

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1355

**KLASIFIKACIJA PATOGENOSTI MUTACIJA GENA BRCA1 I
BRCA2 METODAMA STROJNOG UČENJA**

Dora Bilić-Pavlinović

Zagreb, lipanj 2026.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1355

**KLASIFIKACIJA PATOGENOSTI MUTACIJA GENA BRCA1 I
BRCA2 METODAMA STROJNOG UČENJA**

Dora Bilić-Pavlinović

Zagreb, lipanj 2026.

DIPLOMSKI ZADATAK br. 1355

Pristupnica: **Dora Bilić-Pavlinović (0036538609)**

Studij: Računarstvo

Profil: Znanost o podacima

Mentor: prof. dr. sc. Alan Jović

Zadatak: **Klasifikacija patogenosti mutacija gena BRCA1 i BRCA2 metodama strojnog učenja**

Opis zadatka:

Geni BRCA1 i BRCA2 imaju ključnu ulogu u održavanju stabilnosti genoma i popravku oštećenja DNA, dok njihove mutacije predstavljaju jedan od najznačajnijih čimbenika nasljednog rizika za razvoj karcinoma dojke i jajnika. U ovom radu potrebno je primijeniti i vrednovati metode strojnog učenja za preciznu klasifikaciju mutacija unutar navedenih gena u klinički relevantne kategorije: patogene ili benigne. Središnji problem istraživanja je automatizacija analize varijanti nepoznatog značaja (VUS) čija interpretacija izravno utječe na dijagnostičke postupke. U istraživanju se trebaju koristiti podaci integrirani iz javnih baza BRCA Exchange i NCBI (ClinVar), uz poseban naglasak na bioinformatičke značajke poput vjerojatnosti poremećaja splajsinga (splice scores). U okviru rada potrebno je usporediti performanse algoritama logističke regresije, stroja potpornih vektora, slučajne šume i XGBoosta. Analizom važnosti pojedinih značajki potrebno je utvrditi koliki je doprinos specifičnih bioloških i statističkih parametara točnosti predviđanja. Konačni modeli trebaju se vrednovati uobičajenim klasifikacijskim metrikama s naglaskom na postizanju visoke osjetljivosti u prepoznavanju štetnih mutacija.

Rok za predaju rada: 26. lipnja 2026.

Zahvala

Ovim putem bih htjela iskreno zahvaliti onima koji su bili uz mene kroz proteklih 5 godina. Hvala mojoj obitelji koja me uvijek podržavala i posebno mami koja je bila tu da me tješi i bodri kada je bilo najteže. Također se želim zahvaliti prijateljima koji su mi bili glavni poticaj za učenje i održavali me budnom kroz mnoge čitave noći učenja. Na kraju bih se iskreno htjela zahvaliti i svome mentoru, Alanu Joviću za velikodušnu pomoć i posvećenost kroz seminarske radove i ovaj diplomski rad.

Sadržaj

Sažetak.....	1
Summary.....	2
Uvod	3
1. Biološka pozadina	5
1.1. Osnovni pojmovi genetike.....	5
1.2. Uloga gena BRCA1 i BRCA2 u očuvanju genomske stabilnosti.....	6
1.3. Tipovi mutacija i funkcionalne posljedice.....	7
1.4. Zašto su mutacije <i>missense</i> u fokusu računalnog modeliranja?	9
1.5. Patogene varijante i cjeloživotni rizik	9
1.6. Problem varijanti nepoznatog značenja i uloga računalnih metoda	10
1.7. Računalni prediktori i uloga bioinformatičkih značajki	10
2. Podaci i metodologija	12
2.1. Izvori podataka	12
2.1.1. BRCA Exchange (Skup A).....	12
2.1.2. ClinVar (Skup B).....	12
2.1.3. Usporedba i metodološka povezanost izvora	14
2.2. Inženjerstvo značajki i priprema podataka	15
2.2.1. Definicija i priprema ciljne varijable.....	15
2.2.2. Genomska i strukturna obilježja varijanti.....	16
2.2.3. Integracija populacijskih frekvencija i <i>in silico</i> prediktora (gnomAD v4.1)	17
2.2.4. Problem nedostajućih podataka i odabir značajki	19
2.2.5. Sprječavanje cirkularnosti i curenja podataka	20
2.2.6. Specifični podskup: Inženjerstvo značajki za podskup varijanti <i>missense</i> ..	21
2.3. Postavljanje modela i hiperparametri	24

2.3.1.	Podjela podataka i rješavanje disbalansa klasa	24
2.3.2.	Sprječavanje curenja podataka	24
2.3.3.	Arhitektura i hiperparametri algoritama strojnoga učenja.....	25
2.3.4.	Metoda unakrsne validacije	25
3.	Pregled modela strojnog učenja.....	26
3.1.	Logistička regresija.....	26
3.2.	Metoda XGBoost.....	27
3.3.	Slučajna šuma	28
3.4.	Stroj potpornih vektora.....	29
4.	Rezultati i evaluacija modela.....	32
4.1.	Dizajn istraživanja	32
4.2.	Prvi stupanj: Globalna evaluacija na cjelokupnom skupu BRCA Exchange	33
4.3.	Drugi stupanj: Specijalizirana analiza mutacija <i>missense</i> skupa A.....	35
4.4.	Treći stupanj: Modeliranje i unakrsna validacija na neovisnom skupu ClinVar.	37
4.5.	Četvrti stupanj: Vanjska validacija na razriješenim varijantama VUS	39
5.	Rasprava	41
5.1.	Ograničenja analize na miješanim skupovima podataka.....	41
5.2.	Usporedba linearnih i nelinearnih modela.....	41
5.3.	Klinička primjenjivost i analiza pogrešaka	42
5.4.	Ograničenja istraživanja	43
5.5.	Usporedba sa srodnim istraživanjima.....	43
	Zaključak	47
	References	49
	Skraćenice.....	52
	Izjava o korištenju umjetne inteligencije.....	53

Sažetak

Točna klinička interpretacija genetskih varijanti gena *BRCA1* i *BRCA2* ključna je za procjenu nasljednog rizika od karcinoma dojke i jajnika. Značajan izazov u praksi predstavljaju varijante nepoznatog značenja (engl. *Variants of Uncertain Significance*, VUS), posebice mutacije krivog smisla (engl. *missense*), čiji funkcionalni učinak nije moguće odrediti jednostavnim pravilima ranog prekida translacije. U ovom radu razvijen je i vrednovan cjevovod strojnog učenja za binarnu klasifikaciju patogenosti ovih genetskih varijanti integracijom podataka iz baza BRCA Exchange i ClinVar s populacijskim frekvencijama i naprednim *in silico* prediktorima iz baze gnomAD v4.1. Uspoređene su performanse logističke regresije, stroja potpornih vektora, slučajne šume i XGBoosta kroz četiri stupnja evaluacije. Rezultati pokazuju da nelinearni ansambli stabala odluke znatno nadmašuju linearne modele na podskupu mutacija krivog smisla, pri čemu XGBoost ostvaruje ROC-AUC od 0,9502 na unakrsnoj validaciji te točnost od 91,3 % u vanjskoj reklasifikaciji VUS varijanti, što je usporedivo s rezultatima iz srodnih istraživanja. Predloženi pristup demonstrira primjenjivost strojnog učenja kao potencijalno korisnog pomoćnog alata za kliničku trijažu i prioritizaciju genetskih varijanti.

Ključne riječi: *BRCA1*, *BRCA2*, strojno učenje, XGBoost, slučajna šuma, varijante nepoznatog značenja (VUS), mutacije *missense*, gnomAD, bioinformatika.

Summary

Accurate clinical interpretation of genetic variants in *BRCA1* and *BRCA2* genes is essential for assessing hereditary breast and ovarian cancer risks. A major challenge in clinical genetics arises from Variants of Uncertain Significance (VUS), particularly missense mutations, whose functional impact cannot be readily determined by simple loss-of-function rules. In this thesis, a comprehensive machine learning pipeline was developed and evaluated for binary pathogenicity classification of these genetic variants by integrating data from BRCA Exchange and ClinVar repositories with population frequencies and advanced *in silico* predictors from gnomAD v4.1. The performance of Logistic Regression, Support Vector Machines, Random Forest, and XGBoost was benchmarked across a four-stage evaluation design. The results demonstrate that non-linear tree-based ensemble models significantly outperform linear baselines on the missense subset, with XGBoost achieving an ROC-AUC of 0.9502 in cross-validation and a 91.3% accuracy in external reclassification of resolved VUS variants, which is comparable to results of related research. The proposed pipeline highlights the effectiveness of machine learning as a supportive decision-aid tool for variant triage and prioritization.

Keywords: *BRCA1*, *BRCA2*, machine learning, XGBoost, Random Forest, Variants of Uncertain Significance (VUS), missense mutations, gnomAD, bioinformatics.

Uvod

Razvojem i širom primjenom tehnologija sekvenciranja nove generacije (engl. *Next-Generation Sequencing*, NGS), u suvremenoj onkologiji i medicinskoj genetici naglo je porasla količina dostupnih genomskih podataka. Iako je sekvenciranje postalo znatno brže i pristupačnije, točna interpretacija biološkog i kliničkog značaja identificiranih varijanti i dalje predstavlja velik izazov. Taj je problem posebno izražen kod gena *BRCA1* i *BRCA2*, čije mutacije su vezane za rizik razvoja karcinoma dojke i jajnika, ali i drugih oblika raka poput karcinoma prostate i gušterače [1] [2] [3].

Geni *BRCA1* i *BRCA2* djeluju kao tumorski supresori uključeni u popravak dvolančanih prekida DNA. Patogene varijante povezane su s visokim cjeloživotnim rizikom od razvoja malignih bolesti (pri čemu rizik za karcinom dojke doseže do 87 % za nositelje mutacije u genu *BRCA1*), dok benigne varijante predstavljaju neutralni genetski polimorfizam koji ne narušava funkciju proteina. Međutim, više od trećine varijanti u kliničkim bazama podataka trenutno je klasificirano kao varijante nepoznatog značenja (engl. *Variants of Uncertain Significance*, VUS). Postojanje takvih varijanti otežava donošenje kliničkih odluka o profilaktičkim zahvatima (npr. mastektomiji i ooforektomiji) ili primjeni ciljane terapije [1].

Tradicionalna klasifikacija prema službenim smjernicama zahtijeva opsežne epidemiološke, obiteljske i laboratorijske funkcionalne podatke. Kako laboratorijska validacija zahtijeva mnogo vremena i resursa, metode strojnog učenja postaju ključan alat za automatiziranu i bržu procjenu patogenosti [1] [3].

Cilj ovog diplomskog rada jest razviti, vrednovati i usporediti modele strojnog učenja za binarnu klasifikaciju mutacija gena *BRCA1* i *BRCA2* u patogene i benigne varijante. Istraživanje se temelji na podacima preuzetima iz javnih repozitorija ClinVar (NCBI) i BRCA Exchange. Poseban naglasak stavljen je na integraciju sveobuhvatnog skupa bioinformatičkih značajki, od populacijskih frekvencija i agregiranih *in silico* prediktora štetnosti (CADD, REVEL, PolyPhen-2) do alata za predikciju poremećaja prekrajanja (SpliceAI, Pangolin).

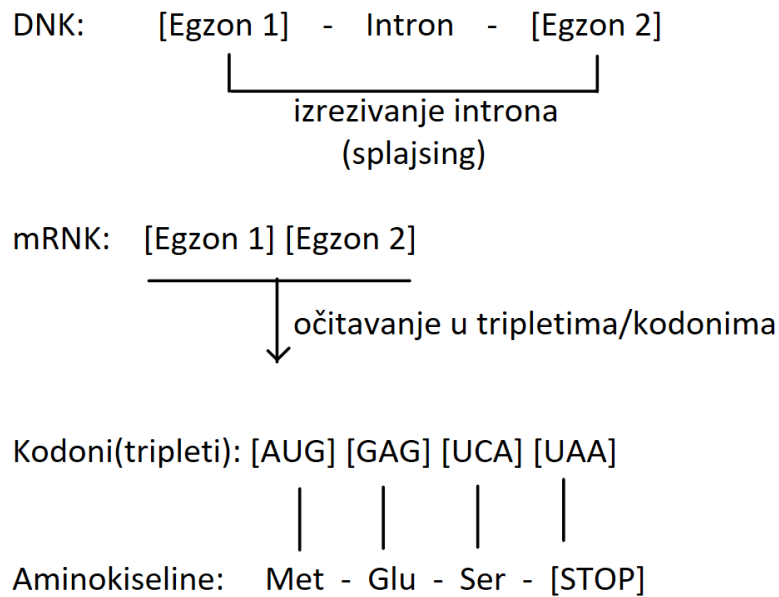
U radu su odabrana, primijenjena i uspoređena četiri algoritma različite složenosti: logistička regresija, stroj potpornih vektora (SVM), slučajna šuma (engl. *Random Forest*) i ekstremno pojačavanje modela gradijentom (*XGBoost*). Pripremom specifičnih značajki i analizom njihove važnosti u modelima utvrđeno je u kojoj mjeri pojedini biološki prediktori i populacijske frekvencije doprinose točnosti klasifikacije. Modeli su evaluirani primjenom unakrsne validacije i metrika prilagođenih neuravnoteženim klasama.

1. Biološka pozadina

1.1. Osnovni pojmovi genetike

Radi boljeg razumijevanja tematike ovog rada, najprije će se definirati temeljni biološki pojmovi i mehanizmi relevantni za interpretaciju varijanti.

- **DNK i nukleotidi:** Molekula DNK nosi genetsku uputu organizma, a sastoji se od dvaju komplementarnih lanaca građenih od četiriju različitih nukleotidnih baza: adenina (A), citozina (C), gvanina (G) i timina (T). [4]
- **Geni, egzoni i introni:** Gen je funkcionalni odsječak DNK koji nosi uputu za sintezu proteina. U ljudskom genomu geni se sastoje od:
 - **Egzona:** kodirajućih regija koje sadrže informaciju za sintezu proteina.
 - **Introna:** nekodirajućih regija smještenih između egzona. Tijekom prijepisa genetske informacije u glasničku RNK (mRNK), introni se uklanjaju procesom prekrajanja ili splajsinga (engl. *splicing*), čime se egzoni povezuju u neprekinuti kodirajući slijed [4].
- **Kodoni (tripleti) i pretvaranje u aminokiseline:** Genetski kod čita se u grupama od po tri uzastopna nukleotida (kodonima) na molekuli mRNK. Za razliku od molekule DNK u kojoj se nalazi timin (T), u molekuli RNK timin je zamijenjen bazom uracilom (U). Od ukupno $4^3 = 64$ moguća kodona, 61 kodon kodira jednu od 20 standardnih aminokiselina dok preostala 3 stop-kodona (UAA, UAG i UGA) ne kodiraju aminokiseline, već služe kao signal za završetak translacije i otpuštanje sintetiziranog proteina (4), što je shematski prikazano na Slici 1.1.



Slika 1.1 Prikaz toka sinteze proteina

1.2. Uloga gena **BRCA1** i **BRCA2** u očuvanju genomske stabilnosti

Geni *BRCA1* (*Breast Cancer 1*) i *BRCA2* (*Breast Cancer 2*) tumorski su supresori čiji proteini sudjeluju u popravku oštećenja DNA i održavanju stabilnosti genoma [5].

Gen *BRCA1* smješten je na kromosomu 17, sastoji se od 24 egzona te kodira protein od 1863 aminokiseline. U njegovoj strukturi ključne su dvije domene:

- N-terminalna domena (pozicije 1–109): veže se na protein BARD1 i sudjeluje u ranoj signalizaciji oštećenja DNA.
- C-terminalna domena (pozicije 1650–1863): omogućuje interakciju s proteinima koji reguliraju stanični odgovor na oštećenja genoma. [5] [6]

Gen *BRCA2* nalazi se na kromosomu 13, obuhvaća 27 egzona te kodira protein od 3418 aminokiselina. Ključne regije proteina su:

- BRC regija (egzon 11): sadrži osam ponavljajućih motiva koji izravno vežu protein RAD51, neophodan za popravak DNA.

- Domena za vezanje DNA (C-terminalni dio): omogućuje vezanje proteina *BRCA2* na jednolančanu DNA [5] [6].

Oštećenja u obliku dvolančanih prekida DNA (engl. *Double-Strand Breaks*, DSB) stanica može popraviti dvama mehanizmima: nehomolognim spajanjem krajeva (engl. *Non-Homologous End Joining*, NHEJ) i homolognom rekombinacijom (engl. *Homologous Recombination*, HR) [7].

Dok je NHEJ sklon pogreškama jer spaja krajeve bez predloška, HR predstavlja precizan mehanizam koji koristi intaktnu sestrinsku kromatidu kao kalup. Proteini *BRCA1* i *BRCA2* ključni su za funkciju HR-a: *BRCA1* sudjeluje u prepoznavanju oštećenja i pripremi krajeva DNA, dok *BRCA2* usmjerava *RAD51* na mjesto prekida. Gubitkom funkcije bilo kojeg od ovih proteina mehanizam HR-a zakazuje, a stanica prelazi na NHEJ, što dovodi do nakupljanja mutacija i povećanog rizika od maligne transformacije [6].

1.3. Tipovi mutacija i funkcionalne posljedice

Ovisno o utjecaju na kodon i sintezu proteina, mutacije u kodirajućim regijama dijele se u dvije osnovne skupine:

1) Trunkirajuće mutacije (engl. *Loss-of-Function*)

Trunkirajuće mutacije uzrokuju rani prekid sinteze ili nastanak skraćenog proteina. U ovu skupinu ubrajaju se:

- Besmislene mutacije (engl. *Nonsense mutations*): Mutacija zamjenjuje jedan nukleotid unutar kodona tako da se kodon koji je kodirao aminokiselinu pretvara u jedan od tri stop-kodona. Ribosom pri translaciji uočava rani stop-kodon i prekida daljnu sintezu proteina. Ovo rezultira nepotpunim proteinom koji gubi svoju normalnu funkciju.
- Mutacije s pomakom okvira čitanja (engl. *Frameshift mutations*): Nastaju ubacivanjem ili brisanjem nukleotida (iznimka je kada je broj dodanih/brisanih nukleotida djeljiv s 3). Budući da se mRNK čita u grupama od po tri nukleotida, pomak od samo jednog mjesta mijenja značenje svih sljedećih kodona naniže. To dovodi do ugradnje pogrešnih aminokiselina te većinom uzrokuje pojavu ranog stop-kodona ubrzo nakon mjesta mutacije.

- Kanonske mutacije mjesta prekrajanja (engl. *canonical splice site*): Mutacije na rubovima introna (na njihovom početku ili kraju). Zbog oštećenja tih rubnih mjesta stanica ne može pravilno prepoznati gdje intron počinje i završava, pa greškom izreže koristan egzon (engl. *exon skipping*) ili ostavi dio introna unutar mRNK (engl. *intron retention*), što narušava slijed čitanja. [8]

Stanica posjeduje zaštitni mehanizam NMD (engl. *Nonsense-Mediated mRNA Decay*) koji prepoznaje prisutnost ranih stop-kodona u mRNK lancu. Ako NMD prepozna rani stop-kodon, označava tu mRNK za potpunu razgradnju prije nego što započne translacija. Time se sprječava stvaranje potencijalno štetnih skraćenih proteina. Međutim, ako mutacija izbjegne NMD razgradnju, nastaje skraćeni protein kojemu nedostaju ključne funkcionalne domene [8] [4].

2) Mutacije krivog smisla (engl. *Missense mutations*)

Mutacija krivog smisla (dalje: mutacija *missense*) nastaje zamjenom pojedinačnog nukleotida u kodonu, što uzrokuje ugradnju drugačije aminokiseline u polipeptidni lanac, dok okvir čitanja i duljina proteina ostaju nepromijenjeni.

- Izvorni kodon: [GAG] → Glutaminska kiselina (Glu)
- Mutirani kodon: [GTG] → Valin (Val)

Učinak mutacije *missense* ovisi o poziciji unutar proteina i fizikalno-kemijskim svojstvima zamijenjene aminokiseline:

- Benigni učinak: javlja se ako se zamjena dogodi u varijabilnoj/neosjetljivoj regiji ili ako nova aminokiselina ima slična svojstva (naboj, polarnost, hidrofobnost).
- Patogeni učinak: nastaje ako se zamjena dogodi unutar ključne funkcionalne domene ili ako znatna promjena svojstava aminokiseline naruši prostornu strukturu, stabilnost proteina ili interakciju s drugim proteinima u mehanizmu homologne rekombinacije [8].

1.4. Zašto su mutacije *missense* u fokusu računalnog modeliranja?

Većina istraživanja i prediktivnih modela u bioinformatici usmjerena je isključivo na klasifikaciju varijanti *missense*. Izdvajanje podskupa mutacija *missense* temeljeno je na dva ključna razloga:

Sprječavanje pristranosti modela: Trunkirajuće mutacije uzrokuju znatne strukturne promjene i u pravilu su patogene (gotovo 100 % slučajeva). Uključivanje takvih varijanti u skup za učenje omogućilo bi modelima postizanje visokih metrika oslanjanjem isključivo na binarni indikator prisutnosti ranog stop-kodona. Takav model ne bi naučio suptilne biokemijske zakonitosti potrebne za točnu klasifikaciju varijanti *missense* kod kojih je duljina proteina očuvana [9].

ACMG/AMP smjernice i kriterij PVS1: Prema smjernicama Američkog koledža za medicinsku genetiku i genomiku (ACMG) i Udruženja za molekularnu patologiju (AMP), trunkirajuće mutacije u genima *BRCA1* i *BRCA2* izravno zadovoljavaju najjači kriterij patogenosti (PVS1). Za njih je klinička interpretacija izravna i ne zahtijeva primjenu prediktivnih modela. Za varijante *missense* pravilo PVS1 ne vrijedi, pa se za njihovu procjenu patogenosti moraju koristiti i slabiji neizravni dokazi (uključujući računalne predikcije) [8].

1.5. Patogene varijante i cjeloživotni rizik

Mutacije u genima *BRCA1* i *BRCA2* nasljeđuju se autosomno dominantno, što znači da srodnici u prvom stupnju (djeca, braća, sestre) imaju 50 % vjerojatnosti nositeljstva.

Nositeljstvo patogene mutacije značajno povećava cjeloživotni rizik od razvoja karcinoma:

- Karcinom dojke: U općoj populaciji cjeloživotni rizik iznosi oko 12 %. Kod nositeljica patogene mutacije u genu *BRCA1* rizik iznosi 65 % - 72 %, a za *BRCA2* 45 % - 69 % [10, 11].

- Karcinom jajnika: U općoj populaciji cjeloživotni rizik iznosi 1.2%. Kod nositeljica mutacije u genu *BRCA1* rizik raste na 39 % - 44 %, a za *BRCA2* na 11 % - 17 % [10, 11].
- Povezani karcinomi: Povišen je i rizik od razvoja karcinoma gušterače, prostate te muškog raka dojke [5, 11].

1.6. Problem varijanti nepoznatog značenja i uloga računalnih metoda

Unatoč napretku u kliničkom sekvenciranju i prikupljanju podataka u javnim repozitorijima, značajan udio identificiranih varijanti u genima *BRCA1* i *BRCA2* i dalje ostaje u kategoriji varijanti nepoznatog značenja. Za ove VUS varijante ne postoji dovoljno epidemioloških, obiteljskih ili funkcionalnih dokaza kako bi ih se pouzdano svrstalo u patogene ili benigne [1].

Khandakji i Mifsud (2022) ističu da varijante *missense* nepoznatog značenja čine približno 40–50 % svih registriranih varijanti *missense* u genima *BRCA1* i *BRCA2* u bazama poput ClinVara [1]. Nemogućnost pouzdane interpretacije ovih varijanti stvara veliku nesigurnost u kliničkom odlučivanju jer se pacijentima ne mogu sa sigurnošću preporučiti profilaktički kirurški zahvati (poput mastektomije ili ooforektomije) niti usmjerene ciljane terapije (1). Budući da su *in vitro* eksperimentalna ispitivanja iznimno skupa i dugotrajna, primjena metoda strojnog učenja predstavlja ključni pristup za automatiziranu procjenu njihove patogenosti [1].

1.7. Računalni prediktori i uloga bioinformatičkih značajki

Službene smjernice ACMG-a priznaju predikcije računalnih (*in silico*) alata kao pomoćne dokaze u klasifikaciji varijanti.

Suvremeni algoritmi strojnog učenja i računalni prediktori transformiraju biološke i genomske informacije o mutacijama *missense* u numeričke značajke koje se primarno dijele u nekoliko ključnih kategorija:

- **Evolucijska konzerviranost:** temelji se na pretpostavci da su aminokiselinske pozicije i nukleotidi sačuvani kroz evoluciju od ključne važnosti za funkciju proteina. Mjere poput ocjena PhyloP [12] kvantificiraju stopu evolucijske konzerviranosti na temelju višestrukog poravnanja genoma, pri čemu pozitivne vrijednosti ukazuju na pročišćavajuću selekciju i visoku očuvanost nukleotida.
- **Pojedinačni alati i fizikalno-kemijski indeksi:** klasični prediktori poput SIFT-a [13] i PolyPhen-2 [14] procjenjuju utjecaj aminokiselinske zamjene kombinirajući evolucijski profil s fizikalno-kemijskim svojstvima (promjena naboja, polarnosti i hidrofobnosti). Pritom SIFT računa vjerojatnost tolerancije na skali od 0 do 1 uz invertiranu logiku (vrijednosti $\leq 0,05$ označavaju štetne mutacije), dok PolyPhen-2 generira ocjene u rasponu od 0 do 1 gdje veće vrijednosti upućuju na oštećenje funkcije proteina.
- **Agregirani (*ensemble*) prediktori:** okviri poput REVEL-a [15] i CADD-a [16] integriraju ocjene više pojedinačnih alata i genomskih anotacija u jedinstvenu ocjenu štetnosti, čime se smanjuje šum pojedinačnih modela. REVEL je specijaliziran za rijetke varijante *missense* s kontinuiranim rasponom [0,1], dok CADD izražava štetnost kroz normalizirani PHRED-skor gdje vrijednost iznad 20 označava varijantu među 1 % najštetnijih mogućih promjena u humanom genomu.
- **Predikcija utjecaja na izrezivanje introna (splajsing):** pojedine mutacije *missense* unutar egzona mogu stvoriti nova ili narušiti postojeća mjesta izrezivanja. Specijalizirani modeli temeljeni na dubokom učenju, poput alata SpliceAI [17] i Pangolin [18], procjenjuju vjerojatnost poremećaja donorskih i akceptorskih mjesta prekrajanja kroz ocjene u rasponu [0,1], čime omogućuju detekciju patogenih varijanti koje narušavaju obradu mRNK unatoč očuvanoj duljini polipeptidnog lanca.

Integracija ovako strukturiranog skupa značajki unutar modela strojnog učenja omogućuje točniju predikciju i pruža temelj za sustavnu reklasifikaciju VUS varijanti [1, 8].

2. Podaci i metodologija

2.1. Izvori podataka

U ovom radu korištena su dva primarna izvora kliničkih i genomskih podataka o varijantama gena *BRCA1* i *BRCA2*: **BRCA Exchange (Skup A)** i **ClinVar (Skup B)** (obogaćen podacima iz populacijske baze gnomAD v4.1).

2.1.1. BRCA Exchange (Skup A)

Baza BRCA Exchange (verzija v3.139.0) objedinjuje podatke o varijantama iz više repozitorija [19], uključujući konzorcij ENIGMA, ClinVar, LOVD/exLOVD, BIC te populacijske baze (1000 Genomes, ExAC, ESP).

Sirovi skup podataka sadrži ukupno 17.770 varijanti opisanih s 89 atributa (8.578 za *BRCA1* i 9.192 za *BRCA2*). Za definiranje binarne ciljne varijable korištena je isključivo oznaka *Clinical_significance_ENIGMA* jer ENIGMA konzorcij provodi stručnu klasifikaciju prema strogim kriterijima uz najvišu razinu dokaza.

Od ukupnog broja analiziranih zapisa, ENIGMA konsenzusnu oznaku *Benign* (0) ili *Pathogenic* (1) nosi 3.284 instanci (985 benignih i 2.299 patogenih; 1.530 za *BRCA1* i 1.754 za *BRCA2*). Taj podskup čini primarni skup za učenje i testiranje (Skup A). Unutar njega, varijanti *missense* ima točno 196. Relativno mali udio klasificiranih varijanti *missense* posljedica je konzervativnih kriterija konzorcija ENIGMA, koji većinu takvih mutacija zadržava u kategoriji VUS uslijed nedostatka nedvojbenih eksperimentalnih dokaza.

U sirovom obliku, baza ima nisku popunjenost pojedinih značajki: populacijske frekvencije iz ExAC-a dostupne su za 19,4 % varijanti, iz baze 1000 Genomes za 32,5 %, dok su klasični *in silico* prediktori (PolyPhen, SIFT) popunjeni za samo 27,3 % zapisa.

2.1.2. ClinVar (Skup B)

Baza ClinVar (NCBI) javni je repozitorij kliničkih klasifikacija varijanti koje prijavljuju dijagnostički laboratoriji, istraživačke skupine i stručni paneli [20]. Korištenje ClinVara kao drugog skupa podataka omogućuje evaluaciju modela na znatno većem broju varijanti i

provjeru robusnosti u realnim, šumovitijim uvjetima gdje podaci potječu od heterogenih podnositelja.

Za potrebe istraživanja preuzet je cjeloviti tablični izvoz baze (variant_summary.txt.gz) filtriran na simbole gena *BRCA1* i *BRCA2*, što je dalo 75.132 primjeraka. Radi usklađenosti s referentnom bazom gnomAD v4.1, zadržani su isključivo unosi preslikani na referentni sklop GRCh38. Nakon uklanjanja 308 zapisa s dupliciranim genomskim koordinatama, dobiven je početni skup od 36.863 jedinstvene varijante (15.303 za *BRCA1* i 21.560 za *BRCA2*).

Svaki zapis u bazi ClinVar opisan je dvjema ključnim, međusobno neovisnim dimenzijama:

1. Razina pouzdanosti priloženih dokaza (*ReviewStatus*): sustav vrednovanja koji kvantificira stupanj recenzije i konsenzusa među prijaviteljima, a formalno se preslikava u ocjenu zvjezdica (od 0 do 3 zvjezdice). Distribucija svih početnih 37.171 GRCh38 unosa prema ovom kriteriju prikazana je u Tablici 2.1.
2. Klinički značaj varijante (*ClinicalSignificance*): stvarna prijavljena interpretacija patogenosti varijante (npr. *Pathogenic*, *Benign*, *Uncertain significance*, *Likely benign* i dr.).

Tablica 2.1 Distribucija varijanti u bazi ClinVar prema razini dokaza (engl. *review status*)

Status pregleda (<i>Review status</i>)	Sustav zvjezdica	Broj varijanti
<i>Reviewed by expert panel</i> (stručni panel)	3 ★★★	7.590
<i>Criteria provided, multiple submitters, no conflicts</i> (bez konflikata)	2 ★★	6.368
<i>Criteria provided, single submitter</i> (jedan podnositelj)	1 ★	11.768
<i>Criteria provided, conflicting classifications</i> (konfliktni navodi)*	1* ★	8.659
Bez navedenih kriterija / bez klasifikacije	0	2.786

* Napomena: Konfliktne klasifikacije formalno nose 1 zvjezdicu, ali predstavljaju izravno neslaganje među podnositeljima te su izbačene pri definiranju ciljne varijable.

Kako bi se spriječio unos šuma u oznakama (*label noise*), odabir instanci za učenje modela nije proveden pukim filtriranjem po broju zvjezdica, već kombinacijom strogih semantičkih kriterija iz stupca ClinicalSignificance i eliminacijom konfliktnih unosa:

- U skup su uvršteni isključivo jednoznačni unosi s nedvojbenim statusom: strogo Pathogenic (pozitivna klasa 1) ili strogo Benign (negativna klasa 0).
- Sve varijante s nejasnim, kompozitnim ili prijelaznim oznakama — uključujući *Uncertain significance* (VUS), *Likely pathogenic*, *Likely benign*, unose s kosom crtom (npr. *Pathogenic/Likely pathogenic*) te varijante bez priložene klasifikacije isključene su iz primarnog skupa za učenje.
- Time su automatski eliminirani i svi konflikti zapisi (*Criteria provided, conflicting classifications*), jer oni po definiciji sadrže oprečne navode različitih laboratorija.

U konačnom klasificiranom Skupu B, odabrane varijante potječu isključivo iz kategorija s usklađenim dokazima (kategorije s 1, 2 ili 3 zvjezdice bez konflikata).

Primjenom ovog strogog kriterija čistoće ciljne varijable, konačni Skup B obuhvaća točno 10.063 varijante (1.529 benignih i 8.534 patogene; 4.343 za *BRCA1* i 5.720 za *BRCA2*).

Unutar ovog filtriranog skupa nalazi se točno 443 mutacije *missense* s jednoznačnim statusom (Pathogenic ili Benign). Pod pojmom „nedvojben klinički status” podrazumijeva se da su sve odabrane varijante *missense* u stupcu ClinicalSignificance nosile isključivo čistu oznaku patogenosti ili benignosti, bez konfliktnih tvrdnji među podnositeljima i bez svrstavanja u prijelaznu kategoriju VUS.

2.1.3. Usporedba i metodološka povezanost izvora

Dva korištena izvora imaju komplementarnu ulogu u eksperimentima: Skup A (ENIGMA) služi kao visoko pouzdan i stručno verificiran skup za primarni razvoj modela, dok Skup B (ClinVar) pruža znatno veći obujam podataka za provjeru generalizacije na širem i heterogenijem uzorku. Usporedni pregled svojstava prikazan je u tablici 2.2 .

Tablica 2.1.2 Usporedni pregled baza BRCA Exchange (Skup A) i ClinVar (Skup B)

Svojstvo	BRCA Exchange (Skup A)	ClinVar (Skup B)
Sirovi broj varijanti	17.770	36.863
Klasificirani podskup (Ciljna varijabla)	3.284	10.063
Omjer klasa (Benigno : Patogeno)	30% : 70%	15% : 85%
Izvor i pouzdanost labela	ENIGMA konsenzus (stručni panel)	Heterogeni podnosiitelji (0–3 zvjezdice)
Klasificirane varijante <i>missense</i>	196	443
Primarni <i>in silico</i> prediktori (sirovo)	Niska pokrivenost (PolyPhen/SIFT ~27%)	Bez primarnih prediktora zbog čega se obogaćuje s dodatnim podacima (gnomAD, vidjeti u nastavku)

2.2. Inženjerstvo značajki i priprema podataka

Inženjerstvo značajki (engl. *feature engineering*) važan je korak dubinske analize podataka u kojem se sirovi genomski, proteinski i populacijski podaci transformiraju u strukturirane numeričke attribute pogodne za algoritme strojnog učenja, uz očuvanje biološkog konteksta i sprječavanje curenja podataka (engl. *data leakage*).

2.2.1. Definicija i priprema ciljne varijable

Radi uklanjanja šuma u oznakama (*label noise*), problem je postavljen kao binarna klasifikacija uz strogi kriterij odabira:

- **Klasa 1 (Patogena):** varijante službeno označene isključivo kao *Pathogenic*.
- **Klasa 0 (Benigna):** varijante službeno označene isključivo kao *Benign*.

Varijante nepoznatog značenja (VUS), unosi s konfliktnim tvrdnjama te granične kategorije (*Likely pathogenic*, *Likely benign*) namjerno su izostavljeni iz skupa za učenje. Time se osigurava visoka pouzdanost i čistoća ciljne varijable.

Distribucija uzoraka u pripremljenim skupovima iznosi:

- **Skup A (BRCA Exchange):** 3.284 varijante (2.299 patogenih / 70,0 % i 985 benignih / 30,0 %).
- **Skup B (ClinVar):** Nakon deduplikacije prema jedinstvenim koordinatama (*Kromosom : Pozicija : Referentni alel : Alternativni alel*), dobiveno je 10.063 klasificiranih varijanti (8.534 patogene / 84,8 % i 1.529 benignih / 15,2 %).

2.2.2. Genomska i strukturna obilježja varijanti

Iz sirovih genomskih i proteinskih HGVS zapisa izveden je osnovni skup značajki:

1. **Tip genomske varijante (variant_type):** određen usporedbom duljina referentnog (ref_length) i alternativnog (alt_length) alela:
 - *Single Nucleotide Variant* (SNV): $\text{len(Ref)} = 1 \wedge \text{len(Alt)} = 1$
 - *Insertion*: $\text{len(Ref)} < \text{len(Alt)}$
 - *Deletion*: $\text{len(Ref)} > \text{len(Alt)}$
 - *Multi-Nucleotide Variant* (MNV): $\text{len(Ref)} = \text{len(Alt)} > 1$

Razlika u duljini alela zabilježena je kao značajka variant_length_diff (alt_length – ref_length).

2. **Tip proteinske posljedice (protein_change_type):** izveden parsiranjem HGVS zapisa proteinske promjene (polja HGVS_Protein ili Name). Zbog specifičnosti nomenklature, parsiranje je provedeno strogo definiranim redoslijedom:
 - Mutacija s pomakom okvira čitanja (engl. *frameshift*): prepoznaje se po oznaci fs (budući da ovaj HGVS zapis često sadrži i poziciju novog stop-kodona, npr. p.Lys6ArgfsTer7, provjera za fs mora prethoditi provjeri za stop-kodon kako se varijanta ne bi pogrešno klasificirala kao besmislena).
 - Besmislena mutacija (engl. *nonsense*): prepoznaje se po oznaci * ili Ter (uveden preuranjeni stop-kodon bez pomaka okvira čitanja).

- Delecija / Insercija / Duplikacija (engl. *deletion* / *insertion* / *duplication*): prepoznaju se po oznakama del, ins, odnosno dup.
- Mutacija krivog smisla (engl. *missense*): supstitucija jedne aminokiseline, pri čemu se podržavaju zapisi sa zagradama (uobičajeno u BRCA Exchange, npr. p.(Arg18His)) i bez unutarnjih zagrada (uobičajeno u ClinVar, npr. p.Cys61Gly).
- Sinonimna ili tiha mutacija (engl. *synonymous*): prepoznaje se po oznaci =.
- Nepoznato ili nekodirajuće (engl. *unknown*): varijante bez izravnog zapisa aminokiselinske promjene (intronske i nekodirajuće regije, regije prekrajanja).

3. **Genomski kontekst:** indikatorske binarne varijable is_brca1 i is_brca2 te numerička oznaka kromosoma chr_numeric (17 za *BRCA1*, 13 za *BRCA2*).

2.2.3. Integracija populacijskih frekvencija i *in silico* prediktora (gnomAD v4.1)

Pregled i prednosti baze gnomAD v4.1

Genome Aggregation Database (gnomAD) trenutačno je najveći i najreferentniji javno dostupni resurs za populacijsku genomiku [21]. Verzija **gnomAD v4.1** obuhvaća sekvencirane podatke preko 807.000 pojedinaca usklađenih na GRCh38 referentni genom, pri čemu integrira opće populacijske alelne frekvencije i bogat spektar modernih *in silico* alata za procjenu štetnosti varijanti.

Za razliku od ranijih verzija koje su bile ograničene na manje uzorke, starije sklopove genoma (GRCh37/hg19) ili isključivo na jednostavne populacijske frekvencije, cjevovod gnomAD v4.1 pruža objedinjen pristup najvažnijim prediktorima:

- **Populacijska dinamika i selekcijski pritisak:** obuhvaća opće alelne frekvencije (gnomad_af) i maksimalne filtrirane alelne frekvencije s 95 % pouzdanosti (gnomad_faf95_max), koje predstavljaju temelj za primjenu službenih kriterija benignosti ACMG/ClinGen (BA1/BS1).

- **Prediktori za zamjene aminokiselina (alati *missense*):** integrira ansambl i pojedinačne alate visoke razlučivosti poput REVEL-a (gnomad_revel_max), PolyPhen-2 (gnomad_polyphen_max) i SIFT-a (gnomad_sift_max).
- **Opći *in silico* prediktori i očuvanost:** obuhvaća ocjene štetnosti CADD (gnomad_cadd) te ocjene evolucijske konzerviranosti nukleotida PhyloP kroz višestruko poravnanje 100 vrsta kralježnjaka (gnomad_phylop).
- **Prediktori poremećaja prekrajanja (*splice site*):** sadrži napredne modele temeljene na dubokom učenju kao što su SpliceAI (gnomad_spliceai_ds_max) i Pangolin (gnomad_pangolin_largest_ds).

Pristup integraciji vanjskih podataka u ovom radu metodološki je podijeljen prema primarnom izvoru podataka:

- **Skup A (BRCA Exchange):** Zadržava izvorno dostupne podatke (1000 Genomes, ExAC, PolyPhen, SIFT). Podaci iz baze gnomAD v4.1 namjerno nisu spajani kako bi se spriječila cirkularnost i curenje oznaka (*label leakage*) — ENIGMA konzorcij u sklopu baze BRCA Exchange već implicitno koristi populacijske frekvencije pri ručnom kuriranju i dodjeli Bayesovih vjerojatnosti patogenosti.
- **Skup B (ClinVar):** Budući da ClinVar agregira neovisne i heterogene podneske različitih laboratorija bez ujednačenog Bayesovog modela, provedeno je programsko spajanje s gnomAD v4.1 GraphQL API-jem preko jedinstvenih genomskih ključeva (*Kromosom : Pozicija : Referentni alel : Alternativni alel*).

Za razliku od tradicionalnih repozitorija gdje su *in silico* ocjene primarno ograničene na supstitucije *missense*, gnomAD v4.1 omogućuje dohvat općih značajki (CADD, PhyloP, SpliceAI i Pangolin) za cjelokupan spektar varijanti u genu, uključujući intronske regije i područja prekrajanja.

Metodološko i empirijsko opravdanje integracije na podskupu *missense*

Nužnost integracije podataka gnomAD najizraženija je na podskupu varijanti *missense* u skupu ClinVar. Sa stajališta inženjerstva značajki, unutar podskupa *missense* cjelokupan skup osnovnih genomskih obilježja gubi diskriminativnu moć. Varijable poput tipa varijante (*variant_type* = SNV), duljine alela (*ref_length* = 1, *alt_length* = 1) ili tipa proteinske promjene (*protein_change_type* = *missense*) postaju konstante za sve uzorke.

Posljedično, bez integracije prediktora iz gnomAD-a, jedino preostalo informativno obilježje na ovom podskupu bila bi indikatorska varijabla gena (is_brca1 u odnosu na is_brca2). Uključivanjem naprednih *in silico* ocjena i populacijskih frekvencija modelima je omogućen uvid u funkcionalne posljedice i evolucijski kontekst svake pojedine zamjene aminokiseline.

2.2.4. Problem nedostajućih podataka i odabir značajki

Rukovanje nedostajućim vrijednostima predstavlja značajan metodološki izazov u pripremi genomskih podataka. Udio nedostajućih vrijednosti znatno varira ovisno o tipu varijable i njezinu biološkom kontekstu:

- **Kategoričke i izravno izvedene genomske značajke** (ref_length, protein_change_type, is_brca1, chr_numeric itd.) nemaju nedostajućih vrijednosti (0,0 % praznina u oba skupa podataka).
- **Populacijske frekvencije i *in silico* prediktori** odlikuju se visokim udjelom nedostajućih podataka (između 85 % i 96 %). Primjerice, unutar strogo klasificiranog skupa ClinVar:
 - Opći prediktori i frekvencije (gnomad_af, gnomad_cadd, gnomad_spliceai_ds_max, gnomad_phylop) nedostaju u približno 85,9 % zapisa.
 - Prediktori specifični isključivo za varijante *missense* (gnomad_revel_max, gnomad_sift_max, gnomad_polyphen_max) nedostaju u čak 96,8 % slučajeva na razini cjelokupnog skupa.

Visok udio nedostajućih vrijednosti u ovim varijablama nije posljedica tehničkih nedostataka pri ekstrakciji podataka, već proizlazi iz bioloških i metodoloških zakonitosti:

1. **Rijetkost patogenih varijanti:** Visoko penetrantne i štetne mutacije izrazito su rijetke u općoj populaciji, zbog čega se prirodno ne pojavljuju u referentnim populacijskim bazama zdravih pojedinaca (poput gnomAD-a ili 1000 Genomesa), pa im je alelna frekvencija nedefinirana.
2. **Usko definirana domena alata:** Pojedinačni prediktori poput PolyPhen-2, SIFT-a ili REVEL-a algoritamski su dizajnirani isključivo za procjenu utjecaja zamjene jedne aminokiseline. Za trunkirajuće (besmislene, s pomakom okvira čitanja),

sinonimne ili nekodirajuće varijante ove vrijednosti biološki i matematički ne postoje.

Odabir značajki za opći model:

Kako bi se izbjegla distorzija u modelima osjetljivima na masovnu zamjenu nedostajućih podataka, za razvoj općeg modela klasifikacije cjelokupnog spektra varijanti postavljen je formalni prag od **maksimalno 50 % nedostajućih vrijednosti** za zadržavanje pojedine značajke.

Nakon filtriranja varijabli prema ovom kriteriju i primjene kodiranja *one-hot* na kategoričke varijable (variant_type i protein_change_type, uz parametar drop_first=True radi izbjegavanja multikolinearnosti), konačni skup za opći model na Skupu A (BRCA Exchange) obuhvaća **14 značajki**:

- **Genomski kontekst:** is_brca1, is_brca2, chr_numeric
- **Duljine alela:** ref_length, alt_length, variant_length_diff
- **Tip proteinske promjene (one-hot):** protein_change_type_frameshift, protein_change_type_missense, protein_change_type_nonsense, protein_change_type_synonymous, protein_change_type_unknown
- **Tip varijante (one-hot):** variant_type_SNV, variant_type_deletion, variant_type_insertion

2.2.5. Sprječavanje cirkularnosti i curenja podataka

Tijekom odabira značajki i pripreme podataka posebna je pažnja posvećena eliminaciji cirkularnosti i metodološke pristranosti.

U bazi BRCA Exchange dio varijanti dobiva konsenzusnu ENIGMA klasifikaciju primjenom multifaktorijskog Bayesovog modela. Taj model izravno agregira obiteljsku segregaciju (Segregation_LR_exLOVD), istovremeno pojavljivanje s patogenim mutacijama *in trans* (Co_occurrence_LR_exLOVD) te izračunate posteriorne vjerojatnosti (Posterior_probability_exLOVD). Uključivanje takvih atributa u skup za učenje omogućilo bi modelima rekonstrukciju samog pravila dodjele oznake (engl. *label reconstruction*) umjesto prepoznavanja bioloških i molekularnih mehanizama patogenosti. Zbog toga su sva Bayesova polja iz repozitorija exLOVD, kao i izvedene binarne zastavice o postojanju funkcionalnih dokaza (has_functional_evidence), strogo isključeni iz skupa značajki.

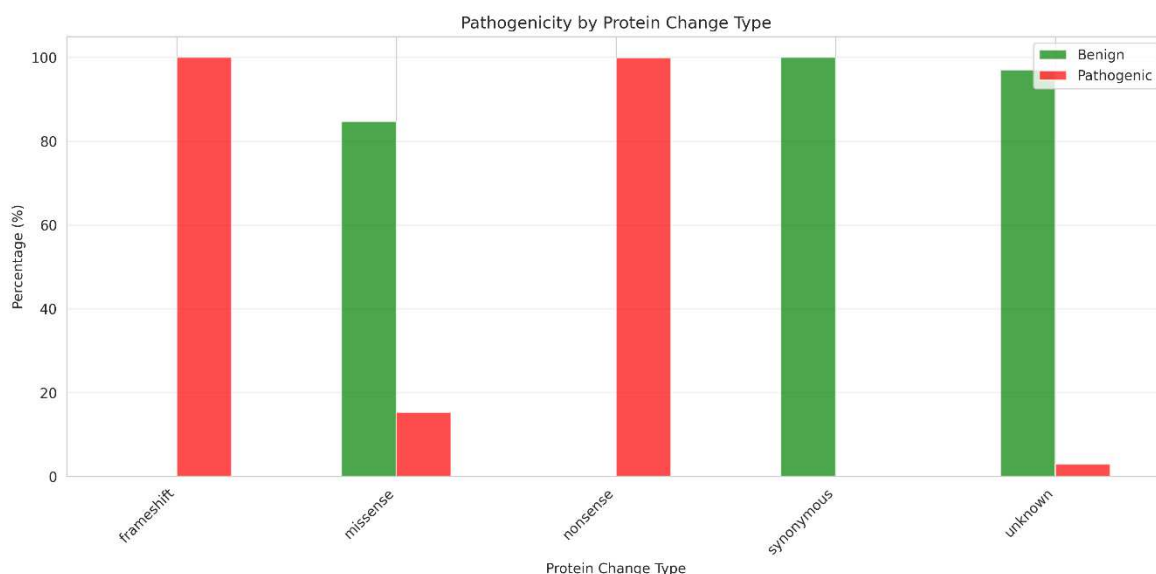
U skupu ClinVar zadržane su populacijske frekvencije i s njima povezani prediktori kako bi se modelirali realni uvjeti kliničke interpretacije, uz uvažavanje činjenice da niske populacijske frekvencije čine standardni pomoćni kriterij prema ACMG/ClinGen smjernicama (BA1/BS1).

Kako bi se spriječilo curenje informacija tijekom cjelokupnog cjevovoda predobrade podataka:

- **Izolirana podjela podataka:** Podjela na skup za učenje (80 %) i testiranje (20 %) provedena je stratificirano prije bilo kakve statističke transformacije ili obrade podataka.
- **Zasebno prilagođavanje transformacija:** Imputacija nedostajućih vrijednosti medijanom (`SimpleImputer`) i standardizacija značajki (`StandardScaler`) prilagođene su (`fit`) isključivo na skupu za učenje, a zatim primijenjene (`transform`) na testni skup i pojedinačne validacijske pregrade unutar unakrsne validacije.

2.2.6. Specifični podskup: Inženjerstvo značajki za podskup varijanti *missense*

Analizom kliničkog značaja u ovisnosti o tipu proteinske promjene (Slika 2.1) uočava se jasna biološka pravilnost koja opravdava izdvajanje varijanti *missense* u zaseban analitički fokus.



Slika 2.1 Distribucija patogenosti varijanti u bazi podataka BRCA Exchange prema tipu promjene na razini proteina.

Sve varijante s pomakom okvira čitanja (*frameshift*) i besmislene varijante (*nonsense*) u bazi podataka klasificirane su kao 100 % patogene uslijed ranog prekida translacije i potpunog gubitka funkcije proteina. S druge strane, sinonimne varijante (*synonymous*) u potpunosti pripadaju benignoj klasi jer ne mijenjaju primarni slijed aminokiselina.

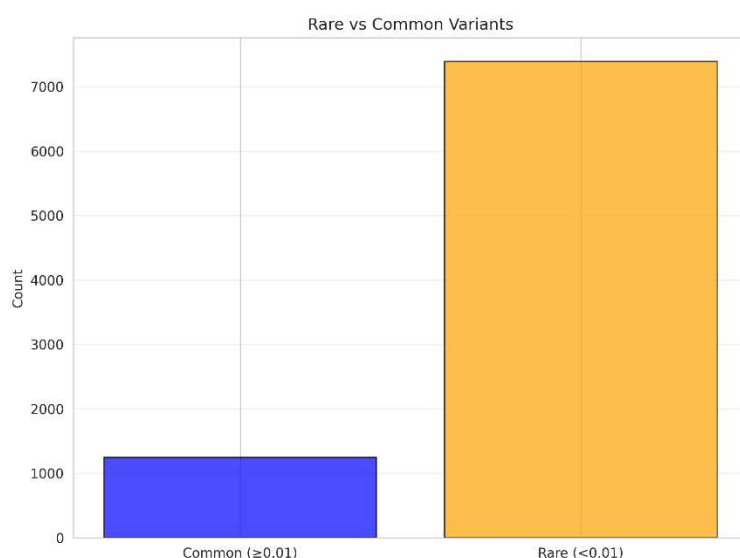
Jedina kategorija koja pokazuje izrazitu heterogenost jesu varijante *missense*, gdje približno 85 % varijanti pripada benignoj klasi, a oko 15 % patogenoj. Dok se klasifikacija na cjelokupnom skupu podataka svodi na prepoznavanje tipa mutacije (pri čemu modeli trivijalno reproduciraju ACMG PVS1 kriterij za trunkirajuće mutacije), stvarni bioinformatički izazov leži u razlikovanju patogenih od benignih mutacija *missense* na temelju suptilnih biokemijskih i evolucijskih obilježja.

Unutar izoliranog podskupa *missense*, opće genomske značajke (poput `ref_length`, `alt_length` ili `variant_type_SNV`) postaju neinformativne konstante jer su sve varijante *missense* jednonukleotidne supstitucije (SNV) duljine 1. S druge strane, alati za procjenu funkcionalnog učinka aminokiselinske zamjene (*in silico* prediktori) unutar ovog definiranog podskupa postižu visoku pokrivenost (od 72 % do 100 %).

Zbog toga je za analizu mutacija *missense* definiran specijalizirani skup od **10 značajki** koji obuhvaća:

- **Genomski kontekst:** indikator gena (`is_brca1`)

- **Populacijske frekvencije:** opća alelna frekvencija (gnomad_af) i maksimalna filtrirana alelna frekvencija (gnomad_faf95_max)
- **Prediktore za zamjene aminokiselina:** ocjene alata REVEL (gnomad_revel_max), PolyPhen-2 (gnomad_polyphen_max) i SIFT (gnomad_sift_max)
- **Opću štetnost i evolucijsku konzerviranost:** ocjena CADD (gnomad_cadd) i stupanj očuvanosti PhyloP (gnomad_phylop)
- **Prediktore poremećaja splajsinga:** ocjena SpliceAI (gnomad_spliceai_ds_max) i ocjena Pangolin (gnomad_pangolin_largest_ds)



Slika 2.2 Raspodjela rijetkih i čestih varijanti u skupu podataka BRCA Exchange prema pragu populacijske frekvencije.

U skladu sa smjernicama ACMG/AMP, populacijska frekvencija predstavlja jedan od najpouzdanijih filtara za utvrđivanje benignosti varijanti. Iz raspodjele prikazane na Slici 2.2 vidljivo je da velika većina patogenih varijanti u analiziranom skupu pripada izrazito rijetkim promjenama s maksimalnom zabilježenom frekvencijom alela manjom od 0,01 (AF < 1 %).

Varijante koje se javljaju s višom frekvencijom u općoj populaciji redovito su benigne, budući da visoko penetrantne patogene mutacije u genima *BRCA1* i *BRCA2* podliježu snažnoj negativnoj selekciji. Ovaj uvid potvrđuje opravdanost uključivanja maksimalne populacijske frekvencije kao ključnog negativnog prediktora patogenosti u cjevovodu za učenje modela, čime se omogućuje precizna evaluacija na biološki najzahtjevnijem podskupu bez distorzije uzrokovane prisutnošću očitih trunkirajućih mutacija.

2.3. Postavljanje modela i hiperparametri

Za potrebe učenja i testiranja modela strojnog učenja definiran je jedinstven cjevovod (engl. *machine learning pipeline*) koji objedinjuje korake predobrade, skaliranja i optimizacije modela.

2.3.1. Podjela podataka i rješavanje disbalansa klasa

Stratificirana podjela: Skupovi podataka podijeljeni su na skup za učenje (80 %) i testni skup (20 %). Podjela je provedena stratificirano u odnosu na ciljnu varijablu kako bi se očuvao jednak omjer patogenih i benignih varijanti u oba podskupa (`random_state=42`).

Balansiranje težina klasa: S obzirom na izrazitu neuravnoteženost između broja patogenih i benignih varijanti, u svim je modelima primijenjena penalizacija pogrešaka na manjinskoj klasi. Za modele implementirane kroz biblioteku `scikit-learn` korišten je ugrađeni parametar `class_weight='balanced'`. Za algoritam XGBoost izračunat je parametar `scale_pos_weight` koji predstavlja egzaktni omjer broja negativnih i pozitivnih instanci u skupu za učenje.

2.3.2. Sprječavanje curenja podataka

Arhitektura cjevovoda (engl. *pipeline*): Kako bi se spriječilo curenje statističkih informacija iz preklopa unakrsne validacije (engl. *cross-validation folds*) i testnog skupa u fazu učenja, cjelokupna predobrada podataka i modeliranje integrirani su u jedinstvenu arhitekturu cjevovoda.

Zamjena nedostajućih vrijednosti i skaliranje: Nedostajuće vrijednosti zamijenjene su medijanom odgovarajuće značajke primjenom razreda `SimpleImputer`. Za modele logističke regresije i stroja potpornih vektora dodatno je primijenjena standardizacija vrijednosti (z-skor pomoću razreda `StandardScaler`). Svi parametri zamjene vrijednosti i skaliranja prilagođavaju se (metoda `fit`) isključivo na podacima za učenje unutar svakog pojedinog preklopa unakrsne validacije (odnosno na skupu za učenje pri fiksnoj podjeli), a potom se transformacija (metoda `transform`), bez ponovnog izračuna statistika, primjenjuje na ispitni preklap unakrsne validacije i testni skup.

2.3.3. Arhitektura i hiperparametri algoritama strojnoga učenja

U radu su korištena četiri algoritma s fiksnim hiperparametrima:

- **Logistička regresija:** Maksimalni broj iteracija optimizacijskog postupka postavljen je na `max_iter=1000` radi osiguranja konvergencije algoritma na standardiziranim podacima.
- **Slučajna šuma:** Ansambl je konfiguriran na 100 stabala odluke (`n_estimators=100`). Radi kontrole složenosti modela i sprječavanja prenaučivosti, maksimalna dubina stabla ograničena je na `max_depth=10`, uz minimalno 5 uzoraka za grananje čvora (`min_samples_split=5`) i minimalno 2 uzorka u krajnjem listu (`min_samples_leaf=2`).
- **XGBoost:** Korišteno je 100 stabala (`n_estimators=100`) uz maksimalnu dubinu `max_depth=6` i stopu učenja $\eta = 0,1$ (`learning_rate=0.1`). Kao kriterij optimizacije korištena je funkcija logaritamskog gubitka (`eval_metric='logloss'`). Dodatna stohastička regularizacija postignuta je nasumičnim odabirom 80 % uzoraka (`subsample=0.8`) i 80 % značajki (`colsample_bytree=0.8`) pri izgradnji svakog stabla.
- **SVM:** Upotrijebljen je uz radijalnu baznu funkciju (kernel RBF) i parametar regularizacije $C = 1,0$. Širina jezgre određena je automatskim skaliranjem (`gamma='scale'`). Uključen je parametar `probability=True` kako bi se omogućio izračun vjerojatnosnih procjena potrebnih za ROC-AUC i PR-AUC analizu.

2.3.4. Metoda unakrsne validacije

S obzirom na ograničeni broj potvrđenih varijanti *missense*, procjena stabilnosti i performansi modela provedena je primjenom stratificirane 5-struke unakrsne validacije (engl. *5-fold Stratified Cross-Validation*).

Ovaj postupak osigurava iterativno ispitivanje kroz cijeli skup podataka uz očuvanje izvornog omjera klasa u svakom preklopu. Konačni rezultati prikazani su kroz srednju vrijednost i standardnu devijaciju ocjenjivanih metrika kroz svih pet preklopa.

3. Pregled modela strojnog učenja

Klasifikacija patogenosti genetskih mutacija u ovom se radu provodi nad strukturiranim tabličnim podacima dobivenim iz bioinformatičkih repozitorija i populacijskih baza. Kako bi se usporedili različiti pristupi učenju granice odluke u prostoru genomskih i biokemijskih značajki, odabrana su četiri modela: od jednostavnijih parametarskih modela s linearnom granicom do nelinearnih ansambala na bazi stabala odluke te modela utemeljenih na teoriji statističkog učenja i maksimizaciji margine.

3.1. Logistička regresija

Logistička regresija predstavlja temeljni parametarski algoritam za zadatke binarne klasifikacije. U ovom radu binarna logistička regresija služi kao osnovni model (*baseline*) za procjenu vjerojatnosti da pojedina mutacija u genima *BRCA1* i *BRCA2* bude patogena ($y = 1$) ili benigna ($y = 0$).

Model pretpostavlja da je logit vjerojatnosti linearan u odnosu na ulazne značajke. Vjerojatnost pripadnosti pozitivnoj klasi računa se primjenom logističke (sigmoidalne) funkcije $\sigma(z)$ nad linearnom kombinacijom ulaznih varijabli :

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (3.1)$$

gdje je:

$$z = \mathbf{w}^T \mathbf{x} + b = w_1 x_1 + w_2 x_2 + \dots + w_d x_d + b \quad (3.2)$$

U ovom izrazu $\mathbf{w} \in \mathbb{R}^d$ je vektor težinskih koeficijenata pridruženih značajkama, $\mathbf{x} \in \mathbb{R}^d$ je ulazni vektor bioinformatičkih značajki pojedine varijante, a $b \in \mathbb{R}$ je slobodni član (*bias*) [22].

Glavna prednost logističke regresije u bioinformatički leži u transparentnosti i izravnoj interpretabilnosti parametara. Svakoj značajki pridružen je koeficijent w_j , pri čemu eksponent koeficijenta (e^{w_j}) odgovara omjeru šansi (engl. *Odds Ratio*, OR) (12). Pozitivna vrijednost koeficijenta znači da porast vrijednosti te značajke povećava šansu da je varijanta patogena, dok negativan predznak upućuje na zaštitni, odnosno benigni karakter. Na taj način model omogućuje izravnu provjeru usklađenosti naučenih težina s biološkim znanjem (npr. očekuje se da visoka alelna frekvencija u populaciji ima negativan koeficijent i smanjuje vjerojatnost patogenosti).

Osim jednostavne interpretacije, algoritam se brzo uči, lako konvergira i pokazuje stabilno ponašanje na manjim skupovima podataka, što je čest slučaj kod filtriranih podskupova genetskih varijanti. Ipak, osnovno ograničenje modela je pretpostavka o linearnoj granici odlučivanja. Logistička regresija ne može samostalno uhvatiti nelinearne interakcije između različitih alata (npr. situaciju gdje je varijanta patogena samo ako su i CADD i SpliceAI istovremeno iznad određenog praga), zbog čega u radu primarno služi kao referentna točka za usporedbu sa složenijim, nelinearnim algoritmima.

3.2. Metoda XGBoost

Ekstremno pojačavanje modela gradijentom (engl. *Extreme Gradient Boosting*, (i dalje) XGBoost) napredni je algoritam strojnog učenja temeljen na tehnici pojačavanja modela stabala odluke gradijentom. Za razliku od metoda paralelnog združivanja, pojačavanje model gradijentom gradi sekvencijalni ansambl u kojem se stabla odluke dodaju jedno za drugim, pri čemu svako novo stablo uči na rezidualima, odnosno pogreškama koje su napravila prethodna stabla.

Proces dodavanja stabala u svakom koraku t prikazuje se izrazom:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(\mathbf{x}_i) \quad (3.3)$$

pri čemu su:

- $\hat{y}_i^{(t)}$ predikcija za instancu i nakon dodavanja t -tog stabla (koje je optimirano tako da minimizira zadanu funkciju gubitka),
- $f_t(\mathbf{x}_i)$ predikcija t -tog stabla za ulazni vektor $\mathbf{x}_i$,
- η stopa učenja— udio kojim svako novo stablo pridonosi ukupnoj predikciji i kontrolira brzinu konvergencije.

Konačna predikcija ansambla sastavljenog od ukupno T stabala izračunava se na sljedeći način:

$$\hat{y}_i = \sum_{t=1}^T \eta f_t(x_i) \quad (3.4)$$

Kod binarne klasifikacije ovaj se konačni zbroj prevodi u kalibriranu vjerojatnost patogenosti primjenom standardne sigmoidalne funkcije $\sigma(\hat{y}_i)$ [23].

U analizi mutacija gena *BRCA1* i *BRCA2*, XGBoost je posebno pogodan jer efikasno obrađuje tablične podatke s izraženim razlikama u distribucijama značajki. Algoritam aproksimira funkciju gubitka pomoću Taylorovog razvoja drugog reda, što osigurava brzu i stabilnu optimizaciju. Također, u samu funkciju cilja ugrađena je L_1 i L_2 regularizacija složenosti stabala (visine stabla i vrijednosti težina u listovima), što sprječava model da memorira šum u podacima. U kontekstu genetskih značajki, XGBoost je u stanju automatski modelirati složene nelinearne međuovisnosti (primjerice, kombinaciju očuvanosti regije i promjene naboja aminokiseline) bez potrebe za ručnim kreiranjem interakcijskih varijabli.

3.3. Slučajna šuma

Slučajna šuma (engl. *Random Forest*) je nelinearni ansambl algoritam temeljen na tehnici paralelnog združivanja modela (engl. *Bagging — Bootstrap Aggregating*). Za razliku od sekvencijalnog pristupa kod XGBoosta, slučajna šuma paralelno gradi veći broj međusobno neovisnih stabala odluke. Svako stablo u šumi uči se na zasebnom *bootstrap* uzorku,

odnosno podskupu podataka dobivenom slučajnim odabirom primjeraka iz skupa za učenje s ponavljanjem.

Kako bi se smanjila međusobna koreliranost stabala i osigurala što veća raznolikost ansambla, algoritam pri svakom grananju čvora ne pretražuje cijeli prostor značajki, već bira optimalnu podjelu unutar slučajno odabranog podskupa atributa (uobičajeno $\sqrt{d}$, gdje je d ukupan broj značajki).

Konačna odluka ansambla donosi se većinskim glasanjem ili usrednjavanjem vjerojatnosnih procjena svih izgrađenih stabala [24]:

$$P(y = 1|\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B P_b(y = 1|\mathbf{x}) \quad (3.5)$$

gdje B označava ukupan broj stabala u šumi, a $P_b(y = 1|\mathbf{x})$ vjerojatnost pozitivne klase dodijeljenu od strane b -tog stabla.

Glavna teorijska prednost slučajne šume u odnosu na pojedinačna stabla leži u drastičnom smanjenju varijance modela bez povećanja pristranosti (engl. *bias*). Slučajna šuma pokazuje izrazitu robusnost prema šumu i stršecim vrijednostima (engl. *outlier*), a njezina struktura ne zahtijeva prethodno skaliranje ili normalizaciju numeričkih varijabli. Također, model pruža prirodan način procjene relativne važnosti značajki (engl. *Feature Importance*) praćenjem prosječnog smanjenja Ginijeve nečistoće kroz sva grananja u šumi, što daje izravan uvid u to koji bioinformatički prediktori nose najveću diskriminativnu moć pri razdvajanju patogenih od benignih varijanti.

3.4. Stroj potpornih vektora

Stroj potpornih vektora (engl. *Support Vector Machine*, (i dalje) SVM) deterministički je klasifikacijski algoritam utemeljen na principu strukturne minimizacije rizika i teoriji statističkog učenja. Osnovni cilj SVM-a je u prostoru značajki pronaći optimalnu hiperravninu koja razdvaja točke dviju klasa uz maksimalnu marginu, odnosno najveću udaljenost od hiperravnine do najbližih primjeraka skupa za učenje, koji se nazivaju potpornim vektorima.

U idealnom, linearno odvojivom slučaju s tvrdom marginom (engl. *hard margin*), hiperravnina odluke definirana je relacijom:

$$\mathbf{w}^T \mathbf{x} + b = 0 \quad (3.6)$$

uz uvjet ispravnog razdvajanja svih instanci skupa:

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad \forall i \in \{1, \dots, n\} \quad (3.7)$$

gdje $\mathbf{w}$ predstavlja vektor težina koji definira orijentaciju hiperravnine, b je slobodni član, $\mathbf{x}_i$ ulazni vektor značajki, a $y_i \in \{-1, +1\}$ binarna oznaka pripadnosti klasi.

Budući da su u realnim genomskim podacima klase redovito preklapajuće i opterećene šumom, koristi se formulacija s mekom marginom (engl. *soft margin*). Uvode se nenegativne varijable popuštanja $\xi_i \geq 0$ koje dopuštaju određenu razinu pogreške ili ulazak točaka unutar same margine:

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i \quad (3.8)$$

Optimizacijski zadatak svodi se na minimizaciju konveksne funkcije:

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \quad (3.9)$$

pri čemu regularizacijski parametar $C > 0$ kontrolira kompromis između maksimizacije širine margine i penalizacije točaka koje krše uvjete razdvajanja.

Za rješavanje nelinearno odvojivih problema primjenjuje se metoda jezgrenih funkcija (engl. *kernel trick*), koja implicitno preslikava podatke u višedimenzionalni prostor u kojem postaje moguće linearno razdvajanje. U ovom radu koristi se radijalna bazna funkcija (engl. *Radial Basis Function*, RBF) [22]:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (3.10)$$

gdje parametar $\gamma > 0$ određuje radijus utjecaja pojedinačnog potpornog vektora na oblik granice odluke. SVM s jezgrom RBF vrlo je efikasan na podacima manje i srednje veličine te pokazuje visoku otpornost na prenaučavanje u višedimenzionalnom prostoru, no računski je zahtjevniji pri izračunu vjerojatnosti te zahtijeva obavezno prethodno skaliranje svih ulaznih značajki.

4. Rezultati i evaluacija modela

U ovom su poglavlju prikazani rezultati eksperimenata provedenih u četiri uzastopna stupnja. Cilj ovakvog strukturiranja bio je sustavno ispitati ponašanje modela: od osnovne globalne klasifikacije svih tipova mutacija, preko specijalizirane analize varijanti *missense*, do provjere generalizacije na neovisnim podacima i mogućnosti primjene na varijantama nepoznatog značenja (VUS).

4.1. Dizajn istraživanja

Eksperimentalni postupak podijeljen je u četiri stupnja kako bi se postupno izolirali specifični bioinformatički izazovi:

1. Globalna evaluacija na skupu BRCA Exchange: Modeliranje cjelokupnog spektra varijanti (trunkirajuće, mutacije krivog smisla, tihe i nekodirajuće) korištenjem osnovnih genomskih značajki. Svrha ovog stupnja bila je provjeriti sposobnost modela da prepoznaju opće biološke zakonitosti (poput utjecaja ranog stop-kodona).
2. Analiza mutacija *missense* u skupu BRCA Exchange: Izolacija podskupa varijanti *missense* s oznakama ENIGMA. Ovaj stupanj poslužio je za procjenu ponašanja modela kada se uklone trivijalno rješive trunkirajuće mutacije.
3. Modeliranje i unakrsna validacija na bazi ClinVar: Učenje i 5-struka unakrsna validacija na znatno većem podskupu jednoznačno klasificiranih varijanti *missense* uz integraciju proširenog skupa *in silico* i populacijskih značajki iz baze gnomAD v4.1.
4. Vanjska validacija na razriješenim VUS varijantama: Evaluacija modela učenog na bazi ClinVar na neovisnom testnom skupu povijesnih VUS varijanti iz baze BRCA Exchange koje su u međuvremenu eksperimentalno razriješene i reklasificirane.

4.2. Prvi stupanj: Globalna evaluacija na cjelokupnom skupu BRCA Exchange

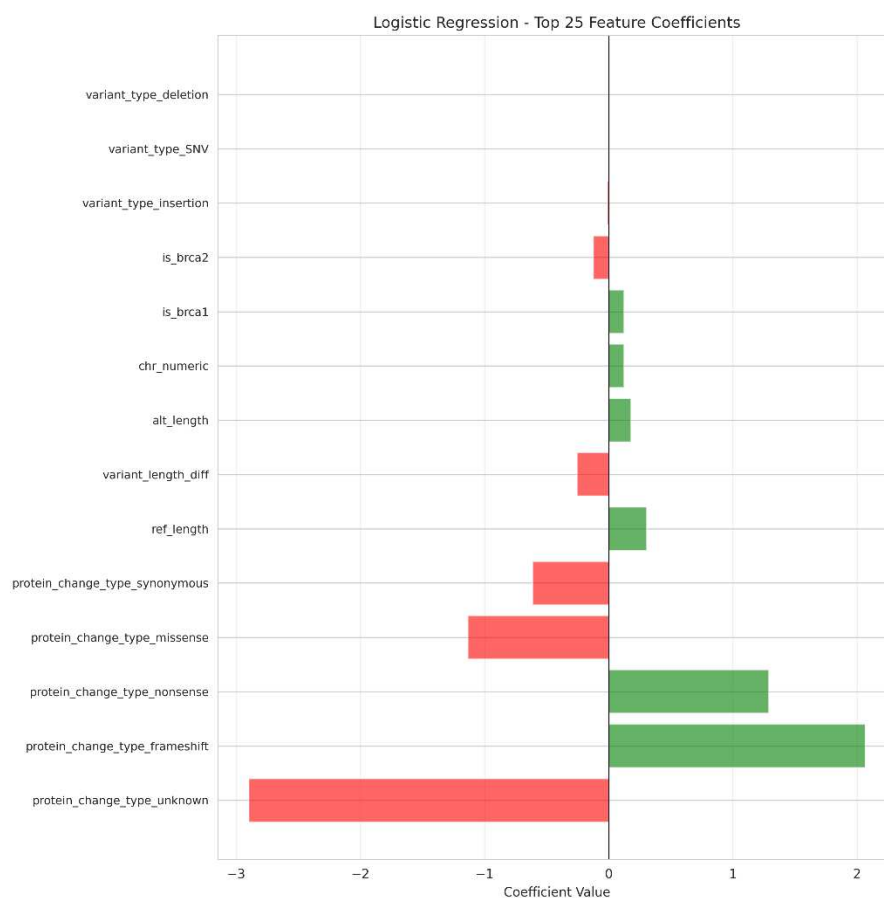
U početnoj fazi istraživanja, modeli strojnog učenja učeni su i testirani na cjelokupnom klasificiranom skupu BRCA Exchange (Skup A), koji obuhvaća ukupno 3.284 varijante (2.299 patogenih i 985 benignih). Evaluacija na izdvojenom testnom skupu pokazala je iznimno visoke performanse za sva četiri algoritma, pri čemu su logistička regresija, slučajna šuma, XGBoost i SVM ostvarili identičnu točnost od 98,63 % (Tablica 4.1).

Tablica 4.1 Performanse modela na testnom dijelu cjelokupnog skupa BRCA Exchange (Stupanj 1)

Algoritam	Točnost (Accuracy)	Preciznost (Precision)	Odziv (Recall)	F1-mjera	ROC- AUC	PR-AUC
Logistička regresija	0.9863	1.0000	0.9804	0.9901	0.9931	0.9971
Slučajna šuma	0.9863	1.0000	0.9804	0.9901	0.9929	0.9971
XGBoost	0.9863	1.0000	0.9804	0.9901	0.9927	0.9970
SVM	0.9863	1.0000	0.9804	0.9901	0.9927	0.9968

Međutim, detaljniji uvid u strukturu podataka i mehanizam donošenja odluka otkriva da je ovakva klasifikacija u velikoj mjeri trivijalna. Čak 68 % promatranog skupa čine trunkirajuće mutacije (besmislene mutacije i mutacije s pomakom okvira čitanja), za koje modeli vrlo lako usvajaju izravno pravilo patogenosti. Zbog te linearne odvojivosti i

najjednostavniji linearni model (logistička regresija) postiže gotovo maksimalne metrike



uspješnosti.

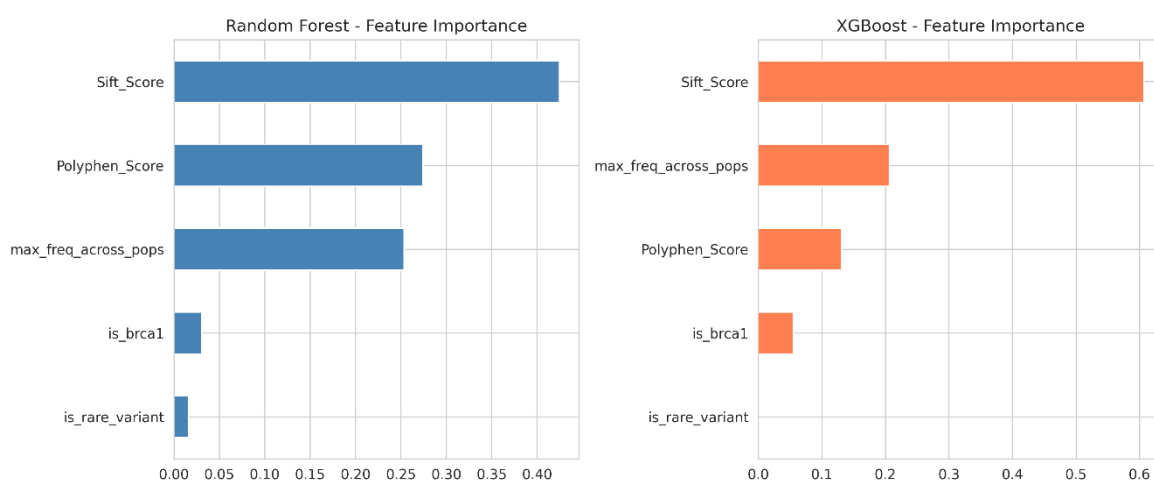
Slika 3.1 Standardizirani koeficijenti modela logističke regresije treniranog na cjelokupnom skupu podataka BRCA Exchange.

Iako modeli učeni na cjelokupnom skupu postižu prividno visoku točnost i metriku ROC-AUC iznad 0,99, analiza standardiziranih koeficijenata logističke regresije (Slika 4.1) potvrđuje da se odluke donose gotovo isključivo na temelju kategoričkih indikatora tipa proteinske promjene.

Visoki pozitivni koeficijenti za značajke `protein_change_type_frameshift` (+2,07) i `protein_change_type_nonsense` (+1,29) dominantno usmjeravaju predikciju prema patogenoj klasi, dok značajke poput `protein_change_type_unknown` (-2,90) i `protein_change_type_missense` (-1,13) djeluju kao snažni prediktori benignosti. Ovi rezultati potvrđuju da globalni modeli ne uče složena biokemijska i strukturna svojstva pojedinih varijanti, već se primarno oslanjaju na binarnu odvojivost trunkirajućih mutacija. To empirijski opravdava nužnost prelaska na specijaliziranu analizu isključivo varijanti *missense* u nastavku istraživanja.

4.3. Drugi stupanj: Specijalizirana analiza mutacija *missense* skupa A

U drugom stupnju modeli su evaluirani isključivo na podskupu od 196 varijanti *missense* iz baze BRCA Exchange (30 patogenih i 166 benignih). Kako bi se osigurala stroga izolacija i spriječilo umjetno poboljšanje rezultata, ovaj skup podataka analiziran je bez spajanja s naprednim frekvencijama gnomAD. Zbog relativno male veličine uzorka, primijenjena je stratificirana 5-struka unakrsna validacija (engl. *5-fold Stratified CV*) radi pouzdane procjene prosječnih performansi i varijance modela.



Slika 4.2 Relativna važnost značajki za modele slučajne šume (*Random Forest*) i XGBoost na podskupu varijanti *missense*.

Dodatno je provedena ugniježdjena unakrsna validacija za model slučajne šume (vanjska petlja 5-struka, unutarnja 3-struka, prostor pretrage: $\text{max_depth} \in \{3, 5, 7\}$, $\text{min_samples_leaf} \in \{2, 4, 8\}$, $\text{n_estimators} \in \{100, 200\}$) kako bi se provjerilo doprinosi li sustavna optimizacija hiperparametara poboljšanju performansi. Optimirani model postigao je $\text{ROC-AUC} = 0,950 \pm 0,054$, statistički ekvivalentno modelu s ručno odabranim, regulariziranim hiperparametrima ($\text{ROC-AUC} = 0,953 \pm 0,050$), što potvrđuje da dodatna optimizacija na ovako malom uzorku ($n = 196$) ne donosi mjerljivu korist i nosi rizik prenaučivosti.

Kada se cjevovod strojnog učenja primijeni isključivo na podskup mutacija *missense* (Slika 4.2), dolazi do bitne promjene u strukturi važnosti prediktora. Modeli temeljeni na stablima odluke više se ne mogu oslanjati na kategorički tip mutacije (jer su sve varijante SNV-ovi

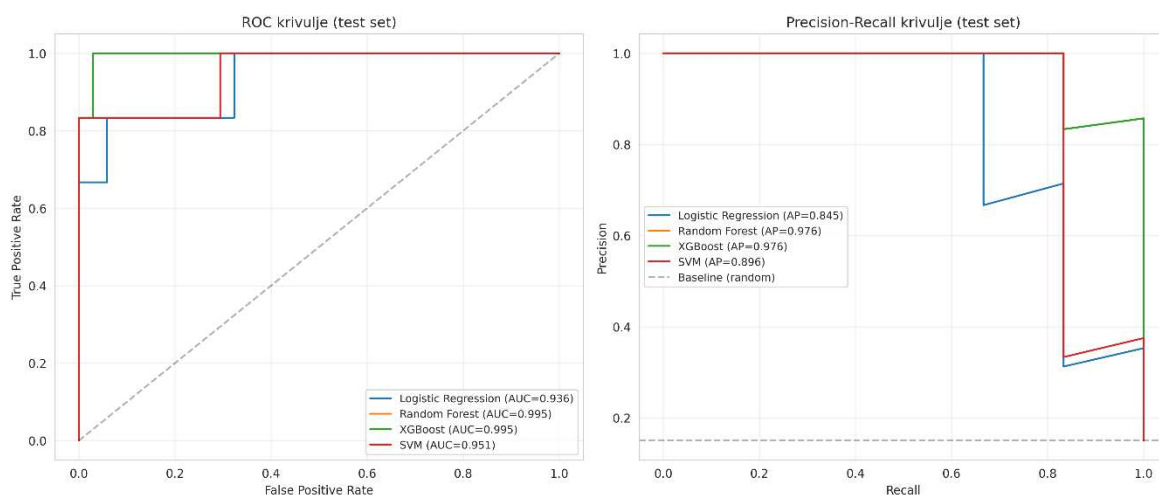
koji mijenjaju jednu aminokiselinu), već primarni fokus prebacuju na *in silico* alate za predviđanje funkcionalnog utjecaja zamjene aminokiselina.

Kod oba modela ansambala (Random Forest i XGBoost), ocjene Sift_Score i Polyphen_Score generiraju preko 60 % ukupne važnosti značajki, dok značajan doprinos daje i maksimalna populacijska frekvencija (max_freq_across_pops). Ovaj pomak u važnosti prediktora pokazuje da su modeli usmjerili proces učenja na prepoznavanje biokemijskih poremećaja u strukturi proteina i evolucijsku konzerviranost pozicija na kojima se mutacije javljaju.

Izolacija varijanti *missense* dovela je do jasne diferencijacije među algoritmima. Modeli s linearnom granicom ili distancijskim pristupom pokazali su izražen pad točnosti: logistička regresija i SVM ostvarili su prosječnu točnost od oko 78,56 % uz nisku preciznost za patogenu klasu ($\approx 0,41$). S druge strane, nelinearni ansambli stabala odluke pokazali su znatno veću sposobnost prepoznavanja suptilnih bioloških obrazaca, pri čemu su slučajna šuma i XGBoost postigli prosječnu točnost iznad 90 % te znatno višu F1-mjeru (Tablica 4.2).

Tablica 4.1.2 Performanse modela na testnom dijelu missense podskupa BRCA Exchange (stupanj 2)

Algoritam	Točnost (CV mean)	Std	Preciznost	Odziv	F1	ROC-AUC	Std
Logistička regresija	0,7856	0,0478	0,4143	0,9000	0,5640	0,9123	0,0461
Slučajna šuma	0,9132	0,0476	0,6925	0,8667	0,7598	0,9531	0,0495
XGBoost	0,9183	0,0498	0,7362	0,7667	0,7469	0,9506	0,0615
SVM	0,7855	0,0581	0,4151	0,8667	0,5572	0,9023	0,0526



Slika 4.3 Krivulje ROC i *Precision-Recall* modela strojnog učenja na testnom podskupu missense (BRCA Exchange).

Diskriminativna moć i stabilnost ispitivanih algoritama na testnom skupu varijanti *missense* vizualizirane su pomoću krivulja ROC i *Precision-Recall* (slika 4.3). Modeli ansambala stabala odluke pokazuju prednost u odnosu na linearne i distancijske modele: slučajna šuma i XGBoost ostvaruju površinu ispod ROC krivulje od ROC-AUC = 0,995, dok na Precision-Recall krivulji postižu prosječnu preciznost od PR-AUC = 0,976.

S obzirom na izražen klasni debalans prisutan unutar varijanti *missense* (gdje benigne varijante čine veliku većinu), visoka površina ispod krivulje PR potvrđuje da modeli ansambala zadržavaju visoku razinu preciznosti bez značajnog kompromitiranja odziva, odnosno uspješno identificiraju patogene slučajeve uz nizak udio lažno pozitivnih ishoda.

4.4. Treći stupanj: Modeliranje i unakrsna validacija na neovisnom skupu ClinVar

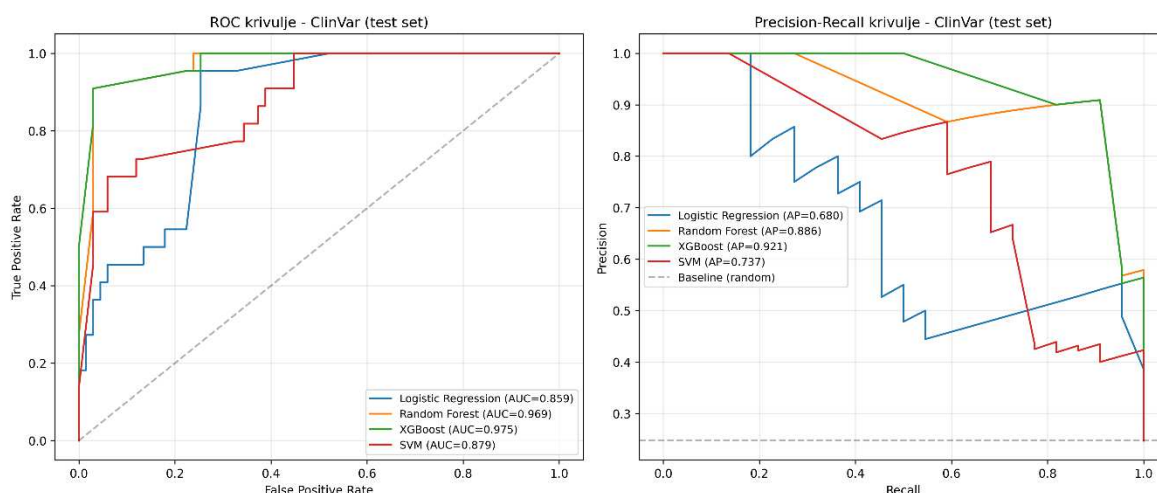
Kao provjera generalizacijske sposobnosti modela, ista je metodologija primijenjena na neovisnom i strogo filtriranom podskupu varijanti *missense* iz baze ClinVar (skup B), koji obuhvaća 443 varijante (108 patogenih i 335 benignih). Ovaj skup podataka sadrži veću razinu šuma uslijed heterogenih izvora prijava, no obogaćen je modernim *in silico* prediktorima (CADD, REVEL, PhyloP, SpliceAI, Pangolin) te populacijskim frekvencijama iz baze gnomAD v4.1.

Rezultati stratificirane 5-struke unakrsne validacije potvrdili su trendove uočene u drugom stupnju. Relativni odnos među algoritmima ostao je dosljedan i na ovom neovisnom skupu podataka: ansambl stabala odluke ponovno su ostvarili znatno bolje rezultate u odnosu na linearne modele i SVM. Model XGBoost ostvario je najvišu prosječnu točnost (86,67 %) i ROC-AUC vrijednost (0,9502), uz visoku razinu odziva za patogenu klasu (Tablica 4.3).

Tablica 4.3 Prosječne performanse modela iz 5-struke unakrsne validacije na skupu ClinVar *missense*

Algoritam	Točnost (CV mean)	Odziv (CV mean)	ROC-AUC (CV mean)
Logistička regresija	0,7382 ± 0,0460	0,8065 ± 0,0587	0,8071 ± 0,0345
Slučajna šuma	0,8397 ± 0,0236	0,9082 ± 0,0949	0,9482 ± 0,0271
XGBoost	0,8667 ± 0,0417	0,8541 ± 0,1048	0,9502 ± 0,0277
SVM	0,7516 ± 0,0152	0,7697 ± 0,0959	0,8525 ± 0,0348

Kako bi se provjerila otpornost razvijenih modela na neviđenim podacima i uklonila mogućnost prilagodbe specifičnostima baze BRCA Exchange, provedena je evaluacija na izdvojenom testnom skupu iz baze ClinVar uz pridružene gnomAD v4.1 značajke (slika 4.4).



Slika 4.4 Krivulje ROC i *Precision-Recall* modela na izdvojenom testnom skupu baze ClinVar.

Rezultati na izdvojenom testnom skupu potvrđuju visoku diskriminativnu moć predloženog cjevovoda. Model XGBoost na testnom skupu postiže ROC-AUC = 0,975 i PR-AUC = 0,921, dok slučajna šuma ostvaruje ROC-AUC = 0,969. Postizanje stabilnih metrika na potpuno neovisnom izvoru podataka s novim prediktorima (poput ocjena CADD i konzerviranosti PhyloP) sugerira dobra svojstva generalizacije modela unutar promatranog skupa značajki.

4.5. Četvrti stupanj: Vanjska validacija na razriješenim varijantama VUS

Završni dio evaluacije obuhvatio je primjenu modela učenih na skupu od 196 varijanti ENIGMA na vanjski testni skup. Analizirane su varijante koje u bazi BRCA Exchange imaju status varijanti nepoznatog značenja (VUS). Usporedbom s bazom ClinVar identificirane su 103 varijante (75 benignih i 28 patogenih) koje posjeduju nedvosmisleni, strogu kliničku oznaku stručnih panela (3 zvjezdice). Pojam vanjske validacije u ovom se kontekstu odnosi na strogu instancijsku izolaciju skupa podataka — evaluirane varijante VUS nisu bile dio procesa učenja niti optimizacije modela. Ipak, s obzirom na to da baze BRCA Exchange i ClinVar dijele dio podataka i anotacijskih izvora, ovu validaciju treba promatrati kao provjeru generalizacije na neviđenim instancama unutar srodnog ekosustava podataka, a ne kao prospektivno ispitivanje na potpuno neovisnoj populaciji.

Na ovom vanjskom testnom skupu nelinearni modeli ostvarili su visoku točnost reklasifikacije. XGBoost i Random Forest točno su predvidjeli status za oko 90 % varijanti, pri čemu je XGBoost ostvario ukupnu točnost od 91,3 % (Tablica 4.4).

Tablica 4.4 Performanse modela u reklasifikaciji varijanti VUS na vanjskom testnom skupu s nedvojbenim ClinVar statusom

Algoritam	Ukupna točnost	Točno prepoznate benigne (od 75)	Točno prepoznate patogene (od 28)
Random Forest	92/103 (89,3 %)	69/75 (92,0 %)	23/28 (82,1 %)
XGBoost	94/103 (91,3 %)	71/75 (94,7 %)	23/28 (82,1 %)

Logistička regresija	73/103 (70,9 %)	49/75 (65,3 %)	24/28 (85,7 %)
SVM	71/103 (68,9 %)	48/75 (64,0 %)	23/28 (82,1 %)

Linearni modeli (logistička regresija i SVM) postigli su nešto viši odziv za patogenu klasu, no uz znatno veći broj lažno pozitivnih ishoda na benignoj klasi, što je rezultiralo znatno nižom ukupnom točkom točnosti ($\approx 70\%$) u odnosu na ansamble stabala.

U kliničkom kontekstu, lažno negativne pogreške (FN) nose najveći rizik jer patogenu mutaciju proglašavaju benignom, što može dovesti do izostanka preventivnog praćenja. U ovom je testiranju pri korištenju modela slučajne šume, zabilježeno 5 takvih slučajeva (4,9 % ukupnog skupa). Detaljniji uvid u te varijante pokazao je da se radi o zamjenama s graničnim biokemijskim ocjenama štetnosti kod kojih su populacijske frekvencije izrazito niske, što otežava jednoznačno razdvajanje bez dodatnih funkcionalnih eksperimenata.

Dobiveni rezultati pokazuju da model uspješno prepoznaje veliku većinu razriješenih slučajeva te može poslužiti kao pomoćni sustav za trijažu i prioritizaciju varijanti VUS u kliničkim genetskim laboratorijima.

5. Rasprava

Rezultati dobiveni u ovom istraživanju pružaju uvid u izazove i mogućnosti primjene algoritama strojnog učenja u kliničkoj genetici, s posebnim naglaskom na klasifikaciju mutacija gena *BRCA1* i *BRCA2*. Eksperimentalna analiza pokazuje da se težina predviđanja patogenosti bitno mijenja ovisno o molekularnom tipu promatrane genetske varijante.

5.1. Ograničenja analize na miješanim skupovima podataka

U prvom stupnju istraživanja primjena modela na cjelokupnom klasificiranom skupu rezultirala je točnošću od 98,63 % na testnom podskupu, neovisno o odabranom algoritmu. Međutim, ovako visoke performanse, uključujući i one najjednostavnijeg linearnog modela, posljedica su same distribucije skupa u kojem dominiraju trunkirajuće mutacije (mutacije s pomakom okvira čitanja i besmislene mutacije).

Budući da takve promjene dovode do preuranjenog prekida translacije i izravnog gubitka funkcije proteina, njihova je patogenost linearno odvojiva i jednostavna za prepoznavanje. Ovim rezultatom je potvrđena početna pretpostavka da evaluacija na miješanim podacima može stvoriti lažan dojam visoke prediktivne moći, prikrivajući stvarni bioinformatički problem — klasifikaciju mutacija *missense* kod kojih je duljina proteina očuvana.

5.2. Usporedba linearnih i nelinearnih modela

Stvarna provjera generalizacijskih sposobnosti provedena je na izoliranim podskupovima varijanti *missense* iz baza BRCA Exchange i ClinVar. Rezultati su pokazali da jednostavni linearni i distancijski modeli (logistička regresija i SVM) gube na točnosti i preciznosti kada se suoče s biokemijski suptilnim promjenama unutar aminokiselinskog slijeda. S druge strane, oba nelinearna ansambla temeljena na stablima odluke (slučajna šuma i XGBoost) ostvarila su usporedive, međusobno statistički ekvivalentne performanse (ROC-AUC $\approx$ 0,95), zadržavši visoku stabilnost s točnošću iznad 90 % na unakrsnoj validaciji.

Ovakva diferencijacija potvrđuje da su odnosi između bioloških prediktora izrazito nelinearni te da je sama klasa modela temeljenih na stablima odluke znatno prikladnija za ovaj problem. Model logističke regresije u velikoj se mjeri oslanjao na pojedinačne *in silico*

alate (poput ocjene SIFT s koeficijentom -2,14). Nasuprot tome, analiza važnosti značajki pokazala je da ansambli stabala najveći dio informacija crpe iz zajedničke integracije ocjena štetnosti (gnomad_cadd) i populacijskih frekvencija (gnomad_faf95_max, gnomad_af).

Prednost ovih modela leži u mogućnosti prepoznavanja složenih pragova odlučivanja. Primjerice, varijanta s visokom ocjenom štetnosti može biti klasificirana kao benigna ako je prisutna u dovoljno visokoj frekvenciji unutar zdrave populacije. Također, prediktori utjecaja na izrezivanje introna (gnomad_spliceai_ds_max) ostvarili su mjerljiv doprinos, što potvrđuje spoznaje iz literature da dio varijanti *missense* svoju patogenost duguje narušavanju mehanizama prekrajavanja mRNK na spojevima egzona i introna.

5.3. Klinička primjenjivost i analiza pogrešaka

Potvrda praktične vrijednosti modela dobivena je u četvrtom stupnju, gdje su modeli primijenjeni na varijante nepoznatog značenja (VUS) iz baze BRCA Exchange, čiji je klinički status provjeren kroz konsenzusne oznake stručnih panela u bazi ClinVar. Model slučajne šume točno je predvidio status za 92 od 103 ispitane varijante (89,3 %), dok je XGBoost postigao točnost od 91,3 %, što ukazuje na to da predloženi cjevovod posjeduje dobra generalizacijska svojstva i može poslužiti kao pomoćni *in silico* dokaz prema ACMG smjernicama.

Ipak, detaljna raščlamba pogrešaka nalaže oprez pri razmatranju ovakvih sustava u kliničkoj praksi. Kod modela slučajne šume zabilježeno je 11 pogrešnih klasifikacija: 6 lažno pozitivnih (benigne varijante klasificirane kao patogene) i 5 lažno negativnih (patogene varijante klasificirane kao benigne). U medicinskoj dijagnostici lažno negativni rezultati nose znatno veći rizik jer pacijenti s patogenom mutacijom ostaju bez potrebnog pojačanog praćenja i profilakse. Činjenica da su pogrešno klasificirane varijante imale visoku razinu dokaza u bazi ClinVar sugerira da modeli nailaze na ograničenja kada patogenost proizlazi iz specifičnih bioloških mehanizama koje opći *in silico* prediktori ne uspijevaju u potpunosti obuhvatiti.

Dobiveni rezultati pokazuju da računski modeli ne mogu u potpunosti zamijeniti laboratorijske funkcionalne testove (engl. *functional assays*) i procjenu kliničkih genetičara, ali predstavljaju pouzdan alat za prioritizaciju varijanti i usmjeravanje daljnjih eksperimentalnih ispitivanja.

5.4. Ograničenja istraživanja

Pri interpretaciji rezultata ovog rada potrebno je uvažiti nekoliko metodoloških ograničenja:

1. **Veličina uzorka:** Podskup varijanti *missense* iz baze BRCA Exchange u drugom stupnju obuhvaća 196 instanci, od čega je samo 30 patogenih. Zbog malog broja pozitivnih primjera, rezultati unakrsne validacije pokazuju određenu osjetljivost na podjelu podataka unutar pojedinih pregrada, što se očituje u standardnoj devijaciji točnosti od oko 5 % (0,0476).
2. **Kontrola prenaučivosti:** Primjena nelinearnih modela poput slučajne šume s maksimalnom dubinom stabla 10 (`max_depth=10`) na manjem skupu podataka nosi teoretski rizik memoriranja šuma. Taj je rizik u radu kontroliran korištenjem regularizacije i plićih stabala kod XGBoost modela (`max_depth=6`). Stabilni rezultati postignuti na neovisnom ClinVar skupu ($N = 443$) i vanjskom skupu VUS ($N = 103$) potvrđuju da su ugrađene tehnike regularizacije uspješno spriječile prenaučavanje.
3. **Opseg varijanti:** Istraživanje je fokusirano na točkaste mutacije unutar gena *BRCA1* i *BRCA2*, dok veće strukturne promjene (velike delecije, insercije i duplikacije) zahtijevaju drugačije bioinformatičke alate i izlaze iz okvira ovog modela.

5.5. Usporedba sa srodnim istraživanjima

Kako bi se objektivno procijenila točnost i stvarni doprinos predloženog modela temeljenog na stablima odluke, dobivene rezultate potrebno je usporediti s postojećim istraživanjima u kliničkoj genetici i bioinformatici. Pregled literature pokazuje dvije jasne skupine radova: istraživanja koja modele testiraju na mješovitom skupu svih tipova varijanti te radove usmjerene isključivo na mutacije krivog smisla (*missense*), koje je biološki znatno teže klasificirati. Ova podjela izravno određuje visinu postignutih metrika i primjenjivost modela u praksi, što se vidi iz usporedbe u tablicama 5.1 i 5.2 u nastavku.

Table 5.1 Usporedba performansi modela na mješovitim skupovima varijanti gena BRCA1 i BRCA2

Istraživanje	Izvor podataka i modeli	Podskup varijanti	Točnost (Accuracy)	ROC-AUC
Khandakji i Mifsud (2022) [1]	BRCA Exchange (BRCA2); XGBoost	Svi tipovi varijanti	~99,0 %	~0,99
Karalidou i sur. (2022) [3]	ClinVar (BRCA1/BRCA2); Random Forest, SVM, MLP	Svi tipovi varijanti	98,1 %	—
Ovaj rad (Stupanj 1)	BRCA Exchange (BRCA1/BRCA2); XGBoost, RF, SVM, LR	Svi tipovi varijanti	98,6 %	0,993

Modeli učeni na mješovitim podacima redovito postižu vrlo visoku točnost, najčešće 98 % - 99 %. Primjerice, Khandakji i Mifsud (2022) [1] su pomoću XGBoost-a na podacima iz baze BRCA Exchange postigli točnost od oko 99 % za varijante gena BRCA2. Slično tome, alat MARGINAL (Karalidou i sur., 2022) [3], temeljen na slučajnim šumama i SVM-u, na bazi ClinVar ostvario je točnost od 98,1 %. Isti je rezultat potvrđen i u prvom stupnju ovog istraživanja, gdje su svi testirani algoritmi postigli točnost iznad 98 %.

Međutim, ovako visoki rezultati uglavnom su posljedica neuravnoteženosti podataka. U mješovitim skupovima dominiraju trunkirajuće mutacije, prvenstveno delecije s pomakom okvira čitanja i besmislene mutacije. One uvode preuranjeni stop-kodon, što dovodi do razgradnje mRNK ili stvaranja nefunkcionalnog proteina. Zbog toga su gotovo uvijek patogene, pa model zapravo uči jednostavno pravilo koje odgovara kriteriju gubitka funkcije (PVS1) prema smjernicama ACMG-a. Iako su matematički točni, takvi modeli ne donose novu dijagnostičku vrijednost jer je patogenost trunkirajućih mutacija jasna i bez strojnog učenja.

Table 5.2 Usporedba performansi modela na specijaliziranim podskupovima varijanti *missense*

Istraživanje	Izvor podataka i modeli	Podskup varijanti	Točnost (Accuracy)	ROC-AUC
Hart i sur. (2020) [25]	Eksperimentalni eseji (BRCA1/BRCA2); Random Forest	Isključivo <i>missense</i>	90,0 % – 93,4 %	0,90 – 0,92
Wu i sur. (2025) [2]	ClinVar / VEP; Vektorska ugniježđenja (LLM)	Reklasifikacija <i>missense</i> VUS	~91,0 %	—
Ovaj rad (Stupanj 2)	BRCA Exchange (BRCA1/BRCA2); XGBoost (5-struka CV)	Isključivo <i>missense</i>	91,8 %	0,951
Ovaj rad (Stupanj 3)	ClinVar + gnomAD (BRCA1/BRCA2); XGBoost (5-struka CV)	Isključivo <i>missense</i>	86,7 %	0,950
Ovaj rad (Stupanj 4)	ClinVar / BRCA Exchange; XGBoost	Vanjska validacija <i>missense</i> VUS	91,3 %	—

U literaturi je primjetan nedostatak radova koji modele uče isključivo na varijantama *missense* izravno preuzetim iz baze BRCA Exchange. Razlog tome leži u konzervativnim kriterijima konzorcija ENIGMA koji većinu takvih mutacija zadržavaju u kategoriji VUS, zbog čega baza sadrži vrlo malen broj klasificiranih varijanti *missense* (u analiziranom skupu svega 196 instanci, od čega samo 30 patogenih). Iz tog razloga, srodna istraživanja modele uče ili na mješovitim skupovima cjelokupnog spektra mutacija ili se za analizu zamjena *missense* okreću većim repozitorijima poput ClinVara te rezultatima eksperimentalnih funkcionalnih eseja.

Pravi izazov nastaje tek pri klasifikaciji mutacija *missense*. Budući da se kod njih duljina proteina ne mijenja i nema očitih strukturnih oštećenja, modeli moraju prepoznati složenije signale poput biokemijskih svojstava, utjecaja na prekrajanje i evolucijske očuvanosti pozicija. Na ovom području referentan je rad Harta i suradnika (2020) [25] s alatom BRCA-ML. Koristeći rezultate funkcionalnih testova i slučajne šume, postigli su ROC-AUC od 0,92 za BRCA1 i 0,90 za BRCA2 isključivo na varijantama *missense*.

Usporedbe radi, model XGBoost razvijen u drugom stupnju ovog rada na podacima iz baze BRCA Exchange postiže točnost od 91,8 posto uz ROC-AUC od 0,951, što se podudara s rezultatima alata BRCA-ML. Primjena modela na bazu ClinVar, koja sadrži više šuma, smanjila je točnost na očekivanih 86,7 %, ali je ROC-AUC ostao stabilan na 0,950.

Važan metodološki problem pri radu s računalnim prediktorima jest rizik od cirkularnosti i curenja podataka. Kako navode Grimm i suradnici (2015) [9], problem nastaje ako se modeli testiraju na varijantama čija je izvorna klinička oznaka djelomično donesena upravo pomoću tih istih alata. Taj je rizik stvaran jer službene smjernice ACMG-a dopuštaju korištenje alata poput REVEL-a, CADD-a ili PolyPhen-a kao dokaza patogenosti (PP3) ili benignosti (BP4), a preporuke ClinGena za neke alate omogućuju i višu razinu dokazne snage. Ako se to ne uzme u obzir, model neće naučiti biološke uzorke, nego će samo prekopirati logiku postojećih alata, što dovodi do nerealno dobrih rezultata.

U ovom je radu taj problem riješen strogim odabirom podataka za učenje. Skup za učenje ograničen je isključivo na varijante oko kojih postoji jasan stručni konsenzus, dok su varijante nepoznatog značenja (VUS) i vjerojatno patogene varijante izostavljene. Time je osigurano da se model uči na varijantama čiji se status temelji na pouzdanim obiteljskim i eksperimentalnim podacima. Zahvaljujući tome, korištenje prediktora i alelnih frekvencija iz baze gnomAD v4.1 pruža stvaran prediktivni signal, bez rizika od cirkularnosti.

Pouzdanost ovakvog pristupa potvrđuju rezultati vanjske validacije. Testiranjem modela na skupu od 103 povijesne varijante — koje u trenutku učenja nisu imale definiran status, već su tek naknadno razriješene u bazi ClinVar — XGBoost je postigao točnost od 91,3 posto. Taj je rezultat usporediv s novijim pristupima u literaturi; primjerice, Wu i suradnici (2025) [2] razvili su model VUS.Life temeljen na jezičnim modelima koji postiže točnost od oko 91 % na istom zadatku reklasifikacije varijanti nepoznatog značenja.

Rezultati pokazuju da pažljivo odabrane značajke iz populacijskih baza daju pouzdan signal za klasifikaciju. Modeli poput stabala s pojačavanjem gradijenta uspješno prepoznaju nelinearne biološke obrasce, a uz kontrolu cirkularnosti mogu poslužiti kao koristan alat za analizu i prioritizaciju nerazjašnjenih varijanti u svakodnevnoj dijagnostici.

Zaključak

Klasifikacija patogenosti varijanti gena BRCA1 i BRCA2 pokazala se zadatkom čija težina gotovo u potpunosti ovisi o tipu promatrane mutacije. Za trunkirajuće mutacije problem je, u skladu s ACMG PVS1 kriterijem, već rutinski riješen, jer i najjednostavniji model ovdje postiže gotovo savršenu točnost budući da se patogenost svodi na prepoznavanje ranog stop-kodona. Stvarni izazov, kao što je u ovom radu pokazano izolacijom podskupa *missense*, leži u razlikovanju patogenih od benignih zamjena pojedinačnih aminokiselina, gdje razlika između klasa postaje suptilna i ovisi o kombinaciji evolucijske konzerviranosti, fizikalno-kemijskih svojstava i populacijske učestalosti.

Na tom težem podskupu oba nelinearna modela ansambala (slučajna šuma i XGBoost) ostvarila su usporedive i međusobno statistički ekvivalentne performanse, dosljedno nadmašivši logističku regresiju i SVM i na podacima BRCA Exchange ($n = 196$) i na ClinVar ($n = 443$), s ROC-AUC vrijednostima iznad 0,95. Analiza važnosti značajki pokazala je da uspješnost ansambala stabala odluke proizlazi iz kombinirane integracije SIFT i PolyPhen ocjena uz populacijsku frekvenciju na manjem skupu, odnosno iz ocjena CADD i frekvencija gnomAD uz doprinos prediktora SpliceAI na većem skupu ClinVar.

Vanjska validacija na 103 varijante VUS naknadno razriješene u ClinVaru dala je najkonkretniji pokazatelj primjenjivosti: slučajna šuma je točno klasificirala 92, a XGBoost 94 varijante. Analiza pogrešaka pritom je važnija od same točnosti, jer je model slučajne šume počinio 5 lažno negativnih klasifikacija, što u kliničkom kontekstu znači 5 potencijalno patogenih varijanti pogrešno označenih kao benigne. To ograničenje treba uzeti ozbiljno: uz uzorak od svega 196, odnosno 443 varijanti *missense*, dobiveni rezultati pokazuju da je pristup obećavajući, ali ih ne treba tumačiti kao dokaz da je problem klasifikacije BRCA1/BRCA2 varijanti *missense* riješen. Za konačnu potvrdu bilo bi potrebno ponoviti evaluaciju na većim, neovisno prikupljenim skupovima te usporediti rezultate s postojećim *in silico* alatima (REVEL, CADD) na istim varijantama.

Unatoč tim ograničenjima, cjevovod razvijen u ovom radu pokazuje da kombinacija populacijskih frekvencija i *in silico* prediktora, uz pažljivu kontrolu curenja podataka i cirkularnosti, može poslužiti kao koristan pomoćni alat za prioritizaciju varijanti VUS prema smjernicama ACMG-a. Ne kao zamjena za funkcionalne testove ili stručnu kliničku

procjenu, već kao dodatni izvor dokaza koji može usmjeriti ograničene laboratorijske resurse prema varijantama gdje je daljnje ispitivanje najpotrebnije.

References

- [1] M. N. Khandakji i B. Mifsud, »Gene-specific machine learning model to predict the pathogenicity of BRCA2 variants,« *Frontiers in Genetics*, svez. 13, 2022.
- [2] J. Wu, M. Muriello, D. G. Basel i X. Gai, *Predicting Genetic Variant Pathogenicity Using Vector Embeddings*, medRxiv, 2025.
- [3] V. Karalidou, D. Kalfakakou, A. Papathanasiou, F. Fostira i G. K. Matsopoulos, »MARGINAL: An Automatic Classification of Variants in BRCA1 and BRCA2 Genes Using a Machine Learning Model,« *Biomolecules*, svez. 12, br. 11, 2022.
- [4] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts i P. Walter, *Molecular Biology of the Cell*, New York: Garland Science, 2017.
- [5] S. A. Narod i W. D. Foulkes, »BRCA1 and BRCA2: 1994 and beyond,« *Nature Reviews Cancer*, svez. 4, br. 9, 2004.
- [6] R. Roy, J. Chun i S. N. Powell, »BRCA1 and BRCA2: different roles in a common pathway of genome protection,« *Nature Reviews Cancer*, svez. 11, br. 1, 2011.
- [7] C. J. Lord i A. Ashworth, »BRCAness revisited,« *Nature Reviews Cancer*, svez. 16, br. 2, 2016.
- [8] S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding i H. L. Rehm, »Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology,« *Genetics in Medicine*, svez. 17, br. 5, 2015.
- [9] D. G. Grimm, C. A. Damen, A. Karch i P. M. Krawitz, »The evaluation of tools for predicting the impact of missense variants is affected by computational biases,« *Human Mutation*, svez. 36, br. 5, 2015.

- [10] A. Antoniou, P. D. P. Pharoah, S. Narod i D. F. e. a. Easton, »Average Risks of Breast and Ovarian Cancer Associated with BRCA1 or BRCA2 Mutations Detected in Case Series Unselected for Family History,« *The American Journal of Human Genetics*, svez. 72, br. 5, 2003.
- [11] K. B. Kuchenbaecker, J. L. Hopper, D. R. Barnes, K. A. Phillips, T. M. Mooij, M.-J. Roos-Blom, S. Jervis, A. C. Antoniou i e. al., »Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers,« *JAMA*, svez. 317, br. 23, 2017.
- [12] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom i A. Siepel, »Detection of nonneutral substitution rates on mammalian phylogenies,« *Genome Research*, svez. 20, br. 1, 2010.
- [13] P. C. Ng i S. Henikoff, »SIFT: Predicting amino acid changes that affect protein function,« *Nucleic Acids Research*, svez. 31, br. 13, 2003.
- [14] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov i S. R. Sunyaev, »A method and server for predicting damaging missense mutations,« *Nature Methods*, svez. 7, 2010.
- [15] N. M. Ioannidis, J. H. Jung, V. Pejaver, M. Mort, D. N. Cooper, C. Wang, M. Y. Han, S. Batzoglu i e. al., »REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants,« *The American Journal of Human Genetics*, svez. 99, br. 4, 2016.
- [16] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper i J. Shendure, »A general framework for estimating the relative pathogenicity of human genetic variants,« *Nature Genetics*, svez. 46, br. 3, 2014.
- [17] K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J. F. McRae, S. F. Darbandi, S. M. Mehrabian i e. al., »Predicting Splicing from Primary Sequence with Deep Learning,« *Cell*, svez. 176, br. 3, 2019.
- [18] T. Zeng i Y. I. Li, »Predicting RNA splicing from DNA sequence using Pangolin,« *Genome Biology*, svez. 23, br. 1, 2022.

- [19] M. S. Cline, R. G. Liao, M. T. Parsons, B. Paten, F. Alquaddoomi i e. al., »BRCA Exchange: A collaborative resource and community for the sharing of BRCA1 and BRCA2 variant information,« *PLOS Genetics*, svez. 14, br. 12, 2018.
- [20] M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipothu, R. Davidson, J. Hart, B. L. Kattman i e. al., »ClinVar: improving access to variant interpretations and supporting evidence,« *Nucleic Acids Research*, svez. 46, br. D1, 2018.
- [21] K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi i e. al., »The mutational constraint spectrum quantified from variation in 141,456 humans,« *Nature*, svez. 581, br. 7809, 2020.
- [22] J. Šnajder i B. D. Bašić, *Strojno učenje*, Zagreb: Fakultet elektrotehnike i računarstva, 2014.
- [23] T. Chen i C. Guestrin, »XGBoost: A Scalable Tree Boosting System,« u *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
- [24] L. Breiman, »Random Forests,« *Machine Learning*, svez. 45, 2 May 2001.
- [25] S. N. Hart, E. C. Polley, H. Shimelis, S. Yadav i F. J. Couch, »Prediction of the functional impact of missense variants in BRCA1 and BRCA2 with BRCA-ML,« *npj Breast Cancer*, svez. 6, br. 1, 2020.

Skraćenice

5. ACMG – *American College of Medical Genetics and Genomics* (Američki koledž za medicinsku genetiku i genomiku).
6. AMP – *Association for Molecular Pathology* (Udruženje za molekularnu patologiju).
7. BRCA1 – *Breast Cancer 1*.
8. BRCA2 – *Breast Cancer 2*.
9. CV – *Cross-Validation* (Unakrsna validacija).
10. DNK / DNA – Deoksiribonukleinska kiselina.
11. DSB – *Double-Strand Breaks* (Dvolančani prekidi DNA).
12. FN – *False Negative* (Lažno negativne pogreške).
13. HR – *Homologous Recombination* (Homologna rekombinacija).
14. MNV – *Multi-Nucleotide Variant*.
15. mRNK – Glasnička RNK.
16. NGS – *Next-Generation Sequencing* (Sekvenciranje nove generacije).
17. NHEJ – *Non-Homologous End Joining* (Nehomologno spajanje krajeva).
18. NMD – *Nonsense-Mediated mRNA Decay*.
19. OR – *Odds Ratio* (Omjer šansi).
20. PR-AUC – Površina ispod *Precision-Recall* krivulje.
21. RF – *Random Forest* (Slučajna šuma).
22. ROC-AUC – Površina ispod ROC krivulje.
23. SNV – *Single Nucleotide Variant*.
24. SVM – *Support Vector Machine* (Stroj potpornih vektora).
25. VUS – *Variants of Uncertain Significance* (Varijante nepoznatog značenja).
26. XGBoost – *Extreme Gradient Boosting* (Ekstremni gradijentni potisak).

Izjava o korištenju umjetne inteligencije

Potvrđujem da sam tijekom izrade ovog završnog rada koristila umjetnu inteligenciju u skladu s Politikom primjerenog korištenja umjetne inteligencije na Fakultetu elektrotehnike i računarstva, te da umjetna inteligencija nije zamijenila moje kritičko razmišljanje niti moj doprinos.