--Max	Min--Max	Min--Max	Min--Max	Min--Max
Eksperimentalni skup									
PB	0.57--0.91	0.00--1.00	0.20--1.00	0.18--0.68	0.05--0.99	0.09--0.99	0.00--1.00	0.14--0.99	0.00--1.00
PB2	0.50--1.00	0.50--1.00	0.43--1.00	0.43--0.79	0.00--1.00	0.33--1.00	0.05--1.00	0.23--1.00	0.18--1.00
PB3	0.32--1.00	0.32--1.00	0.00--1.00	0.00--0.91	0.05--1.00	0.00--1.00	0.05--1.00	0.23--1.00	0.05--1.00
PB+PB2	0.45--1.00	0.50--1.00	0.43--1.00	0.43--0.79	0.00--1.00	0.33--1.00	0.09--1.00	0.27--1.00	0.14--1.00
PB+PB2+PB3	0.36--1.00	0.36--1.00	0.00--1.00	0.00--0.96	0.05--1.00	0.17--1.00	0.14--1.00	0.27--1.00	0.14--1.00
Prosek makro	0.84	0.79	0.50	0.53	0.54	0.67	0.53	0.70	0.61
Prosek težinski	0.92	0.94	0.79	0.40	0.96	0.96	0.95	0.94	0.96
Prošireni skup									
PB	0.67--0.95	0.37--0.97	0.18--0.99	0.35--0.84	0.42--0.99	0.49--0.99	0.00--1.00	0.58--0.99	0.49--0.99
PB2	0.72--1.00	0.67--1.00	0.72--0.99	0.66--0.84	0.51--1.00	0.77--1.00	0.42--1.00	0.63--0.99	0.58--1.00
PB3	0.70--1.00	0.62--1.00	0.00--1.00	0.72--0.93	0.51--1.00	0.70--0.99	0.47--1.00	0.67--0.98	0.53--1.00
PB+PB2	0.72--1.00	0.67--1.00	0.77--0.99	0.67--0.82	0.51--1.00	0.77--1.00	0.44--1.00	0.63--1.00	0.58--1.00
PB+PB2+PB3	0.70--1.00	0.66--1.00	0.00--1.00	0.63--0.91	0.51--1.00	0.72--1.00	0.47--1.00	0.70--0.99	0.58--1.00
Prosek makro	0.91	0.87	0.59	0.74	0.86	0.88	0.66	0.89	0.88
Prosek težinski	0.94	0.95	0.86	0.74	0.96	0.96	0.89	0.96	0.96

Na osnovu tabela 4.13, 4.14 i 4.15 može se uočiti da se algoritmi ne ponašaju jednako stabilno nad svim skupovima karakteristika. Na oba skupa najbolje zbirne vrednosti, posmatrano kroz red *Prosek makro* u tabeli 4.15, ostvaruju logistička regresija, linearni SVM, KNN, neuronska mreža i XGBoost, dok se drvo odlučivanja i SVM sa RBF jezgrom pokazuju kao najslabiji modeli, sa prosečnim makro F1 vrednostima 0.37 i 0.44 na eksperimentalnom odnosno 0.66 i 0.61 na proširenom skupu. To ukazuje da ovi modeli nisu jednako stabilni kroz različite PB reprezentacije.

Važno je primetiti da visoka preciznost ne znači nužno i dobro ukupno ponašanje modela. Na primer, nasumična šuma na eksperimentalnom skupu ima visok

Tabela 4.15: Upoređivanje F1-mere (eng. *F1-score*) algoritama na eksperimentalnom i proširenom skupu

Karakteristike	LR	LinSVM	RBF	DTC	RF	KNN	NB	NN	XGB
	Min--Max	Min--Max	Min--Max	Min--Max	Min--Max	Min--Max	Min--Max	Min--Max	Min--Max
Eksperimentalni skup									
PB	0.12--0.85	0.00--0.97	0.06--0.46	0.06--0.79	0.09--0.97	0.15--0.98	0.00--0.96	0.21--0.98	0.00--0.97
PB2	0.45--1.00	0.40--1.00	0.33--1.00	0.08--0.72	0.00--1.00	0.47--1.00	0.06--1.00	0.31--0.98	0.29--0.98
PB3	0.37--1.00	0.30--0.96	0.00--0.96	0.00--0.89	0.09--1.00	0.00--1.00	0.08--1.00	0.22--0.94	0.09--0.99
PB+PB2	0.41--1.00	0.40--1.00	0.34--1.00	0.07--0.72	0.00--1.00	0.48--1.00	0.12--1.00	0.40--1.00	0.22--0.98
PB+PB2+PB3	0.33--1.00	0.35--0.96	0.00--0.96	0.00--0.88	0.09--1.00	0.29--1.00	0.21--1.00	0.32--0.96	0.21--0.99
Prosek makro	0.81	0.72	0.44	0.37	0.60	0.71	0.59	0.66	0.67
Prosek težinski	0.94	0.94	0.80	0.47	0.94	0.96	0.93	0.94	0.95
Prošireni skup									
PB	0.34--0.91	0.52--0.96	0.29--0.95	0.25--0.86	0.57--0.97	0.63--0.98	0.00--0.83	0.65--0.97	0.65--0.97
PB2	0.72--1.00	0.73--1.00	0.63--0.99	0.23--0.90	0.68--0.98	0.80--0.99	0.59--0.96	0.75--0.99	0.74--0.99
PB3	0.77--1.00	0.73--0.99	0.00--0.84	0.28--0.91	0.68--0.98	0.72--0.98	0.63--0.98	0.76--0.98	0.69--0.99
PB+PB2	0.72--1.00	0.73--1.00	0.66--0.99	0.23--0.86	0.68--0.98	0.79--0.99	0.61--0.96	0.75--0.99	0.72--0.99
PB+PB2+PB3	0.79--1.00	0.78--0.99	0.00--0.84	0.30--0.90	0.68--0.98	0.78--0.99	0.63--0.97	0.81--0.99	0.74--0.99
Prosek makro	0.89	0.88	0.61	0.66	0.90	0.90	0.71	0.90	0.90
Prosek težinski	0.94	0.95	0.82	0.78	0.95	0.96	0.86	0.96	0.96

prosečan makro rezultat za preciznost, 0.88, ali znatno niži prosečan makro odziv, 0.54. To znači da model, kada neku instancu klasifikuje u određenu klasu, često može biti precizan, ali ne uspeva jednako dobro da pronađe dovoljan broj instanci svih klasa. Zbog toga se kod ovog modela javlja razlika između preciznosti i odziva, što se odražava i na nižu makro F1-meru.

Na eksperimentalnom skupu posebno su izražene razlike između algoritama kod trigramskih i kombinovanih n-gramskih karakteristika. Iako najbolji modeli nad ovim karakteristikama postižu visoke vrednosti, algoritmi poput drveta odlučivanja, nasumičnih šuma, k najbližih suseda i SVM modela sa RBF jezgrom, imaju veoma niske minimalne vrednosti F1-mere po klasama, uključujući i vrednosti jednake nuli. To pokazuje da bogatije n-gramske reprezentacije ne koriste svim algoritmima podjednako. Kod modela koji nisu dovoljno stabilni u prostoru sa velikim brojem karakteristika, pojedine klase ostaju slabo prepoznate.

Na proširenom skupu rezultati se u velikom broju slučajeva poboljšavaju. To se vidi kroz više makro vrednosti preciznosti, odziva i F1-mere, naročito kod logističke regresije, SVM modela, KNN-a, neuronske mreže i XGBoost algoritma. Povećanje broja instanci u manjinskim klasama doprinosi tome da modeli stabilnije uče obrašce iz PB sekvenci. Ipak, ni prošireni skup ne uklanja sve probleme: kod pojedinih algoritama i dalje se javljaju niske minimalne vrednosti po klasama, što pokazuje da konkretno dopunjavanje podataka primenjeno u ovom radu nije bilo dovoljno da svaki model podjednako dobro prepoznaje sve tipove koronavirusa.

Zbog toga se kao najpouzdaniji pokazatelj ukupnog ponašanja modela koristi F1-mera, jer istovremeno uzima u obzir preciznost i odziv. Rezultati iz tabela po-

tvrdju da su linearni modeli, posebno logistička regresija, veoma konkurentni u ovom zadatku, dok složeniji modeli ne daju nužno bolje rezultate. Takođe, prošireni skup podataka omogućava stabilnije korišćenje n-gramskih karakteristika, ali konačna uspešnost i dalje zavisi od izbora algoritma.

4.3 Analiza rezultata po klasama

Zbirne metrike, kao što su tačnost, makro F1-mera i težinska F1-mera, daju opštu sliku o performansama modela, ali ne pokazuju koje klase se prepoznaju uspešno, a kod kojih dolazi do grešaka.

Koliko uspešno modeli prepoznaju pojedinačne klase može se videti uz pomoć klasifikacionih izveštaja i matrica konfuzije. U tabeli 4.16 su prikazani rezultati za dva najbolja modela, odabrana prema makro F1-meri, koja je zbog nebalansiranosti skupa najinformativnija za poređenje. Za eksperimentalni skup to je model logističke regresije nad kombinacijom učestalosti proteinskih blokova i bigrama, dok je za prošireni skup posmatrana logistička regresija nad trigramskim karakteristikama. Detaljni klasifikacioni izveštaji za ostale algoritme i skupove karakteristika dati su u dodatku.

Tabela 4.16: Klasifikacioni izveštaji za najbolje modele na eksperimentalnom i proširenom skupu

Eksperimentalni skup <i>logistička regresija, PB+PB2</i>				Prošireni skup <i>logistička regresija, PB3</i>			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	1.00	1.00	1.00	BatCoV-HKU5	1.00	0.91	0.95
HCoV-229E	1.00	1.00	1.00	HCoV-229E	1.00	1.00	1.00
HCoV-HKU1	1.00	1.00	1.00	HCoV-HKU1	0.99	0.97	0.98
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.97	0.98	0.98
MERS-CoV	0.90	0.82	0.86	MERS-CoV	0.92	0.98	0.95
Porcine-EDV	1.00	0.83	0.91	Porcine-EDV	1.00	1.00	1.00
SARS-CoV	0.37	0.45	0.41	SARS-CoV	0.86	0.70	0.77
SARS-CoV-2	0.99	0.98	0.98	SARS-CoV-2	0.99	0.99	0.99
makro prosek	0.91	0.89	0.89	makro prosek	0.97	0.94	0.95
težinski prosek	0.97	0.97	0.97	težinski prosek	0.98	0.98	0.98

Na eksperimentalnom skupu najbolji model veoma uspešno prepoznaje većinu klasa. Klase BatCoV-HKU5, HCoV-229E, HCoV-HKU1 i HCoV-OC43 imaju vrednosti preciznosti, odziva i F1-mere jednake 1.00, što znači da su sve instance ovih klasa ispravno klasifikovane. Visoke vrednosti ostvarene su i za klase MERS-CoV i Porcine-EDV. Kod klase MERS-CoV preciznost iznosi 0.90, odziv 0.82,

a F1-mera 0.86, dok kod klase Porcine-EDV preciznost iznosi 1.00, odziv 0.83, a F1-mera 0.91. Ipak, rezultate za manje zastupljene klase treba tumačiti oprezno zbog malog broja instanci.

Najizraženiji problem na eksperimentalnom skupu javlja se kod klase SARS-CoV. Za ovu klasu preciznost iznosi 0.37, odziv 0.45, a F1-mera 0.41. To znači da model ne pronalazi dovoljan broj instanci klase SARS-CoV, a istovremeno deo instanci koje predviđa kao SARS-CoV pripada drugim klasama. Ovakav rezultat ukazuje da se SARS-CoV najteže razdvaja od ostalih klasa, posebno od klase SARS-CoV-2.

Na proširenom skupu uočava se poboljšanje rezultata kod većine klasa. Klase HCoV-229E i Porcine-EDV dostižu vrednosti svih metrika jednake 1.00. Visoke vrednosti ostvaruju i klase HCoV-HKU1 i HCoV-OC43, čije F1-mere iznose 0.98, kao i klasa MERS-CoV, kod koje F1-mera iznosi 0.95. Klasa BatCoV-HKU5 takođe ima visoku F1-meru od 0.95, uz preciznost 1.00 i odziv 0.91. Ovi rezultati ukazuju da dodavanje predviđenih struktura doprinosi boljem prepoznavanju manje zastupljenih tipova koronavirusa.

Najveće poboljšanje u odnosu na eksperimentalni skup vidi se kod klase SARS-CoV. F1-mera za ovu klasu raste sa 0.41 na 0.77, dok preciznost raste sa 0.37 na 0.86. Ipak, odziv za SARS-CoV ostaje niži od odziva većine ostalih klasa i iznosi 0.70, što znači da model i dalje ne uspeva da pronade sve instance ove klase. Drugim rečima, predikcije klase SARS-CoV postaju znatno pouzdanije, ali deo stvarnih instanci SARS-CoV i dalje biva klasifikovan kao neka druga klasa.

Tabela 4.17: Matrice konfuzije za najbolje modele na eksperimentalnom i proširenom skupu

Eksperimentalni skup									Prošireni skup								
	BHKU5	229E	HKU1	OC43	MERS	PEDV	SARS	SARS2		BHKU5	229E	HKU1	OC43	MERS	PEDV	SARS	SARS2
BHKU5	5	0	0	0	0	0	0	0	BHKU5	29	0	0	0	3	0	0	0
229E	0	7	0	0	0	0	0	0	229E	0	57	0	0	0	0	0	0
HKU1	0	0	23	0	0	0	0	0	HKU1	0	0	71	2	0	0	0	0
OC43	0	0	0	14	0	0	0	0	OC43	0	0	1	63	0	0	0	0
MERS	0	0	0	0	9	0	0	2	MERS	0	0	0	0	60	0	1	0
PEDV	0	0	0	0	0	5	1	0	PEDV	0	0	0	0	0	56	0	0
SARS	0	0	0	0	0	0	10	12	SARS	0	0	0	0	0	0	30	13
SARS2	0	0	0	0	1	0	16	922	SARS2	0	0	0	0	2	0	4	933

Napomena: Redovi predstavljaju stvarne klase, a kolone predviđene klase. Korišćene su skraćenice: BHKU5 – BatCoV-HKU5, PEDV – Porcine-EDV, SARS2 – SARS-CoV-2.

Matrice konfuzije prikazane u tabeli 4.17 potvrđuju da se na eksperimentalnom skupu najveći broj grešaka javlja između klasa SARS-CoV i SARS-CoV-2. Od 22 instance klase SARS-CoV, 10 je ispravno klasifikovano, dok je 12 instanci klasifikovano kao SARS-CoV-2. Sa druge strane, od 939 instanci klase SARS-CoV-2, 922

su ispravno klasifikovane, dok je 16 instanci klasifikovano kao SARS-CoV, a jedna kao MERS-CoV. Ovaj rezultat pokazuje da model najteže razdvaja klase SARS-CoV i SARS-CoV-2, što se može povezati sa njihovom biološkom i strukturnom srodnošću.

Na proširenom skupu broj grešaka između klasa SARS-CoV i SARS-CoV-2 se smanjuje, ali problem ne nestaje u potpunosti. Od 43 instance klase SARS-CoV, 30 je ispravno klasifikovano, dok je 13 instanci klasifikovano kao SARS-CoV-2. Kod klase SARS-CoV-2, 933 instance su ispravno klasifikovane, dok su 4 instance klasifikovane kao SARS-CoV, a 2 kao MERS-CoV. Dakle, dopunjavanje skupa predviđenim strukturama doprinosi boljem razdvajanju ove dve klase, ali SARS-CoV ostaje najproblematičnija klasa.

Pored grešaka između klasa SARS-CoV i SARS-CoV-2, na proširenom skupu se javlja i nekoliko grešaka kod drugih klasa. Klasa BatCoV-HKU5 ima 29 ispravno klasifikovanih instanci od ukupno 32, dok su 3 instance klasifikovane kao MERS-CoV. Kod klase HCoV-HKU1, 71 od 73 instance su ispravno klasifikovane, dok su 2 instance klasifikovane kao HCoV-OC43. Kod klase HCoV-OC43, 63 od 64 instance su ispravno klasifikovane, a jedna je klasifikovana kao HCoV-HKU1. Ove greške su znatno ređe od grešaka kod klase SARS-CoV, ali pokazuju da se određeno preklapanje javlja i između drugih srodnih ili strukturno sličnih tipova koronavirusa.

Ovakav rezultat je značajan jer pokazuje da predviđene strukture mogu biti korisne u slučajevima kada za pojedine tipove koronavirusa ne postoji dovoljno eksperimentalno određenih struktura. U kontekstu ovog rada prošireni skup nije u potpunosti uklonio problem nebalansiranosti i sličnosti pojedinih klasa, ali je omogućio stabilnije učenje obrazaca kod manje zastupljenih tipova virusa. Time se otvara prostor za buduće eksperimente u kojima bi se dodatnim podacima i pažljivijim izborom struktura moglo dodatno unaprediti prepoznavanje problematičnih klasa.

Rezultati iz tabela u dodatku pokazuju da se sličan problem kod klase SARS-CoV javlja i kod drugih algoritama. Na eksperimentalnom skupu ova klasa često ima nisku F1-meru, naročito kod nasumične šume, XGBoost algoritma, naivnog Bajesovog klasifikatora i drveta odlučivanja. Kod pojedinih modela preciznost za ovu klasu može biti relativno visoka, ali je odziv nizak, što znači da model retko predviđa klasu SARS-CoV i da veliki broj njenih instanci svrstava u druge klase, najčešće u SARS-CoV-2.

Drvo odlučivanja se pokazuje kao jedan od najslabijih modela i u analizi zbirnih metrika i u analizi po klasama. U dodatnim tabelama vidi se da kod ovog algoritma pojedine klase imaju izrazito niske vrednosti preciznosti, odziva ili F1-mere, što potvrđuje da jedno stablo odlučivanja nije dovoljno stabilno za ovaj skup podataka. Kod nasumične šume i XGBoost algoritma težinske vrednosti mogu biti visoke, ali rezultati po klasama pokazuju da se greške i dalje koncentrišu u manjinskim klasama, posebno kod klase SARS-CoV.

Sa druge strane, logistička regresija i linearni SVM pokazuju stabilnije ponašanje po klasama. Logistička regresija daje najbolji rezultat na oba skupa podataka, dok linearni SVM u više slučajeva ostvaruje rezultate bliske najboljem modelu. To ukazuje da složeniji modeli ne moraju biti pogodniji za ovaj problem, naročito kada je skup podataka nebalansiran i kada pojedine klase imaju mali broj instanci.

Na osnovu celokupne analize po klasama može se zaključiti da proširivanje skupa podataka predviđenim strukturama doprinosi stabilnijem prepoznavanju manjinskih klasa, ali ne uklanja sve izvore grešaka. Poboljšanje je posebno vidljivo kod klase SARS-CoV, koja ipak ostaje najteža za klasifikaciju i na proširenom skupu. To ukazuje da su lokalne strukturne karakteristike spajk-proteina virusa SARS-CoV i SARS-CoV-2 međusobno sličnije nego kod većine drugih parova razmatranih klasa. Zbog toga analiza po klasama predstavlja važnu dopunu zbirnim metrikama: ona pokazuje ne samo koji model ima najbolji ukupni rezultat, već i gde se greške zaista javljaju i koje klase ostaju najteže za razdvajanje.

Značajnost atributa

Kao dopuna prethodnoj analizi najboljih modela, ispitano je koji atributi najviše doprinose njihovim odlukama. Pošto su se u oba slučaja kao najbolji pokazali modeli logističke regresije, značaj atributa procenjen je na osnovu apsolutnih vrednosti naučenih koeficijenata. Kod višeklasne klasifikacije svaki atribut ima koeficijent za svaku klasu, pa je za svaki atribut izračunata prosečna apsolutna vrednost koeficijenta kroz sve klase. Dobijene vrednosti su zatim normalizovane u odnosu na najveću vrednost u okviru istog modela, tako da relativni značaj 1 označava najznačajniji atribut za posmatrani model.

U tabeli 4.18 prikazano je po deset najznačajnijih atributa za najbolji model na eksperimentalnom i na proširenom skupu. Kolona sa najizraženijom klasom označava klasu za koju dati atribut ima najveću apsolutnu vrednost koeficijenta.

Tabela 4.18: Najznačajniji atributi za najbolje modele prema apsolutnim vrednostima koeficijenata logističke regresije

Eksperimentalni skup, PB+PB2				Prošireni skup, PB3			
Tip	Obrazac	Rel. značaj	Najizr. klasa	Tip	Obrazac	Rel. značaj	Najizr. klasa
PB2	jh	1.000	SARS-CoV-2	PB3	fln	1.000	SARS-CoV-2
PB2	ee	0.978	SARS-CoV-2	PB3	dee	0.962	SARS-CoV-2
PB2	ab	0.968	SARS-CoV-2	PB3	abg	0.940	BatCoV-HKU5
PB2	bf	0.906	SARS-CoV	PB3	iab	0.920	Porcine-EDV
PB2	jb	0.848	SARS-CoV-2	PB3	abj	0.912	MERS-CoV
PB2	jd	0.822	BatCoV-HKU5	PB3	nmj	0.908	SARS-CoV-2
PB2	bp	0.811	SARS-CoV	PB3	dce	0.884	SARS-CoV-2
PB2	af	0.809	SARS-CoV	PB3	ded	0.879	SARS-CoV-2
PB2	hj	0.781	SARS-CoV	PB3	jhi	0.877	HCoV-OC43
PB2	cc	0.740	SARS-CoV-2	PB3	ilm	0.869	HCoV-OC43

Na eksperimentalnom skupu gde je korišćena kombinacija učestalosti pojedinačnih proteinskih blokova i bigrama, očekivano je da se među deset najznačajnijih atributa nalaze samo bigrami. Ovakav rezultat je u skladu sa prethodnim analizama, u kojima se pokazalo da same učestalosti pojedinačnih blokova ne nose dovoljno informacija kao n-gramske reprezentacije.

Kod proširenog skupa najbolji rezultat postignut je trigramskom reprezentacijom. Pošto je proširivanjem skupa povećan broj instanci u manjinskim klasama, model je mogao uspešnije da koristi složenije karakteristike, odnosno obrasce od tri uzastopna proteinska bloka. To ukazuje da dodatne instance mogu doprineti stabilnijem učenju lokalnih PB obrazaca koji nisu dovoljno pouzdani kada su pojedine klase predstavljene malim brojem primera.

Može se primetiti i da se pojedini obrasci delimično ponavljaju kroz obe analize. U tabeli 4.18 vidi se da je bigram *ab* među najznačajnijim atributima na eksperimentalnom skupu, dok se na proširenom skupu među najznačajnijim trigramima pojavljuju obrasci *abg*, *iab* i *abj*. To ukazuje na to da obrasci koji uključuju prelaz *ab* imaju važnu ulogu u odlukama oba modela, ali u proširenom skupu dolaze do izražaja u širem trigramskom kontekstu. Slično tome, na proširenom skupu se više puta pojavljuju trigrami koji uključuju blokove *d* i *e*, kao što su *dee*, *dce* i *ded*, što ukazuje da su određeni lokalni obrasci PB sekvence posebno informativni za razlikovanje klasa.

Kolona sa najizraženijom klasom pokazuje da se najznačajniji atributi ne vezuju samo za većinsku klasu SARS-CoV-2. Na eksperimentalnom skupu deo značajnih bigrama najizraženiji je za SARS-CoV i BatCoV-HKU5, dok se na proširenom skupu među najizraženijim klasama pojavljuju i BatCoV-HKU5, Porcine-EDV,

MERS-CoV i HCoV-OC43. To znači da izdvojeni PB n-grami ne doprinose isključivo prepoznavanju dominantne klase, već učestvuju i u razlikovanju više različitih tipova koronavirusa.

Glava 5

Zaključak

U ovom radu opisana je klasifikacija tipova koronavirusa na osnovu karakteristika spajk-proteina predstavljenih pomoću proteinskih blokova. Najpre su uvedeni osnovni biološki pojmovi, strukturni alfabeti, a posebno je obrađen strukturni alfabet proteinski blokovi. Zatim je opisan način formiranja skupa podataka, pripreme PB sekvenci i izdvajanja karakteristika. Nakon toga, prikazani su korišćeni algoritmi istraživanja podataka i izvršeno je poređenje njihovih rezultata na eksperimentalnom i proširenom skupu podataka.

Rezultati su pokazali da reprezentacija proteina pomoću proteinskih blokova može biti korisna za klasifikaciju tipova koronavirusa. To se posebno vidi kod PB n-gramskih karakteristika, koje su dale znatno bolje rezultate od samih učestalosti pojedinačnih proteinskih blokova. Na eksperimentalnom skupu najbolji rezultat ostvaren je kombinacijom učestalosti PB i bigrama, sa makro F1-merom od 0.86. Na proširenom skupu najbolji rezultati ostvareni su trigramskim i kombinovanim n-gramskim karakteristikama, pri čemu je logistička regresija u oba slučaja ostvarila vrednost od 0.94. Najbolje rezultate nisu dale same učestalosti pojedinačnih proteinskih blokova, već n-gramske karakteristike, posebno bigrami, trigrami i njihove kombinacije. Time je potvrđeno da za razlikovanje tipova koronavirusa nije značajna samo zastupljenost pojedinačnih lokalnih strukturnih obrazaca, već i njihov redosled u PB sekvenci. Na eksperimentalnom skupu najbolji balans dale su bigramske karakteristike i kombinacija učestalosti PB sa bigramima, dok su se na proširenom skupu kao najuspešnije pokazale trigramske i kombinovane n-gramske reprezentacije.

Kao najuspešniji i najstabilniji model pokazala se logistička regresija. Na eksperimentalnom skupu, u kombinaciji sa učestalostima proteinskih blokova i bi-

gramima, ostvarila je vrednosti makro preciznosti, makro odziva i makro F1-mere od 0.87, 0.87 i 0.86, dok je na proširenom skupu, uz trigramske karakteristike, ostvarila vrednosti 0.96, 0.93 i 0.94. SVM modeli su u više slučajeva ostvarili rezultate bliske logističkoj regresiji, posebno linearni SVM na eksperimentalnom skupu i RBF SVM kod pojedinih reprezentacija na proširenom skupu. Ipak, složeniji modeli nisu nužno davali bolje rezultate. Kod RBF SVM-a sa trigramskim i kombinovanim n-gramskim karakteristikama uočeno je izraženo preprilagođavanje, dok su drvo odlučivanja i pojedini nelinearni modeli imali slabiju stabilnost u analizi po klasama.

Poređenjem eksperimentalnog i proširenog skupa uočeno je da dodavanje struktura predviđenih pomoću AlphaFold2/ColabFold alata doprinosi stabilnijem prepoznavanju manje zastupljenih klasa. Najveće poboljšanje primećeno je kod klase SARS-CoV, čija je F1-mera porasla sa 0.41 na eksperimentalnom skupu na 0.77 na proširenom skupu. Ipak, ova klasa je ostala najteža za klasifikaciju, pre svega zbog čestog mešanja sa klasom SARS-CoV-2. To ukazuje da su u korišćenoj PB reprezentaciji lokalni strukturni obrasci spajk-proteina ova dva tipa koronavirusa teže razdvojivi u odnosu na ostale razmatrane klase.

Jedan od pravaca daljeg istraživanja bio bi proširivanje skupa podataka novim eksperimentalno određenim strukturama, posebno za klase koje su u ovom radu bile slabije zastupljene. Pored toga, bilo bi korisno detaljnije ispitati kvalitet i uticaj predviđenih struktura, kao i primeniti dodatne metode za izdvajanje karakteristika iz PB sekvenci, na primer ponavljajuće strukturne obrasce. Takođe, buduća istraživanja mogla bi uključiti i druge proteine koronavirusa ili kombinovati PB karakteristike sa aminokiselinskim karakteristikama.

Na osnovu celokupne analize može se zaključiti da proteinski blokovi predstavljaju dobru osnovu za strukturnu reprezentaciju spajk-proteina u zadatku klasifikacije tipova koronavirusa. Rezultati rada pokazuju da strukturni alfabeti mogu imati značajnu primenu u bioinformatičkoj analizi virusnih proteina i da predstavljaju koristan pravac za dalja istraživanja u oblasti klasifikacije proteinskih struktura.

Glava 6

Dodatne tabele rezultata po klasama

U ovom dodatku prikazani su klasifikacioni izveštaji po klasama za sve korišćene skupove karakteristika i sve razmatrane algoritme. Svaka tabela prikazuje rezultate za eksperimentalni i prošireni skup podataka.

Logistička regresija (LR)

Tabela 6.1: Rezultati klasifikacije modela LR korišćenjem PB na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.50	0.60	0.55	BatCoV-HKU5	0.55	0.81	0.66
HCoV-229E	0.36	0.57	0.44	HCoV-229E	0.87	0.95	0.91
HCoV-HKU1	0.69	0.87	0.77	HCoV-HKU1	0.71	0.71	0.71
HCoV-OC43	0.35	0.57	0.43	HCoV-OC43	0.57	0.67	0.61
MERS-CoV	0.77	0.91	0.83	MERS-CoV	0.59	0.79	0.68
Porcine-EDV	0.31	0.67	0.42	Porcine-EDV	0.75	0.95	0.83
SARS-CoV	0.07	0.68	0.12	SARS-CoV	0.22	0.79	0.34
SARS-CoV-2	0.98	0.74	0.85	SARS-CoV-2	0.98	0.79	0.87
makro prosek	0.50	0.70	0.55	makro prosek	0.65	0.81	0.70
težinski prosek	0.93	0.74	0.82	težinski prosek	0.88	0.79	0.82

Tabela 6.2: Rezultati klasifikacije modela LR korišćenjem PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	1.00	1.00	1.00	BatCoV-HKU5	0.91	0.91	0.91
HCoV-229E	1.00	1.00	1.00	HCoV-229E	1.00	1.00	1.00
HCoV-HKU1	1.00	1.00	1.00	HCoV-HKU1	0.97	0.95	0.96
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.94	0.98	0.96
MERS-CoV	0.82	0.82	0.82	MERS-CoV	0.89	0.97	0.93
Porcine-EDV	1.00	0.83	0.91	Porcine-EDV	1.00	1.00	1.00
SARS-CoV	0.41	0.50	0.45	SARS-CoV	0.72	0.72	0.72
SARS-CoV-2	0.99	0.98	0.98	SARS-CoV-2	0.99	0.98	0.99
makro prosek	0.90	0.89	0.90	makro prosek	0.93	0.94	0.93
težinski prosek	0.97	0.97	0.97	težinski prosek	0.97	0.97	0.97

Tabela 6.3: Rezultati klasifikacije modela LR korišćenjem PB3 na eksperimentalnom skupu

Eksperimentalni skup			
Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.83	1.00	0.91
HCoV-229E	1.00	0.57	0.73
HCoV-HKU1	0.96	1.00	0.98
HCoV-OC43	1.00	1.00	1.00
MERS-CoV	0.90	0.82	0.86
Porcine-EDV	1.00	1.00	1.00
SARS-CoV	0.44	0.32	0.37
SARS-CoV-2	0.98	0.99	0.99
makro prosek	0.89	0.84	0.85
težinski prosek	0.97	0.97	0.97

Tabela 6.4: Rezultati klasifikacije modela LR korišćenjem PB+PB2 na proširenom skupu

Prošireni skup			
Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.91	0.91	0.91
HCoV-229E	1.00	1.00	1.00
HCoV-HKU1	0.96	0.95	0.95
HCoV-OC43	0.94	0.97	0.95
MERS-CoV	0.89	0.97	0.93
Porcine-EDV	1.00	0.98	0.99
SARS-CoV	0.72	0.72	0.72
SARS-CoV-2	0.99	0.98	0.99
makro prosek	0.93	0.93	0.93
težinski prosek	0.97	0.97	0.97

Tabela 6.5: Rezultati klasifikacije modela LR korišćenjem PB+PB2+PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.83	1.00	0.91	BatCoV-HKU5	1.00	0.91	0.95
HCoV-229E	0.83	0.71	0.77	HCoV-229E	1.00	1.00	1.00
HCoV-HKU1	0.82	1.00	0.90	HCoV-HKU1	0.97	0.93	0.95
HCoV-OC43	0.82	1.00	0.90	HCoV-OC43	0.93	0.98	0.95
MERS-CoV	0.83	0.91	0.87	MERS-CoV	0.92	0.98	0.95
Porcine-EDV	1.00	1.00	1.00	Porcine-EDV	1.00	1.00	1.00
SARS-CoV	0.30	0.36	0.33	SARS-CoV	0.91	0.70	0.79
SARS-CoV-2	0.98	0.97	0.98	SARS-CoV-2	0.99	0.99	0.99
makro prosek	0.80	0.87	0.83	makro prosek	0.96	0.94	0.95
težinski prosek	0.96	0.95	0.96	težinski prosek	0.98	0.98	0.98

Linearni SVM (LinSVM)

Tabela 6.6: Rezultati klasifikacije modela LinSVM korišćenjem PB na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.42	1.00	0.59	BatCoV-HKU5	0.73	0.50	0.59
HCoV-229E	0.20	0.29	0.24	HCoV-229E	0.93	0.95	0.94
HCoV-HKU1	0.59	0.83	0.69	HCoV-HKU1	0.70	0.78	0.74
HCoV-OC43	0.44	0.57	0.50	HCoV-OC43	0.76	0.64	0.69
MERS-CoV	0.67	0.73	0.70	MERS-CoV	0.68	0.75	0.71
Porcine-EDV	0.44	0.67	0.53	Porcine-EDV	0.86	0.91	0.89
SARS-CoV	0.00	0.00	0.00	SARS-CoV	0.89	0.37	0.52
SARS-CoV-2	0.97	0.96	0.97	SARS-CoV-2	0.95	0.97	0.96
makro prosek	0.47	0.63	0.53	makro prosek	0.81	0.73	0.76
težinski prosek	0.92	0.93	0.92	težinski prosek	0.90	0.90	0.90

Tabela 6.7: Rezultati klasifikacije modela LinSVM korišćenjem PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	1.00	1.00	1.00	BatCoV-HKU5	0.90	0.81	0.85
HCoV-229E	1.00	0.71	0.83	HCoV-229E	1.00	1.00	1.00
HCoV-HKU1	0.96	1.00	0.98	HCoV-HKU1	0.96	0.92	0.94
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.92	0.95	0.94
MERS-CoV	0.75	0.82	0.78	MERS-CoV	0.84	0.97	0.90
Porcine-EDV	0.83	0.83	0.83	Porcine-EDV	1.00	1.00	1.00
SARS-CoV	0.33	0.50	0.40	SARS-CoV	0.81	0.67	0.73
SARS-CoV-2	0.99	0.98	0.98	SARS-CoV-2	0.99	0.99	0.99
makro prosek	0.86	0.86	0.85	makro prosek	0.93	0.91	0.92
težinski prosek	0.97	0.96	0.97	težinski prosek	0.97	0.97	0.97

Tabela 6.8: Rezultati klasifikacije modela LinSVM korišćenjem PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.75	0.60	0.67	BatCoV-HKU5	0.87	0.62	0.73
HCoV-229E	1.00	0.57	0.73	HCoV-229E	1.00	0.98	0.99
HCoV-HKU1	0.61	1.00	0.75	HCoV-HKU1	0.97	0.95	0.96
HCoV-OC43	0.64	1.00	0.78	HCoV-OC43	0.91	0.97	0.94
MERS-CoV	0.50	0.91	0.65	MERS-CoV	0.83	0.98	0.90
Porcine-EDV	0.55	1.00	0.71	Porcine-EDV	0.89	1.00	0.94
SARS-CoV	0.29	0.32	0.30	SARS-CoV	0.83	0.67	0.74
SARS-CoV-2	0.98	0.95	0.96	SARS-CoV-2	0.99	0.99	0.99
makro prosek	0.66	0.79	0.69	makro prosek	0.91	0.90	0.90
težinski prosek	0.95	0.93	0.93	težinski prosek	0.96	0.96	0.96

Tabela 6.9: Rezultati klasifikacije modela LinSVM korišćenjem PB+PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	1.00	1.00	1.00	BatCoV-HKU5	0.90	0.81	0.85
HCoV-229E	1.00	0.86	0.92	HCoV-229E	1.00	1.00	1.00
HCoV-HKU1	0.96	1.00	0.98	HCoV-HKU1	0.96	0.92	0.94
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.92	0.95	0.94
MERS-CoV	0.75	0.82	0.78	MERS-CoV	0.84	0.97	0.90
Porcine-EDV	0.83	0.83	0.83	Porcine-EDV	1.00	1.00	1.00
SARS-CoV	0.33	0.50	0.40	SARS-CoV	0.81	0.67	0.73
SARS-CoV-2	0.99	0.97	0.98	SARS-CoV-2	0.99	0.99	0.99
makro prosek	0.86	0.87	0.86	makro prosek	0.93	0.91	0.92
težinski prosek	0.97	0.96	0.97	težinski prosek	0.97	0.97	0.97

Tabela 6.10: Rezultati klasifikacije modela LinSVM korišćenjem PB+PB2+PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.60	0.60	0.60	BatCoV-HKU5	0.95	0.66	0.78
HCoV-229E	1.00	0.57	0.73	HCoV-229E	1.00	0.98	0.99
HCoV-HKU1	0.55	1.00	0.71	HCoV-HKU1	0.96	0.93	0.94
HCoV-OC43	0.58	1.00	0.74	HCoV-OC43	0.89	0.97	0.93
MERS-CoV	0.53	0.91	0.67	MERS-CoV	0.85	0.98	0.91
Porcine-EDV	0.50	1.00	0.67	Porcine-EDV	0.89	1.00	0.94
SARS-CoV	0.33	0.36	0.35	SARS-CoV	0.86	0.72	0.78
SARS-CoV-2	0.98	0.94	0.96	SARS-CoV-2	0.99	0.99	0.99
makro prosek	0.63	0.80	0.68	makro prosek	0.92	0.90	0.91
težinski prosek	0.94	0.93	0.93	težinski prosek	0.97	0.97	0.97

SVM sa RBF jezgrom

Tabela 6.11: Rezultati klasifikacije modela SVM-a sa RBF jezgrom korišćenjem PB na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.08	1.00	0.15	BatCoV-HKU5	0.62	0.66	0.64
HCoV-229E	0.19	0.86	0.31	HCoV-229E	1.00	0.89	0.94
HCoV-HKU1	0.41	0.52	0.46	HCoV-HKU1	0.72	0.18	0.29
HCoV-OC43	0.16	0.64	0.26	HCoV-OC43	0.55	0.75	0.64
MERS-CoV	0.14	0.45	0.21	MERS-CoV	0.83	0.57	0.68
Porcine-EDV	0.11	0.67	0.18	Porcine-EDV	0.93	0.89	0.91
SARS-CoV	0.03	0.86	0.06	SARS-CoV	0.91	0.47	0.62
SARS-CoV-2	0.98	0.20	0.34	SARS-CoV-2	0.92	0.99	0.95
makro prosek	0.26	0.65	0.25	makro prosek	0.81	0.68	0.71
težinski prosek	0.92	0.25	0.33	težinski prosek	0.88	0.88	0.87

Tabela 6.12: Rezultati klasifikacije modela SVM-a sa RBF jezgrom korišćenjem PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	1.00	0.60	0.75	BatCoV-HKU5	0.97	0.88	0.92
HCoV-229E	1.00	0.43	0.60	HCoV-229E	1.00	0.98	0.99
HCoV-HKU1	1.00	1.00	1.00	HCoV-HKU1	0.99	0.99	0.99
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.98	0.98	0.98
MERS-CoV	0.75	0.82	0.78	MERS-CoV	0.91	0.97	0.94
Porcine-EDV	1.00	0.67	0.80	Porcine-EDV	1.00	0.98	0.99
SARS-CoV	0.24	0.55	0.33	SARS-CoV	0.55	0.72	0.63
SARS-CoV-2	0.98	0.96	0.97	SARS-CoV-2	0.98	0.97	0.98
makro prosek	0.87	0.75	0.78	makro prosek	0.92	0.93	0.93
težinski prosek	0.96	0.94	0.95	težinski prosek	0.97	0.96	0.96

Tabela 6.13: Rezultati klasifikacije modela SVM-a sa RBF jezgrom korišćenjem PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.00	0.00	0.00	BatCoV-HKU5	0.00	0.00	0.00
HCoV-229E	0.00	0.00	0.00	HCoV-229E	1.00	0.19	0.32
HCoV-HKU1	1.00	0.17	0.30	HCoV-HKU1	1.00	0.16	0.28
HCoV-OC43	1.00	0.07	0.13	HCoV-OC43	1.00	0.08	0.14
MERS-CoV	1.00	0.09	0.17	MERS-CoV	1.00	0.08	0.15
Porcine-EDV	0.00	0.00	0.00	Porcine-EDV	1.00	0.11	0.19
SARS-CoV	0.67	0.09	0.16	SARS-CoV	1.00	0.05	0.09
SARS-CoV-2	0.92	1.00	0.96	SARS-CoV-2	0.73	1.00	0.84
makro prosek	0.57	0.18	0.21	makro prosek	0.84	0.21	0.25
težinski prosek	0.90	0.92	0.89	težinski prosek	0.79	0.74	0.65

Tabela 6.14: Rezultati klasifikacije modela SVM-a sa RBF jezgrom korišćenjem PB+PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	1.00	0.80	0.89	BatCoV-HKU5	0.97	0.88	0.92
HCoV-229E	1.00	0.43	0.60	HCoV-229E	1.00	0.98	0.99
HCoV-HKU1	1.00	1.00	1.00	HCoV-HKU1	0.99	0.99	0.99
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.98	0.98	0.98
MERS-CoV	0.75	0.82	0.78	MERS-CoV	0.91	0.97	0.94
Porcine-EDV	1.00	0.67	0.80	Porcine-EDV	1.00	0.98	0.99
SARS-CoV	0.25	0.55	0.34	SARS-CoV	0.58	0.77	0.66
SARS-CoV-2	0.98	0.96	0.97	SARS-CoV-2	0.99	0.97	0.98
makro prosek	0.87	0.78	0.80	makro prosek	0.93	0.94	0.93
težinski prosek	0.96	0.95	0.95	težinski prosek	0.97	0.97	0.97

Tabela 6.15: Rezultati klasifikacije modela SVM-a sa RBF jezgrom korišćenjem PB+PB2+PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.00	0.00	0.00	BatCoV-HKU5	0.00	0.00	0.00
HCoV-229E	0.00	0.00	0.00	HCoV-229E	1.00	0.12	0.22
HCoV-HKU1	1.00	0.13	0.23	HCoV-HKU1	1.00	0.12	0.22
HCoV-OC43	0.00	0.00	0.00	HCoV-OC43	1.00	0.05	0.09
MERS-CoV	0.00	0.00	0.00	MERS-CoV	1.00	0.08	0.15
Porcine-EDV	0.00	0.00	0.00	Porcine-EDV	1.00	0.11	0.19
SARS-CoV	0.67	0.09	0.16	SARS-CoV	1.00	0.05	0.09
SARS-CoV-2	0.92	1.00	0.96	SARS-CoV-2	0.73	1.00	0.84
makro prosek	0.32	0.15	0.17	makro prosek	0.84	0.19	0.23
težinski prosek	0.88	0.92	0.88	težinski prosek	0.78	0.73	0.64

Metod k najbližih suseda (KNN)

Tabela 6.16: Rezultati klasifikacije modela KNN korišćenjem PB na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.50	0.40	0.44	BatCoV-HKU5	0.79	0.59	0.68
HCoV-229E	1.00	0.43	0.60	HCoV-229E	1.00	0.88	0.93
HCoV-HKU1	0.84	0.91	0.88	HCoV-HKU1	0.83	0.85	0.84
HCoV-OC43	0.80	0.57	0.67	HCoV-OC43	0.79	0.77	0.78
MERS-CoV	0.71	0.45	0.56	MERS-CoV	0.78	0.84	0.81
Porcine-EDV	0.80	0.67	0.73	Porcine-EDV	0.98	0.96	0.97
SARS-CoV	0.50	0.09	0.15	SARS-CoV	0.88	0.49	0.63
SARS-CoV-2	0.96	0.99	0.98	SARS-CoV-2	0.96	0.99	0.98
makro prosek	0.76	0.56	0.63	makro prosek	0.88	0.80	0.83
težinski prosek	0.94	0.95	0.94	težinski prosek	0.93	0.93	0.93

Tabela 6.17: Rezultati klasifikacije modela KNN korišćenjem PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.62	1.00	0.77	BatCoV-HKU5	0.90	0.81	0.85
HCoV-229E	1.00	0.43	0.60	HCoV-229E	1.00	0.93	0.96
HCoV-HKU1	1.00	1.00	1.00	HCoV-HKU1	1.00	0.90	0.95
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.90	1.00	0.95
MERS-CoV	0.88	0.64	0.74	MERS-CoV	0.88	0.95	0.91
Porcine-EDV	1.00	0.33	0.50	Porcine-EDV	1.00	0.93	0.96
SARS-CoV	0.67	0.36	0.47	SARS-CoV	0.82	0.77	0.80
SARS-CoV-2	0.98	0.99	0.99	SARS-CoV-2	0.98	0.99	0.99
makro prosek	0.89	0.72	0.76	makro prosek	0.94	0.91	0.92
težinski prosek	0.97	0.97	0.97	težinski prosek	0.97	0.97	0.97

Tabela 6.18: Rezultati klasifikacije modela KNN korišćenjem PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.71	1.00	0.83	BatCoV-HKU5	0.88	0.94	0.91
HCoV-229E	1.00	0.43	0.60	HCoV-229E	1.00	0.91	0.95
HCoV-HKU1	1.00	0.96	0.98	HCoV-HKU1	0.95	0.81	0.87
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.92	0.89	0.90
MERS-CoV	0.88	0.64	0.74	MERS-CoV	0.98	0.87	0.92
Porcine-EDV	0.00	0.00	0.00	Porcine-EDV	0.98	0.91	0.94
SARS-CoV	0.50	0.36	0.42	SARS-CoV	0.75	0.70	0.72
SARS-CoV-2	0.97	0.99	0.98	SARS-CoV-2	0.96	0.99	0.98
makro prosek	0.76	0.67	0.69	makro prosek	0.93	0.88	0.90
težinski prosek	0.95	0.96	0.96	težinski prosek	0.95	0.95	0.95

Tabela 6.19: Rezultati klasifikacije modela KNN korišćenjem PB+PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.71	1.00	0.83	BatCoV-HKU5	0.87	0.81	0.84
HCoV-229E	1.00	0.43	0.60	HCoV-229E	1.00	0.93	0.96
HCoV-HKU1	0.96	1.00	0.98	HCoV-HKU1	1.00	0.92	0.96
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.91	1.00	0.96
MERS-CoV	0.88	0.64	0.74	MERS-CoV	0.88	0.93	0.90
Porcine-EDV	1.00	0.33	0.50	Porcine-EDV	1.00	0.93	0.96
SARS-CoV	0.73	0.36	0.48	SARS-CoV	0.80	0.77	0.79
SARS-CoV-2	0.98	1.00	0.99	SARS-CoV-2	0.98	0.99	0.99
makro prosek	0.91	0.72	0.76	makro prosek	0.93	0.91	0.92
težinski prosek	0.97	0.97	0.97	težinski prosek	0.97	0.97	0.97

Tabela 6.20: Rezultati klasifikacije modela KNN korišćenjem PB+PB2+PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.71	1.00	0.83	BatCoV-HKU5	0.97	1.00	0.98
HCoV-229E	1.00	0.43	0.60	HCoV-229E	1.00	0.91	0.95
HCoV-HKU1	1.00	1.00	1.00	HCoV-HKU1	0.95	0.96	0.95
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.95	0.92	0.94
MERS-CoV	0.78	0.64	0.70	MERS-CoV	0.97	0.95	0.96
Porcine-EDV	1.00	0.17	0.29	Porcine-EDV	1.00	0.93	0.96
SARS-CoV	0.62	0.36	0.46	SARS-CoV	0.84	0.72	0.78
SARS-CoV-2	0.97	0.99	0.98	SARS-CoV-2	0.98	0.99	0.99
makro prosek	0.89	0.70	0.73	makro prosek	0.96	0.92	0.94
težinski prosek	0.96	0.97	0.96	težinski prosek	0.97	0.97	0.97

Drvo odlučivanja (DTC)

Tabela 6.21: Rezultati klasifikacije modela DTC korišćenjem PB na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.11	0.20	0.14	BatCoV-HKU5	0.24	0.41	0.30
HCoV-229E	0.09	0.29	0.14	HCoV-229E	0.77	0.84	0.81
HCoV-HKU1	0.37	0.57	0.45	HCoV-HKU1	0.31	0.66	0.42
HCoV-OC43	0.11	0.36	0.17	HCoV-OC43	0.52	0.45	0.48
MERS-CoV	0.05	0.18	0.08	MERS-CoV	0.38	0.57	0.45
Porcine-EDV	0.16	0.50	0.24	Porcine-EDV	0.93	0.77	0.84
SARS-CoV	0.03	0.27	0.06	SARS-CoV	0.20	0.35	0.25
SARS-CoV-2	0.95	0.68	0.79	SARS-CoV-2	0.94	0.78	0.86
makro prosek	0.23	0.38	0.26	makro prosek	0.54	0.60	0.55
težinski prosek	0.88	0.65	0.74	težinski prosek	0.81	0.73	0.76

Tabela 6.22: Rezultati klasifikacije modela DTC korišćenjem PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.23	0.60	0.33	BatCoV-HKU5	0.45	0.78	0.57
HCoV-229E	0.21	0.43	0.29	HCoV-229E	0.85	0.82	0.84
HCoV-HKU1	0.54	0.65	0.59	HCoV-HKU1	0.42	0.79	0.55
HCoV-OC43	0.65	0.79	0.71	HCoV-OC43	0.49	0.73	0.59
MERS-CoV	0.18	0.55	0.27	MERS-CoV	0.78	0.69	0.73
Porcine-EDV	0.50	0.50	0.50	Porcine-EDV	0.96	0.84	0.90
SARS-CoV	0.04	0.73	0.08	SARS-CoV	0.13	0.77	0.23
SARS-CoV-2	0.99	0.57	0.72	SARS-CoV-2	0.99	0.66	0.79
makro prosek	0.42	0.60	0.44	makro prosek	0.63	0.76	0.65
težinski prosek	0.93	0.58	0.70	težinski prosek	0.87	0.69	0.75

Tabela 6.23: Rezultati klasifikacije modela DTC korišćenjem PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.00	0.00	0.00	BatCoV-HKU5	0.79	0.72	0.75
HCoV-229E	0.67	0.86	0.75	HCoV-229E	0.90	0.93	0.91
HCoV-HKU1	0.88	0.91	0.89	HCoV-HKU1	0.50	0.77	0.61
HCoV-OC43	0.86	0.43	0.57	HCoV-OC43	0.73	0.84	0.78
MERS-CoV	0.05	0.82	0.09	MERS-CoV	0.74	0.84	0.78
Porcine-EDV	0.01	0.33	0.02	Porcine-EDV	0.93	0.89	0.91
SARS-CoV	0.02	0.59	0.05	SARS-CoV	0.17	0.74	0.28
SARS-CoV-2	0.86	0.05	0.09	SARS-CoV-2	0.98	0.78	0.87
makro prosek	0.42	0.50	0.31	makro prosek	0.72	0.81	0.74
težinski prosek	0.82	0.10	0.11	težinski prosek	0.89	0.79	0.83

Tabela 6.24: Rezultati klasifikacije modela DTC korišćenjem PB+PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.23	0.60	0.33	BatCoV-HKU5	0.43	0.78	0.56
HCoV-229E	0.21	0.43	0.29	HCoV-229E	0.85	0.82	0.84
HCoV-HKU1	0.54	0.65	0.59	HCoV-HKU1	0.38	0.79	0.52
HCoV-OC43	0.65	0.79	0.71	HCoV-OC43	0.49	0.73	0.59
MERS-CoV	0.19	0.55	0.28	MERS-CoV	0.79	0.67	0.73
Porcine-EDV	0.36	0.67	0.47	Porcine-EDV	0.96	0.79	0.86
SARS-CoV	0.04	0.64	0.07	SARS-CoV	0.14	0.72	0.23
SARS-CoV-2	0.99	0.57	0.72	SARS-CoV-2	0.98	0.67	0.80
makro prosek	0.40	0.61	0.43	makro prosek	0.63	0.75	0.64
težinski prosek	0.93	0.57	0.69	težinski prosek	0.87	0.69	0.75

Tabela 6.25: Rezultati klasifikacije modela DTC korišćenjem PB+PB2+PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.00	0.00	0.00	BatCoV-HKU5	0.70	0.66	0.68
HCoV-229E	0.67	0.86	0.75	HCoV-229E	0.90	0.91	0.90
HCoV-HKU1	0.81	0.96	0.88	HCoV-HKU1	0.47	0.70	0.56
HCoV-OC43	1.00	0.71	0.83	HCoV-OC43	0.72	0.83	0.77
MERS-CoV	0.02	0.82	0.05	MERS-CoV	0.70	0.84	0.76
Porcine-EDV	0.60	0.50	0.55	Porcine-EDV	0.91	0.89	0.90
SARS-CoV	0.02	0.59	0.05	SARS-CoV	0.20	0.63	0.30
SARS-CoV-2	0.95	0.04	0.08	SARS-CoV-2	0.97	0.82	0.89
makro prosek	0.51	0.56	0.40	makro prosek	0.70	0.78	0.72
težinski prosek	0.91	0.10	0.11	težinski prosek	0.88	0.81	0.84

Nasumične šume (RF)

Tabela 6.26: Rezultati klasifikacije modela RF korišćenjem PB na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.33	0.20	0.25	BatCoV-HKU5	0.85	0.69	0.76
HCoV-229E	1.00	0.14	0.25	HCoV-229E	0.98	0.88	0.93
HCoV-HKU1	0.63	0.83	0.72	HCoV-HKU1	0.77	0.78	0.78
HCoV-OC43	0.80	0.29	0.42	HCoV-OC43	0.80	0.69	0.74
MERS-CoV	0.67	0.36	0.47	MERS-CoV	0.84	0.84	0.84
Porcine-EDV	1.00	0.33	0.50	Porcine-EDV	0.98	0.91	0.94
SARS-CoV	1.00	0.05	0.09	SARS-CoV	0.90	0.42	0.57
SARS-CoV-2	0.95	0.99	0.97	SARS-CoV-2	0.94	0.99	0.97
makro prosek	0.80	0.40	0.46	makro prosek	0.88	0.77	0.81
težinski prosek	0.94	0.94	0.92	težinski prosek	0.92	0.92	0.92

Tabela 6.27: Rezultati klasifikacije modela RF korišćenjem PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.75	0.60	0.67	BatCoV-HKU5	0.90	0.88	0.89
HCoV-229E	0.00	0.00	0.00	HCoV-229E	0.98	0.89	0.94
HCoV-HKU1	1.00	0.96	0.98	HCoV-HKU1	0.97	0.96	0.97
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.97	0.97	0.97
MERS-CoV	0.89	0.73	0.80	MERS-CoV	0.96	0.89	0.92
Porcine-EDV	1.00	0.17	0.29	Porcine-EDV	1.00	0.93	0.96
SARS-CoV	1.00	0.05	0.09	SARS-CoV	1.00	0.51	0.68
SARS-CoV-2	0.96	1.00	0.98	SARS-CoV-2	0.96	1.00	0.98
makro prosek	0.82	0.56	0.60	makro prosek	0.97	0.88	0.91
težinski prosek	0.96	0.96	0.95	težinski prosek	0.97	0.96	0.96

Tabela 6.28: Rezultati klasifikacije modela RF korišćenjem PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	1.00	0.60	0.75	BatCoV-HKU5	1.00	0.94	0.97
HCoV-229E	1.00	0.43	0.60	HCoV-229E	1.00	0.91	0.95
HCoV-HKU1	1.00	1.00	1.00	HCoV-HKU1	0.99	0.96	0.97
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.97	0.98	0.98
MERS-CoV	0.90	0.82	0.86	MERS-CoV	0.95	0.97	0.96
Porcine-EDV	1.00	0.33	0.50	Porcine-EDV	1.00	0.93	0.96
SARS-CoV	1.00	0.05	0.09	SARS-CoV	1.00	0.51	0.68
SARS-CoV-2	0.97	1.00	0.98	SARS-CoV-2	0.96	1.00	0.98
makro prosek	0.98	0.65	0.72	makro prosek	0.98	0.90	0.93
težinski prosek	0.97	0.97	0.96	težinski prosek	0.97	0.97	0.97

Tabela 6.29: Rezultati klasifikacije modela RF korišćenjem PB+PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.67	0.40	0.50	BatCoV-HKU5	1.00	0.84	0.92
HCoV-229E	0.00	0.00	0.00	HCoV-229E	1.00	0.89	0.94
HCoV-HKU1	1.00	0.96	0.98	HCoV-HKU1	0.97	0.97	0.97
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.98	0.97	0.98
MERS-CoV	0.88	0.64	0.74	MERS-CoV	0.93	0.92	0.93
Porcine-EDV	1.00	0.17	0.29	Porcine-EDV	1.00	0.89	0.94
SARS-CoV	1.00	0.05	0.09	SARS-CoV	1.00	0.51	0.68
SARS-CoV-2	0.96	1.00	0.98	SARS-CoV-2	0.96	1.00	0.98
makro prosek	0.81	0.53	0.57	makro prosek	0.98	0.88	0.92
težinski prosek	0.95	0.96	0.94	težinski prosek	0.96	0.96	0.96

Tabela 6.30: Rezultati klasifikacije modela RF korišćenjem PB+PB2+PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	1.00	0.40	0.57	BatCoV-HKU5	1.00	0.84	0.92
HCoV-229E	1.00	0.14	0.25	HCoV-229E	1.00	0.91	0.95
HCoV-HKU1	1.00	0.96	0.98	HCoV-HKU1	0.97	0.97	0.97
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.97	0.97	0.97
MERS-CoV	0.89	0.73	0.80	MERS-CoV	0.93	0.93	0.93
Porcine-EDV	1.00	0.33	0.50	Porcine-EDV	1.00	0.93	0.96
SARS-CoV	1.00	0.05	0.09	SARS-CoV	1.00	0.51	0.68
SARS-CoV-2	0.96	1.00	0.98	SARS-CoV-2	0.96	1.00	0.98
makro prosek	0.98	0.58	0.65	makro prosek	0.98	0.88	0.92
težinski prosek	0.96	0.96	0.95	težinski prosek	0.97	0.97	0.96

Naivni Bajesov klasifikator (NB)

Tabela 6.31: Rezultati klasifikacije modela NB korišćenjem PB na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.00	0.00	0.00	BatCoV-HKU5	0.00	0.00	0.00
HCoV-229E	0.00	0.00	0.00	HCoV-229E	0.00	0.00	0.00
HCoV-HKU1	0.00	0.00	0.00	HCoV-HKU1	1.00	0.04	0.08
HCoV-OC43	0.00	0.00	0.00	HCoV-OC43	0.33	0.02	0.03
MERS-CoV	0.00	0.00	0.00	MERS-CoV	0.00	0.00	0.00
Porcine-EDV	0.00	0.00	0.00	Porcine-EDV	0.00	0.00	0.00
SARS-CoV	0.00	0.00	0.00	SARS-CoV	0.00	0.00	0.00
SARS-CoV-2	0.91	1.00	0.96	SARS-CoV-2	0.71	1.00	0.83
makro prosek	0.11	0.12	0.12	makro prosek	0.26	0.13	0.12
težinski prosek	0.84	0.91	0.87	težinski prosek	0.58	0.71	0.60

Tabela 6.32: Rezultati klasifikacije modela NB korišćenjem PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.67	0.40	0.50	BatCoV-HKU5	0.92	0.75	0.83
HCoV-229E	1.00	0.43	0.60	HCoV-229E	1.00	0.89	0.94
HCoV-HKU1	1.00	0.87	0.93	HCoV-HKU1	0.93	0.53	0.68
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.80	0.77	0.78
MERS-CoV	0.89	0.73	0.80	MERS-CoV	0.93	0.82	0.87
Porcine-EDV	1.00	0.50	0.67	Porcine-EDV	1.00	0.89	0.94
SARS-CoV	0.10	0.05	0.06	SARS-CoV	1.00	0.42	0.59
SARS-CoV-2	0.96	0.99	0.97	SARS-CoV-2	0.92	1.00	0.96
makro prosek	0.83	0.62	0.69	makro prosek	0.94	0.76	0.82
težinski prosek	0.94	0.95	0.95	težinski prosek	0.92	0.92	0.91

Tabela 6.33: Rezultati klasifikacije modela NB korišćenjem PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	1.00	0.40	0.57	BatCoV-HKU5	0.95	0.62	0.75
HCoV-229E	1.00	0.43	0.60	HCoV-229E	1.00	0.89	0.94
HCoV-HKU1	1.00	0.91	0.95	HCoV-HKU1	0.95	0.85	0.90
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.92	0.92	0.92
MERS-CoV	0.82	0.82	0.82	MERS-CoV	0.86	0.92	0.89
Porcine-EDV	1.00	0.50	0.67	Porcine-EDV	1.00	0.93	0.96
SARS-CoV	0.25	0.05	0.08	SARS-CoV	1.00	0.47	0.63
SARS-CoV-2	0.96	0.99	0.98	SARS-CoV-2	0.95	1.00	0.98
makro prosek	0.88	0.64	0.71	makro prosek	0.96	0.83	0.87
težinski prosek	0.95	0.96	0.95	težinski prosek	0.95	0.95	0.95

Tabela 6.34: Rezultati klasifikacije modela NB korišćenjem PB+PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.67	0.40	0.50	BatCoV-HKU5	0.93	0.78	0.85
HCoV-229E	1.00	0.43	0.60	HCoV-229E	1.00	0.89	0.94
HCoV-HKU1	1.00	0.87	0.93	HCoV-HKU1	0.93	0.53	0.68
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.80	0.75	0.77
MERS-CoV	0.80	0.73	0.76	MERS-CoV	0.96	0.82	0.88
Porcine-EDV	1.00	0.50	0.67	Porcine-EDV	1.00	0.89	0.94
SARS-CoV	0.18	0.09	0.12	SARS-CoV	1.00	0.44	0.61
SARS-CoV-2	0.96	0.99	0.97	SARS-CoV-2	0.92	1.00	0.96
makro prosek	0.83	0.63	0.69	makro prosek	0.94	0.76	0.83
težinski prosek	0.94	0.95	0.95	težinski prosek	0.92	0.92	0.91

Tabela 6.35: Rezultati klasifikacije modela NB korišćenjem PB+PB2+PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	1.00	0.40	0.57	BatCoV-HKU5	0.96	0.78	0.86
HCoV-229E	1.00	0.43	0.60	HCoV-229E	1.00	0.89	0.94
HCoV-HKU1	1.00	0.96	0.98	HCoV-HKU1	0.97	0.78	0.86
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.89	0.92	0.91
MERS-CoV	0.82	0.82	0.82	MERS-CoV	0.93	0.92	0.93
Porcine-EDV	1.00	0.50	0.67	Porcine-EDV	1.00	0.91	0.95
SARS-CoV	0.43	0.14	0.21	SARS-CoV	1.00	0.47	0.63
SARS-CoV-2	0.97	0.99	0.98	SARS-CoV-2	0.95	1.00	0.97
makro prosek	0.90	0.65	0.73	makro prosek	0.96	0.83	0.88
težinski prosek	0.96	0.96	0.96	težinski prosek	0.95	0.95	0.95

Neuronska mreža (NN)

Tabela 6.36: Rezultati klasifikacije modela NN korišćenjem PB na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.67	0.40	0.50	BatCoV-HKU5	0.82	0.72	0.77
HCoV-229E	0.50	0.14	0.22	HCoV-229E	0.94	0.89	0.92
HCoV-HKU1	0.80	0.87	0.83	HCoV-HKU1	0.76	0.82	0.79
HCoV-OC43	0.70	0.50	0.58	HCoV-OC43	0.81	0.72	0.76
MERS-CoV	0.86	0.55	0.67	MERS-CoV	0.84	0.80	0.82
Porcine-EDV	1.00	0.67	0.80	Porcine-EDV	0.95	0.93	0.94
SARS-CoV	0.50	0.14	0.21	SARS-CoV	0.74	0.58	0.65
SARS-CoV-2	0.96	0.99	0.98	SARS-CoV-2	0.96	0.99	0.97
makro prosek	0.75	0.53	0.60	makro prosek	0.85	0.81	0.83
težinski prosek	0.94	0.95	0.94	težinski prosek	0.93	0.93	0.93

Tabela 6.37: Rezultati klasifikacije modela NN korišćenjem PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	1.00	0.40	0.57	BatCoV-HKU5	0.96	0.78	0.86
HCoV-229E	1.00	0.43	0.60	HCoV-229E	0.98	0.98	0.98
HCoV-HKU1	1.00	0.96	0.98	HCoV-HKU1	0.97	0.93	0.95
HCoV-OC43	0.93	1.00	0.97	HCoV-OC43	0.93	0.97	0.95
MERS-CoV	0.73	0.73	0.73	MERS-CoV	0.88	0.97	0.92
Porcine-EDV	1.00	0.50	0.67	Porcine-EDV	1.00	0.93	0.96
SARS-CoV	0.50	0.23	0.31	SARS-CoV	0.93	0.63	0.75
SARS-CoV-2	0.97	0.99	0.98	SARS-CoV-2	0.98	0.99	0.99
makro prosek	0.89	0.65	0.73	makro prosek	0.95	0.90	0.92
težinski prosek	0.96	0.96	0.96	težinski prosek	0.97	0.97	0.97

Tabela 6.38: Rezultati klasifikacije modela NN korišćenjem PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.36	1.00	0.53	BatCoV-HKU5	0.88	0.88	0.88
HCoV-229E	0.75	0.43	0.55	HCoV-229E	0.98	0.98	0.98
HCoV-HKU1	0.40	1.00	0.57	HCoV-HKU1	0.90	0.95	0.92
HCoV-OC43	0.52	1.00	0.68	HCoV-OC43	0.88	0.91	0.89
MERS-CoV	0.71	0.91	0.80	MERS-CoV	0.91	0.95	0.93
Porcine-EDV	0.20	0.67	0.31	Porcine-EDV	0.98	0.96	0.97
SARS-CoV	0.22	0.23	0.22	SARS-CoV	0.88	0.67	0.76
SARS-CoV-2	0.98	0.90	0.94	SARS-CoV-2	0.98	0.98	0.98
makro prosek	0.52	0.77	0.57	makro prosek	0.92	0.91	0.91
težinski prosek	0.93	0.89	0.90	težinski prosek	0.96	0.96	0.96

Tabela 6.39: Rezultati klasifikacije modela NN korišćenjem PB+PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	1.00	1.00	1.00	BatCoV-HKU5	0.90	0.84	0.87
HCoV-229E	1.00	0.43	0.60	HCoV-229E	1.00	0.96	0.98
HCoV-HKU1	0.96	1.00	0.98	HCoV-HKU1	0.96	0.95	0.95
HCoV-OC43	1.00	1.00	1.00	HCoV-OC43	0.94	0.97	0.95
MERS-CoV	0.80	0.73	0.76	MERS-CoV	0.92	0.93	0.93
Porcine-EDV	1.00	0.50	0.67	Porcine-EDV	1.00	0.96	0.98
SARS-CoV	0.75	0.27	0.40	SARS-CoV	0.93	0.63	0.75
SARS-CoV-2	0.97	0.99	0.98	SARS-CoV-2	0.98	1.00	0.99
makro prosek	0.94	0.74	0.80	makro prosek	0.95	0.91	0.93
težinski prosek	0.97	0.97	0.96	težinski prosek	0.97	0.97	0.97

Tabela 6.40: Rezultati klasifikacije modela NN korišćenjem PB+PB2+PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.22	0.80	0.35	BatCoV-HKU5	0.90	0.84	0.87
HCoV-229E	0.30	0.43	0.35	HCoV-229E	1.00	0.96	0.98
HCoV-HKU1	0.82	1.00	0.90	HCoV-HKU1	0.92	0.95	0.93
HCoV-OC43	0.47	1.00	0.64	HCoV-OC43	0.88	0.91	0.89
MERS-CoV	0.75	0.82	0.78	MERS-CoV	0.89	0.95	0.92
Porcine-EDV	0.50	1.00	0.67	Porcine-EDV	0.98	0.98	0.98
SARS-CoV	0.40	0.27	0.32	SARS-CoV	0.97	0.70	0.81
SARS-CoV-2	0.98	0.94	0.96	SARS-CoV-2	0.98	0.99	0.99
makro prosek	0.56	0.78	0.62	makro prosek	0.94	0.91	0.92
težinski prosek	0.94	0.92	0.93	težinski prosek	0.97	0.97	0.97

XGBoost (XGB)

Tabela 6.41: Rezultati klasifikacije modela XGB korišćenjem PB na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.00	0.00	0.00	BatCoV-HKU5	0.74	0.72	0.73
HCoV-229E	0.00	0.00	0.00	HCoV-229E	1.00	0.89	0.94
HCoV-HKU1	0.95	0.78	0.86	HCoV-HKU1	0.83	0.75	0.79
HCoV-OC43	0.83	0.36	0.50	HCoV-OC43	0.83	0.75	0.79
MERS-CoV	0.50	0.18	0.27	MERS-CoV	0.85	0.84	0.84
Porcine-EDV	1.00	0.33	0.50	Porcine-EDV	0.96	0.91	0.94
SARS-CoV	0.33	0.05	0.08	SARS-CoV	0.95	0.49	0.65
SARS-CoV-2	0.94	1.00	0.97	SARS-CoV-2	0.95	0.99	0.97
makro prosek	0.57	0.34	0.40	makro prosek	0.89	0.79	0.83
težinski prosek	0.91	0.94	0.92	težinski prosek	0.93	0.93	0.93

Tabela 6.42: Rezultati klasifikacije modela XGB korišćenjem PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.50	0.20	0.29	BatCoV-HKU5	0.86	0.78	0.82
HCoV-229E	1.00	0.57	0.73	HCoV-229E	1.00	0.95	0.97
HCoV-HKU1	0.96	0.96	0.96	HCoV-HKU1	0.96	0.99	0.97
HCoV-OC43	1.00	0.93	0.96	HCoV-OC43	0.98	0.97	0.98
MERS-CoV	0.70	0.64	0.67	MERS-CoV	0.86	0.93	0.90
Porcine-EDV	1.00	0.67	0.80	Porcine-EDV	1.00	0.96	0.98
SARS-CoV	0.67	0.18	0.29	SARS-CoV	1.00	0.58	0.74
SARS-CoV-2	0.97	1.00	0.98	SARS-CoV-2	0.98	1.00	0.99
makro prosek	0.85	0.64	0.71	makro prosek	0.96	0.90	0.92
težinski prosek	0.96	0.97	0.96	težinski prosek	0.97	0.97	0.97

Tabela 6.43: Rezultati klasifikacije modela XGB korišćenjem PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.75	0.60	0.67	BatCoV-HKU5	0.91	0.91	0.91
HCoV-229E	1.00	0.71	0.83	HCoV-229E	1.00	0.98	0.99
HCoV-HKU1	0.92	1.00	0.96	HCoV-HKU1	0.95	0.96	0.95
HCoV-OC43	0.92	0.79	0.85	HCoV-OC43	0.95	0.97	0.96
MERS-CoV	0.80	0.73	0.76	MERS-CoV	0.93	0.93	0.93
Porcine-EDV	1.00	0.83	0.91	Porcine-EDV	1.00	0.98	0.99
SARS-CoV	1.00	0.05	0.09	SARS-CoV	0.96	0.53	0.69
SARS-CoV-2	0.97	1.00	0.99	SARS-CoV-2	0.98	1.00	0.99
makro prosek	0.92	0.71	0.76	makro prosek	0.96	0.91	0.93
težinski prosek	0.97	0.97	0.96	težinski prosek	0.97	0.97	0.97

Tabela 6.44: Rezultati klasifikacije modela XGB korišćenjem PB+PB2 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	0.50	0.20	0.29	BatCoV-HKU5	0.86	0.75	0.80
HCoV-229E	1.00	0.57	0.73	HCoV-229E	1.00	0.91	0.95
HCoV-HKU1	0.96	0.96	0.96	HCoV-HKU1	0.97	0.97	0.97
HCoV-OC43	1.00	0.93	0.96	HCoV-OC43	0.97	0.97	0.97
MERS-CoV	0.70	0.64	0.67	MERS-CoV	0.85	0.93	0.89
Porcine-EDV	1.00	0.67	0.80	Porcine-EDV	0.98	0.96	0.97
SARS-CoV	0.60	0.14	0.22	SARS-CoV	0.96	0.58	0.72
SARS-CoV-2	0.97	1.00	0.98	SARS-CoV-2	0.97	1.00	0.99
makro prosek	0.84	0.64	0.70	makro prosek	0.95	0.89	0.91
težinski prosek	0.96	0.96	0.96	težinski prosek	0.97	0.97	0.96

Tabela 6.45: Rezultati klasifikacije modela XGB korišćenjem PB+PB2+PB3 na eksperimentalnom i proširenom skupu

Eksperimentalni skup				Prošireni skup			
Tip virusa	Prec.	Odziv	F1	Tip virusa	Prec.	Odziv	F1
BatCoV-HKU5	1.00	0.40	0.57	BatCoV-HKU5	0.90	0.88	0.89
HCoV-229E	1.00	0.71	0.83	HCoV-229E	1.00	0.96	0.98
HCoV-HKU1	0.92	0.96	0.94	HCoV-HKU1	0.96	0.99	0.97
HCoV-OC43	1.00	0.86	0.92	HCoV-OC43	0.98	0.98	0.98
MERS-CoV	0.82	0.82	0.82	MERS-CoV	0.93	0.92	0.93
Porcine-EDV	1.00	0.83	0.91	Porcine-EDV	1.00	0.96	0.98
SARS-CoV	0.50	0.14	0.21	SARS-CoV	1.00	0.58	0.74
SARS-CoV-2	0.97	1.00	0.99	SARS-CoV-2	0.98	1.00	0.99
makro prosek	0.90	0.71	0.77	makro prosek	0.97	0.91	0.93
težinski prosek	0.96	0.97	0.96	težinski prosek	0.97	0.97	0.97

Literatura

- [1] NCBI Protein. on-line at: <https://www.ncbi.nlm.nih.gov/protein/>, Pristupljeno: jun 2026.
- [2] RCSB Protein Data Bank. on-line at: <https://www.rcsb.org/>, Pristupljeno: jun 2026.
- [3] Analytics Vidhya. Insight into Support Vector Machine along with Code. <https://www.analyticsvidhya.com/blog/2021/04/insight-into-svm-support-vector-machine-along-with-code/>.
- [4] Jonathan Barnoud, Hubert Santuz, Pierrick Craveur, Agnel Praveen Joseph, Vincent Jallu, Alexandre G. de Brevern, and Pierre Poulain. Pbxplore: a tool to analyze local protein structure and deformability with protein blocks. *PeerJ*, 5:e4013, 2017. on-line at: <https://peerj.com/articles/4013/>.
- [5] Leo Breiman. Random forests. *Machine Learning*, 45:5--32, 2001. <https://doi.org/10.1023/A:1010933404324>.
- [6] Engelbert Buxbaum. *Fundamentals of Protein Structure and Function*. Springer International Publishing, Cham, 2 edition, 2015. <https://link.springer.com/book/10.1007/978-3-319-19920-7>.
- [7] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785--794, 2016. <https://doi.org/10.1145/2939672.2939785>.
- [8] Mary Ann Clark, Matthew Douglas, and Jung Choi. Biology 2e: 3.4 proteins. OpenStax, 2018. <https://openstax.org/books/biology-2e/pages/3-4-proteins>.

- [9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273--297, 1995. <https://link.springer.com/article/10.1007/BF00994018>.
- [10] Thomas. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21--27, 1967. <https://isl.stanford.edu/~cover/papers/transIT/0021cove.pdf>.
- [11] Jie Cui, Fang Li, and Zheng-Li Shi. Origin and evolution of pathogenic coronaviruses. *Nature Reviews Microbiology*, 17(3):181--192, 2019. <https://www.nature.com/articles/s41579-018-0118-9>.
- [12] Alexandre G. de Brevern, Catherine Etchebest, and Serge Hazout. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins: Structure, Function, and Genetics*, 41(3):271--287, 2000. <https://pubmed.ncbi.nlm.nih.gov/11025540/>.
- [13] Anthony R. Fehr and Stanley Perlman. Coronaviruses: An overview of their replication and pathogenesis. In *Coronaviruses: Methods and Protocols*, volume 1282 of *Methods in Molecular Biology*, pages 1--23. Springer, 2015. <https://pmc.ncbi.nlm.nih.gov/articles/PMC4369385/>.
- [14] GeeksforGeeks. Neural Network Node. <https://www.geeksforgeeks.org/deep-learning/neural-network-node/>.
- [15] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2 edition, 2009. <https://hastie.su.domains/ElemStatLearn/>.
- [16] Yuan. Huang, Chan. Yang, Xin-feng. Xu, Wei. Xu, and Shu-wen. Liu. Structural and functional properties of sars-cov-2 spike protein: potential antiviral drug development for covid-19. *Acta Pharmacologica Sinica*, 41:1141--1149, 2020. <https://www.nature.com/articles/s41401-020-0485-4>.
- [17] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429--449, 2002. <https://sci-hub.box/10.3233/ida-2002-6504>.

- [18] Agnel Praveen Joseph, Garima Agarwal, Shailesh Mahajan, and et al. A short survey on protein blocks. *Biophysical Reviews*, 2(3):137--147, 2010. <https://pmc.ncbi.nlm.nih.gov/articles/PMC3124139/>.
- [19] John Jumper, Richard Evans, Alexander Pritzel, and et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583--589, 2021. <https://www.nature.com/articles/s41586-021-03819-2>.
- [20] Fang Li. Structure, function, and evolution of coronavirus spike proteins. *Annual Review of Virology*, 3:237--261, 2016. <https://www.annualreviews.org/doi/10.1146/annurev-virology-110615-042301>.
- [21] Mirjana M. Maljković, Nenad S. Mitić, and Alexandre G. de Brevern. Prediction of structural alphabet protein blocks using data mining. *Biochimie*, 197:74--85, 2022. <https://www.sciencedirect.com/science/article/abs/pii/S0300908422000293>.
- [22] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. ColabFold: making protein folding accessible to all. *Nature Methods*, 19:679--682, 2022. <https://www.nature.com/articles/s41592-022-01488-1>.
- [23] Mladen Nikolić and Predrag Janičić. *Veštačka inteligencija*. Matematički fakultet, Beograd, 5 edition, 2025. Pristupljeno: jul 2026.
- [24] Bernard Offmann and Alexandre G. de Brevern. A 25-year journey with protein blocks: Unveiling the versatility of a structural alphabet. *Biochimie*, 239:58--71, 2025. <https://www.sciencedirect.com/science/article/pii/S0300908425001750>.
- [25] OpenStax. Biology for AP Courses: 3.4 proteins. <https://openstax.org/books/biology-ap-courses/pages/3-4-proteins>.
- [26] Sergey Ovchinnikov, Milot Mirdita, and Martin Steinegger. AlphaFold2_batch.ipynb. ColabFold GitHub repository. https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/batch/AlphaFold2_batch.ipynb?authuser=0#scrollTo=k0blAo-xetgx, Pristupljeno: jun 2026.

- [27] scikit-learn developers. Decision Trees. <https://scikit-learn.org/stable/modules/tree.html>. Pristupljeno: jul 2026.
- [28] scikit-learn developers. DecisionTreeClassifier. <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>. Pristupljeno: jul 2026.
- [29] scikit-learn developers. Ensemble methods. <https://scikit-learn.org/stable/modules/ensemble.html>. Pristupljeno: jul 2026.
- [30] scikit-learn developers. KNeighborsClassifier. <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>. Pristupljeno: jul 2026.
- [31] scikit-learn developers. Linear Models. https://scikit-learn.org/stable/modules/linear_model.html. Pristupljeno: jul 2026.
- [32] scikit-learn developers. LogisticRegression. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. Pristupljeno: jul 2026.
- [33] scikit-learn developers. MultinomialNB. https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html. Pristupljeno: jul 2026.
- [34] scikit-learn developers. Naive Bayes. https://scikit-learn.org/stable/modules/naive_bayes.html. Pristupljeno: jul 2026.
- [35] scikit-learn developers. Nearest Neighbors. <https://scikit-learn.org/stable/modules/neighbors.html>. Pristupljeno: jul 2026.
- [36] scikit-learn developers. Neural network models (supervised). https://scikit-learn.org/stable/modules/neural_networks_supervised.html. Pristupljeno: jul 2026.
- [37] scikit-learn developers. RandomForestClassifier. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>. Pristupljeno: jul 2026.
- [38] scikit-learn developers. Support Vector Machines. <https://scikit-learn.org/stable/modules/svm.html>. Pristupljeno: jul 2026.

- [39] Madhumati Sevvana, Thomas Klose, and Michael G. Rossmann. Principles of virus structure. In *Encyclopedia of Virology*. Elsevier, 2021. <https://pmc.ncbi.nlm.nih.gov/articles/PMC7567489/#s0015>.
- [40] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. *Introduction to Data Mining*. Pearson, 2 edition, 2019. <https://www.pearson.com/en-us/subject-catalog/p/introduction-to-data-mining/P2000000003204>.
- [41] XGBoost developers. XGBoost Documentation. <https://xgboost.readthedocs.io/>. Pristupljeno: jul 2026.

Biografija autora

Nikola Stojanović rođen 11.09.2000. u Čačku. Osnovnu školu „Branislav Petrović” i matematički smer Čačanske gimnazije završava kao nosilac obe Vukove diplome. Nakon završene srednje škole upisao je osnovne akademske studije na Matematičkom fakultetu Univerziteta u Beogradu, smer Informatika. Osnovne studije je završio 2024. godine sa prosekom 8,40 i time stekao zvanje Diplomirani informatičar. Po završetku osnovnih, upisao je master studije na istom fakultetu, takođe na smeru Informatika.

Napomena o upotrebi alata zasnovanih na veštačkoj inteligenciji

Tokom izrade ovog rada korišćen je alat *ChatGPT* kao pomoćni alat. Alat je korišćen prvenstveno za podršku u razvoju programske implementacije, uključujući organizaciju i unapređenje delova programskog koda, pomoć pri otklanjanju grešaka i pronalaženju odgovarajućih implementacionih rešenja. Predlozi dobijeni korišćenjem alata zasnovanog na veštačkoj inteligenciji nisu preuzimani bez dodatne provere. Ispravnost programskog koda proveravana je pokretanjem implementiranih funkcionalnosti, kontrolom dobijenih izlaza i poređenjem rezultata sa očekivanim ponašanjem. Teorijska objašnjenja i metodološke tvrdnje proveravani su na osnovu stručne literature navedene u bibliografiji, kao i na osnovu stvarno dobijenih rezultata eksperimenata. Svi delovi uključeni u rad dodatno su provereni, korigovani i prilagođeni temi master rada.