

UNIVERZITET U BEOGRADU
BIOLOŠKI FAKULTET

Dejan J. Šorgić

Prediktivno modelovanje u forenzičkoj genetici:
klasifikacija etničke pripadnosti i prepoznavanje
incesta na osnovu autozomnih *STR* profila

doktorska disertacija

Beograd, 2026

UNIVERSITY OF BELGRADE
FACULTY OF BIOLOGY

Dejan J. Šorgić

Predictive Modeling in Forensic Genetics: Ethnic
Affiliation Classification and Incest Detection
Based on Autosomal STR Profiles

Doctoral Dissertation

Belgrade, 2026

Mentor:

dr Dušan Keckarević, vanredni profesor
Univerzitet u Beogradu – Biološki fakultet

Članovi komisije:

dr Milica Keckarević–Marković, vanredni profesor
Univerzitet u Beogradu – Biološki fakultet

dr Mihailo Jelić, vanredni profesor
Univerzitet u Beogradu – Biološki fakultet

dr Maja Živković, naučni savetnik
Univerzitet u Beogradu – Institut za nuklearne nauke „Vinča“, Institut od nacionalnog
značaja za Republiku Srbiju

Datum odbrane: _____

Zahvalnica

Posebnu zahvalnost dugujem svom mentoru, prof. dr Dušanu Keckareviću, na stručnom vođenju i podršci tokom izrade ove disertacije, ali i tokom deset godina našeg poznanstva i saradnje. Zahvalan sam mu što je prepoznao moj potencijal i preporučio me za posao koji je usmerio moj dalji profesionalni razvoj. Hvala mu što je na ovom putu bio pouzdan vodič, ali mi je ostavio dovoljno prostora da pronađem sopstveni put.

Zahvalnost dugujem i dr Aleksandri Stefanović, sa kojom već deset godina delim svakodnevni rad i profesionalne izazove. Svojim znanjem i podrškom usmeravala je moj profesionalni razvoj od samih početaka, a posebno sam joj zahvalan na komentarima i sugestijama tokom izrade i pisanja ove teze.

Zahvaljujem se prof. dr Milici Keckarević-Marković, prof. dr Mihailu Jeliću i dr Maji Živković na pažljivom čitanju disertacije, korisnim savetima i sugestijama koje su mi pružili tokom njene izrade. Svojim zapažanjima ukazali su mi na pojedine aspekte rada koje prethodno nisam sagledao u dovoljnoj meri i time značajno doprineli njegovom unapređenju i konačnom oblikovanju.

Zahvalnost dugujem i svojim roditeljima i sestri, koji su mi tokom celog života bili oslonac. Njihova vera pratila me je kroz sve važne životne odluke. Verujem da im ovaj trenutak znači koliko i meni.

Najiskreniju zahvalnost dugujem svojoj supruzi Ivani što smo, uz dvoje dece i dva doktorata koja su nastajala gotovo istovremeno, uspeali da sačuvamo razum. Hvala ti na podršci koju mi pružaš svih ovih godina.

Za Lanu i Galu, moje najvažnije životno dostignuće. Voli vas tata. Uvek verujte u sebe.

Prediktivno modelovanje u forenzičkoj genetici: klasifikacija etničke pripadnosti i prepoznavanje incesta na osnovu autozomnih *STR* profila

Sažetak

Cilj disertacije bio je razvoj i evaluacija modela mašinskog učenja za predikciju porekla i detekciju incestuoznih odnosa na osnovu autozomnih *STR* profila.

Za predikciju porekla generisano je 360000 sintetičkih profila na osnovu učestalosti alela 29 lokusa iz četiri referentne grupe: afroameričke, azijske, hispanoameričke i severnoameričke populacije evropskog porekla (SA–EP). Među poređenim algoritmima, *XGBoost* je postigao najvišu tačnost: 94,24 % za četiri referentne grupe, 99,06 % nakon isključenja hispanoameričke grupe i 98,57 % nakon njenog objedinjavanja sa SA–EP grupom. Tačnost klasifikacije nije se znatno povećavala kada je broj profila po grupi povećan sa 10000 na 90000. Najinformativniji lokusi za predikciju porekla bili su *Penta E*, *FL3A01* i *SE33*. Kombinacije od 21 lokusa koje odgovaraju kompletima GlobalFiler i Investigator 24plex postigle su tačnost u klasifikaciji od 88,28 % za četiri, odnosno 96,30 % za tri populacione grupe.

Za detekciju incesta generisani su parovi profila majke i deteta u scenarijima neincestuoznog očinstva i incestuoznih odnosa otac–ćerka i brat–sestra. *CatBoost* je postigao tačnost od 96,94 % pri analizi 29 lokusa i 95,00 % pri analizi 21 lokusa. Prisustvo jednostepenih mutacija nije imalo uticaj na tačnost predikcije, koja je iznosila 95,02 %. Model zasnovan na učestalostima alela populacije Saudijske Arabije postigao je tačnost od 94,64 %, a model zasnovan na modifikovanim učestalostima alela za Saudijsku Arabiju 94,93 %. Model zasnovan isključivo na jedinstvenim *STR* profilima dece postigao je tačnost od 90,37 %. Sva tri modela zasnivala su se na analizi 21 lokusa.

Rezultati su pokazali da modeli mašinskog učenja mogu da detektuju informacije iz standardnih *STR* profila za oba analizirana zadatka.

Ključne reči: predikcija porekla, *STR* profili, mašinsko učenje, forenzička genetika, analiza srodstva, detekcija incesta

Naučna oblast: Biologija

Uža naučna oblast: Genetika

Predictive Modeling in Forensic Genetics: Ethnic Affiliation Classification and Incest Detection Based on Autosomal STR Profiles

Abstract

The aim of this dissertation was to develop and evaluate machine learning models for ancestry prediction and the detection of incestuous relationships based on autosomal STR profiles.

For ancestry prediction, 360,000 synthetic profiles were generated using allele frequencies at 29 loci from four reference groups: African American, Asian, Hispanic American, and Caucasian. Among the algorithms compared, XGBoost achieved the highest accuracy: 94.24% for the four reference groups, 99.06% after excluding the Hispanic American group, and 98.57% after merging it with the NA–EA group. Classification accuracy did not increase substantially when the number of profiles per group was increased from 10,000 to 90,000. The most informative loci for ancestry prediction were Penta E, F13A01, and SE33. The 21-locus combinations corresponding to the GlobalFiler and Investigator 24plex kits achieved classification accuracies of 88.28% for four and 96.30% for three population groups, respectively.

For incest detection, pairs of mother–child profiles were generated under scenarios involving non-incestuous paternity and father–daughter and brother–sister incest. CatBoost achieved an accuracy of 96.94% when analysing 29 loci and 95.00% when analysing 21 loci. The presence of single-step mutations had no effect on predictive accuracy, which was 95.02%. The model based on allele frequencies from the Saudi Arabian population achieved an accuracy of 94.64%, while the model based on modified Saudi Arabian allele frequencies achieved an accuracy of 94.93%. A model based solely on the unique STR profiles of the children achieved an accuracy of 90.37%. All three models were based on the analysis of 21 loci.

The results demonstrated that machine learning models can detect information from standard STR profiles that is relevant to both tasks investigated.

Key words: ancestry prediction, *STR* profiles, machine learning, forensic genetics, kinship analysis, incest detection

Scientific field: Biology

Scientific subfield: Genetics

Lista skraćenica

ABO – *ABO blood group system* (ABO sistem krvnih grupa)

AdaBoost – *Adaptive Boosting* (adaptivno pojačavanje)

AI – *Artificial Intelligence* (veštačka inteligencija)

AIM – *Ancestry-Informative Marker* (marker informativan za poreklo)

AIM-SNP – *Ancestry-Informative Marker Single Nucleotide Polymorphism* (jednonukleotidni polimorfizam informativan za poreklo)

AIM-STR – *Ancestry-Informative Marker Short Tandem Repeat* (kratki tandemski ponovak informativan za poreklo)

ROC AUC – *Area Under the Receiver Operating Characteristic Curve* (površina ispod ROC krive)

BGA – *Biogeographical Ancestry* (biogeografsko poreklo)

bp – *base pairs* (bazni parovi)

CE – *Capillary Electrophoresis* (kapilarna elektroforeza)

CII – *Combined Incest Index* (kombinovani indeks incesta)

CoAIMs – *Continental Ancestry-Informative Markers* (markeri informativni za kontinentalno poreklo)

CODIS – *Combined DNA Index System* (kombinovani sistem DNK indeksa)

CPI – *Combined Paternity Index* (kombinovani indeks očinstva)

DF – *Data Frame* (tabelarna struktura podataka)

DNK – *Deoxyribonucleic Acid* (dezoksiribonukleinska kiselina)

DT – *Decision Trees* (stabla odlučivanja)

DTNB – *Decision Table/Naive Bayes* (algoritam zasnovan na tabeli odlučivanja i naivnom Bajesovom klasifikatoru)

ENFSI – *European Network of Forensic Science Institutes* (Evropska mreža forenzičkih instituta)

EPG – *Electropherogram* (elektroforegram)

ESS – *European Standard Set* (evropski standardni set lokusa)

F – *Inbreeding Coefficient* (koeficijent inbridinga)

F1 – *F1 score* (F1 skor)

FBI – *Federal Bureau of Investigation* (Federalni istražni biro)

F_{ST} – *Fixation Index* (indeks genetičke diferencijacije između populacija)

GDA – *Genetic Distance Algorithm* (algoritam genetičke distance)

GWAS – *Genome-Wide Association Study* (studija asocijacije na nivou čitavog genoma)

HARE – *Harmonized Ancestry and Race/Ethnicity* (harmonizovano poreklo i rasa/etnicitet)

Hd – *Defence Hypothesis* (hipoteza odbrane)

HGDP – *Human Genome Diversity Project* (Projekat ljudske genetičke raznovrsnosti)

HGDP–CEPH – *Human Genome Diversity Project – Centre d'Étude du Polymorphisme Humain* (Projekat ljudske genetičke raznovrsnosti – CEPH panel)

Hp – *Prosecution Hypothesis* (hipoteza tužilaštva)

HUGO – *Human Genome Organisation* (Organizacija za ljudski genom)

HWE – *Hardy–Weinberg Equilibrium* (Hardi–Vajnbergova ravnoteža)

IBD – *Identical by Descent* (identično po poreklu)

IBS – *Identical by State* (identično po stanju)

IGSR – *International Genome Sample Resource* (Međunarodni resurs genomskih uzoraka)

II – *Incest Index* (indeks incesta)

KNN – *K–Nearest Neighbors* (algoritam K–najbližih suseda)

LIME – *Local Interpretable Model–Agnostic Explanations* (lokalno interpretabilna objašnjenja nezavisna od modela)

LightGBM – *Light Gradient Boosting Machine* (laka mašina gradijentnog pojačavanja)

LR – *Likelihood Ratio* (indeks verodostojnosti)

MCC – *Matthews Correlation Coefficient* (Matthewsov koeficijent korelacije)

MCMC – *Markov Chain Monte Carlo* (Monte Karlo metode zasnovane na Markovljevim lancima)

MDS – *Multidimensional Scaling* (multidimenzionalno skaliranje)

ML – *Machine Learning* (mašinsko učenje)

MLP – *Multi–Layer Perceptron* (višeslojni perceptron)

MLR – *Multinomial Logistic Regression* (multinomijalna logistička regresija)

MPS – *Massively Parallel Sequencing* (masivno paralelno sekvenciranje)

mtDNK – *mitochondrial DNA* (mitohondrijska DNK)

NB – *Naive Bayes* (naivni Bajesov klasifikator)

NDIS – *National DNA Index System* (Nacionalni sistem DNK indeksa)

NGM – *Next Generation Multiplex* (multipleks nove generacije)

NN – *Unknown Person / Unidentified Person* (nepoznato lice)

NOC – *Number of Contributors* (broj doprinosilaca)

NPV – *Negative Predictive Value* (negativna prediktivna vrednost)

NRY – *Non–Recombining Region of the Y Chromosome* (nerekombinjući region Y–hromozoma)

PCA – *Principal Component Analysis* (analiza glavnih komponenti)

PCoA – *Principal Coordinates Analysis* (analiza glavnih koordinata)

PCR – *Polymerase Chain Reaction* (polimerazna lančana reakcija)

PI – *Paternity Index* (indeks očinstva)

PPE – *Power of Paternity Exclusion* (snaga isključivanja očinstva)

PPV – *Positive Predictive Value* (pozitivna prediktivna vrednost)

RF – *Random Forest* (slučajna šuma)

ROC – *Receiver Operating Characteristic* (karakteristika operativnog prijema; ROC kriva)

ROH – *Runs of Homozygosity* (kontinualni segmenti homozigotnosti)

SA-EP – *severnoamerička populacija evropskog porekla (North American population of European ancestry)*

SHAP – *SHapley Additive exPlanations* (Shapley-jeva aditivna objašnjenja)

SNP – *Single Nucleotide Polymorphism* (jednonukleotidni polimorfizam)

SSM – *Slipped Strand Mismatching* (pogrešno uparivanje usled klizanja DNK lanca)

STR – *Short Tandem Repeats* (kratki tandemski ponovci)

SVM – *Support Vector Machines* (mašine potpornih vektora)

VNTR – *Variable Number of Tandem Repeats* (varijabilan broj tandemskih ponovaka)

XGBoost – *Extreme Gradient Boosting* (ekstremno gradijentno pojačavanje)

X-STR – *X-chromosomal Short Tandem Repeats* (X-hromozomski kratki tandemski ponovci)

Y-STR – *Y-chromosomal Short Tandem Repeats* (Y-hromozomski kratki tandemski ponovci)

SADRŽAJ

I UVOD.....	1
1.1. Forenzička genetika – istorijski razvoj, značaj i savremene primene.....	1
1.1.1. Serološka era i pre–DNK biološka identifikacija	1
1.1.2. Početak DNK profilisanja: <i>VNTR</i> /minisateliti i prve sudske primene.....	1
1.1.3. <i>PCR</i> revolucija i prelaz ka rutinskoj genotipizaciji (multiplex i <i>CE</i>)	1
1.2. Autozomni <i>STR</i> markeri – biološka osnova, polimorfizam i forenzički potencijal.....	1
1.2.1. <i>STR</i> lokusi: definicija, raspodela u genomu i tipovi ponovaka.....	1
1.2.2. Mehanizmi nastanka varijabilnosti i demografski signal <i>STR</i> markera.....	2
1.2.3. Forenzički potencijal i rutinska analiza: multiplex <i>PCR</i> i kapilarna elektroforeza	2
1.2.4. DNK registri i standardizovana razmena podataka: <i>CODIS</i> i <i>ESS</i>	3
1.2.5. Masivno paralelno sekvenciranje i <i>STR</i> varijante zasnovane na sekvenci.....	3
1.3. Osnovni principi populacione genetike kao osnova za analizu genetičke raznovrsnosti i predikciju porekla	4
1.3.1. Populaciona genetika u forenzičkom kontekstu: učestalosti i interpretacija	4
1.3.2. Valundov efekat i posledice populacione heterogenosti u forenzičkim proračunima	4
1.3.3. Neadaptivni procesi u evoluciji varijabilnosti	4
1.3.4. <i>IBD/IBS</i> i autozigoznost: značaj za procenu srodstva i ukrštanja u srodstvu	5
1.3.5. Globalna genetička raznovrsnost i prelaz ka prediktivnim modelima.....	5
1.4. Genetička struktura savremenih populacija – kontinent, subkontinent i etničke grupe.....	5
1.4.1. Pojam genetičke strukture i nivoi analize	5
1.4.2. Kontinentalna struktura i globalni referentni projekti.....	6
1.4.3. Subkontinentalna i regionalna struktura: klasteri i kontinumi	6
1.5. Biogeografsko poreklo – koncept, pristupi i izazovi	7
1.5.1. Definicija biogeografskog porekla i konceptualna ograničenja.....	7
1.5.2. Mešanje između populacija i genetički kontinuumi	7
1.5.3. Referentne populacije i baze genetičkih podataka u analizi biogeografskog porekla	8
1.6. Markeri za predikciju porekla	8
1.6.1. Autozomni markeri informativni za poreklo	8
1.6.2. Uniparentalni markeri: Y–hromozom i mitohondrijska DNK.....	9
1.6.3. Algoritmi za klasifikaciju biogeografskog porekla.....	10
1.7. <i>STR</i> markeri kao alat za predikciju porekla – prednosti, razvoj pristupa i ograničenja.....	11
1.7.1. Rani pokušaji i dokaz koncepta	11
1.7.2. Bajesovski modeli populacionog grupisanja i klasifikacije.....	11

1.7.3. Proširenje rezolucije: selekcija markera i pristupi mašinskog učenja.....	12
1.8. Konsangvinitet, inbriding i incest – genetičke i biološke osnove.....	13
1.8.1. Konsangvinitet i inbriding: pojmovna i populaciono–genetička distinkcija	13
1.8.2. Inbriding koeficijent (F), <i>IBD</i> i autozigoznost.....	13
1.8.3. Incest kao ekstremni oblik inbridinga.....	13
1.8.4. Biološke posledice inbridinga.....	14
1.8.5. Inbriding depresija	14
1.8.6. Manifestacija incesta u autozomnim <i>STR</i> profilima	14
1.9. Forenzički izazovi u detekciji incestuoznih odnosa.....	15
1.9.1. Preklapanje genetičkih signala: incest nasuprot populacionom inbridingu.....	15
1.9.2. Dokumentovani forenzički slučajevi kao ilustracija metodoloških problema	16
1.10. Indeks verodostojnosti kao pristup u forenzičkoj genetici: primena i ograničenja u srodničkim i incestuoznim scenarijima.....	16
1.10.1. Indeks verodostojnosti kao osnovni okvir interpretacije DNK dokaza	16
1.10.2. Indeks očinstva kao poseban oblik <i>LR</i> pristupa	17
1.10.3. Ograničenja <i>LR</i> interpretacije u incestuoznim scenarijima	17
1.10.4. Tehnička ograničenja <i>LR</i> modela.....	18
1.10.5. Metodološki jaz kao osnova za primenu mašinskog učenja	18
1.11. Primena savremenih <i>AI/ML</i> metoda u forenzičkoj genetici.....	18
1.11.1. Kategorije i metode u mašinskom učenju	18
1.11.2. Testirani algoritmi.....	19
1.11.3. Primene <i>AI/ML</i> u savremenoj forenzičkoj genetici.....	21
1.11.4. Prednosti modela <i>XGBoost</i> i <i>CatBoost</i> u obradi forenzičkih genetičkih podataka	22
1.12. Istraživački problem.....	23
Predmet naučnog istraživanja	24
HIPOTEZE I CILJEVI ISTRAŽIVANJA	25
II MATERIJAL I METODE.....	27
2.1 Pristup istraživanju.....	27
2.1.1. Izvori podataka.....	27
2.1.2. Generisanje <i>STR</i> profila na osnovu učestalosti alela	27
2.1.3. Generisanje <i>STR</i> profila na osnovu ukrštanja	28
2.1.4. Implementacija jednostepene mutacije	28
2.1.5. Ispitivanje homozigotnosti pri detekciji incesta.....	28
2.1.6. Ispitivanje lokusa sa identičnim heterozigotnim genotipovima majke i deteta	29
2.1.7. Računanje sličnosti <i>STR</i> profila majke i deteta.....	29

2.1.8. Podele baza podataka	29
2.1.9. Testiranje modela	29
2.1.10. Optimizacija hiperparametara	30
2.1.11. Evaluacija performansi modela	30
2.1.12. Analiza važnosti pojedinačnih lokusa	31
2.1.13. Vizualizacija genetičke diferencijacije (<i>PCA</i>)	31
2.1.14. Uticaj selekcije lokusa i veličine baze podataka na predikciju porekla	31
2.2. Specifičnosti analize	31
2.2.1. Predikcija porekla	31
2.2.2. Detekcija incestuoznih odnosa	31
2.2.3. Dostupnost podataka i skripti	32
III REZULTATI	33
3.1. Rezultati studije za predikciju porekla	33
3.1.1. Testiranje <i>ML</i> modela za predikciju porekla	33
3.1.2. Optimizacija hiperparametara <i>XGBoost</i> modela	33
3.1.3. Model predikcije na osnovu četiri osnovne populacije	34
3.1.4. Ocena informativnosti pojedinačnih lokusa za predikciju porekla	36
3.1.5. Model za klasifikaciju tri grupe (hispanoamerički poduzorak isključen)	36
3.1.6. Model za klasifikaciju tri grupe (objedinjene SA–EP i hispanoamerička populacija)	38
3.1.7. <i>PCA</i> analiza	39
3.1.8. Performanse modela klasifikacije sa promenljivim veličinama uzoraka	40
3.1.9. Uticaj broja i kombinacije lokusa na performanse modela	41
3.2. Rezultati studije detekcije incesta	42
3.2.1. Testiranje <i>ML</i> modela za detekciju incesta	42
3.2.2. Optimizacija hiperparametara <i>CatBoost</i> modela	42
3.2.3. Model za opštu populaciju (varijanta sa 29 lokusa)	43
3.2.4. Genetički indikatori konsangviniteta	45
3.2.5. Model za opštu populaciju (varijanta sa 21 lokusom)	47
3.2.6. Ocena informativnosti pojedinačnih lokusa za detekciju incesta	47
3.2.7. Model sa jednostepenom mutacijom za opštu populaciju (21 lokus)	48
3.2.8. Performanse modela na osnovu jedinstvenih profila dece (21 lokus)	49
3.2.9. <i>PCA</i> analiza	49
3.2.10. Početni model za Saudijsku Arabiju	51
3.2.11. Modifikovani model za Saudijsku Arabiju	52

3.2.12. Višeklasna klasifikacija.....	52
IV DISKUSIJA	54
4.1. Predikcija porekla	54
4.1.1. Tačnost prediktivnog modela pri predikciji porekla	54
4.1.2. Uticaj veličine skupa podataka	55
4.1.3. Broj testiranih lokusa i njihova informativnost.....	55
4.1.4. Praktična primena predikcije porekla u forenzičkim istragama.....	56
4.1.5. Poređenje sa prethodnim studijama	57
4.1.6. Prednosti i metodološka vrednost sintetičkih <i>STR</i> profila	58
4.1.7. Ograničenja i izazovi primene sintetičkih <i>STR</i> profila	59
4.1.8. Uticaj samoprijavljenog porekla na analizu.....	60
4.1.9. Kompleksnost genetičkog nasleđa hispanoameričkih populacija	61
4.2. Analiza incesta	61
4.2.1. Značaj analize incesta	61
4.2.2. Metodološki aspekti genetičke analize incesta	62
4.2.3. Dalji razvoj metodologije za analizu incesta	62
4.2.4. Homozigotnost i predikcija incesta na osnovu jedinstvenih profila dece	63
4.2.5. Identični heterozigotni genotipovi i ukupna sličnost <i>STR</i> profila.....	64
4.2.6. Implementacija jednostepene mutacije	64
4.2.7. Problematika višeklasne klasifikacije	65
4.2.8. Binarna klasifikacija incesta	66
4.2.9. Informativnost lokusa za predikciju incesta.....	67
4.3. Prediktivni modeli u forenzičkoj genetici: prednosti i izazovi u primeni.....	67
4.4. Ograničenja studije i buduće smernice istraživanja	68
4.5. Etika i odgovorna primena prediktivnih modela.....	69
4.6. Naučni doprinos istraživanja.....	70
ZAKLJUČCI.....	71
LITERATURA.....	73
Biografija autora	87
Izjava o autorstvu	88
Izjava o istovetnosti štampane i elektronske verzije doktorskog rada	89
Izjava o korišćenju	90

I UVOD

1.1. Forenzička genetika – istorijski razvoj, značaj i savremene primene

1.1.1. Serološka era i pre-DNK biološka identifikacija

Forenzička genetika se proteklih decenija razvila u zasebnu disciplinu savremenih forenzičkih nauka. Njenom razvoju prethodila je serološka era, krajem 19. i početkom 20. veka, kada su analize krvnih grupa, serumskih proteina i enzima predstavljale osnovu biološkog razlikovanja pojedinaca. Otkriće ABO krvnog sistema od strane *Karla Landsteinera* početkom 20. veka, a potom i potvrda da se krvne grupe nasleđuju, omogućili su ranu primenu krvnih grupa u forenzičkoj praksi, parničnim predmetima i antropološkim studijama (Dean, 2005). Iako je taj serološki sistem obezbeđivao mogućnost isključivanja pojedinaca, njegov ograničeni broj polimorfni varijanti davao je relativno mali broj fenotipova i time značajno ograničavao diskriminatornu moć, zbog čega se serologija u savremenom kontekstu smatra metodom ograničene forenzičke vrednosti (Gefrides i Welch, 2011).

1.1.2. Početak DNK profilisanja: VNTR/minisateliti i prve sudske primene

Alec Jeffreys je 1985. godine identifikovao hipervarijabilne minisatelitske sekvence, poznate i kao VNTR markeri (eng. *Variable Number of Tandem Repeats*), čime je uveden koncept DNK profilisanja (Jeffreys i sar., 1985). Već 1986. godine, DNK kao dokaz je primenjen u krivičnom postupku u slučaju *Regina v. Colin Pitchfork*, kada je analiza minisatelitskih profila omogućila isključenje pogrešno osumnjičenog, povezivanje dva ranije nepovezana slučaja silovanja i ubistva i, konačno, identifikaciju stvarnog počinioaca putem masovnog DNK testiranja lokalne zajednice. Ovaj slučaj se danas navodi kao jedan od prvih značajnih primera primene DNK profilisanja u forenzičkim istragama (Jobling i Gill, 2004). Iako su minisatelitski markeri značajno unapredili mogućnosti biološke identifikacije, njihova rutinska primena bila je ograničena Southern blot metodologijom, potrebom za većim količinama očuvane DNK, te smanjenom upotrebljivošću kod degradovanih tragova (Butler, 2005, 2012).

1.1.3. PCR revolucija i prelaz ka rutinskoj genotipizaciji (multiplex i CE)

Lančana reakcija polimeraze (PCR, eng. *Polymerase Chain Reaction*), razvijena sredinom 1980-ih godina, omogućila je enzimsko umnožavanje DNK i time znatno smanjila potrebnu količinu DNK materijala iz biološkog traga (Saiki i sar., 1985; Mullis i sar., 1986). Na ovo se nadovezuje razvoj simultanog umnožavanja više ciljnih sekvenci molekula DNK, multiplex PCR pristupa (Kimpton i sar., 1993), kao i uvođenje kapilarne elektroforeze (CE, eng. *Capillary Electrophoresis*) kao standardne metode detekcije alela (Butler, 2012). Ključni tehnološki i metodološki pomak u forenzičkoj genetici predstavljalo je otkriće i upotreba kratkih uzastopnih ponovaka (STR, eng. *Short Tandem Repeats*) (Butler, 2012).

1.2. Autozomni STR markeri – biološka osnova, polimorfizam i forenzički potencijal

1.2.1. STR lokusi: definicija, raspodela u genomu i tipovi ponovaka

STR sekvence predstavljaju kratke DNK sekvence, najčešće dužine od dva do šest nukleotida, organizovane u niz uzastopnih ponovaka, pri čemu se aleli na istom STR lokusu razlikuju prema broju ponovaka. Individue se mogu razlikovati prema kombinaciji alela koje poseduju u genotipu, dok se učestalosti pojedinih alela mogu razlikovati među populacijama (Willems i sar., 2014). Najveći deo STR lokusa nalazi se u nekodirajućim regionima genoma, pri čemu hromozom 19 sadrži najveću gustinu ovih ponavljajućih sekvenci (Fan i Chu, 2007). STR sekvence javljaju se približno na svakih

15 kb ljudskog genoma i organizovane su kao di-, tri-, tetra-, penta- ili heksanukleotidni ponovci (Butler, 2005). U ljudskom genomu najčešće su zastupljeni dinukleotidni *STR* ponovci, ali se u forenzičkoj praksi dominantno koriste tetranukleotidni lokusi (Alketbi, 2023). Tetranukleotidni *STR* ponovci omogućavaju stabilniju amplifikaciju jer je kod njih manji rizik od izostanka umnožavanja alela (tzv. ispadanja alela) i manji udeo artefaktnih *stater* alela (alela manjih ili većih od originalnog alela nastalih usled proklizavanja nastajućeg ili matričnog lanca prilikom umnožavanja DNK) u poređenju sa dinukleotidnim ponovcima, čime se olakšava interpretacija DNK profila (Brookes i sar., 2012; Butler, 2012).

1.2.2. Mehanizmi nastanka varijabilnosti i demografski signal *STR* markera

Promene u broju ponovaka stvaraju nove alelne varijante, zbog čega *STR* lokuse odlikuje visok stepen polimorfnosti. Biološka osnova njihove varijabilnosti zasniva se na replikacionoj nestabilnosti (*SSM*, eng. *Slipped strand mispairing*), pri kojoj tokom replikacije DNK dolazi do klizanja jednog DNK lanca u odnosu na drugi, što rezultuje dodavanjem ili gubitkom jednog ili više ponovaka (Levinson i Gutman, 1987). Relativno visoka stopa mutacije, reda veličine 10^{-3} do 10^{-4} , obezbeđuje ovim markerima izrazitu varijabilnost, ali i sposobnost da reflektuju demografske procese koji su oblikovali savremene ljudske populacije (Ellegren, 2004). *STR* lokusi se u populaciono-genetičkim analizama najčešće posmatraju kao selektivno neutralni markeri. Na toj pretpostavci zasniva se njihova primena u procenama vremena divergencije populacija, efektivne veličine populacije i obima protoka gena. Direktna uticaj selekcije na njihovu varijabilnost smatra se, u proseku, slabijim nego kod kodirajućih sekvenci (Ellegren, 2004).

1.2.3. Forenzički potencijal i rutinska analiza: multiplex *PCR* i kapilarna elektroforeza

Visoka informativnost i jednostavno umnožavanje autozomnih *STR* markera doveli su do njihove široke primene u forenzičkoj genetici tokom poslednjih decenija. U poređenju sa *VNTR* markerima, čiji aleli mogu dostići 400–1000 baznih parova (bp, eng. *base pairs*), veličine *STR* alela se uobičajeno kreću između 100 i 400 bp, što povećava verovatnoću uspešne amplifikacije i nakon degradacije DNK molekula (Butler, 2005). Prilikom odabira *STR* lokusa za forenzičke panele vodi se računa da lokusi budu genetički nezavisni, odnosno da se nalaze na različitim hromozomima ili da se nalaze na dovoljno udaljenim pozicijama istog hromozoma, kako bi se isključio uticaj genetičke vezanosti na statističku interpretaciju rezultata (Butler, 2012). Standardizacijom forenzičkih *STR* panela omogućene su razmene podataka između DNK registara i uporedivost statističke interpretacije. Kombinovanjem većeg broja visoko informativnih lokusa postiže se veoma niska verovatnoća slučajnog podudaranja, reda veličine 10^{-17} i niža, zbog čega se *STR* analiza standardno koristi za individualnu identifikaciju (Butler, 2015b).

Polimorfizam se odnosi na broj i raspodelu alela na određenom lokusu u populaciji, pri čemu su informativniji oni lokusi koji imaju veći broj alela i ravnomerniju raspodelu njihovih učestalosti (Butler, 2012, 2015a). Uočena heterozigotnost predstavlja učestalost heterozigotnih genotipova u analiziranom uzorku, dok očekivana heterozigotnost izražava verovatnoću da će nasumično izabrana individua na datom lokusu imati dva različita alela, pri čemu njena vrednost zavisi od broja alela i raspodele njihovih učestalosti u populaciji (Butler, 2015a). Lokusi sa većim stepenom heterozigotnosti proizvode veći broj mogućih genotipskih kombinacija i omogućavaju veće razlikovanje pojedinaca, zbog čega se primenjuju u forenzičkoj identifikaciji (Butler, 2012). Ta sposobnost se kvantifikuje kroz moć diskriminacije, koja izražava verovatnoću da će dva nasumično izabrana, nesrodna pojedinca imati različite genotipove na istom lokusu (Butler, 2015a). Suprotan parametar predstavlja verovatnoća slučajnog podudaranja, odnosno verovatnoća da dve nesrodne osobe iz iste populacije imaju isti genotip na datom lokusu (Butler, 2015a). Zbog toga se u forenzičkim panelima prednost daje lokusima kod kojih je moć diskriminacije visoka, a verovatnoća

slučajnog podudaranja niska, jer takvi lokusi najviše doprinose ukupnoj individualizacionoj vrednosti DNK profila (Butler, 2012; Moretti i sar., 2016).

1.2.4. DNK registri i standardizovana razmena podataka: *CODIS* i *ESS*

Razvojem nacionalnih i međunarodnih DNK registara omogućeni su poređenje genetičkih profila i standardizovana razmena podataka između laboratorija. U Sjedinjenim Američkim Državama, nakon pilot-projekta sprovedenog 1996–1997. godine, uspostavljen je Kombinovani sistem DNK indeksa (*CODIS*, eng. *Combined DNA Index System*), zasnovan na skupu od 13 osnovnih autozomnih *STR* lokusa. Ovi lokusi su odabrani zbog svoje visoke polimorfnosti, pri čemu su lokusi *FGA*, *D18S51* i *D21S11* pokazali naročito visok diskriminatorni potencijal, dok su *TPOX*, *CSF1PO* i *TH01* među manje informativnim lokusima (Butler, 2012). Tokom 2010–ih godina *CODIS* panel proširen je radi povećanja diskriminatorne moći i omogućavanja međunarodne kompatibilnosti, te je od 1. januara 2017. godine u rutinskoj upotrebi prošireni skup od ukupno 20 osnovnih *CODIS* lokusa, implementiran u sve validirane *PCR* komplete hemikalija prihvaćene za korišćenje u Nacionalnom sistemu DNK indeksa (*NDIS*, eng. *National DNA Index System*) (Butler, 2012; FBI, 2024). Paralelno sa razvojem američkog sistema, u Evropi je uspostavljen Evropski standardni set lokusa (*ESS*, eng. *European Standard Set*), koji je u svojoj inicijalnoj verziji od 2001. godine obuhvatao sedam autozomnih *STR* lokusa (*FGA*, *TH01*, *vWA*, *D3S1358*, *D8S1179*, *D18S51* i *D21S11*). Usled potrebe da se dodatno smanji verovatnoća slučajnog podudaranja u okviru međudržavne razmene DNK profila, *ESS* panel je 2009. godine proširen za pet dodatnih lokusa (*D10S1248*, *D22S1045*, *D2S441*, *D12S391* i *DIS1656*), čime je uspostavljen standard od ukupno 12 autozomnih *STR* lokusa, koji i danas predstavlja skup neophodan za međunarodnu razmenu DNK profila u okviru Evropske unije (Schneider, 2009; Butler, 2012; ENFSI, 2023). Iako *ESS* definiše osnovni standardni skup, savremena praksa podrazumeva analizu većeg broja lokusa kroz komercijalne multipleks sisteme, dok pojedini nacionalni DNK registri uključuju i dodatne lokuse, poput *SE33*, koji je zbog polimorfnosti prvo bio prepoznat i uključen u nemački nacionalni DNK registar (Butler, 2012).

1.2.5. Masivno paralelno sekvenciranje i *STR* varijante zasnovane na sekvenci

Masivno paralelno sekvenciranje (*MPS*, eng. *Massively Parallel Sequencing*) omogućilo je detaljniju analizu varijabilnosti *STR* markera. Za razliku od kapilarne elektroforeze, koja razlikuje alele isključivo po dužini, *MPS* omogućava detekciju varijanti ponavljajućih motiva i sekvenci koje ih ograničavaju (eng. *flanking*), čime se povećava informativnost svakog lokusa. Na taj način, isti nominalni (dužinski definisan) *STR* alel može biti razložen na više različitih varijanti koje nose dodatnu diskriminatornu informaciju (Gettings i sar., 2015). Više studija kvantifikovalo je porast broja detektovanih alela pri primeni *MPS* tehnologije. Novroski i sar. (2016) pokazali su da analiza sekvenci 27 autozomnih *STR* lokusa rezultuje uočavanjem 746 različitih alela, u poređenju sa 357 alela definisanih isključivo na osnovu dužina fragmenata, što predstavlja više nego dvostruko povećanje broja alela. Sličan obrazac zabeležen je i u preglednoj studiji autora Gettings i sar. (2015), gde je pokazano da pojedini visoko polimorfni lokusi pokazuju povećanje broja alela pri analizi sekvenci. Na primer, lokus *SE33* pokazuje porast sa 50 dužinski definisanih alela na 152 alela definisana na osnovu sekvence, odnosno broj detektovanih alela povećan je 3 puta. Kod lokusa *D12S391* zabeležen je porast sa 24 na 84 alela, što odgovara povećanju od 3,5 puta.

Gettings i sar. (2018) zabeležili su populaciono specifične učestalosti određenih motiva na *D3S1358* lokusu u afroameričkoj populaciji, kao i na lokusima *D4S2408* i *D16S539* u azijskoj populaciji. Ovo ukazuje da profili dobijeni analizom čitavih sekvenci mogu nositi dodatni signal populacione diferencijacije, koji nije evidentan pri analizi klasičnim *CE* pristupom.

1.3. Osnovni principi populacione genetike kao osnova za analizu genetičke raznovrsnosti i predikciju porekla

1.3.1. Populaciona genetika u forenzičkom kontekstu: učestalosti i interpretacija

Populaciona genetika predstavlja teorijski i empirijski okvir za proučavanje raspodele genetičke varijabilnosti unutar i između populacija, kroz analizu učestalosti alela i genotipova, kao i demografskih i evolucionih procesa koji oblikuju njihovu dinamiku kroz generacije (Charlesworth i Charlesworth, 2017).

Jedan od temeljnih koncepata populacione genetike jeste Hardy–Vajnbergova ravnoteža (*HWE*, eng. *Hardy–Weinberg equilibrium*), koja opisuje teorijsku osnovu genetičke ravnoteže u idealizovanoj populaciji. Takva populacija je efektivno velika, sa nasumičnim ukrštanjem (panmiksijom), a mutacije i selekcija se mogu zanemariti, te ne dovode do promena učestalosti alela. U tim uslovima, učestalosti genotipova mogu se izračunati na osnovu učestalosti alela i ostaju konstantne nakon jedne generacije panmiktickog ukrštanja (Mayo, 2008).

U praksi, *HWE* se koristi kao referentna („nulta“) pretpostavka koja podrazumeva statističku nezavisnost alela unutar genotipa, dok se odstupanja tumače kao posledica narušavanja te pretpostavke. Takva odstupanja mogu nastati usled populacionih procesa kao što su genetički drift u populacijama ograničene veličine, ukrštanja u srodstvu, stratifikacije populacije i mešanja između populacija (eng. *admixture*; u daljem tekstu: admiksija) (Trikalinos i sar., 2006).

1.3.2. Valundov efekat i posledice populacione heterogenosti u forenzičkim proračunima

Valundov efekat (eng. *Wahlund effect*) javlja se kada se više genetički diferenciranih subpopulacija posmatra kao jedinstvena celina, što dovodi do smanjenja ukupne heterozigotnosti i povećanja homozigotnosti u odnosu na očekivanja prema *HWE* modelu, čak i ako je svaka subpopulacija pojedinačno u ravnoteži (Garnier–Géré i Chikhi, 2013; De Meeûs, 2018). U forenzičkom kontekstu, pogrešno pretpostavljanje populacione homogenosti može dovesti do pristrasnosti u proceni verovatnoće slučajnog poklapanja DNK profila, zbog čega je potrebno primeniti odgovarajuće učestalosti alela i korektivne faktore (Garnier–Géré i Chikhi, 2013).

1.3.3. Neadaptivni procesi u evoluciji varijabilnosti

Ljudske populacije sa izraženim osnivačkim efektima i genetičkim driftom mogu se analizirati na primeru istorijski endogamnih jevrejskih zajednica. Analize na nivou čitavog genoma pokazale su povećano deljenje segmenata DNK identičnih po poreklu (*IBD*, eng. *Identical by Descent*) unutar pojedinih jevrejskih populacija, pri čemu je ono bilo naročito izraženo kod libijskih, džerbskih i tuniskih Jevreja. Opšti stepen deljenja unutar analiziranih jevrejskih populacija bio je uporediv sa onim koji se očekuje između četvrtih i petih rođaka. Analize glavnih komponenti (*PCA*, eng. *Principal Component Analysis*) pokazale su formiranje centralnog genetičkog klastera koji je obuhvatao jevrejske populacije Evrope, Severne Afrike i Bliskog istoka, uz dodatno razdvajanje pojedinačnih zajednica u subklustere. U pojedinim jevrejskim populacijama zabeležena je i povećana učestalost određenih autozomno recesivnih monogenskih poremećaja, uključujući Tay–Sachsovu i Gaucherovu bolest. Iako su kao mogući mehanizmi povećane učestalosti pojedinih mutacija razmatrani i selekcija i genetički drift, raspodela aškenaskih osnivačkih mutacija i rezultati testova neutralnosti podržali su drift više nego selekciju kao objašnjenje njihovih povećanih učestalosti (Ostrer i Skorecki, 2013).

1.3.4. IBD/IBS i autozigoznost: značaj za procenu srodstva i ukrštanja u srodstvu

Razlikovanje alela identičnih po poreklu (*IBD*) i alela identičnih po stanju (*IBS*, eng. *Identical by State*) relevantno je za analizu predikcije porekla i ukrštanja u srodstvu. *IBD* aleli predstavljaju varijante koje dele zajedničko poreklo u odnosu na referentnu populaciju, odnosno potiču od istog alela prisutnog kod zajedničkog pretka, dok *IBS* aleli označavaju alele koji su međusobno jednaki po stanju (npr. veličini ili sekvenci), ali bez implikacije o zajedničkom poreklu (Powell i sar., 2010).

U velikim, približno panmiksičkim populacijama, podudarnost alela između nepovezanih individua najčešće je posledica *IBS*-a, dok u genetički izolovanim, osnivačkim ili inbredovanim populacijama raste verovatnoća da aleli budu identični po poreklu, usled povećanog deljenja zajedničkih predaka u skorijoj prošlosti (Powell i sar., 2010). U tom smislu, i individue koje se u populaciono–genetičkom smislu smatraju „nepovezanim“ mogu deliti određeni stepen *IBD*-a, budući da su svi članovi populacije, posmatrani dovoljno unazad u vremenu, genealoški povezani, pri čemu razlike među parovima individua proizilaze iz različite bliskosti tog zajedničkog porekla (Powell i sar., 2010).

Specifičan oblik *IBD*-a predstavlja autozigotnost, kod koje su oba alela u homozigotnom genotipu identična po poreklu i potiču od istog pretka, najčešće kao posledica nedavnog ukrštanja u srodstvu ili brakova unutar ograničenog kruga srodnika (Powell i sar., 2010). Ovaj fenomen se koristi za kvantitativnu procenu ukrštanja u srodstvu na nivou genoma.

1.3.5. Globalna genetička raznovrsnost i prelaz ka prediktivnim modelima

Savremeni projekti sekvenciranja i mapiranja genetičke varijabilnosti, poput velikih međunarodnih referentnih studija ljudske genetičke raznovrsnosti, omogućili su detaljan uvid u globalnu raspodelu genetičkih varijanti u ljudskom genomu. Rezultati ovih istraživanja pokazuju da se najveći deo ukupne genetičke varijabilnosti, 93–95 %, nalazi unutar populacija, dok razlike između glavnih populacionih grupa doprinose svega 3–5 %. Ipak, uprkos relativno malom udelu međupopulacione varijabilnosti, postoje stabilni obrasci u raspodeli učestalosti alela i strukturi genotipova koji odražavaju dugotrajne demografske procese i mogu omogućiti razlikovanje populacija (Rosenberg i sar., 2002).

1.4. Genetička struktura savremenih populacija – kontinent, subkontinent i etničke grupe

1.4.1. Pojam genetičke strukture i nivoi analize

Pojam genetičke strukture populacije odnosi se na obrazac raspodele učestalosti alela unutar i između populacija, koji odražava istoriju razdvajanja, migracija i mešanja između populacija. U opštem populaciono–genetičkom smislu, obrasci genetičke strukture ne moraju odražavati savremeni protok gena, već mogu odražavati posledice istorijskih događaja, kao što su pleistocenske glacijacije i postglacijalne ekspanzije areala, čiji se efekti mogu zadržati u genetičkom zapisu populacija tokom dugih vremenskih perioda (Bossart i Pashley Prowell, 1998). Analize multilokusnih genotipova pokazale su da je genetička diferencijacija na globalnom, kontinentalnom nivou izražena i relativno stabilna, dok se na regionalnom i lokalnom nivou uočavaju postepeni prelazi u učestalostima alela i delimična pripadnost individua različitim genetičkim klasterima. Ovakvi obrasci tumače se kao posledica kontinuiranog protoka gena između susednih populacija i istorijskih procesa njihovog mešanja, a ne kao dokaz genetičkih granica između regionalnih i lokalnih populacija (Rosenberg i sar., 2002).

1.4.2. Kontinentalna struktura i globalni referentni projekti

Na globalnom nivou, ljudska genetička raznovrsnost pokazuje izraženu strukturu koja približno odgovara kontinentalnim regionima. Analize zasnovane na podacima Projekta ljudske genetičke raznovrsnosti – Centra za proučavanje ljudskog polimorfizma (*HGDP–CEPH*, eng. *Human Genome Diversity Project–Centre d’Étude du Polymorphisme Humain*) pokazale su prepoznatljive obrasce genetičke sličnosti unutar ovih regiona, nastale akumulacijom suptilnih razlika u učestalostima alela velikog broja lokusa (Aneli i sar., 2022).

Dodatno, poređenje podataka *HGDP–CEPH* i Projekta 1000 genoma pokazalo je da *HGDP* sadrži značajan broj alternativnih alela koji, iako česti ili čak veoma česti u pojedinim populacijama, nisu prisutni u Projektu 1000 Genoma, uključujući približno milion varijanti sa učestalošću $\geq 20\%$, oko 100.000 varijanti sa učestalošću $\geq 50\%$, kao i oko 1000 varijanti fiksiranih u barem jednoj populaciji, pri čemu su Afrika, Okeanija i Amerika regioni sa najvećim udelom takvih varijanti (Bergström i sar., 2020). Međutim, nijedna varijanta nije pronađena kao fiksirana na nivou čitavog kontinenta ili velike geografske oblasti, što potvrđuje da je većina zajedničke genetičke varijacije globalno deljena, dok se kontinentalna struktura prvenstveno manifestuje kroz razlike u učestalostima alela, a ne kroz postojanje alela specifičnih na nivou kontinenta (Bergström i sar., 2020).

1.4.3. Subkontinentalna i regionalna struktura: klasteri i kontinumi

Na subkontinentalnom i regionalnom nivou, genetička struktura postaje finija. U Evropi se uočavaju regionalne razlike u genetičkom sastavu populacija koje odražavaju različite udele glavnih predačkih komponenti. Udeo rane farmerske komponente bliskoistočnog porekla pokazuje izražen geografski gradijent, krećući se od približno 30 % u baltičkom regionu do oko 90 % u mediteranskim populacijama, dok severne populacije imaju veći udeo zapadnoevropske lovačko–sakupljačke komponente u poređenju sa južnim delovima Evrope (Lazaridis i sar., 2014).

Slična slojevitost uočava se i u Aziji, naročito u jugoistočnoj Aziji, gde genetički obrasci odražavaju dugotrajnu demografsku dinamiku i višestruke migracione talase povezane sa širenjem poljoprivrede iz istočne Azije, interakcijama sa autohtonim lovačko–sakupljačkim populacijama i kasnijim regionalnim migracijama (Lipson i sar., 2018). Ovakvi primeri subkontinentalne i regionalne genetičke diferencijacije ukazuju da se genetička varijabilnost u ljudskim populacijama raspoređuje kontinuirano, zbog čega je potrebno razgraničiti biološke i sociokulturne pojmove koji se koriste u opisivanju ljudske raznolikosti.

Na najužem nivou, etnička pripadnost (eng. *Ethnicity*) posmatra se kao sociokulturni konstrukt (jezik, religija, običaji i samoidentifikacija), koji ne mora imati direktnu vezu sa genetičkim kategorijama (Rebbeck i Sankar, 2005). Nasuprot tome, poreklo (eng. *Ancestry*) je pojam usmeren na genetičku varijabilnost i istorijske migracije, tj. na biološko nasleđe i raspodelu genetičkih varijanti unutar i između populacija (Rebbeck i Sankar, 2005).

U izvornom skupu podataka koji su objavili Hill i sar. (2013), anonimni uzorci donora bili su raspoređeni u četiri kategorije prema samoprijavljenoj etničkoj pripadnosti: afroameričku, azijsko–američku, hispanoameričku i kategoriju *Caucasian*. Steffen i sar. (2017) naknadno su objavili revidirane učestalosti alela, koje su korišćene u ovom istraživanju, a koje su dostupne na internet stranici kompanije Promega. Navedene kategorije u disertaciji nisu bile tretirane kao pojedinačne etničke zajednice, već kao referentne populacione grupe definisane učestalostima alela iz izvornih skupova podataka. Kategorija *Caucasian* označena je, radi veće terminološke preciznosti, kao severnoamerička populacija evropskog porekla (SA–EP).

Shodno tome, u daljem tekstu korišćen je termin „poreklo“, „biogeografsko poreklo“ (*BGA*, eng. *biogeographical ancestry*) i „pripadnost referentnoj populacionoj grupi“. Termin „etnička pripadnost“ zadržan je u zvaničnom naslovu disertacije. Prediktivni modeli razvijeni u okviru ove disertacije nisu određivali etničku pripadnost pojedinca kao sociokulturni identitet, već su klasifikovali sintetičke *STR* profile u jednu od četiri referentne populacione grupe.

1.5. Biogeografsko poreklo – koncept, pristupi i izazovi

1.5.1. Definicija biogeografskog porekla i konceptualna ograničenja

Koncept biogeografskog porekla zasniva se na ideji da genetički sastav osobe odražava istorijske veze između njenih predaka i određenih geografskih regiona, jer su u genomu savremenih populacija sačuvani tragovi migracija, izolacija i mešanja tokom milenijuma (Rosenberg i sar., 2002). Biogeografsko poreklo u genetičkom smislu ne predstavlja etnički identitet niti kulturnu pripadnost (Royal i sar., 2010), već se procenjuje iz statistički prepoznatljivih obrazaca učestalosti alela koji se javljaju u određenim populacionim grupama (McNevin, 2020). U forenzičkom kontekstu, *BGA* se koristi kao dopunski analitički parametar, koji može pružiti orijentacionu informaciju o poreklu nepoznate osobe, posebno u situacijama kada direktna identifikacija nije moguća (Devesse i sar., 2023).

U predikciji porekla genetički podaci koriste se za prepoznavanje pripadnosti populaciji, dok su izbor markera, adekvatnost referentnih uzoraka i modela za poređenje presudni za validnost zaključaka (McNevin, 2020).

1.5.2. Mešanje između populacija i genetički kontinuumi

Mešanje između populacija može otežati određivanje populacione pripadnosti. Problem je izražen u populacijama sa složenom demografskom istorijom, u kojima je savremena genetička varijabilnost oblikovana višestrukim migracionim tokovima i njihovim međusobnim mešanjem. Primer genetički heterogene populacije jeste Latinska Amerika, čiji je savremeni genetički sastav posledica mešanja evropskih, autohtonih američkih i afričkih populacija tokom istorije (Ruiz–Linares i sar., 2014).

Ovakva genetička struktura utiče na modele za predikciju biogeografskog porekla. U admiksovanim populacijama, pristupi koji pretpostavljaju postojanje jasno definisanih referentnih klastera mogu dovesti do problematičnih interpretacija. Umesto jedinstvene klasifikacije porekla, savremeni modeli predikcije treba da budu sposobni da identifikuju i kvantifikuju relativne doprinose različitih predačkih komponenti. Admiksija je česta pojava u savremenim ljudskim populacijama, a genetička varijacija se manifestuje u obliku gradijenata, a ne strogo razdvojenih populacionih grupa, što predstavlja izazov za procenu porekla zasnovanu na panelima sa ograničenim brojem markera informativnih za poreklo (Phillips i sar., 2007; Novembre i sar., 2008; Kidd i sar., 2014).

Dodatni konceptualni i interpretativni izazov predstavlja razlikovanje biogeografskog porekla, kao statističkog genetičkog signala zasnovanog na obrascima učestalosti, i etničke pripadnosti, koja predstavlja sociokulturnu kategoriju. Iako se ovi pojmovi u određenim kontekstima delimično preklapaju, genetičke analize ne opisuju kulturološki aspekt ili samoidentifikaciju pojedinca, već probabilističke obrasce sličnosti sa referentnim populacijama. Studije sprovedene u Latinskoj Americi jasno pokazuju da je samopercepcija ispitanika često pristrasna i snažno oblikovana fenotipskim i socijalnim faktorima (Ruiz–Linares i sar., 2014). Ista studija ukazuje da se kategorije samopercipirane etničke/rasne pripadnosti i fenotipski indikatori ne mogu pouzdano tretirati kao direktan ekvivalent genetičke procene porekla, jer se genetičke komponente značajno razlikuju između zemalja i unutar istih kategorija.

1.5.3. Referentne populacije i baze genetičkih podataka u analizi biogeografskog porekla

Dodatno ograničenje u proceni biogeografskog porekla proizlazi iz strukture i karakteristika referentnih populacija i baza genetičkih podataka.

Referentne populacije omogućavaju predikciju biogeografskog porekla poređenjem genotipa nepoznatog donora sa skupovima genotipova poznatog porekla (Cheung i sar., 2018). U savremenoj forenzičkoj genetici, referentni genotipovi najčešće se preuzimaju iz javno dostupnih baza genetičkih podataka koje obuhvataju globalne i regionalne populacione skupove sa različitim stepenom geografske pokrivenosti i brojem ispitanika po populaciji (McNevin, 2020), uključujući resurse poput *HGDP-CEPH* baze, *IGSR* baze (eng. *International Genome Sample Resource*, ranije 1000 Genomes Project) i *HUGO Pan-Asian SNP* baze (eng. *Human Genome Organisation Pan-Asian SNP Consortium*) (Cheung i sar., 2018). Iako baze sa širokom geografskom zastupljenošću nude bolju ukupnu reprezentativnost, one sadrže ograničen broj uzoraka po pojedinačnim subpopulacijama, dok su kontinentalne kategorizacije same po sebi delimično arbitrarne, naročito u geografskim zonama između Evrope i Azije (McNevin, 2020). Dodatni metodološki problem predstavlja činjenica da značajan deo populacija u ovim bazama pokazuje različite stepene genetičke admiksije, uprkos formalnoj klasifikaciji i samodeklaraciji porekla, što može dovesti do sistematskih grešaka u treniranju i primeni klasifikacionih algoritama (Cheung i sar., 2018). Referentne populacije treba da budu reprezentativne za ciljne biogeografske grupe, jer neadekvatan izbor skupa za treniranje povećava rizik od pogrešne klasifikacije, naročito kod admiksivanih individua (Cheung i sar., 2018; McNevin, 2020). Ograničenja postojećih referentnih populacija ogledaju se u nemogućnosti da se pouzdano razlikuju pojedinci sa mešovitim poreklom od onih čiji genetički profil potiče iz jedne, ali u referentnom skupu nereprezentovane populacije (Cheung i sar., 2018). Iako bi referentne populacije izvedene iz drevnih ljudskih ostataka mogle pružiti veću genetičku „čistoću“, njihova primena je ograničena praktičnim i metodološkim problemima, uključujući ograničenu dostupnost uzoraka, nedostatak pouzdanih podataka o poreklu i rizik od kontaminacije (Cheung i sar., 2018). U forenzičkom kontekstu, naročito u heterogenim sredinama, ova ograničenja mogu umanjiti tačnost *BGA* predikcije, pa je poznavanje strukture referentnih baza i njihovih ograničenja potrebno radi sprečavanja preterano sigurnih zaključaka (Cheung i sar., 2018; McNevin, 2020).

1.6. Markeri za predikciju porekla

Predikcija biogeografskog porekla u kontekstu skorašnje demografske istorije populacija može se zasnivati na analizi autozomnih *IBD* haplotipskih segmenata. Tokom generacija, rekombinacija deli nasledne haplotipove, tako da udaljenje od zajedničkog pretka rezultuje kraćim *IBD* segmentima, dok duži segmenti ukazuju na skorašnje zajedničko poreklo. Zbog toga dužina *IBD* segmenata predstavlja informativan izvor podataka, omogućavajući rekonstrukciju migracija i populacionih procesa (Ralph i Coop, 2013).

1.6.1. Autozomni markeri informativni za poreklo

Termin markeri informativni za poreklo (*AIMs*, eng. *ancestry informative markers*) se odnosi na autozomne markere koji istovremeno odražavaju i maternalni i paternalni genetički doprinos nepoznatom profilu (Rosenberg i sar., 2003). U praktičnom smislu, *AIMs* se identifikuju kao lokusi čije se učestalosti alela razlikuju između populacija, pri čemu je geografska raspodela genetičkih polimorfizama u velikoj meri povezana sa obrascima ljudske istorije, migracija i populacione strukture (Tvedebrink, 2022).

Stepen diferencijacije između populacija često se kvantifikuje indeksom F_{ST} , koji predstavlja meru razlika u učestalostima alela između populacija i može imati vrednosti od 0, kada razlike ne postoje, do 1, kada su različiti aleli fiksirani u različitim populacijama (Myles i sar., 2009). U tom

smislu, *SNP* lokusi sa visokim F_{ST} vrednostima smatraju se posebno informativnim za detekciju populacione strukture i koriste se kao *AIMs* (Myles i sar., 2009). U forenzičkom kontekstu, *AIM* paneli moraju biti dovoljno osetljivi za uzorke sa minimalnom količinom biološkog materijala, zbog čega se biraju relativno mali, ciljano selektovani setovi markera (Phillips, 2015).

Klasični autozomni *STR* lokusi su, međutim, dizajnirani da maksimizuju varijabilnost između pojedinaca unutar iste populacije, što rezultuje visokom heterozigotnošću i velikom diskriminativnom moći na nivou pojedinca. Iako se zbog niskih F_{ST} vrednosti smatraju slabije informativnim za populacionu strukturu, empirijske analize pokazuju da nizak F_{ST} kod visoko heterozigotnih lokusa ne mora odražavati odsustvo informacije o poreklu. Algee-Hewitt i sar. (2016) dokazuju da je F_{ST} matematički ograničen kod lokusa sa visokom heterozigotnošću, zbog čega čak i markeri sa izraženim populacionim signalom mogu imati niske F_{ST} vrednosti. Posledično, kod forenzičkih *STR* lokusa populaciona informacija se pre svega ispoljava kroz obrasce učestalosti alela, uključujući raspodelu retkih alela, dok F_{ST} predstavlja ograničenu meru njihove informativnosti o poreklu. Za markere kod kojih varijabilnost nastaje stepenastim promenama broja ponovaka, kakvi su *STR* lokusi, kao dopuna F_{ST} koristi se i R_{ST} statistika, koja uzima u obzir razlike u veličini alela, odnosno broju ponovaka između alela (Slatkin, 1995).

Iako su *STR* markeri standard u forenzičkoj identifikaciji i testovima srodstva, u kontekstu predikcije porekla najčešće se koriste *SNP* paneli, pa se i komercijalne forenzičke analize porekla dominantno oslanjaju na pažljivo odabrane setove autozomnih *AIM-SNP* (Tvedebrink, 2022). Prednost *SNP* markera ogleda se u njihovoj brojnosti, ravnomernoj raspodeli duž genoma i izuzetno niskoj mutacionoj stopi, što ih čini stabilnim kroz veliki broj generacija i pogodnim za analizu populacionih obrazaca koji odražavaju dugoročnu demografsku istoriju (Phillips, 2015; Goli i sar., 2024). Pored *SNP*-ova, kao potencijalni *AIMs* mogu se koristiti i mikrohplotipovi (grupe blisko lociranih *SNP*-ova) i inserciono-delecioni polimorfizmi (indel), što proširuje spektar dostupnih markera za predikciju porekla (Tvedebrink, 2022).

Rezultati su uslovljeni izborom markera uključenih u analizu: paneli dizajnirani da imaju veću rezoluciju za određene (sub)populacione strukture mogu imati ograničenu diskriminativnu moć u drugim regionima, pa se zaključci o poreklu moraju interpretirati u okviru referentnih populacija i konkretnog *AIM* skupa (Tvedebrink, 2022).

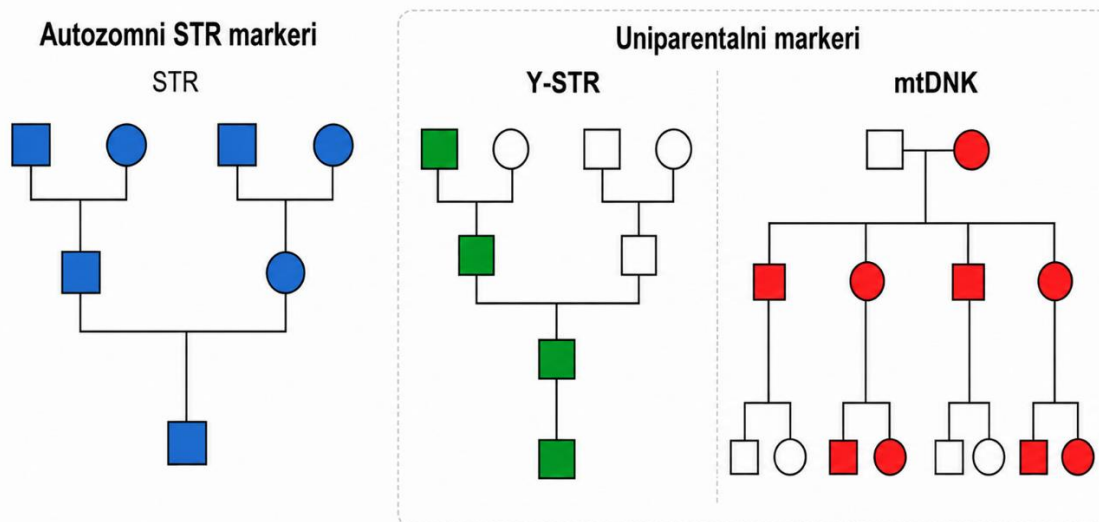
1.6.2. Uniparentalni markeri: Y-hromozom i mitohondrijska DNK

Pre uvođenja autozomnih *AIMs*, u forenzičkoj genetici su se široko koristili tzv. uniparentalni markeri, odnosno markeri koji se nasleđuju po muškoj ili ženskoj liniji (Shriver i Kittles, 2004). Ovoj grupi pripadaju Y-hromozomski markeri sa nerekombinujućeg dela Y-hromozoma (*STR*), koji su prisutni isključivo kod muškaraca i nasleđuju se isključivo paternalnom linijom (Jobling i Tyler-Smith, 2003), kao i mitohondrijska DNK (mtDNK), koja se nalazi kod oba pola, ali se prenosi isključivo maternalnom linijom (Behar i sar., 2012). Ovi regioni predstavljaju segmente genoma koji se, zbog odsustva rekombinacije, prenose kroz paternalne (*STR*) odnosno maternalne (mtDNK) linije, čime su pogodni za praćenje jedne genealogijske grane kroz veći broj generacija (Calafell i Larmuseau, 2017). Kombinacija materinskog nasleđivanja i odsustva rekombinacije kod mtDNK omogućava rekonstrukciju genealoških odnosa i praćenje razlika u učestalostima matrilinearnih klada između populacija na kontinentalnim i regionalnim nivoima (Kivisild, 2015).

Nasuprot tome, kao što je navedeno, autozomni markeri predstavljaju kombinovani signal svih predaka, sa najjačim doprinosom najskorijih generacija, što ih čini pogodnijim za individualnu predikciju porekla (Phillips, 2015). Halder i sar. (2008) navode da autozomni markeri omogućavaju merenje admiksije „unutar individue“ kroz doprinose svih predaka, a ne samo dela linija. Informativni signal se kod autozomnih markera sa porastom broja mejoza progresivno smanjuje. Kod udaljenijeg

srodstva (iznad šestog stepena), čak i analize zasnovane na čitavom genomu pate od smanjene preciznosti usled rekombinacije, što predstavlja jedno od ograničenja autozomnih pristupa u genealoškom smislu (Calafell i Larmuseau, 2017).

Ova razlika između uniparentalnog i autozomnog nasleđivanja grafički je prikazana na Slici 1, gde se jasno uočava da Y-hromozom i mtDNK prate paternalne, odnosno maternalne linije, dok autozomna DNK predstavlja kombinovani genetički doprinos svih predaka u prethodnim generacijama.



Slika 1. Obrasci nasleđivanja genetičkih markera. Prikazani su obrasci nasleđivanja autozomnih i uniparentalnih markera. Autozomni *STR* markeri se nasleđuju od oba roditelja i kroz rekombinaciju predstavljaju kombinovani doprinos većeg broja predaka. Nasuprot tome, Y-hromozom se prenosi po paternalnoj liniji i odražava jednu naslednu liniju. Mitohondrijska DNK se prenosi maternalno, pri čemu je majka prenosi na svu decu. Kvadrati označavaju muškarce, krugovi žene, dok plava, zelena i crvena boja označavaju autozomne *STR* markere, Y-*STR* markere i mitohondrijsku DNK.

1.6.3. Algoritmi za klasifikaciju biogeografskog porekla

Predikcija biogeografskog porekla, pored izbora genetičkih markera, zavisi i od algoritama za klasifikaciju genotipova (McNevin, 2020). Ovi algoritmi mogu se svrstati u tri šire kategorije: Bajesove, metode zasnovane na distanci i regresione metode (Cheung i sar., 2017). Pored ovih pristupa, koriste se i multidimenzionalne metode (*MDS*, eng. *multidimensional scaling*), poput analize glavnih komponenti (*PCA*) i analize glavnih koordinata (*PCoA*, eng. *principal coordinates analysis*), koje služe za prostornu reprezentaciju genetičkih distanci između genotipova (McNevin, 2020).

Bajesovi klasifikatori koriste informacije o referentnim populacijama, izražene kroz učestalosti genotipova i alela, kako bi ažurirali *a priori* pretpostavke o pripadnosti pojedinca određenoj biogeografskoj grupi i izračunali posteriorne verovatnoće pripadnosti. Najpoznatiji predstavnik ovog pristupa, *STRUCTURE*, opisan kao Bajesovski metod modelskog klasterovanja, prvobitno je razvijen za populacionu analizu, ali je kasnije prilagođen za *BGA* analizu, pri čemu se genotipovi dodeljuju jednoj ili više populacija definisanih brojem *K*, uz pretpostavke *HWE* ravnoteže i odsustva vezanosti lokusa (Cheung i sar., 2017; McNevin, 2020).

Generički Bajesov klasifikator koristi genotipske učestalosti iz skupa za treniranje, u ponderisanoj i neponderisanoj varijanti u odnosu na veličinu referentnih populacija. Neponderisana

varijanta podrazumeva jednak doprinos svih populacija nezavisno od broja uzoraka, dok ponderisana varijanta uzima u obzir njihovu relativnu zastupljenost, čime se omogućava ispitivanje uticaja neravnomernog uzorkovanja na tačnost klasifikacije. Ovaj pristup ne zahteva procenu učestalosti alela pod pretpostavkom *HWE*, što ga čini metodološki jednostavnijim za primenu (Cheung i sar., 2017).

Pristup zasnovan na distanci oslanja se na kvantifikaciju razlika u učestalosti genotipova između populacija i pokazuje osetljivost na admiksivane populacije. Kao regresioni pristup, multinomijalna logistička regresija omogućava višeklasnu klasifikaciju, ali se pokazala problematičnom u prisustvu nedostajućih genotipova, koji se kodiraju kao posebna kategorija i dovode do povećanog broja pogrešnih klasifikacija (Cheung i sar., 2017).

Uporedna analiza performansi zasnovana na *ROC AUC* pokazala je da su Bajesovski pristupi, naročito *STRUCTURE* i generički Bajesov klasifikator, imali najvišu diskriminatornu sposobnost i stabilnost kako za kompletne tako i za parcijalne genotipove. Nasuprot tome, algoritam genetičke distance (*GDA*) i naročito multinomijalna logistička regresija (*MLR*) pokazuju smanjenu pouzdanost kod admiksivanih uzoraka, zbog čega autori zaključuju da generički Bajesov klasifikator predstavlja praktičnu i lako primenljivu alternativu programu *STRUCTURE* u *BGA* analizi (Cheung i sar., 2017).

Metode poput *PCA* zasnivaju se na numeričkoj reprezentaciji genotipova i mogu da očuvaju genetičke distance pri direktnom numeričkom kodiranju bialelnih markera, dok kod multialelnih varijanti, poput trialelnih *SNP*-ova i mikrohplotipova, izbor numeričkog kodiranja može dovesti do iskrivljenja genetičkih distanci, ukoliko se ne primene odgovarajuće pomoćne promenljive. Za razliku od toga, *PCoA* kao ulaz koristi matricu genetičkih distanci između individua, čime omogućava analizu genotipova sa više alelnih varijanti. Ipak, budući da se prostorna reprezentacija ograničava na dve ili tri dimenzije, deo ukupne genetičke varijabilnosti ostaje nepredstavljen, zbog čega *MDS* metode imaju pre svega eksplorativnu i ilustrativnu, a ne striktno klasifikacionu ulogu (McNevin, 2020).

1.7. STR markeri kao alat za predikciju porekla – prednosti, razvoj pristupa i ograničenja

1.7.1. Rani pokušaji i dokaz koncepta

Rani pokušaji primene *STR* markera u predikciji porekla bili su ograničeni brojem dostupnih lokusa i skromnim statističkim pristupima. Lowe i sar. (2001) među prvima su pokazali da već i mali panel od šest autozomnih *STR* lokusa (*HUMVWA31/A*, *HUMTH01*, *HUMFIBRA*, *D8S1179*, *D21S11* i *D18S51*) može da pruži diskriminativnu moć pri razlikovanju pet etničkih grupa u Ujedinjenom Kraljevstvu. Iako ograničene preciznosti, studija autora Lowe i sar. (2001) pokazala je da forenzički *STR* profili mogu nositi informaciju o populacionom poreklu, uprkos tome što nisu dizajnirani u tu svrhu.

1.7.2. Bajesovski modeli populacionog grupisanja i klasifikacije

Pritchard i sar. (2000) uveli su probabilistički pristup zasnovan na modelima populacionog grupisanja (engl. *model-based clustering*). Ovaj pristup opisuje populacije kao statističke entitete definisane učestalostima alela, pri čemu se genotipski podaci pojedinaca koriste za procenu verovatnoće pripadnosti različitim populacionim komponentama. Metodologija omogućava procenu proporcija admiksije, odnosno procenata genoma koji potiču iz različitih populacionih izvora. Korišćenjem Markovljevih lanaca Monte Karlo algoritama (MCMC, eng. *Markov Chain Monte Carlo*) moguće je dobiti probabilističke procene populacione pripadnosti i proporcija admiksije, što ovakve analize čini primenljivim u heterogenim populacijama.

Na temelju ovog metodološkog napretka razvila su se detaljnija istraživanja primene *STR* markera u predikciji porekla. Graydon i sar. (2009) testirali su performanse 15 autozomnih *STR* lokusa koristeći Bajesov klasifikacioni pristup na kombinaciji stvarnih i simuliranih DNK profila. Uključene populacije Han Kineza, Japanaca, Korejaca, američkih belaca i indijske autohtone grupe Lodha omogućile su procenu modela u demografski raznolikim kontekstima. Rezultati su potvrdili da panel standardnih forenzičkih markera može efikasnije razlikovati genetički udaljenije populacione grupe kada se primeni Bajesovski klasifikacioni pristup zasnovan na populacionim učestalostima alela.

1.7.3. Proširenje rezolucije: selekcija markera i pristupi mašinskog učenja

S ciljem unapređenja predikcije porekla, Londin i sar. (2010) predložili su panel od 36 mikrosatelitnih markera informativnih za kontinentalno poreklo (*CoAIMs*), selektovanih na osnovu visokih vrednosti genetičke diferencijacije (F_{ST}). Rezultati su pokazali da *CoAIMs* panel može jasno razlikovati pet globalnih populacionih regiona (Evropa/Srednji istok, Istočna Azija, Afrika, Amerike i Okeanija), koristeći i *PCA* i Bajesovo klasterovanje u softveru *STRUCTURE*.

Pereira i sar. (2011) razvili su onlajn platformu „*PopAffiliator*“, zasnovanu na mašinskom učenju i *DTNB* algoritmu implementiranom u *Weka* softveru. Ovaj model, treniran na oko 15000 DNK profila, omogućio je klasifikaciju u tri šire populacione kategorije (evroazijsku, istočnoazijsku i podsaharsko afričku). *PopAffiliator* pokazao je mogućnost kombinovanja *STR* markera i ML algoritama u rutinskoj primeni.

Phillips i sar. (2013) su razvili 12-*STR* multipleks zasnovan na markerima informativnim za poreklo (*AIM-STR*), selektovanim iz većeg skupa tetranukleotidnih lokusa. Primenom prilagođenog Bajesovog klasifikatora *Snipper* autori su pokazali da takvi markeri mogu uspešno doprineti proceni biogeografskog porekla, posebno kada se kombinuju sa *SNP* markerima, uz dodatnu prednost mogućnosti detekcije mešovityh DNK tragova.

Algee-Hewitt i sar. (2016) analizirali su informativnost forenzičkih *CODIS* lokusa u predikciji biogeografskog porekla. Analizirali su 792 mikrosatelitska markera kod 978 osoba iz 53 globalne populacije, poredeći informativnost *CODIS* lokusa sa slučajno odabranim tetranukleotidnim markerima. Uprkos tome što su *CODIS* markeri izabrani zbog svoje pogodnosti za individualnu identifikaciju, a ne zbog populacione diferencijacije, rezultati su pokazali da oni nose upotrebljiv signal za diferencijaciju velikih kontinentalnih grupa.

Devesse i sar. (2023) primenili su *MPS* tehnologiju za analizu 26 autozomnih *STR* lokusa u pet populacionih grupa. Rezultati su pokazali da podaci zasnovani na čitavoj sekvenci *STR* alela omogućavaju bolju diferencijaciju populacija od klasičnog pristupa zasnovanog samo na dužini alela.

Navedeni radovi prikazuju razvoj primene autozomnih *STR* markera u *BGA* analizama. Uprkos ograničenjima, poput niže rezolucije u odnosu na *SNP* panele, osetljivosti na admiksiju i zavisnosti od referentnih baza podataka, *STR* markeri ostaju primenljivi zbog svoje dostupnosti, standardizovanosti i kompatibilnosti sa rutinskim forenzičkim protokolima. Njihov signal o poreklu, u kombinaciji sa naprednim statističkim modelima i algoritmima mašinskog učenja, omogućava razvoj prediktivnih metoda zasnovanih na podacima koji se redovno generišu u forenzičkim laboratorijama.

1.8. Konsangvinitet, inbriding i incest – genetičke i biološke osnove

1.8.1. Konsangvinitet i inbriding: pojmovna i populaciono–genetička distinkcija

Konsangvinitet označava postojanje genealoške veze između dve osobe, odnosno činjenicu da one dele jednog ili više zajedničkih predaka u relativno skoroj prošlosti. Inbriding, sa druge strane, predstavlja reproduktivnu posledicu takve genealoške povezanosti i odnosi se na ukrštanje između srodnih individua unutar populacije. Iako se ovi termini u literaturi koriste sinonimno, njihovo razgraničenje ima značaj u populacionoj i forenzičkoj genetici. Konsangvinitet je deskriptivna kategorija koja se odnosi na stepen srodstva, dok je inbriding proces koji direktno utiče na genetičku strukturu potomstva i raspodelu alela u populaciji (Bittles, 2003).

U populaciono–genetičkom smislu, inbriding predstavlja odstupanje od panmiktikog ukrštanja i dovodi do povećanja homozigotnosti u odnosu na očekivanja prema *HWE*. Ovo ne mora nužno biti posledica incestnih veza, već se može javiti i u populacijama sa malom efektivnom veličinom, izraženim osnivačkim efektima ili dugotrajnim socijalnim praksama ukrštanja unutar zajednica (Tadmouri i sar., 2009). U forenzičkom kontekstu potrebno je razlikovati biološke posledice inbridinga na nivou populacije od obrazaca koji nastaju kao posledica incestuoznih odnosa.

1.8.2. Inbriding koeficijent (F), *IBD* i autozigoznost

Kvantifikacija inbridinga u populacionoj genetici zasniva se na koeficijentu inbridinga (F), koji predstavlja verovatnoću da su dva alela na istom lokusu u jednoj individui *IBD* aleli. Vrednost F zavisi od genealoškog odnosa roditelja i povećava se sa stepenom njihove srodnosti. Na primer, potomstvo iz odnosa roditelj–dete ili brat–sestra karakteriše teorijski $F \approx 0,25$, dok je kod potomstva brata i sestre od strica, ujaka ili tetke $F \approx 0,0625$ (Falconer i Mackay, 1996).

Povećanje koeficijenta inbridinga direktno dovodi do porasta autozigoznosti, koja predstavlja poseban oblik *IBD*–a i nosi jasnu genealošku informaciju o srodničkom ukrštanju. Za razliku od homozigotnosti koja može nastati i kao posledica slučajnog ukrštanja u panmiktikoj populaciji, autozigoznost je specifično povezana sa nedavnim zajedničkim precima roditelja i odražava strukturu srodstva unutar porodice ili populacije (McQuillan i sar., 2008).

Na nivou čitavog genoma, autozigoznost se ispoljava kroz duge, kontinualne segmente homozigotnosti, poznate kao *ROH* (eng. *Runs of Homozygosity*), čiji broj, dužina i raspored zavise od stepena srodstva roditelja i vremena koje je proteklo do zajedničkog pretka. Kraći i brojniji *ROH* segmenti ukazuju na starije i udaljenije inbriding događaje, dok dugi *ROH* segmenti predstavljaju signal nedavnog inbridinga (McQuillan i sar., 2008). Zbog toga se analiza *ROH* segmenata koristi za procenu intenziteta inbridinga i genetičkog opterećenja populacije, uključujući povećan rizik od ispoljavanja štetnih recesivnih varijanti.

1.8.3. Incest kao ekstremni oblik inbridinga

Incest predstavlja ekstremni oblik inbridinga i odnosi se na reproduktivne odnose bliskih srodnika, kao što su odnosi roditelj–dete i brat–sestra. Ovo za posledicu ima redukciju heterozigotnosti, povećani udeo *IBD* alela i pojavu *ROH* segmenata u genomu potomstva (Falconer i Mackay, 1996; McQuillan i sar., 2008). Reproductivni odnosi između bliskih srodnika zabeleženi su i u istorijskim populacijama na osnovu nalaza arheološko–genetičkih istraživanja. DNK analiza na mumificiranim ostacima iz kasne 18. egipatske dinastije omogućila je rekonstrukciju pedigree koji ukazuje da su roditelji faraona Tutankamona bili brat i sestra, što je u skladu sa poznatim praksama dinastičkog konsangviniteta u starom Egiptu (Hawass i sar., 2010).

1.8.4. Biološke posledice inbridinga

Populacije sa izraženim praksama srodničkog ukrštanja, poput onih na Bliskom istoku, Južnoj Aziji i Severnoj Africi, pokazuju povećanu učestalost autozomno–recesivnih poremećaja. Ovaj obrazac se tumači kao direktna posledica povećanog ispoljavanja štetnih recesivnih varijanti koje u uslovima visoke heterozigotnosti ostaju fenotipski „skrivené“ (Bittles, 2003; Tadmouri i sar., 2009). Poremećaji poput *Bardet–Biedl* sindroma, *Meckel–Gruber* sindroma, osteopetroze i kongenitalne hloridne dijareje često se navode kao klasični primeri genetičkog opterećenja povezanog sa inbridingom (Tadmouri i sar., 2009).

Dodatno, studije na izolovanim populacijama pružaju dokaze o posledicama inbridinga. U populaciji Amiša iz okruga Lancaster u SAD, prosečan udeo autozigoznosti u genomu iznosi oko 3,7 %, dok je u ranije analiziranom uzorku osoba evropskog porekla iz baze UK Biobank iznosio približno 0,4 %. Lynch i sar. (2023) pokazali su da su autozgotni regioni u ovoj populaciji povezani sa promenama biohemijskih parametara, uključujući nivoe bilirubina i *HbA1c*.

Biološke posledice ukrštanja u srodstvu ne ogledaju se samo u povećanoj učestalosti monogenjskih poremećaja već i u promenama kvantitativnih osobina. McQuillan i sar. (2012) pokazali su obrnutu povezanost između genomske homozigotnosti i telesne visine i procenili da kod potomstva blisko srodnih roditelja smanjenje visine može biti do 3 cm, što ukazuje na efekat recesivnih varijanti na poligenjske osobine.

1.8.5. Inbriding depresija

Kumulativni efekti ovakvih ukrštanja objedinjeni su pojmom inbriding depresije, koji označava smanjenje biološke sposobnosti potomstva blisko srodnih jedinki, prvenstveno u pogledu preživljavanja i reproduktivnog uspeha. Inbriding povećava homozigotnost i posledično smanjuje heterozigotnost u genomu potomstva. Time se povećava verovatnoća da štetni recesivni ili delimično recesivni aleli budu prisutni u homozigotnom stanju i da se ispolje njihovi negativni efekti. Nezavisno od toga, i samo smanjenje heterozigotnosti doprinosi inbriding depresiji, gubitkom prednosti heterozigota na lokusima na kojima heterozigotni genotip ima veću biološku sposobnost od oba homozigotna genotipa (Charlesworth i Willis, 2009; Kristensen i sar., 2010). Istraživanja zasnovana na „omičkim“ pristupima dodatno pokazuju da je inbriding povezan sa promenama u ekspresiji gena i u ćelijskim procesima uključenim u odgovor na stres i starenje, što ukazuje da se posledice inbridinga ispoljavaju i na nivou regulacije genoma, a ne isključivo kroz ispoljavanje pojedinačnih recesivnih alela (Kristensen i sar., 2010).

Evropska dinastija Habsburg predstavlja istorijski primer inbriding depresije. Tokom više generacija u ovoj lozi zabeležen je porast koeficijenta inbridinga, koji je kod Čarlsa II od Španije dostigao vrednost $F = 0,254$, što odgovara autozigoznosti od približno 25 % autozomnog genoma, iako su njegovi roditelji bili ujak i sestrčina. Ovakav ekstremno visok stepen inbridinga bio je povezan sa izraženim zdravstvenim i razvojnim poremećajima, uključujući impotenciju i infertilitet (Alvarez i sar., 2009). U širem interpretativnom smislu, Čarls II se može posmatrati kao žrtva dugotrajne dinastičke politike očuvanja nasledstva unutar porodice, prakse koja je na kraju doprinela genetičkom kolapsu ove vladarske loze u Španiji (Alvarez i sar., 2009).

1.8.6. Manifestacija incesta u autozomnim STR profilima

Na nivou autozomnih DNK markera, efekti incesta manifestuju se kroz niz specifičnih obrazaca koji odstupaju od očekivanja panmiksičkog ukrštanja. Uporedne analize incestuoznih i neincestuoznih slučajeva pokazuju da potomci incestuoznih odnosa češće ispoljavaju povišene nivoe homozigotnosti (Wenk i sar., 1994; Corach i sar., 2003). Karakterističan signal predstavlja i pojava

heterozigotnih podudarnosti, pri čemu dete i roditelj dele oba alela na lokusu u heterozigotnom stanju, sa učestalošću većom od očekivane u odnosima roditelj–dete u kojima roditelji deteta nisu međusobno srodni. Ovakvi obrasci, zajedno sa povećanim udelom homozigotnih genotipova na lokusima, uočeni su sa većom učestalošću u incestuoznim slučajevima, što ukazuje na njihovu relevantnost u detekciji incestuoznih odnosa (Corach i sar., 2003).

Iako su povišeni nivoi homozigotnosti znatno češći u incestuoznim slučajevima, homozigotnost sama po sebi nije dovoljan dokaz incesta (Corach i sar., 2003). Kombinacija raspodele homozigotnosti, prisustva heterozigotnih podudarnosti između roditelja i deteta i njihovih obrazaca duž više lokusa, kao i ukupna sličnost *STR* profila, predstavlja osnovu za razvoj pristupa sposobnih da razlikuju incestuozne i neincestuozne scenarije u forenzičkoj praksi (Tamura i sar., 2000; Corach i sar., 2003; Wenk, 2008).

Dodatna složenost javlja se u slučajevima kada potomci potiču iz populacija sa izraženim konsangvinitetom, čak i u odsustvu incestuoznih odnosa. Primer uticaja konsangviniteta na autozomne *STR* profile prikazan je u studiji Bandah–Rozenfeld i sar. (2025), gde je pokazano da povišen stepen autozigoznosti može dovesti do izuzetno visoke sličnosti genetičkih profila između rođene braće. U opisanom slučaju, dva brata su imala potpuno identične DNK profile pri analizi 15 autozomnih *STR* lokusa. Proširenjem analize na 21 autozomni *STR* lokus utvrđeno je podudaranje na 18 lokusa, dok su na tri dodatna lokusa zabeležene razlike između braće. Autori procenjuju da je verovatnoća potpunog podudaranja na prvobitnih 15 lokusa približno 1 prema 6,8 miliona kada roditelji nisu međusobno srodni, dok se povećava na približno 1 prema 1,2 miliona kada oba roditelja potiču iz konsangvinitetnih brakova. Pri korišćenju proširenog panela verovatnoća potpunog podudaranja profila braće procenjena je na približno 1 prema 6 milijardi.

1.9. Forenzički izazovi u detekciji incestuoznih odnosa

1.9.1. Preklapanje genetičkih signala: incest nasuprot populacionom inbridingu

Preklapanje genetičkih signala između incestuoznih odnosa i populacija sa povišenim stepenom inbridinga predstavlja metodološki izazov u detekciji incesta. U oba slučaja uočavaju se povećana homozigotnost i smanjena alelna varijabilnost. Međutim, dok su ovi obrasci u populacionom inbridingu rezultat dugotrajnih demografskih procesa, poput ograničene efektivne veličine populacije i istorijski stabilnih obrazaca srodničkog ukrštanja, u slučaju incesta oni nastaju naglo, kao posledica jedne ili nekoliko ekstremnih reproduktivnih veza (Kirin i sar., 2010; Pemberton i sar., 2012; Khubrani i sar., 2019).

U forenzičkom kontekstu razlikovanje ova dva scenarija neposredno utiče na interpretaciju. Potomci iz populacija sa izraženim istorijskim inbridingom mogu pokazivati povišene nivoe homozigotnosti i u odsustvu incestuoznih odnosa, što stvara rizik od lažno pozitivnih interpretacija ukoliko se povećana homozigotnost posmatra bez uvida u širi populacioni okvir.

Dodatnu složenost uvodi Valundov efekat koji, iako ne implicira homozigotnost na nivou pojedinca, može uticati na procene učestalosti alela i time indirektno uticati na statističke modele korišćene pri interpretaciji genetičkih profila (Balding i Nichols, 1994; National Research Council, 1996).

Razlikovanje konsangviniteta, populacionog inbridinga i incesta potrebno je za interpretaciju *STR* profila u forenzičkoj genetici. Analitičke metode zato treba da integrišu više genetičkih signala istovremeno i da pri tome kontrolišu uticaj populacione strukture.

1.9.2. Dokumentovani forenzički slučajevi kao ilustracija metodoloških problema

Objavljeni forenzički slučajevi ilustruju ograničenja standardnih metoda u detekciji incestuoznih odnosa. Canturk i sar. (2016) opisali su slučaj u kojem su tri biološka brata bila osumnjičena za očinstvo nad detetom njihove sestre. Početne duo analize, koje su obuhvatile poređenje profila deteta sa profilom svakog od navodnih očeva bez uključivanja profila majke, nisu omogućile isključenje nijednog od braće. Tek nakon proširenja analize na veći broj *STR* lokusa i uključivanja majke u analizu, jedan od braće pouzdano je identifikovan kao biološki otac.

Sličan problem opisan je i u radu Cerri i sar. (2006), u kojem je genetički skrining deteta jedne maloletne majke na retku mutaciju povezanu sa cističnom fibrozom inicijalno ukazao na mogućnost incestuoznog očinstva. Otac maloletne majke isključen je kao mogući biološki otac njenog deteta na osnovu nepoklapanja na dva *STR* lokusa. Međutim, analiza sprovedena u programu Familias pokazala je da je najverovatniji biološki otac deteta brat maloletne majke, što je naknadno potvrđeno njegovim priznanjem.

Macan i sar. (2003) predstavili su slučaj očinstva brata nad decom njegove sestre, gde je analiza 17 autozomnih *STR* lokusa pokazala da su oba deteta nasledila sve obligatorne alele od navodnog oca, rezultujući visokom verovatnoćom očinstva. Cosentino i sar. (2015) opisali su slučaj i skrenuli pažnju da kombinacija autozomnih i kratkih tandemskih ponovaka na X hromozomu (*X-STR*, eng. *X-chromosomal Short Tandem Repeat*) omogućava razdvajanje dve alternativne hipoteze, očinstvo oca ili brata, u kontekstu incestuoznog odnosa.

U studiji autora Emre i sar. (2015) pregledano je 1.200 slučajeva spornog očinstva i 650 slučajeva seksualnog zlostavljanja, evidentiranih u Savetu za sudsku medicinu Turske u periodu od 2011. do 2014. godine. Među pregledanim predmetima identifikovano je 50 slučajeva incesta, od kojih je 31 rezultirao trudnoćom. U tri od ta 31 slučaja javile su se teškoće pri isključivanju genetički srodnih navodnih očeva. U dva slučaja analiza 15 *STR* lokusa nije omogućila njihovo razlikovanje, pa je diskriminacija postignuta tek povećanjem broja analiziranih *STR* lokusa. Ovi nalazi ukazuju na teškoće standardne analize očinstva u incestuoznim slučajevima, usled povećanog deljenja alela između blisko srodnih osoba (Emre i sar., 2015).

Standardni kriterijumi zasnovani na pojedinačnim nepoklapanjima, broju deljenih alela ili jednostavnim pragovima statističke značajnosti ne mogu pouzdano razlikovati incest od populacionog inbridinga, niti razdvojiti više alternativnih srodničkih hipoteza u kompleksnim porodičnim strukturama. Navedena ograničenja stvaraju interpretativni rizik u forenzičkoj praksi, jer pogrešna klasifikacija srodničkih odnosa može imati ozbiljne pravne i etičke posledice. Poseban izazov u incestuoznim scenarijima predstavlja pravilno definisanje hipoteza i genealoškog modela pri izračunavanju indeksa verodostojnosti (*LR*, eng. *Likelihood Ratio*).

1.10. Indeks verodostojnosti kao pristup u forenzičkoj genetici: primena i ograničenja u srodničkim i incestuoznim scenarijima

1.10.1. Indeks verodostojnosti kao osnovni okvir interpretacije DNK dokaza

U savremenoj forenzičkoj genetici *LR* predstavlja odnos verovatnoća pod dve suprotstavljene hipoteze. U najopštijem obliku, *LR* se definiše kao odnos verovatnoće dokaza pod hipotezom tužilaštva (H_p), koja podrazumeva da je osoba od interesa biološki povezana sa uzorkom, i verovatnoće istog dokaza pod alternativnom hipotezom odbrane (H_d), prema kojoj doprinos potiče od nepoznate, nepovezane osobe iz populacije (Evett i Weir, 1998).

Prednost *LR* pristupa ogleda se u njegovoj sposobnosti da kvantifikuje težinu dokaza na kontinuiranoj skali, umesto kroz binarne odluke uključivanja ili isključenja, čime se omogućava

nijansirana i transparentna evaluacija genetičkih podudarnosti. Ovakav okvir je postao standard u forenzičkoj genetici i predstavlja osnovu za interpretaciju kako individualnih profila, tako i srodničkih odnosa (National Research Council, 1996).

1.10.2. Indeks očinstva kao poseban oblik *LR* pristupa

U klasičnim slučajevima utvrđivanja očinstva, *LR* pristup se primenjuje kroz indeks očinstva (PI, eng. *Paternity Index*). PI predstavlja odnos verovatnoće posmatranih genetičkih podataka pod hipotezom da je testirani muškarac biološki otac deteta i verovatnoće istih podataka pod alternativnom hipotezom da je biološki otac deteta nasumično izabrani, sa testiranim muškarcem nesrodni muškarac iz referentne populacije. Kombinovanjem *PI* vrednosti za analizirane lokuse dobija se kombinovani indeks očinstva (*CPI*, eng. *Combined Paternity Index*), koji kvantifikuje snagu genetičkih dokaza u odnosu na postavljene hipoteze (Gjertson i sar., 2007).

Za razliku od složenih srodničkih i incestuoznih scenarija, hipoteze u klasičnom očinstvu su jasno definisane, a genealoška struktura jednostavna, što *PI* čini pouzdanim statističkim alatom. Jasnoća hipoteza i ograničen broj mogućih odnosa objašnjavaju zašto *LR* pristup omogućava pouzdanu interpretaciju u standardnim analizama očinstva, dok kod kompleksnijih analiza srodstva definisanje hipoteza i izračunavanje *LR* vrednosti postaje složenije (Gjertson i sar., 2007).

1.10.3. Ograničenja *LR* interpretacije u incestuoznim scenarijima

U incestuoznim scenarijima, primena *LR* pristupa suočava se sa nizom praktičnih problema. U literaturi se ističe da čak i kada se koriste specifični indeksi razvijeni za incest, poput Indeksa incesta (*II*) i Kombinovanog indeksa incesta (*CII*), rezultati mogu ostati nedovoljno diskriminativni, naročito pri ograničenom broju analiziranih lokusa (Wenk, 2008). U takvim slučajevima, interpretativna vrednost *LR* analiza može biti ograničena ukoliko hipoteze ne obuhvate realnu genealošku strukturu posmatranog odnosa.

Pri incestuoznim scenarijima nije jednoznačno odrediti da li H_d treba da podrazumeva nepovezanu osobu iz populacije, bliskog srodnika ili kombinaciju više mogućih odnosa. Različiti izbori hipoteza mogu proizvesti različite *LR* vrednosti, čime se povećava rizik od pogrešnih zaključaka (Gjertson i sar., 2007).

Uticaj izbora alternativne hipoteze na interpretaciju *LR* vrednosti prikazan je u radu autora Morris i sar. (1988), u kojem je uveden avunkularni indeks. Ovaj indeks definisan je kao *LR* kojim se porede hipoteza da je testirani muškarac očevo brat, odnosno stric deteta, i alternativna hipoteza da je testirani muškarac nesrodan detetu. Medijana vrednosti avunkularnog indeksa kod isključenih, uglavnom nesrodnih muškaraca iznosila je približno 0,26. Vrednost avunkularnog indeksa veća od 1 značila je da su posmatrani genetički podaci verovatniji pod hipotezom avunkularnog odnosa nego pod postavljenom alternativnom hipotezom. Ovaj primer pokazao je da vrednost i značenje *LR*-a zavise od hipoteza koje se međusobno porede.

Kasniji pokušaji kvantifikacije incestuoznih odnosa zasnivali su se na povećanoj homozigotnosti i specifičnim obrascima genotipskih poklapanja između majke i deteta. Wenk i sar. (1994) su pokazali da se *VNTR* markeri mogu koristiti za procenu verovatnoće incestuoznog porekla deteta, čime je incest eksplicitno definisan kao poseban statistički problem u forenzičkoj genetici. Tamura i sar. (2000) dodatno su razvili ovaj teorijski okvir formalnim proračunom prelaznih verovatnoća kod multialelnih lokusa.

Relativno male *LR* vrednosti mogu biti dobijene čak i kada je H_p tačna, dok retki događaji mogu proizvesti ekstremno velike *LR* vrednosti, što dovodi do asimetričnih i „teških“ raspodela (Slooten i Egeland, 2014). U incestuoznim scenarijima, ovakvo ponašanje *LR*-a ne ukazuje nužno na

grešku modela, već pokazuje da se distribuciona priroda *LR*-a i izbor hipoteza moraju uzeti u obzir pri interpretaciji.

1.10.4. Tehnička ograničenja *LR* modela

Pored genealoške zavisnosti, *LR* proračuni su osetljivi i na kršenje pretpostavke nezavisnosti lokusa. Delimična vezanost između pojedinih *STR* lokusa može dovesti do izobličenja *LR* vrednosti ukoliko se verovatnoće kombinuju prostim umnožavanjem (O'Connor i Tillmar, 2012). Dodatno, mutacije u pravim srodničkim parovima mogu uticati na *LR* vrednost i moraju se uzeti u obzir primenom odgovarajućeg mutacionog modela (Gjertson i sar., 2007).

Proširenje skupa analiziranih markera uključivanjem *X-STR* i *SNP* pristupa predstavlja dopunu autozomnim analizama. *X-STR* markeri, zahvaljujući specifičnom obrascu nasleđivanja, mogu obezbediti dodatnu diskriminatornu informaciju u scenarijima gde su potencijalni očevi međusobno srodni, što je naročito relevantno u incestuoznim slučajevima (Tillmar i sar., 2017).

1.10.5. Metodološki jaz kao osnova za primenu mašinskog učenja

U kompleksnim srodničkim scenarijima primena *LR* pristupa zahteva da postavljene hipoteze i statistički model odgovaraju razmatranoj genealoškoj strukturi. Povećanje broja lokusa može poboljšati diskriminativnu moć, ali samo po sebi ne otklanja posledice neadekvatno definisanih hipoteza ili zanemarenih zavisnosti između lokusa (Gjertson i sar., 2007; O'Connor i Tillmar, 2012).

Za razliku od *LR* pristupa, zasnovanog na poređenju unapred definisanih hipoteza, algoritmi mašinskog učenja omogućavaju prepoznavanje obrazaca i modelovanje nelinearnih odnosa radi predikcije. Mašinsko učenje ne predstavlja zamenu za *LR* pristup u sudskoj interpretaciji, već komplementarni alat za prepoznavanje atipičnih profila i podršku pri donošenju odluka u kompleksnim forenzičkim scenarijima.

1.11. Primena savremenih *AI/ML* metoda u forenzičkoj genetici

Metode veštačke inteligencije (*AI*, eng. *Artificial intelligence*), uključujući mašinsko učenje (*ML*, eng. *Machine Learning*) i njegove poddiscipline poput dubokog učenja (*DL*, eng. *Deep Learning*), primenjuju se u analizi forenzičkih genetičkih podataka. Tradicionalna forenzička DNK analiza u velikoj meri se oslanjala na interpretaciju elektroforetskih podataka i klasične statističke modele, što interpretaciju čini vremenski zahtevnom i podložnom varijabilnosti, naročito u analizama kompleksnih DNK profila. *AI/ML* metode predstavljaju dopunske alate u savremenoj forenzičkoj genetici. One ne zamenjuju stručnu procenu forenzičkog naučnika, već mogu doprineti automatizovanijoj i doslednijoj obradi genetičkih podataka i proširiti opseg analitičkih zadataka koji se mogu ispitivati (Barash i sar., 2024).

1.11.1. Kategorije i metode u mašinskom učenju

Metode mašinskog učenja najčešće se klasifikuju u četiri osnovne kategorije: nadgledano učenje (eng. *supervised learning*), nenadgledano učenje (eng. *unsupervised learning*), polunadgledano učenje (eng. *semi-supervised learning*) i učenje potkrepljenjem (eng. *reinforcement learning*) (Sarker, 2021).

U nadgledanom učenju, model uči funkciju koja preslikava ulazne podatke u izlaz na osnovu parova ulaz–izlaz, pri čemu se kao najčešći zadaci navode klasifikacija i regresija; cilj je omogućiti predikciju ili dodelu klase novim, nepoznatim primerima (Sarker, 2021). Nasuprot tome, nenadgledano učenje radi nad neoznačenim podacima i koristi se za identifikaciju trendova, strukture

i grupisanja u podacima, pri čemu su tipični zadaci klasterovanje i postupci poput učenja karakteristika i redukcije dimenzionalnosti (Sarker, 2021). Kada su dostupni i označeni i neoznačeni podaci, polunadgledano učenje predstavlja hibridni pristup koji kombinuje ova dva režima, dok učenje potkrepljenjem omogućava modelu da kroz interakciju sa okruženjem i na osnovu sistema nagrada i kazni nauči koje akcije dovode do optimalnog ishoda (Sarker, 2021).

Pored toga, značajan deo pristupa mašinskog učenja zasniva se na metodama udruženih modela (eng. *ensemble methods*), kojima se objedinjuju predikcije više pojedinačnih modela. Kombinovanjem više različitih algoritama ili više instanci istog algoritma treniranih na različitim podskupovima podataka, smanjuje se osetljivost modela na varijacije u podacima i postiže stabilnija predikcija (Yaghoubi i sar., 2024).

ML pristup je često agnostičan u odnosu na proces koji generiše podatke i teži optimizaciji prediktivne tačnosti, umesto parametarskoj proceni probabilističkog modela. ML takođe omogućava efikasno korišćenje visokodimenzionalnih ulaza bez eksplicitnog poznavanja njihove zajedničke raspodele, što je posebno relevantno kako genomske skupove podataka rastu u dimenzionalnosti (Schrider i Kern, 2018).

1.11.2. Testirani algoritmi

Za analizu genetičkih podataka primenjuju se različiti klasifikacioni modeli, pri čemu izbor odgovarajućeg algoritma zavisi od tipa markera, cilja analize (predikcije porekla, klasterovanja, klasifikacije haplogrupe, fenotipske predikcije) i strukture podataka (numeričke ili kategorijske promenljive, prisustva nedostajućih vrednosti u podacima, neuravnoteženih klasa) (Kuhn i Johnson, 2013; Schrider i Kern, 2018; Hancock i Khoshgoufar, 2020).

Mašine potpornih vektora (*SVM*, eng. *Support Vector Machines*) predstavljaju nadgledani algoritam za klasifikaciju koji uči granice između unapred definisanih klasa i omogućava predikciju pripadnosti novih uzoraka. U jednoj studiji *SVM* je treniran na evropskim referentnim uzorcima koristeći koordinate prve dve glavne komponente *PCA* (*PC1* i *PC2*), čime je postignuta klasifikacija evropskog porekla u osam regionalnih klasa, identifikovano neslaganje između samoprijavljenog i genetički određenog porekla (14–23 %) u zavisnosti od korišćenog modela, i ukazano na rane znake genetičke divergencije lokalne populacije evropskih Amerikanaca u Wisconsinu u odnosu na izvorne evropske populacije (Haas i sar., 2013). U multietničkim *GWAS* kohortama Fang i sar. (2019) uveli su *HARE* algoritam, kao kategorijsku varijablu koja harmonizuje genetički procenjeno poreklo i samoprijavljenu rasu/etnicitet, tako što trenira *SVM* da nauči njihovu korespondenciju i zatim kroz pravila dodele identifikuje potencijalno netačne unose, razreši konflikte i doda nedostajuće informacije o etnicitetu, čime se povećava interpretabilnost i detekcija populaciono-specifičnih i heterogenih efekata u *GWAS*-u (Fang i sar., 2019).

Algoritam K-najbližih suseda (*KNN*, eng. *K-Nearest Neighbors*) primenjen je u radu Yuan i sar. (2019) u okviru poređenja više metoda za predikciju porekla na osnovu metilacionih profila placentalnih DNK. *KNN* je pokazao slabije performanse od ostalih ispitivanih algoritama, zbog čega su autori za konačni model *PlaNET* izabrali logističku regresiju sa regularizacijom elastične mreže (eng. *elastic net*). U drugoj studiji autori su primenili algoritam K-najbližih suseda, kao i njegovu podvarijantu zasnovanu na genetičkoj distanci, na sekvence kontrolnog regiona mitohondrijske DNK kako bi pojedince dodelili širokim kontinentalnim populacijama, a zatim i subpopulacionim grupama. Analiza je pokazala izuzetno visoku tačnost klasifikacije na kontinentalnom nivou, koja je dostizala od 82 do 99 % u zavisnosti od K, dok je na nivou subpopulacija iznosila oko 60 %. Ovi rezultati ukazuju da mtDNK podaci nose signal za grubu geografsku diferencijaciju, ali imaju ograničenu rezoluciju za precizniju subpopulacionu klasifikaciju (Yang i sar., 2021).

Naivni Bajesov klasifikator (*NB*, eng. *Naive Bayes*) zasniva se na pretpostavci da su ulazne promenljive međusobno uslovno nezavisne unutar svake klase, i često se koristi kao osnovni

referentni model. U radu autora Polychronopoulos (2012) primenjeno je više metoda mašinskog učenja za predikciju porekla na osnovu varijabilnih regiona mtDNK, pri čemu je kao referentni skup korišćena populaciona baza *FBI*-ja. Naivni Bajesov pristup dao je tačnost klasifikacije od približno 72–76 %, uz slabiju diferencijaciju azijske i hispanske populacije. Slabije performanse ovog modela očekivane su iz bioloških razloga: mitohondrijska DNK ne podleže rekombinaciji i nasleđuje se kao jedinstven haplotip, pa varijante u kontrolnom regionu nisu međusobno nezavisne, već se javljaju u karakterističnim kombinacijama koje definišu haplogrupe. Model koji svaku poziciju tretira kao nezavisnu promenljivu zanemaruje upravo tu haplotipsku strukturu, u kojoj je sadržan najveći deo informacije o poreklu. Iz tog razloga su algoritmi koji ne pretpostavljaju nezavisnost promenljivih, poput *SVM*-a, na istom skupu podataka postigli višu tačnost (Polychronopoulos, 2012). Na performanse su mogli uticati i sastav skupa podataka i način definisanja populacionih grupa.

Odlučujuća stabla (*DT*, eng. *Decision Trees*) predstavljaju hijerarhijske modele koji postupno razdvajaju podatke na osnovu pravila odlučivanja. U radu Hajiloo i sar. (2013) ovaj pristup je primenjen u okviru metode *ETHNOPRED* za predikciju kontinentalnog i subkontinentalnog porekla na osnovu *SNP* podataka iz *HapMap* baze. Rezultati su pokazali da se već sa malim brojem jednostavnih stabala i ograničenim skupom markera može postići visoka tačnost u određivanju kontinentalnog porekla, čak i kada deo genotipskih podataka nedostaje. Daljim proširenjem pristupa omogućena je diferencijacija unutar kontinenata, uključujući razdvajanje evropskih, istočnoazijskih, afričkih i severnoameričkih subpopulacija, kao i pojedinih genetički bliskih afričkih populacija, čime je potvrđena sposobnost modela da prepozna suptilne obrasce populacione strukture. Istovremeno, ograničena uspešnost u razlikovanju veoma srodnih populacija, poput kineskih uzoraka iz različitih geografskih oblasti, ukazuje na prirodna ograničenja ovog pristupa u uslovima minimalne genetičke diferencijacije.

Multinomijalna logistička regresija (*MLR*, eng. *multinomial logistic regression*) predstavlja proširenje binarne logističke regresije koje omogućava rešavanje klasifikacionih problema sa tri ili više kategorija (Kuhn i Johnson, 2013). Primenjena je za predviđanje biogeografskog porekla na osnovu genetičkih markera. Model zasnovan na 142 *AIM-SNP* markera omogućio je diferencijaciju afričkog, evropskog, istočnoazijskog i američko-indijanskog biogeografskog porekla (Cheung i sar., 2017). Slično tome, *MLR* je primenjena za diferencijaciju hispanskog biogeografskog porekla na osnovu 80 *SNP* markera, pri čemu je ostvarena visoka tačnost klasifikacije pojedinaca iz više nacionalnih i regionalnih grupa u referentnom i kombinovanom skupu podataka (Setser i sar., 2020).

Slučajna šuma (*RF*, eng. *Random Forest*) predstavlja metodu zasnovanu na većem broju stabala odlučivanja, pri čemu se kombinuju rezultati više stabala izgrađenih na različitim uzorcima podataka, a konačna odluka donosi većinskim glasanjem. Ovaj neparametarski pristup smanjuje podložnost modela prenaučivosti i omogućava rad sa nedostajućim podacima. U forenzičko-antropološkoj primeni, *RF* je korišćen za klasifikaciju lobanja u referentne grupe američke „crnačke“, američke „belačke“ i jugozapadne hispanske populacije, kombinujući kranimetrijske mere (kontinuirane) i morfoskopske osobine (kategorijske) u jedinstven model, a zaključak je bio da zajedničko modelovanje ova dva tipa podataka povećava tačnost. Pokazano je da ovakav pristup daje stabilnije rezultate u poređenju sa klasičnim diskriminativnim metodama, jer obuhvata širu varijabilnost oblika i morfologije ispitivanih uzoraka (Hefner i sar., 2014).

U radu sa velikim i složenim skupovima podataka primenjuju se i algoritmi koji kombinuju rezultate više pojedinačnih modela, kao što su *LightGBM* (eng. *Light Gradient Boosting Machine*), *AdaBoost* (eng. *Adaptive Boosting*) i ekstremno nasumična stabla (eng. *Extra Trees*), koji dodatno unapređuju sposobnost modela u radu sa velikim i složenim skupovima podataka (Freund i Schapire, 1997; Geurts i sar., 2006; Zhu i sar., 2009; Ke i sar., 2017).

Primena *AI/ML* u forenzičkoj genetici može se razmatrati prema prirodi analitičkog zadatka (klasifikacija, klasterovanje, predikcija) i prema strukturi i ograničenjima dostupnih genetičkih podataka (tip markera, parcijalnost profila, visoka dimenzionalnost).

1.11.3. Primene *AI/ML* u savremenoj forenzičkoj genetici

Savremeni pristupi veštačke inteligencije i mašinskog učenja već se primenjuju u forenzičkoj genetici, posebno za automatizaciju interpretacije (npr. klasifikacija haplogrupa i analiza *STR* profila), uz cilj smanjenja rizika od pogrešne interpretacije. Njihova uloga se posmatra kao podrška postojećim postupcima (brže i konzistentnije odlučivanje, dodatni uvid u obrasce), uz naglašenu potrebu za transparentnošću, interpretabilnošću i primenom pod nadzorom veštaka (Sessa i sar., 2024).

Jedna od primena mašinskog učenja u ovoj oblasti odnosi se na automatsku klasifikaciju haplogrupa na osnovu *Y-STR* ili *mtDNA* podataka. Više studija je pokazalo da različiti *ML* algoritmi mogu sa visokom tačnošću predviđati pripadnost *Y*-haplogrupama iz forenzičkih haplotipova, često uz performanse uporedive sa filogenetskim pristupima, ali uz veću brzinu i manji udeo manuelnog rada (Song i sar., 2019; Sessa i sar., 2024).

Mašinsko učenje primenjuje se i u interpretaciji *STR* podataka, pogotovo pri analizama elektroforegrama *DNK* profila od više doprinosilaca, u kojima se uočava i prisustvo artefakata. Interpretacija *EPG* signala podložna je subjektivnosti i varijabilnosti među analitičarima, posebno kod profila dobijenih iz male količine *DNK*, usled izraženih stohastičkih efekata. Dosadašnja istraživanja pokazala su da neuronske mreže mogu uspešno automatizovati identifikaciju alelnih i nealelnih signala u referentnim profilima, dok noviji radovi demonstriraju da tradicionalni *ML* modeli (npr. *RF*, *SVM*, *MLP*), uključujući i kombinovane pristupe, omogućavaju klasifikaciju signala i u mešanim *EPG* profilima, uz kvantifikovanje nesigurnosti predikcije kroz verovatnoće i pragove odlučivanja. Iako su za sada performanse niže kod kompleksnijih *DNK* profila, ovi pristupi omogućavaju automatizaciju dela postupka i povećavaju transparentnost i ponovljivost interpretacije *STR* podataka (Volgin i sar., 2021; Tan i sar., 2025).

U oblasti probabilističke genotipizacije, savremeni softverski sistemi zasnivaju se prvenstveno na Bajesovim i statističkim modelima sa maksimalnom verovatnoćom koji koriste okvire verovatnoće za evaluaciju *DNK* mešavina i izračunavanje *LR* vrednosti. Kvantitativni (kontinuirani) modeli, poput onih implementiranih u sistemima *STRmix*TM, *EuroForMix* i *DNAStatistX*, koriste informacije o visinama pikova i druge parametre profila kako bi numerički modelovali ponašanje alela, čime se značajno unapređuje dekonvolucija kompleksnih *DNK* profila. Metode mašinskog učenja uglavnom se ne koriste kao zamena za ove probabilističke modele u neposrednom izračunavanju *LR* vrednosti, već se pretežno primenjuju u pomoćnim zadacima, kao što je procena broja doprinosilaca (*NOC*, eng. *Number of Contributors*) (Gill i sar., 2021). Navedeni pristupi još uvek nisu deo rutinske forenzičke prakse, ali se ispituju za automatizaciju pojedinih koraka analize (Sessa i sar., 2024).

Mašinsko učenje primenjuje se u predikciji biogeografskog porekla (Hajiloo i sar., 2013) i fenotipskih osobina nepoznatog donora (Al-Mawlah, 2026). U predikciji fenotipskih osobina, *ML* modeli omogućavaju istovremenu obradu velikog broja markera i modelovanje složenih, nelinearnih odnosa između genotipa i fenotipa (Al-Mawlah, 2026).

Autori de Wit i sar. (2026) predstavili su razvoj i evaluaciju *AI* modela na skupu od 70722 *DNK* profila prikupljenih tokom forenzičkog rada, koji su obuhvatali varijacije u broju doprinosilaca, odnosima komponenti u mešavinama, stepenu degradacije i prisustvu artefakata. Model je dodatno testiran na nezavisnom istraživačkom skupu od 350 mešanih profila sa dva do pet doprinosilaca i pokazao je rezultate u automatskom prepoznavanju alela uporedive sa rezultatima analitičara, iako autori navode da model još nije spreman za rutinsku forenzičku primenu.

1.11.4. Prednosti modela *XGBoost* i *CatBoost* u obradi forenzičkih genetičkih podataka

XGBoost (eng. *Extreme Gradient Boosting*) i *CatBoost* (eng. *Categorical Boosting*) predstavljaju algoritme gradijentno-pojačanih stabala odlučivanja (eng. *gradient-boosted decision trees*) koji se primenjuju u forenzičkoj genetici zbog odnosa između preciznosti, efikasnosti i interpretabilnosti.

XGBoost je skalabilan sistem za pojačavanje stabala (eng. *tree boosting*), dostupan kao paket otvorenog koda (Chen i Guestrin, 2016). Model se trenira dodavanjem velikog broja stabala u složeni model, pri čemu se u svakoj iteraciji uvodi novo stablo koje unapređuje vrednost ciljne funkcije. Visoka efikasnost i primenljivost na velikim skupovima podataka postižu se zahvaljujući optimizacijama kao što su prilagođavanje radu sa retkim i nedostajućim vrednostima, približno učenje zasnovano na ponderisanim kvantilima, kao i unapređeno upravljanje memorijom i obrada podataka koji ne mogu u potpunosti da stanu u radnu memoriju računara (Chen i Guestrin, 2016).

Modeli zasnovani na gradijentnom pojačavanju, poput *XGBoost*, pokazali su se efikasnim u zadacima *BGA* klasifikacije korišćenjem podataka iz 1000 Genomes Project (Phase 3), obuhvatajući 26 populacija raspoređenih u pet glavnih grupa: Afrika, Evropa, Istočna Azija, Južna Azija i Amerika. Selekcija markera sprovedena je primenom *RF* algoritma (strategija „jedan–naspram–ostalih”), nakon čega je za predikciju kontinentalnog porekla primenjen panel od 24 *AIM-SNP*, dok je za finiju podelu unutar Istočne Azije korišćen panel od 34 *AIM-SNP*. Na istočnoazijskom nivou razlikovana su tri populaciona klastera: japanska populacija, Han Kinezi i južno–istočnoazijske populacije. Na ova dva nivoa, modeli zasnovani na *XGBoost*–u postigli su tačnost oko 94 % za kontinentalni nivo i oko 92 % za podelu unutar istočnoazijske grupe (Cai i sar., 2024).

CatBoost, razvijen u kompaniji Yandex, predstavlja metod gradijentnog pojačavanja posebno usmeren na obradu kategorijskih podataka (Hancock i Khoshgoftaar, 2020). Tradicionalni pristupi mašinskog učenja često zahtevaju numeričko preslikavanje kategorijskih promenljivih, što može dovesti do uvođenja neželjenih artefakata ili gubitka informacija. *CatBoost* ovaj problem ublažava ugrađenim mehanizmima za direktnu obradu kategorijskih podataka, pri čemu se kategorijske vrednosti kodiraju na način koji uzima u obzir njihovu povezanost sa ciljnom promenljivom, ali bez korišćenja informacije iz samog uzorka koji se trenutno predviđa. Ovaj pristup, zasnovan na uređenim ciljnim statistikama, značajno smanjuje rizik od curenja informacija iz ciljne promenljive tokom procesa treniranja modela (Hancock i Khoshgoftaar, 2020). Pored toga, *CatBoost* koristi postupak uređenog pojačavanja zasnovan na permutacijama podataka, koji doprinosi smanjenju prekomernog prilagođavanja modela podacima za obuku. Zbog toga je ovaj algoritam naročito pogodan u situacijama u kojima su podaci pretežno kategorijski i gde je poželjno izbeći opsežno ručno prethodno procesovanje ulaznih promenljivih (Hancock i Khoshgoftaar, 2020).

CatBoost je našao primenu pri binarnoj predikciji lokalne tumorske progresije nakon mikrotalasne ablacije kod pacijenata sa ranim stadijumom hepatocelularnog karcinoma, pri čemu je model integrisao kliničke, laboratorijske i tumorske varijable i pokazao najveću diskriminativnu moć u poređenju sa drugim algoritmima mašinskog učenja (Ren i sar., 2023). Shukla i Singh (2024) razvili su *AlzGenPred*, model zasnovan na *CatBoost*, za klasifikaciju gena povezanih sa Alchajmerovom bolešću, koji je na test skupu postigao tačnost od 96,6 %.

U radu autora Almars i sar. (2021) algoritam *CatBoost* primenjen je kao klasifikacioni model na odabranim ekspresijama gena radi razlikovanja benignih i malignih tumora mozga. Nakon izbora najinformativnijih gena postignuta je tačnost klasifikacije od približno 91 %. Autori dodatno ističu da odabrani geni imaju potvrđenu biološku ulogu u ranijim istraživanjima.

Iako su *XGBoost* i *CatBoost* složeni modeli zasnovani na većem broju stabala odlučivanja, njihove odluke se mogu tumačiti primenom naknadnih metoda interpretacije. Metoda *SHAP* (eng. *SHapley Additive exPlanations*) zasniva se na dodeljivanju numeričkog doprinosa svakoj pojedinačnoj osobini za konkretnu predikciju, čime se omogućava kvantitativna procena doprinosa pojedinačnih lokusa ili markera predikciji modela (Lundberg i Lee, 2017). U analizi porekla to može ukazati na lokuse koji su najviše doprineli klasifikaciji pojedinca (Rosenberg i sar., 2003), dok u fenotipskoj predikciji može potvrditi ulogu markera sa poznatom biološkom relevantnošću (npr. varijante u okviru *HERC2* u predikciji boje očiju) (Mengel-From i sar., 2010). Takva usklađenost sa biološkim saznanjima čini rezultate modela pouzdanijim i lakšim za interpretaciju.

Ukratko, *XGBoost* i *CatBoost* predstavljaju alate mašinskog učenja koji se uklapaju u potrebe forenzičke genetike. Njihova sposobnost da modeluju složene nelinearne odnose, rade sa nedostajućim vrednostima i obezbede određeni nivo objašnjivosti čini ih pogodnim kandidatima za primene u kojima je istovremeno potrebna visoka tačnost i metodološka transparentnost.

Istovremeno, implementacija *AI/ML* u forenzičkom okruženju podrazumeva i niz izazova. Validacija mora biti rigorozna: modeli treba da budu provereni na osnovu nezavisnih skupova podataka i u različitim eksperimentalnim uslovima. Neophodna je adekvatna obuka kadra, kako bi analitičari razumeli osnovne principe rada modela, njihove prednosti, ali i ograničenja. Konačno, u pojedinim domenima biće potreban razvoj ili preciziranje standarda i smernica (SWGDM, ENFSI i dr.) koji će definisati uslove primene *AI* alata, način interpretacije rezultata i njihovu prezentaciju na sudu (Sessa i sar., 2024).

AI/ML metode predstavljaju dopunske alate u savremenoj forenzičkoj genetici. One ne zamenjuju stručnu procenu forenzičkog naučnika, već omogućavaju bržu i objektivniju obradu genetičkih dokaza i proširuju opseg analitičkih scenarija koji se mogu rešavati u praksi.

1.12. Istraživački problem

Analiza autozomnih *STR* profila u zadacima procene srodstva i predikcije porekla zasniva se na različitim statističkim okvirima. U analizama srodstva *LR* poredi verovatnoće genetičkih podataka pod unapred definisanim genealoškim hipotezama. Njegova interpretacija u kompleksnim incestuoznim scenarijima zavisi od pravilnog definisanja tih hipoteza, izbora referentne populacije i primenjenog genealoškog modela. Predikcija porekla, nasuprot tome, predstavlja klasifikacioni zadatak u kojem se genotipski profili porede sa unapred definisanim referentnim populacionim grupama.

U ovom istraživanju nadgledano mašinsko učenje primenjeno je kao dopunski prediktivni pristup za prepoznavanje multilokusnih obrazaca u autozomnim *STR* profilima. Modeli su obučavani na profilima sa unapred dodeljenim ciljnim klasama, koje su predstavljale populacione grupe, odnosno kategorije normalnog očinstva i incestuoznog srodstva. Rezultat modela predstavljao je klasifikacionu predikciju pripadnosti jednoj od definisanih klasa, a ne indeks verodostojnosti niti procenu verovatnoće da je određena forenzička hipoteza tačna.

Predmet naučnog istraživanja

Predmet naučnog istraživanja ove doktorske disertacije jeste primena algoritama mašinskog učenja u analizi autozomnih *STR* profila u forenzičkoj genetici radi predikcije porekla i detekcije incestuoznih odnosa.

STR markeri sastoje se od kratkih uzastopnih ponovaka DNK sekvenci i predstavljaju visoko polimorfne sekvence koje se značajno razlikuju među pojedincima i populacijama. Zbog visokog stepena polimorfnosti primenjuju se u identifikaciji pojedinaca u forenzičkoj praksi, kao i u analizama biološkog srodstva.

Istraživanje je obuhvatilo dve glavne oblasti:

1. Predikciju porekla na osnovu autozomnih *STR* profila, pri kojoj je ispitana efikasnost različitih algoritama mašinskog učenja korišćenjem učestalosti alela u različitim populacionim grupama. Potencijalna forenzička primena ove analize odnosi se na kriminalističke istrage u kojima je dostupan *STR* profil NN lica, a ne postoji dovoljno preostalog biološkog materijala za dodatne analize. Predviđanje porekla na osnovu *STR* profila zasniva se na identifikaciji obrazaca karakterističnih za određene populacije i njihovom povezivanju sa određenim geografskim regionima.
2. Detekciju incestuoznih odnosa primenom prediktivnog modela zasnovanog na algoritmima mašinskog učenja, namenjenog razlikovanju incestuoznih i neincestuoznih odnosa na osnovu pojedinačnih *STR* profila, kao i profila majke i deteta. Blisko genetičko srodstvo roditelja povećava verovatnoću da potomak nasledi identične alelne varijante od oba roditelja, što može dovesti do povećane homozigotnosti. Povećana homozigotnost ne mora uvek biti jasno izražena. Pored povećane homozigotnosti, veći broj alela identičnih po poreklu (*IBD*) može dovesti i do većeg broja identičnih heterozigotnih genotipova između majke i deteta. Visoka sličnost *STR* profila srodnih osoba otežava razlikovanje incestuoznih i neincestuoznih odnosa.

U istraživanju su primenjeni algoritmi *XGBoost* za predikciju porekla i *CatBoost* za detekciju incestuoznih odnosa.

HIPOTEZE I CILJEVI ISTRAŽIVANJA

Istraživanje polazi od dve osnovne hipoteze koje se odnose na mogućnosti primene autozomnih *STR* markera i algoritama mašinskog učenja u forenzičkoj genetici:

- H1: Autozomni *STR* markeri sadrže dovoljno populaciono informativnih podataka da se na osnovu njih, primenom algoritama mašinskog učenja, može predvideti poreklo analizirane osobe.
- H2: Autozomni *STR* profili jednog roditelja i deteta sadrže genetičke obrasce koji se, primenom algoritama mašinskog učenja, mogu koristiti za detekciju incestuoznog odnosa između roditelja i deteta.

Radi provere postavljenih hipoteza definisan je opšti cilj istraživanja: razvoj i evaluacija modela zasnovanih na autozomnim *STR* profilima za predikciju pripadnosti definisanim referentnim populacionim grupama i detekciju incestuoznih odnosa primenom algoritama mašinskog učenja.

Specifični ciljevi koji proizlaze iz hipoteze H1

Radi ispitivanja mogućnosti predikcije porekla na osnovu autozomnih *STR* profila, postavljeni su sledeći specifični ciljevi:

1. Ispitati i uporediti performanse algoritama mašinskog učenja *SVM*, *KNN*, Naive Bayes, Decision Tree, multinomijalne logističke regresije i algoritma *XGBoost* radi identifikovanja algoritma sa najboljim performansama u predikciji porekla.
2. Razviti i evaluirati konačni prediktivni model zasnovan na algoritmu sa najboljim performansama za klasifikaciju Afroamerikanaca, Azijata, SA–EP i Hispanoamerikanaca u odgovarajuće populacione grupe.
3. Ispitati uticaj veličine skupa podataka, odnosno broja *STR* profila korišćenih za obučavanje modela, pri skupovima od 90000, 10000, 5000, 1000 i 500 profila po populacionoj grupi, kako bi se odredila najmanja ispitana veličina skupa podataka pri kojoj model postiže stabilne performanse.
4. Proceniti informativnost pojedinačnih lokusa i utvrditi koji lokusi imaju najveći, a koji najmanji doprinos predikciji porekla.
5. Proveriti performanse modela pri korišćenju skupova markera koji odgovaraju komercijalnim *STR* kompletima za amplifikaciju: *Investigator 24plex/GlobalFiler*, *PowerPlex Fusion*, *AmpFI STR Identifier* i *NGM Detect*.
6. Ispitati performanse modela u alternativnim klasifikacionim scenarijima: pri klasifikaciji Afroamerikanaca, Azijata i SA–EP, uz isključivanje hispanoameričke grupe, i pri klasifikaciji Afroamerikanaca, Azijata i objedinjene grupe SA–EP i Hispanoamerikanaca.
7. Priminiti analizu glavnih komponenti radi vizuelizacije genetičkih sličnosti i razlika između analiziranih populacionih grupa i procene stepena njihovog razdvajanja.

Specifični ciljevi koji proizlaze iz hipoteze H2

Radi ispitivanja mogućnosti detekcije incestuoznih odnosa na osnovu autozomnih *STR* profila jednog roditelja i deteta, postavljeni su sledeći specifični ciljevi:

8. Ispitati i uporediti performanse algoritama mašinskog učenja Random Forest, *SVM*, Naive Bayes, *KNN*, *CatBoost*, *XGBoost*, *LightGBM*, *AdaBoost*, Gradient Boosting, Extra Trees i *MLP* radi identifikovanja algoritma sa najboljim performansama u detekciji incestuoznih odnosa.

9. Razviti i evaluirati konačni prediktivni model za detekciju incestuoznih odnosa u opštoj populaciji na osnovu autozomnih *STR* profila majke i deteta sa 29 analiziranih markera.
10. Proceniti performanse modela u različitim populacionim i eksperimentalnim uslovima: u opštoj populaciji pri analizi 21 lokusa; u populaciji Saudijske Arabije sa povišenom stopom homozigotnosti; i u populaciji Saudijske Arabije sa dodatno modifikovanim učestalostima alela.
11. Ispitati uticaj prisustva jednostepenih mutacija *STR* alela na tačnost detekcije incestuoznih odnosa.
12. Ispitati performanse modela u detekciji incestuoznih odnosa na osnovu jedinstvenih profila dece.
13. Analizirati genetičke pokazatelje konsangviniteta, uključujući procenat homozigotnosti u *STR* profilima dece, procenat identičnih heterozigotnih genotipova između majke i deteta i stepen sličnosti *STR* profila majke i deteta.
14. Proceniti informativnost pojedinačnih *STR* lokusa za detekciju incestuoznih odnosa i ispitati uticaj ukupnog broja analiziranih lokusa poređenjem performansi modela pri analizi 21 i 29 lokusa.
15. Primeniti analizu glavnih komponenti radi vizuelizacije genetičkih sličnosti i razlika između analiziranih kategorija srodstva i proceniti mogućnost razlikovanja pojedinačnih oblika incestuoznih odnosa primenom višeklasne klasifikacije.

II MATERIJAL I METODE

2.1 Pristup istraživanju

Istraživanje u okviru ove disertacije sprovedeno je kao *in silico* studija, u kojoj su prediktivni modeli mašinskog učenja primenjeni na generisane *STR* profile radi predikcije porekla i detekcije incestuoznih odnosa. Ovaj odeljak obuhvata izvore podataka, postupke generisanja profila, metode analize podataka i primenjene prediktivne modele.

2.1.1. Izvori podataka

Za potrebe istraživanja korišćene su javno dostupne učestalosti alela iz različitih populacija.

1. Podaci o učestalostima alela za četiri osnovne populacione grupe preuzeti su sa zvanične internet stranice kompanije Promega. Dostupno na (<https://worldwide.promega.com/products/pm/genetic-identity/population-statistics/allele-frequencies/>). Učestalosti su određene na osnovu 1036 ispitanika, od kojih je bilo 342 Afroamerikanca, 361 SA–EP, 236 Hispanoamerikanaca i 97 Azijata. Datoteka je sadržala učestalosti alela za 29 autozomnih *STR* lokusa (*CSF1PO*, *D10S1248*, *D12S391*, *D13S317*, *D16S539*, *D18S51*, *D19S433*, *DIS1656*, *D21S11*, *D22S1045*, *D2S1338*, *D2S441*, *D3S1358*, *D5S818*, *D6S1043*, *D7S820*, *D8S1179*, *F13A01*, *F13B*, *FESFPS*, *FGA*, *LPL*, *Penta C*, *Penta D*, *Penta E*, *SE33*, *TH01*, *TPOX*, *vWA*). Pretprocesuiranje datoteke podrazumevalo je uklanjanje svih dodatnih parametara, osim predstavljenih učestalosti alela. U studiji predikcije porekla korišćene su četiri zasebne datoteke, pri čemu je svaka datoteka odgovarala jednoj populacionoj grupi. U studiji detekcije incesta za opštu populaciju korišćene su objedinjene učestalosti alela određene na osnovu sve četiri populacije (Steffen i sar., 2017).
2. U studiji detekcije incesta korišćene su učestalosti alela za Saudijsku Arabiju, određene na osnovu 523 ispitanika iz pet geografskih regiona. Učestalosti alela utvrđene su za 21 *STR* lokus (*D3S1358*, *vWA*, *D16S539*, *CSF1PO*, *TPOX*, *D8S1179*, *D21S11*, *D18S51*, *D2S441*, *D19S433*, *TH01*, *FGA*, *D22S1045*, *D5S818*, *D13S317*, *D7S820*, *SE33*, *D10S1248*, *DIS1656*, *D12S391* i *D2S1338*) na osnovu amplifikacije kompletom *GlobalFiler*. Autori su zaključili da je deficit heterozigotnosti na svim analiziranim lokusima verovatno povezan sa visokom učestalošću konsangvinitetnih brakova (Khubrani i sar., 2019). Ovi podaci bili su uključeni kako bi se ispitala efikasnost modela u populaciji sa odstupanjem od slučajnog ukrštanja.
3. Dodatna procena uticaja genetičke strukture na efikasnost modela ispitana je tako što su učestalosti alela iz tačke 2. modifikovane uklanjanjem svih alela čije su učestalosti bile manje od 0,003. Na ovaj način uklonjeno je 48 alela, a za preostalih 215 alela izvršena je normalizacija.

2.1.2. Generisanje *STR* profila na osnovu učestalosti alela

Za generisanje *STR* profila iz prethodno opisanih ulaznih datoteka kreirana je skripta u programskom jeziku *Python*, koja je izvršavana u *Google Colabu*. Skripta je nasumično uzorkovala alele na osnovu njihovih učestalosti u odgovarajućoj populaciji i na taj način generisala sintetičke *STR* profile.

Svaki generisani profil sastojao se od 29 ili 21 lokusa, u zavisnosti od korišćenog skupa podataka o učestalostima alela. Svaki lokus bio je predstavljen dvema kolonama, npr. *SE33_1* i *SE33_2*, pri čemu je manji alel uvek dodeljen koloni 1, dok je veći alel dodeljen koloni 2 istog markera.

Za četiri osnovne populacije generisano je po 90000 sintetičkih profila, odnosno 360000 profila ukupno. Svi generisani profili organizovani su u jedinstvenu bazu podataka (eng. *data frame*, *DF*) sa dodatnom kolonom *Ancestry*, koja predstavlja kategorijalnu promenljivu za pripadnost određenoj grupi, u sklopu koje su postojale kategorije: *African American*, *Asian*, *Hispanic*, *Caucasian*.

Na isti način vršeno je i generisanje profila u studiji detekcije incesta, s tim što su generisani profili odgovarali roditeljskim profilima.

2.1.3. Generisanje *STR* profila na osnovu ukrštanja

Inicijalno su generisani *STR* profili na osnovu učestalosti alela za Saudijsku Arabiju i opštu populaciju za 21 i 29 lokusa. Pomoću skripte napisane u *Python*-u profili su nasumično uparivani, nakon čega su generisani profili dece. Na svakom lokusu detetu je nasumično dodeljen po jedan alel iz svakog roditeljskog profila.

Za analizu incestuoznih veza generisani su profili koji simuliraju genetičke obrasce odnosa roditelj–dete i brat–sestra, kao i profili potomaka nastalih incestuoznim vezama. Početni skupovi profila kombinovani su radi formiranja 200000 parova roditelj–dete. Njihovim ukrštanjem generisano je isto toliko profila potomaka za ovu kategoriju incesta. Dodatnih 150000 parova profila brat–sestra korišćeno je za simulaciju ukrštanja i generisanje profila potomaka za ovu kategoriju incesta.

Pošto kategorije nisu sadržale jednak broj profila, za svaku analizu odabrano je po 150 000 *STR* profila iz svake kategorije. Time su klase izjednačene prema veličini i smanjen je mogući uticaj klasne neravnoteže na performanse modela.

Svi generisani profili organizovani su u jedinstvenu bazu podataka. Kolona *Kinship* korišćena je kao kategorijalna ciljna promenljiva i označavala je scenario srodstva dodeljen svakom paru profila. Kategorije su bile *Normal Paternity*, koja je označavala odnos roditelj–dete nastao nasumičnim ukrštanjem; *Normal Sibs*, koja je označavala odnos brat–sestra; *Incest Paternity*, koja je označavala ukrštanje otac–ćerka; i *Incest Sibs*, koja je označavala ukrštanje između brata i sestre.

2.1.4. Implementacija jednostepene mutacije

U ovom delu studije primenjena je mutaciona stopa od 10^{-3} po lokusu (Butler, 2006), kako bi se dodatno simulirala varijabilnost kod sintetički generisanih *STR* profila dece. Skripta napisana u *Python*-u je kod dečjih profila na svakom 2000. mestu vršila nasumičnu promenu alela za jedan ponovak manje ili više u odnosu na onaj koji je definisan na tom mestu.

Stopa mutacije od 10^{-3} definisana je po lokusu, pri čemu je svaki lokus u generisanim profilima predstavljen diploidnim setom alela, odnosno sa dve kolone u matrici podataka. Nasumična promena alela stoga je uvođena na svakom 2000. mestu u matrici, čime je postignuta očekivana učestalost od jedne mutacije na 1000 analiziranih lokusa.

2.1.5. Ispitivanje homozigotnosti pri detekciji incesta

Homozigotnost je ispitana pomoću skripte napisane u *Python*-u. Skripta je za svaki lokus računala homozigotnost tako što je proveravala da li su aleli identični u dvema alelnim kolonama istog lokusa, npr. *2_CSFIPO_1* i *2_CSFIPO_2*. Prefiks 2 ukazuje da se radi o *STR* profilu na drugom mestu, koji odgovara dečjem profilu.

Za svaki lokus zabeležen je status homozigotnosti pojedinačnih profila, nakon čega su podaci grupisani prema kategorijama srodstva. Analizirane su kategorije *Normal Paternity*, *Normal Sibs*,

Incest Paternity i *Incest Sibs*. Za svaki lokus i kategoriju skripta je izračunavala prosečne stope homozigotnosti za svaki lokus unutar svake od kategorija, i izražavala ih u procentima.

2.1.6. Ispitivanje lokusa sa identičnim heterozigotnim genotipovima majke i deteta

Skriptom u programskom jeziku *Python* vršena je provera heterozigotnih alelnih kombinacija identičnih između profila majke i deteta. Ukoliko oba profila dele identični genotip, u novoj koloni se beleži rezultat provere za svaki lokus i pojedinca, a na kraju se računa prosečna stopa lokusa sa identičnim heterozigotnim genotipovima i izražava u procentima za svaki lokus unutar svake od tri testirane kategorije (*Incest Paternity*, *Incest Sibs*, *Normal Paternity*).

2.1.7. Računanje sličnosti *STR* profila majke i deteta

Korišćenjem skripte napisane u programskom jeziku *Python* izračunata je sličnost *STR* profila. Profili su poređeni na svakom lokusu. Kada su oba profila imala identičan homozigotni ili heterozigotni genotip, lokusu je dodeljena ocena 2. Kada su profili delili jedan alel, dodeljena je ocena 1, dok je lokusu bez zajedničkih alela dodeljena ocena 0. Procentualna sličnost izračunata je kao količnik zbira ocena svih analiziranih lokusa i maksimalno mogućeg zbira, odnosno dvostrukog broja analiziranih lokusa, pomnožen sa 100.

Budući da su lokusi na kojima su oba profila imala identičan heterozigotni genotip neposredno doprinosili ukupnoj vrednosti sličnosti ocenom 2, dodatno je izračunata korigovana sličnost profila nakon isključivanja takvih lokusa. Za svaki par profila iz analize su uklonjeni lokusi sa identičnim heterozigotnim genotipovima, nakon čega je zbir ocena na preostalim lokusima podeljen dvostrukim brojem preostalih lokusa i pomnožen sa 100.

Za svaku kategoriju srodstva izračunate su srednja vrednost i standardna devijacija ukupne i korigovane sličnosti *STR* profila.

2.1.8. Podele baza podataka

Baze podataka su u zavisnosti od specifičnosti analize bile podeljene na sledeći način: 80 % do 90 % podataka bilo je namenjeno za trening modela, dok je 5 % do 10 % podataka korišćeno za validaciju i testiranje modela. Kako bi se izbeglo prekomerno prilagođavanje modela na trening podatke, primenjena je desetostruka stratifikovana krosvalidacija, koja omogućava procenu performansi modela na različitim podskupovima podataka. Budući da su svi skupovi poticali iz istog simulacionog okvira, sprovedeni postupci predstavljali su internu evaluaciju modela, a ne spoljašnju validaciju na stvarnim *STR* profilima.

2.1.9. Testiranje modela

U studiji predikcije porekla testirani su sledeći modeli: Support Vector Machine (*SVM*), K-Nearest Neighbors (*KNN*), Naive Bayes, Decision Tree, Linear/Multinomial Logistic Regression (*L/MLR*) i *XGBoost*.

U studiji predikcije incesta testirani su sledeći modeli: Random Forest (*RF*), Support Vector Machine (*SVM*), Naive Bayes, K-Nearest Neighbors (*KNN*), *CatBoost*, *XGBoost*, *LightGBM*, *AdaBoost*, Gradient Boosting, Extra Trees i Multi-layer Perceptron.

Najbolji modeli definisani su na osnovu postignute inicijalne tačnosti prilikom predikcije.

2.1.10. Optimizacija hiperparametara

Za izabrane algoritme na osnovu inicijalno najboljeg učinka izvršena je optimizacija hiperparametara (eng. *grid search*), kako bi se dodatno unapredile osnovne performanse modela.

Za *XGBoost* testirani su sledeći parametri i njihove vrednosti: *learning_rate* (0,01, 0,1, 0,3 i 0,5), *max_depth* (6, 10 i 15), *n_estimators* (100, 200 i 500 stabala), *subsample* (0,5, 0,75 i 1,0) i *colsample_bytree* (0,5, 0,7 i 1,0). U konačnom modelu parametar *gamma* imao je vrednost 0,2, a parametar *min_child_weight* vrednost 5.

Za *CatBoost* testirani su sledeći opsezi: *iterations* (vrednosti 500, 1000 i 1500), *depth* (vrednosti 6, 8 i 10) i *learning_rate* (vrednosti 0,01, 0,05 i 0,1). U konačnom modelu korišćena je *Logloss* funkcija gubitka, uz rani prekid nakon 50 iteracija bez poboljšanja.

2.1.11. Evaluacija performansi modela

Glavni modeli u oba dela studije, *XGBoost* pri predikciji porekla i *CatBoost* pri detekciji incestuoznih odnosa, ocenjeni su kroz standardne parametre, uključujući i:

- Tačnost (eng. *accuracy*): udeo svih ispravno klasifikovanih profila u ukupnom broju analiziranih profila.
- Preciznost (eng. *precision*): udeo tačno pozitivnih predikcija među svim profilima koje je model klasifikovao kao pozitivne. U binarnoj klasifikaciji odgovara pozitivnoj prediktivnoj vrednosti (*PPV*).
- Odziv, odnosno osetljivost (eng. *recall/sensitivity*): udeo stvarno pozitivnih profila koje je model pravilno klasifikovao kao pozitivne.
- F1 skor: harmonijska sredina preciznosti i odziva.
- Površina ispod *ROC* krive (*ROC AUC*, eng. *area under the receiver operating characteristic curve*): mera sposobnosti modela da razlikuje pozitivnu i negativnu klasu kroz različite pragove klasifikacije.
- Uravnotežena tačnost (eng. *balanced accuracy*): aritmetička sredina osetljivosti i specifičnosti u binarnoj, odnosno one-versus-rest proceni.
- Matthewsov koeficijent korelacije (*MCC*): mera povezanosti između stvarnih i predviđenih klasa koja uzima u obzir tačno pozitivne, tačno negativne, lažno pozitivne i lažno negativne rezultate. Vrednosti se kreću od -1 do 1.
- Cohenov kappa koeficijent (eng. *Cohen's kappa*): mera slaganja između stvarnih i predviđenih klasa korigovana za slaganje koje se očekuje slučajno. Vrednosti se kreću od -1 do 1.
- Prevalencija (eng. *prevalence*): udeo stvarno pozitivnih slučajeva u analiziranom skupu. Stopa detekcije (eng. *detection rate*): udeo svih analiziranih slučajeva koji su stvarno pozitivni i pravilno klasifikovani kao pozitivni.
- Prevalencija detekcije (eng. *detection prevalence*): udeo svih analiziranih slučajeva koje je model klasifikovao kao pozitivne.

Dodatno su kreirane matrice konfuzije kako bi se identifikovali obrasci tačnih i pogrešnih klasifikacija, kao i krive preciznost–odziv za detaljniju analizu ponašanja modela u kontekstu specifičnih kategorija.

2.1.12. Analiza važnosti pojedinačnih lokusa

U obe faze istraživanja ispitan je uticaj pojedinačnih lokusa na samu predikciju, kako bi se definisali najinformativniji lokusi. Svakoј koloni dodeljena je vrednost važnosti, koja predstavlja uticaj pojedinačnog alela datog lokusa na preciznost modela.

Kako bi se dobila vrednost uticaja za lokus u celini, a ne za njegove pojedinačne kolone, vrednosti iz kolona _1 i _2 unutar istog lokusa su sabirane, a potom podeljene sa dva, čime je izračunata prosečna vrednost za svaki lokus.

2.1.13. Vizualizacija genetičke diferencijacije (PCA)

Kako bi se ispitala genetička diferencijacija između kategorija profila korišćena je *PCA* analiza. Na ovaj način smanjena je dimenzionalnost podataka, a istovremeno su identifikovani obrasci grupisanja i preklapanja između različitih kategorija.

PCA je sprovedena na kompletnim skupovima podataka, dok je za vizualizaciju nasumično odabran podskup profila (npr. 500 profila po kategoriji) sa ciljem lakše vizualne reprezentacije podataka na grafikonu i njegove interpretacije. Tokom obe faze istraživanja analizirane su prve dve glavne komponente, *PC1* i *PC2*, koje su ujedno i najinformativnije.

2.1.14. Uticaj selekcije lokusa i veličine baze podataka na predikciju porekla

Kod predikcije porekla poseban fokus bio je na proceni uticaja broja lokusa na performanse modela. Za simulaciju realnih scenarija DNK profilisanja korišćene su kombinacije lokusa koje odgovaraju komercijalnim kitovima poput Investigator 24plex/GlobalFiler, PowerPlex Fusion System, NGM Detect i AmpFISTR Identifier. Broj lokusa varirao je od 15 do 29 kako bi se ispitala uslovljenost tačnosti modela brojem testiranih lokusa.

Pored toga, ispitan je uticaj broja korišćenih profila na performanse modela. Sistematskim smanjenjem broja profila po kategorijama od 90000 do 500 ispitivana je tačnost modela i njegova uslovljenost brojem profila. Takođe, cilj je bio da se definiše optimalna donja granica broja profila potrebna za uspešno treniranje i testiranje modela, pogotovo u slučajevima kada populacije imaju složenu genetičku strukturu.

2.2. Specifičnosti analize

2.2.1. Predikcija porekla

Za predikciju porekla vršena su sledeća modelovanja predikcije: 1) klasifikacija četiri osnovne populacije (Afroamerikanci, SA-EP, Azijati, Hispanoamerikanci); 2) klasifikacija tri osnovne populacije (Afroamerikanci, SA-EP, Azijati) i 3) od kojih su dve bile osnovne (Afroamerikanci, Azijati), dok je treću činila objedinjena kategorija na osnovu profila SA-EP i Hispanoamerikanaca.

2.2.2. Detekcija incestuoznih odnosa

Za analizu detekcije incestuoznih odnosa vršene su sledeće predikcije: 1) za opštu populaciju analizom 29 lokusa; 2) za opštu populaciju analizom 21 lokusa; 3) za populaciju Saudijske Arabije analizom 21 lokusa; 4) za populaciju Saudijske Arabije na osnovu modifikovanih učestalosti alela, analizom 21 lokusa; 5) na osnovu jedinstvenih *STR* profila dece analizom 21 lokusa (eng. *offspring-only model*) i 6) višeklasna klasifikacija analizom 21 lokusa.

Model sa 29 lokusa primenjen je isključivo na opštu populaciju, budući da su samo za nju bile dostupne učestalosti alela za svih 29 autozomnih *STR* lokusa (poglavlje 2.1.1). Za populaciju Saudijske Arabije objavljene su učestalosti za 21 lokus obuhvaćen GlobalFiler kompletom hemikalija (Khubrani i sar., 2019), pa je panel od 21 lokusa korišćen kao zajednički skup u svim ostalim analizama, čime je omogućeno direktno poređenje performansi modela između populacija i eksperimentalnih uslova.

2.2.3. Dostupnost podataka i skripti

Formirane baze podataka i skripte u sklopu studije detekcije incesta javno su dostupne na Zenodo repozitorijumu (<https://zenodo.org/records/13287992>), dok su baze za predikciju porekla dostupne pod DOI oznakom 10.1016/j.fsigen.2024.103183.

III REZULTATI

3.1. Rezultati studije za predikciju porekla

3.1.1. Testiranje *ML* modela za predikciju porekla

U Tabeli 1 prikazana je tačnost i F1 vrednost za sve testirane algoritme. Modeli su testirani na bazi podataka sa četiri referentne populacione grupe, pri čemu je zadatak bio klasifikovanje svakog profila u jednu od četiri kategorije.

Tabela 1. Performanse algoritama mašinskog učenja u predikciji porekla

Testirani algoritmi	Tačnost predikcije (%)	F1 vrednost (%)
<i>SVM</i>	80,47	80,29
<i>KNN</i>	70,11	68,97
Naive Bayes	79,08	78,78
Decision Tree	73,28	73,33
<i>MLR</i>	76,45	76,26
<i>XGBoost</i>	93,24	93,23

Model *XGBoost* postigao je najveću tačnost predikcije, 93,24 %, i F1 meru od 93,23 %. Pri početnom poređenju svih modela korišćene su unapred definisane početne vrednosti hiperparametara. Podaci su podeljeni na trening skup, koji je sadržao 80 % profila, validacioni skup sa 10 % profila i interni testni skup sa preostalih 10 % profila.

3.1.2. Optimizacija hiperparametara *XGBoost* modela

Radi optimizacije performansi modela *XGBoost* primenjeno je iterativno testiranje različitih kombinacija hiperparametara metodom pretraživanja mreže (eng. *grid search*). Izabrana kombinacija hiperparametara prikazana je u Tabeli 2.

Tabela 2. Optimizovani hiperparametri *XGBoost* modela

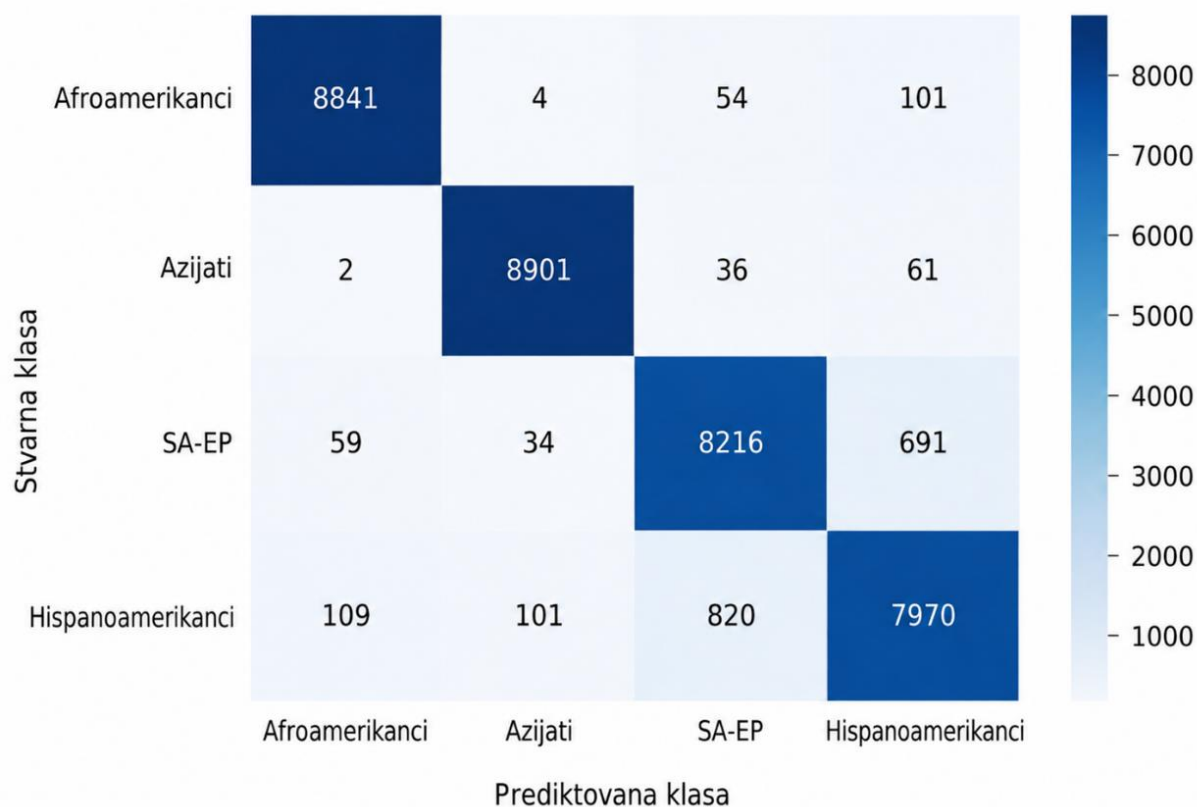
Parametar	Optimizovana vrednost
Metod učenja (eng. <i>booster</i>)	gbtree
Ciljna funkcija (eng. <i>objective</i>)	multi:softprob
Broj klasa (eng. <i>number of classes</i>)	4
Stopa učenja (eng. <i>learning rate/eta</i>)	0,5
Regulacija složenosti modela (eng. <i>gamma</i>)	0,2
Maksimalna dubina stabla (eng. <i>max depth</i>)	6
Minimalna suma Hesijanovih vrednosti u potomku čvora (eng. <i>min child weight</i>)	5
Poduzorak podataka (eng. <i>subsample</i>)	1
Poduzorak karakteristika stabla (eng. <i>colsample_bytree</i>)	0,5

3.1.3. Model predikcije na osnovu četiri osnovne populacije

Model zasnovan na klasifikaciji četiri referentne populacione grupe analizom 29 lokusa postigao je ukupnu tačnost od 94,24 %. Na osnovu matrice konfuzije na Slici 2 možemo da vidimo rezultate klasifikacije koje je model postigao za svaku kategoriju. Dijagonalno postavljena tamnoplava polja pokazuju broj pravilno klasifikovanih profila.

Analizom matrice konfuzije može se utvrditi da je od 9000 profila, najveći broj tačnih klasifikacija postignut za azijsku grupu uzoraka (8901), koju prati afroamerička grupa uzoraka sa 8841 tačnih predikcija. Takođe, zapaža se da je između ove dve grupe postignuto najmanje pogrešnih klasifikacija, sa svega šest pogrešno klasifikovanih profila. Ovaj rezultat pokazuje da je model dobro razlikovao afroameričku i azijsku grupu.

Model je dao 1511 pogrešnih klasifikacija između SA–EP i hispanoameričke grupe uzoraka. SA–EP grupa imala je 784 pogrešno klasifikovana profila, od kojih je 691 profil pripisan hispanoameričkoj populaciji. Hispanoamerička grupa dala je najmanji broj tačnih predikcija sa ispravno klasifikovanih 7970 profila, dok je 820 profila iz ove kategorije model pripisao SA–EP grupi i dodatnih 210 profila afroameričkoj i azijskoj grupi.



Slika 2. Matrica konfuzije za predikciju porekla četiri osnovne populacije analizom 29 lokusa

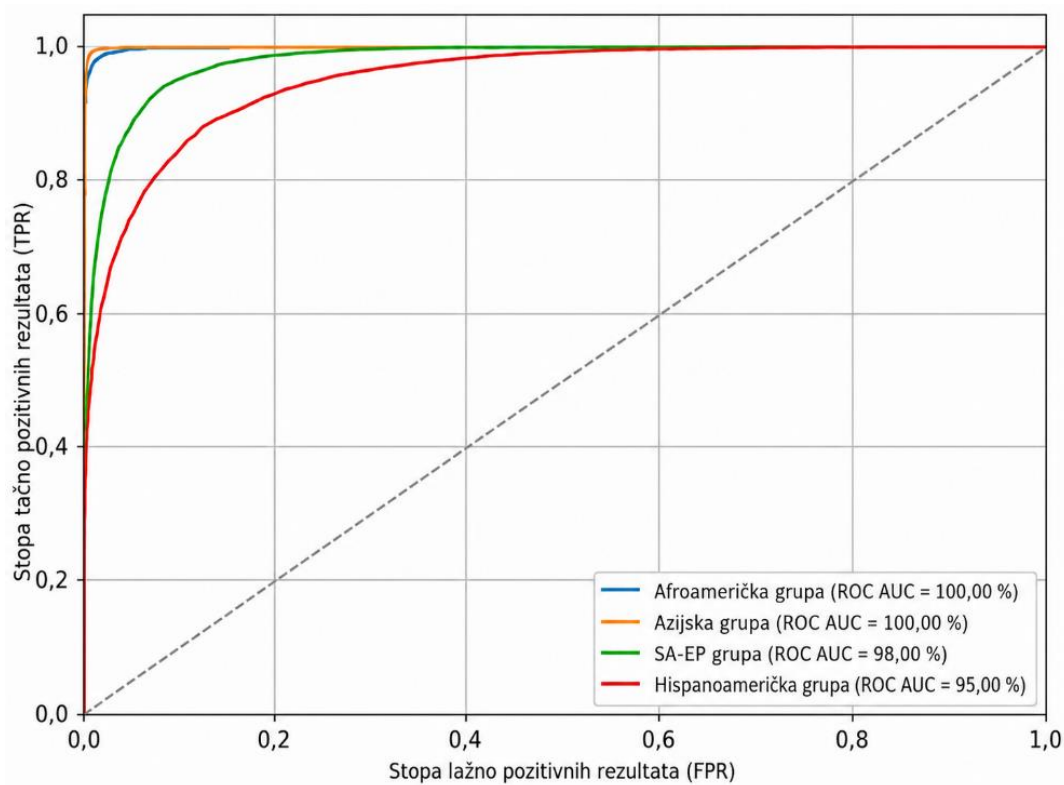
Dodatni statistički parametri modela, uključujući osetljivost, specifičnost, pozitivnu prediktivnu vrednost (*PPV*) i negativnu prediktivnu vrednost (*NPV*), prikazani su u Tabeli 3. Najviše vrednosti osetljivosti, *PPV* i *NPV* zabeležene su za afroameričku i azijsku grupu, dok su niže vrednosti zabeležene za SA–EP i hispanoameričku grupu. Uravnotežena tačnost iznosila je 99,19 % za azijsku grupu, a 92,70 % za hispanoameričku grupu.

Tabela 3. Statistički parametri predikcije porekla prema ispitanim kategorijama

Klasa	Osetljivost (%)	Specifičnost (%)	Pozitivna prediktivna vrednost (%)	Negativna prediktivna vrednost (%)	Prevalencija (%)	Stopa detekcije (%)	Prevalencija detekcije (%)	Uravnotežena tačnost (%)
Afroamerikanci	98,23	99,37	98,11	99,41	25,00	24,56	25,03	98,80
Azijati	98,90	99,49	98,46	99,63	25,00	24,72	25,11	99,19
SA–EP	91,29	96,63	90,03	97,08	25,00	22,82	25,35	93,96
Hispanoamerikanci	88,56	96,84	90,33	96,21	25,00	22,14	24,51	92,70

Rezultati su pokazali više vrednosti osetljivosti, specifičnosti, *PPV* i *NPV* za afroameričku i azijsku grupu nego za SA–EP i hispanoameričku grupu, između kojih je zabeležen veći broj pogrešnih klasifikacija.

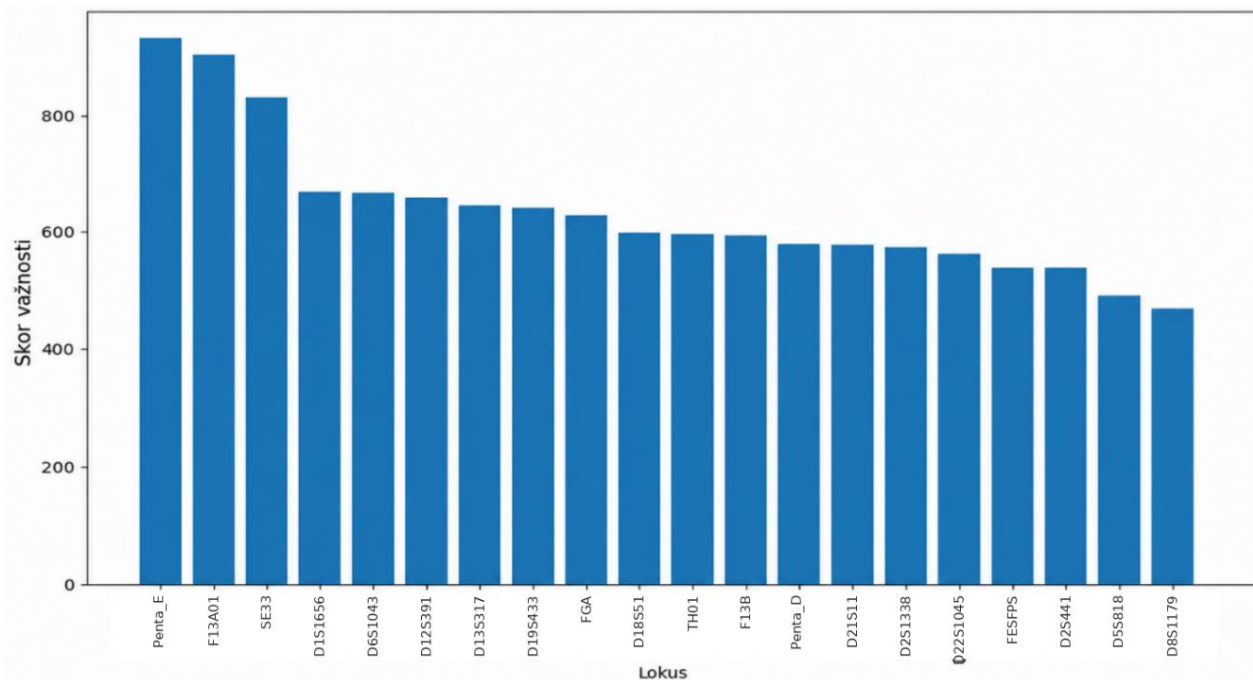
Sposobnost modela da razlikuje klase kroz jedan–protiv–svih (*OVA*) analizu predstavljena je grafikonom *ROC* krive na slici 3. Model je postigao *ROC AUC* vrednosti od 100,00 % za afroameričku i azijsku kategoriju, što ukazuje na visoku diskriminatornu sposobnost modela za ove klase. Za SA–EP grupu dobijena je vrednost od 98,00 %, dok je najniža vrednost od 95,00 % zabeležena za hispanoameričku kategoriju profila.

**Slika 3.** *ROC* krive modela za klasifikaciju četiri referentne populacione grupe.

3.1.4. Ocena informativnosti pojedinačnih lokusa za predikciju porekla

Kako bi se ispitao kumulativni uticaj lokusa, a ne njihov zaseban uticaj, predikcija je vršena na osnovu srednje vrednosti uticaja analiziranih lokusa. Na slici 4 dat je grafikon na kome je predstavljeno 20 najinformativnijih lokusa za predikciju, poređanih od najviše informativnog (krajnje levo) ka najmanje informativnom (krajnje desno).

Kao najinformativniji lokus za predikciju porekla identifikovan je *Penta E*, koga prate *F13A01* i *SE33*, svi sa ocenom značaja većom od 800. U kategoriji sa uticajem većim od 600 našli su se *D1S1656*, *D6S1043*, *D12S391*, *D13S317*, *D19S433* i *FGA* lokus. Kao dva najmanje informativna lokusa na Slici 4 prepoznata su *D5S818* i *D8S1179*. U dodatnoj analizi, nakon isključivanja najmanje informativnih lokusa, ukupna tačnost modela smanjena je za približno jedan procentni poen.

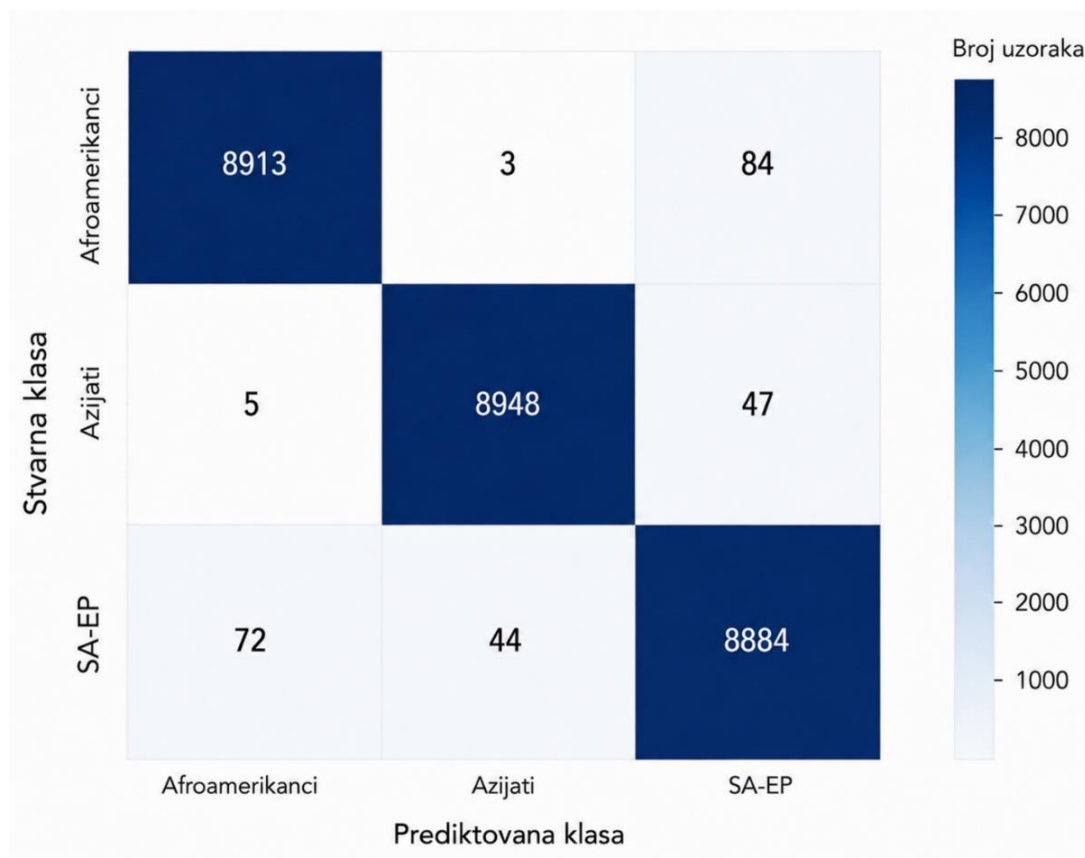


Slika 4. 20 najvažnijih lokusa poređanih prema značaju. Visina stuba kvantifikuje doprinos pojedinačnog lokusa u tačnoj predikciji

3.1.5. Model za klasifikaciju tri grupe (hispanoamerički poduzorak isključen)

Kako je već u prvoj fazi predikcije porekla utvrđeno da model ima poteškoće pri diferencijaciji genetički sličnih populacija, u ovom koraku ispitana je efikasnost modela pri analizi tri osnovne populacije, sa isključenom hispanoameričkom grupom profila. Na ovaj način dobijena je tačnost modela od 99,06 %.

Matrica konfuzije predstavljena je na slici 5, na osnovu koje se zapaža da je najveći broj pravilno klasifikovanih profila, 8948, dobijen za azijsku grupu. Za ovu grupu model je pogrešno klasifikovao 52 profila, od kojih je 5 profila pripisao afroameričkoj grupi i dodatnih 47 SA–EP grupi. Afroamerički poduzorak imao je 8913 tačno klasifikovanih profila, dok su samo 3 profila iz te kategorije dodeljena azijskoj grupi i 84 SA–EP. Najmanje tačnih klasifikacija postignuto je u SA–EP grupi sa 8884 ispravno klasifikovana profila. Iz ove grupe 72 profila pogrešno je prepoznato kao afroamerički, a 44 kao profili iz azijske grupe.



Slika 5. Matrica konfuzije klasifikacije tri populacije (hispanoamerička populacija isključena)

Statistički parametri za ocenu modela predstavljeni su u Tabeli 4. Sposobnost modela da tačno identifikuje pozitivne slučajeve (osetljivost) iznosila je preko 99 % za afroameričku i azijsku grupu profila, dok je kod SA–EP iznosila 98,71 %.

Specifičnost, odnosno udeo stvarno negativnih profila koji su pravilno klasifikovani kao negativni, iznosila je 99,57 % za afroameričku grupu, 99,74 % za azijsku grupu i 99,27 % za SA–EP grupu.

Tabela 4. Parametri određivanja porekla tri ispitane populacije

Klasa	Osetljivost (%)	Specifičnost (%)	Pozitivna prediktivna vrednost (%)	Negativna prediktivna vrednost (%)	Prevalencija (%)	Stopa detekcije (%)	Prevalencija detekcije (%)	Uravnotežena tačnost (%)
Afroamerikanci	99,03	99,57	99,14	99,52	33,33	33,01	33,30	99,30
Azijati	99,42	99,74	99,48	99,71	33,33	33,14	33,31	99,58
SA–EP	98,71	99,27	98,55	99,36	33,33	32,90	33,39	98,99

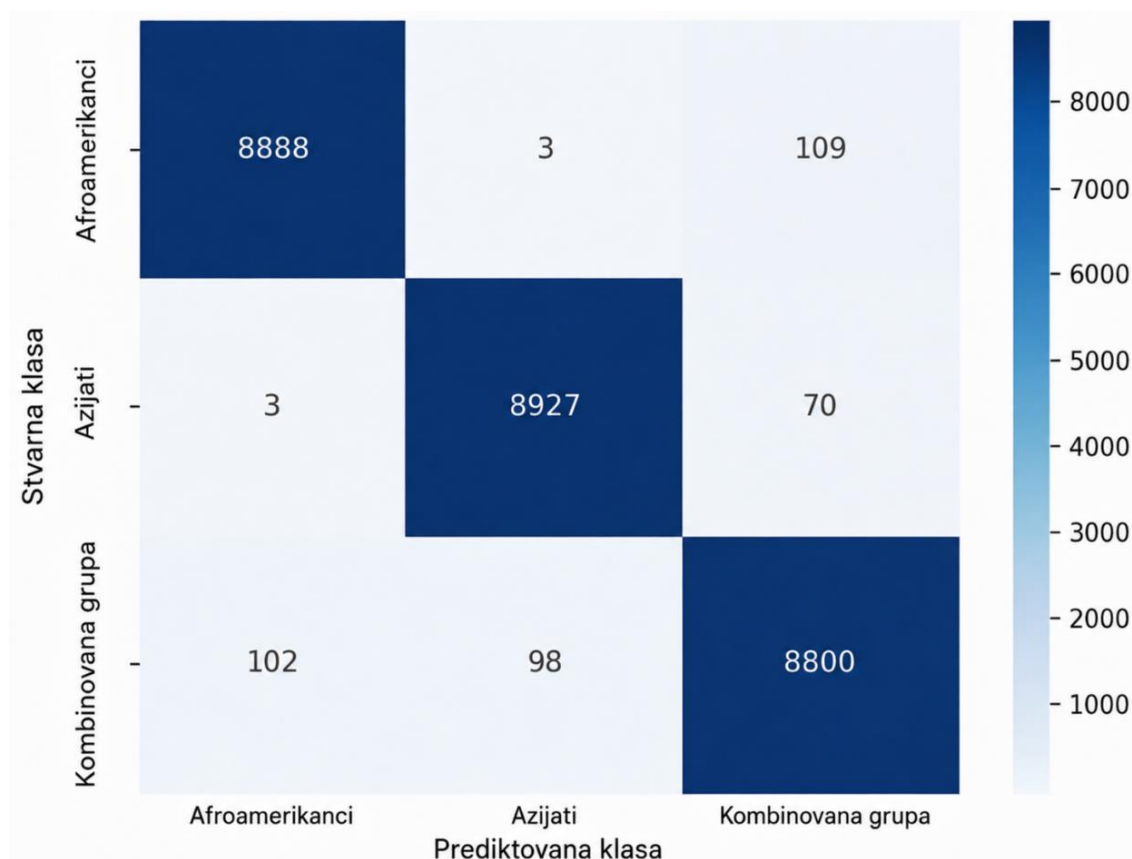
NPV ukazuje na pouzdanost isključivanja određenog porekla, odnosno verovatnoću da profil koji model nije svrstao u datu grupu zaista i ne pripada toj grupi. Ovaj parametar je visok za sve kategorije, a najvišu vrednost, od 99,71 %, ima za azijsku grupu. Na ujednačene stope detekcije ukazuje i ravnomerna distribucija prevalencije, koja je iznosila 33,33 % za svaku grupu.

3.1.6. Model za klasifikaciju tri grupe (objedinjene SA–EP i hispanoamerička populacija)

Sledeći pristup podrazumevao je predikciju afroameričke i azijske grupe kao osnovnih populacija, dok je kao treća korišćena objedinjena (eng. *combined*) kategorija, koja se sastojala od ravnomerno zastupljenih profila hispanoameričke i SA–EP grupe.

Kombinovana kategorija formirana je spajanjem po 45000 profila iz SA–EP i hispanoameričke podgrupe. Ovaj pristup omogućio je bolju klasifikaciju, usled objedinjene kategorije genetički srodnih grupa. Ovim pristupom model je postigao tačnost od 98,57 %.

Uvidom u matricu konfuzije sa Slike 6 uočava se da je model, od 9000 profila unutar svake grupe, ispravno klasifikovao 8888 afroameričkih profila, 8927 azijskih profila i 8800 profila iz kombinovane kategorije. Preciznost modela potvrđuju i pogrešne klasifikacije samo 385 profila.



Slika 6. Matrica konfuzije klasifikacije tri populacije (objedinjena SA–EP i hispanoamerička populacije)

Statistički parametri za ocenu performansi modela prikazani su u tabeli 5. Model je postigao osetljivost od 98,76 % za Afroamerikance, 99,19 % za Azijate i 97,78 % za kombinovanu kategoriju. Specifičnost za sve klase bila je iznad 99 %.

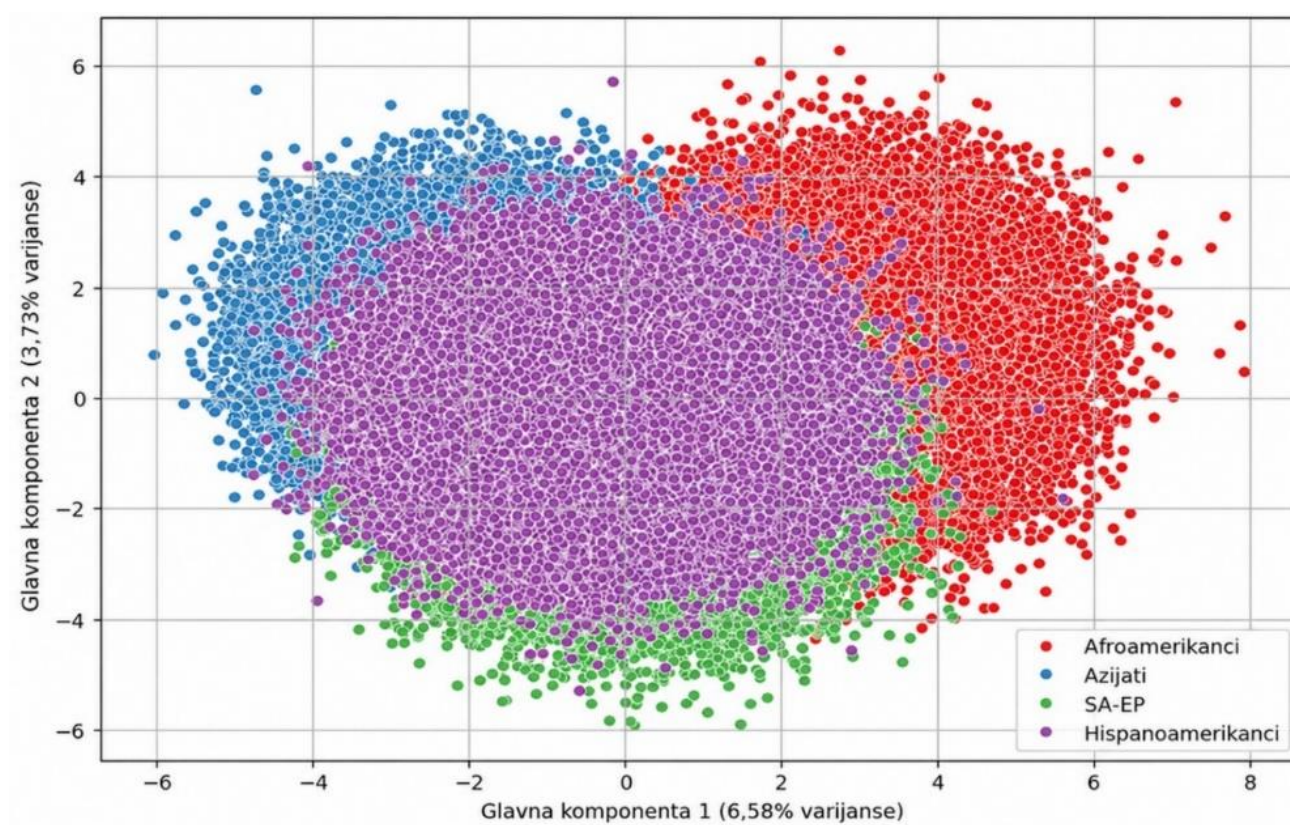
Visoke vrednosti *PPV* i *NPV* ukazale su na visok udeo ispravnih pozitivnih i negativnih predikcija. Uravnotežena tačnost iznosila je 99,09 % za afroameričku grupu, 99,31 % za azijsku grupu i 98,39 % za kombinovanu grupu, pri čemu je najviša vrednost zabeležena za azijsku grupu. Uprkos većoj varijabilnosti u kombinovanoj grupi profila, model je pokazao gotovo identične performanse sa modelom za predikciju na osnovu tri osnovne populacije.

Tabela 5. Performanse modela za klasifikaciju tri kategorije

Klasa	Osetljivost (%)	Specifičnost (%)	Pozitivna prediktivna vrednost (%)	Negativna prediktivna vrednost (%)	Prevalencija (%)	Stopa detekcije (%)	Prevalencija detekcije (%)	Uravnotežena tačnost (%)
Afroamerikanci	98,76	99,42	98,83	99,38	33,33	32,92	33,31	99,09
Azijati	99,19	99,44	98,88	99,59	33,33	33,06	33,44	99,31
Kombinovana grupa	97,78	99,01	98,01	98,89	33,33	32,59	33,26	98,39

3.1.7. PCA analiza

PCA analizom ispitane su prve dve glavne komponente, *PC1* i *PC2*, koje su objasnile 6,58 % i 3,73 % varijabilnosti unutar sve četiri kategorije. Uvidom u PCA dijagram na Slici 7 zapaža se da su afroamerički profili formirali klaster na desnoj strani grafikona, dok su se azijski profili grupisali u gornjem levom uglu. Ovo ukazuje da se radi o genetički udaljenim populacijama, te stoga i ne čudi što su svi testirani modeli vršili najbolju diferencijaciju između ove dve grupe. Dalje se može zapaziti preklapanje između SA–EP i Hispanoamerikanaca, koji su se grupisali u centralnom delu dijagrama. Ovo ukazuje na njihovu genetičku sličnost, koja je razlog slabe klasifikacije između ove dve grupe.

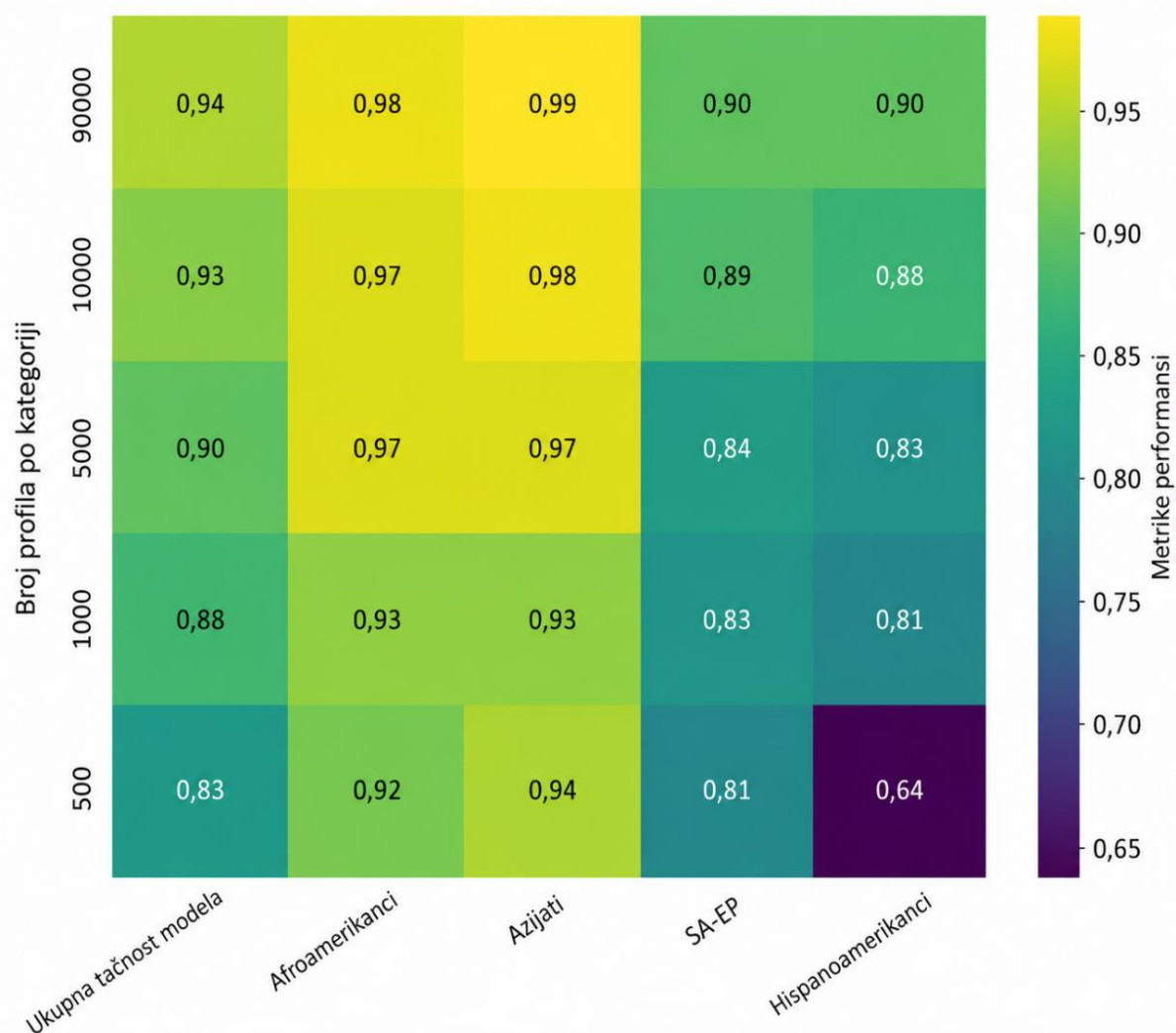


Slika 7. PCA dijagram za četiri populacije. Na x-osi prikazana je prva glavna komponenta (*PC1*), a na y-osi druga glavna komponenta (*PC2*).

3.1.8. Performanse modela klasifikacije sa promenjivim veličinama uzoraka

Ispitana je uslovljenost modela u odnosu na broj profila u sklopu svake populacione grupe. Iterativnim pristupom vršena je predikcija sa 90000, 10000, 5000, 1000 i 500 profila unutar svake grupe. U ovom delu istraživanja baza podataka bila je podeljena na način da je 90 % profila korišćeno za treniranje modela, dok je po 5 % korišćeno za validaciju i testiranje. Razlog ovakve podele leži u tome što, kada se već radi sa manjim brojem profila, treba omogućiti modelu učenje na najvećem mogućem broju podataka. Ukupne tačnosti modela i pojedinačnih kategorija predstavljene su na Slici 8.

U svim ponavljanjima model je dao najvišu tačnost za azijsku grupu, u kojoj je i na osnovu 500 profila uspeo da postigne tačnost od 94 %, dok je analizom svih profila tačnost porasla na 99 %. Sličan trend je zabeležen i unutar afroameričke grupe, s tim što je početna tačnost iznosila 92 %, dok je finalna iznosila 98 %. Ukupna tačnost modela povećala se za 11 procentnih poena, sa 83 % na 94 %, pri klasifikaciji četiri kategorije. Najizraženija promena zabeležena je za hispanoameričku grupu, kod koje se tačnost povećala za 26 procentnih poena. Ovo ukazuje da je modelu bila potrebna veća baza podataka kako bi uspeo da uoči i primeni sve suptilne obrasce koji se javljaju unutar ove grupe. Razlika u tačnosti između modela treniranih na 10000 i na 90000 profila po populacionoj grupi iznosila je oko jednog procentnog poena, čime je 10000 profila po grupi određeno kao tačka zasićenja, odnosno kao najmanja veličina trening skupa pri kojoj model postiže stabilne performanse.



Slika 8. Tačnost ukupnog modela i po kategorijama u odnosu na veličinu skupa podataka

3.1.9. Uticaj broja i kombinacije lokusa na performanse modela

Poslednja faza studije predikcije porekla zasnivala se na analizi različitog broja lokusa i njihovih kombinacija. Iz inicijalno generisanih profila za 29 lokusa, izvršena je ekstrakcija profila koji bi se dobili upotrebom nekog od sledećih amplifikacionih kompleta: *Investigator 24plex/GlobalFiler STR* kit, *PowerPlex Fusion*, *AmpFlSTR Identifiler* i *NGM Detect*.

Informativnost ovih kompleta procenjena je na osnovu tačnosti pri klasifikaciji sve četiri grupe, kao i pri klasifikaciji tri grupe nakon isključenja hispanoameričke populacije. Rezultati ovog dela studije predstavljeni su u Tabeli 6.

Tabela 6. Amplifikacioni kitovi i tačnosti modela pri različitom broju i kombinaciji lokusa

Naziv kompleta hemikalija za amplifikaciju DNK	Tačnost modela (%) za četiri kategorije	Tačnost modela (%) (Hispanoamerikanci isključeni)	Broj testiranih lokusa
<i>Investigator 24plex / GlobalFiler STR</i>	88,28	96,30	21
<i>PowerPlex Fusion System</i>	87,23	95,76	22
<i>AmpFlSTR Identifiler</i>	81,84	92,72	15
<i>NGM Detect</i>	84,97	94,48	16

Kompleti hemikalija poput *Investigator 24plex* i *GlobalFiler STR* pokazali su se najefikasnijima. Budući da obuhvataju iste lokuse, oba kompleta dala su identične tačnosti od 88,28 % za sve četiri grupe i 96,30 % sa isključenom hispanoameričkom grupom. *PowerPlex Fusion* ostvario je slične rezultate, s neznatno manjom tačnošću u odnosu na prethodna dva. Pri predikciji su se najslabije pokazali *AmpFlSTR Identifiler* i *NGM Detect*, koji obuhvataju i najmanji broj analiziranih lokusa.

Ovi rezultati pokazuju da performanse modela zavise i od broja i od informativnosti analiziranih lokusa, pri čemu veći broj lokusa ne dovodi nužno do veće tačnosti klasifikacije. Kompleti sa manjim brojem lokusa mogu ostvariti zadovoljavajuće performanse ukoliko se predikcija odnosi na genetički udaljene grupe.

3.2. Rezultati studije detekcije incesta

3.2.1. Testiranje *ML* modela za detekciju incesta

Kreirane baze podataka sa različitim oblicima srodstva testirane su primenom raznih *ML* algoritama. Cilj je bio ispitati efikasnost modela u binarnoj klasifikaciji incesta, kako bi se identifikovao najefikasniji model za ovaj zadatak. Efikasnost modela ocenjena je, između ostalog, tačnošću i preciznošću, kao i dodatnim parametrima, koji su predstavljeni u Tabeli 7.

Tabela 7. Performanse algoritama mašinskog učenja za detekciju incestuoznih veza

Testirani algoritmi	Tačnost (%)	Preciznost (%)	Osetljivost (%)	F1 vrednost (%)	<i>ROC AUC</i> (%)
Random Forest	78,10	76,97	80,20	78,55	85,85
Support Vector Machine	78,40	78,34	78,50	78,42	87,19
Naive Bayes	74,15	75,23	72,00	73,58	81,71
K-Nearest Neighbors	63,90	65,10	59,90	62,39	68,00
<i>CatBoost</i>	92,35	92,73	91,90	92,31	97,71
<i>XGBoost</i>	90,05	89,54	90,70	90,11	95,52
<i>LightGBM</i>	90,45	90,17	90,80	90,48	96,83
<i>AdaBoost</i>	76,60	76,71	76,40	76,55	84,16
Gradient Boosting	87,80	87,43	88,30	87,86	95,25
Extra Trees	77,40	79,59	73,70	76,53	85,47
Multi-layer Perceptron	75,80	71,32	86,30	78,10	85,77

Zapaža se da su se od svih testiranih algoritama po svojoj efikasnosti izdvojili algoritmi gradijentnog pojačavanja (eng. *gradient boosting*), od kojih je *CatBoost* algoritam postigao tačnost od 92,35 %, F1 skor od 92,31 % i *ROC AUC* od 97,71 %.

3.2.2. Optimizacija hiperparametara *CatBoost* modela

Radi unapređenja performansi *CatBoost* modela izvršena je optimizacija hiperparametara metodom mrežne pretrage. Izabrana kombinacija hiperparametara prikazana je u Tabeli 8.

Tabela 8. Optimizovani hiperparametri *CatBoost* modela

Parametar	Optimizovana vrednost
Broj iteracija (eng. <i>iterations</i>)	1000
Stopa učenja (eng. <i>learning rate</i>)	0,1
Dubina stabla (eng. <i>depth</i>)	10
Funkcija gubitka (eng. <i>loss_function</i>)	Logloss

Optimizacijom hiperparametara postignuta je bolja stabilnost modela i smanjenje prenaučivosti (eng. *overfitting*). *Logloss* je odabran kao odgovarajuća funkcija gubitka za binarnu klasifikaciju, dok je *ROC AUC* korišćen kao metrika diskriminatorne sposobnosti modela. Rani

prekid (eng. *early stopping*) na 50 iteracija omogućio je balans između efikasnosti treniranja i sprečavanja prenaučivosti modela.

3.2.3. Model za opštu populaciju (varijanta sa 29 lokusa)

Nakon podešavanja hiperparametara, *CatBoost* model je pri binarnoj klasifikaciji u opštoj populaciji, na osnovu 29 lokusa, postigao ukupnu tačnost od 96,94 %. Matrica konfuzije predstavljena je na Slici 9, a u nastavku su date *ROC* kriva na Slici 10 i kriva preciznosti i odziva na Slici 11.

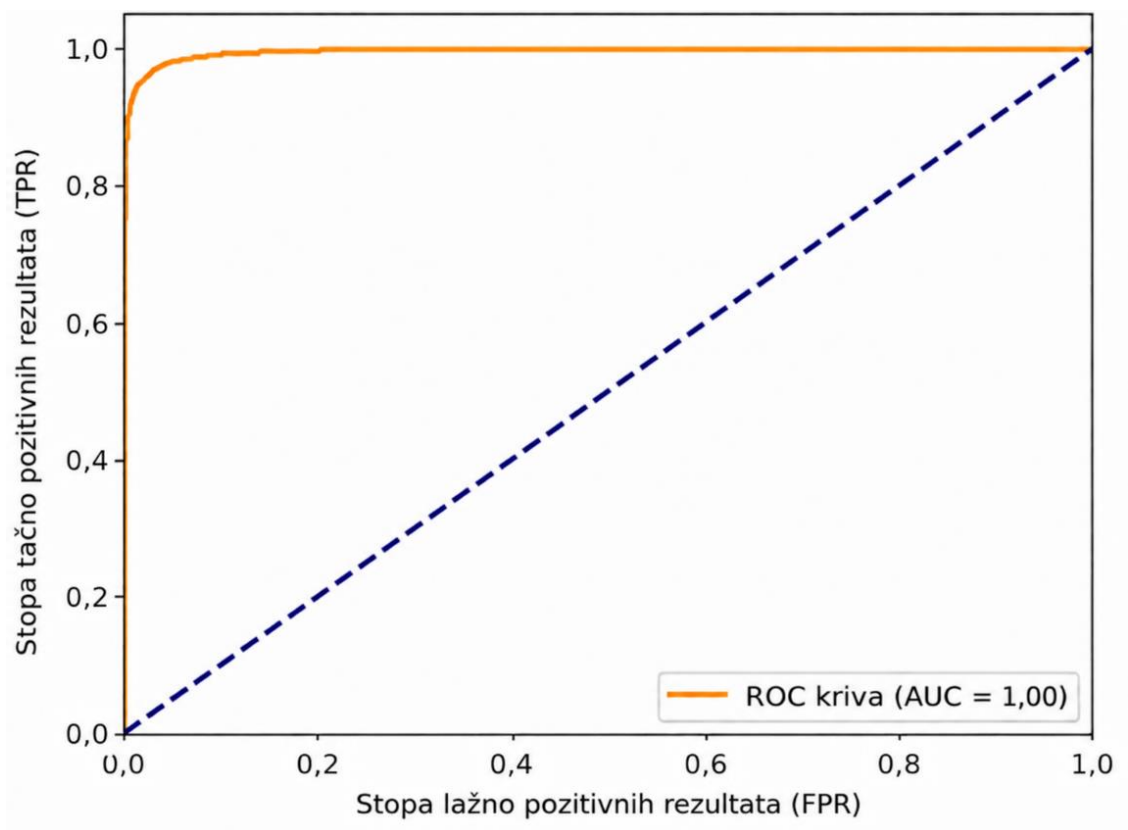


Slika 9. Matrica konfuzije binarne klasifikacije opšte populacije pri analizi 29 lokusa

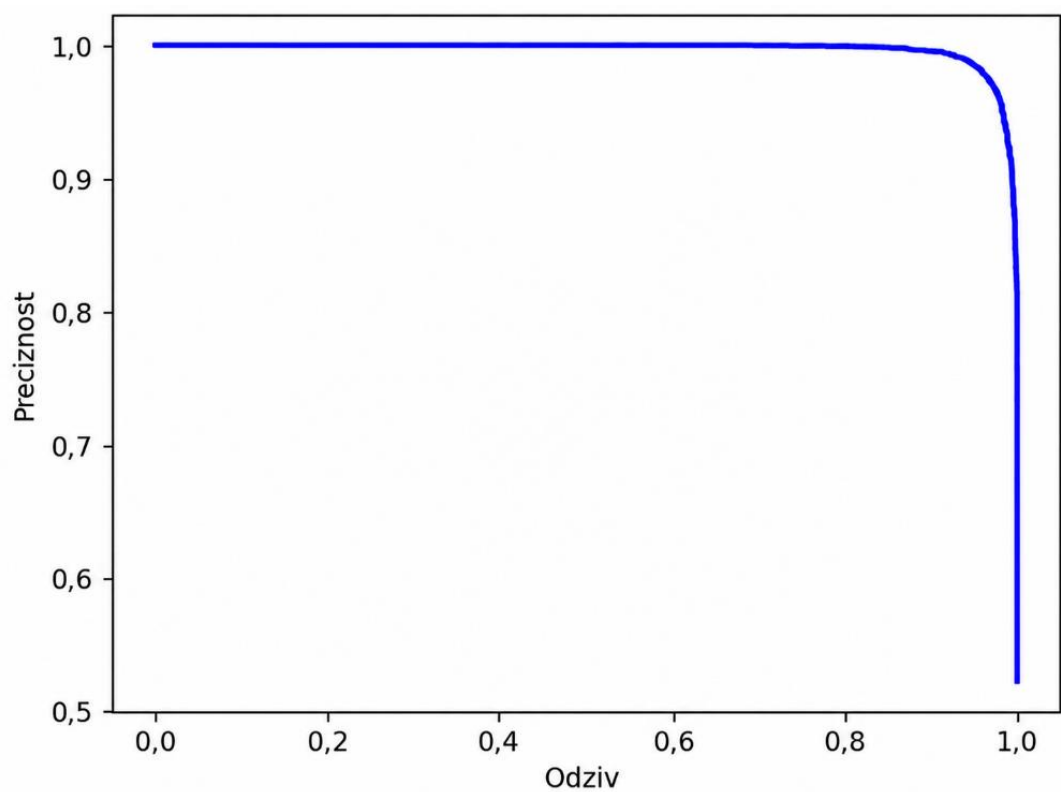
Uvidom u matricu konfuzije zapaža se da je model sa gotovo jednakom uspešnošću pravilno klasifikovao profile majke i deteta nastale nasumičnim ukrštanjem i profile majke i deteta kod kojih je dete posledica konsangviniteta. U tom zadatku dao je 9694 tačne predikcije u kategoriji normalnog očinstva i 9695 u kategoriji incestnog odnosa. Pogrešno je klasifikovao 306 profila kao slučajeve incestuoznog srodstva i 305 profila kao normalno srodstvo. Analizom dodatnih parametara utvrđeno je da su uravnotežena tačnost i *PPV* iznosile 96,94 %, dok su *NPV* i odziv iznosili 96,95 %.

Kriva preciznosti–odziva na Slici 11 pokazala je da je model zadržavao visoku preciznost pri visokim vrednostima odziva. Pri primenjenom klasifikacionom pragu preciznost je iznosila 96,94 %, a odziv 96,95 %, što ukazuje na uravnotežen odnos između lažno pozitivnih i lažno negativnih klasifikacija u internom testnom skupu.

Matthewsov koeficijent korelacije (*MCC*) između stvarnih i predviđenih klasa iznosio je 0,9389. Specifičnost modela iznosila je 96,94 %, što znači da je tačno identifikovao 96,94 % svih negativnih slučajeva. Stopa lažno pozitivnih iznosila je 3,06 %, dok je stopa lažno negativnih iznosila 3,05 %. Cohenov kappa koeficijent, koji odražava stepen slaganja između predviđenih i stvarnih klasifikacija korigovan za slučajno slaganje, iznosio je 0,9389.



Slika 10. ROC kriva modela za opštu populaciju



Slika 11. Kriva preciznosti i odziva

ROC AUC iznosio je 99,62 %, što ukazuje na visoku diskriminatornu sposobnost modela između dve klase.

3.2.4. Genetički indikatori konsangviniteta

Kao prvi indikator konsangviniteta ispitana je sličnost između dva *STR* profila analizom 29 lokusa, u zavisnosti od oblika srodstva. U Tabeli 9 prikazane su srednje vrednosti dobijene analizom 200.000 parova profila u kategorijama normalnog očinstva i scenarija otac–ćerka, odnosno 150.000 parova profila u kategorijama normalnog srodstva brat–sestra i scenarija brat–sestra.

Prosečna ukupna sličnost iznosila je 59,43 % za odnos roditelj–dete u normalnom očinstvu i 69,57 % za odnos majka–dete kada je dete nastalo iz odnosa otac–ćerka, što predstavlja razliku od 10,14 procentnih poena. Za normalan odnos brat–sestra prosečna sličnost iznosila je 61,97 %, a za odnos majka–dete kada je dete nastalo iz odnosa brata i sestre 69,54 %, uz razliku od 7,57 procentnih poena.

Korigovana sličnost iznosila je 52,74 % za normalno očinstvo i 58,10 % za scenario otac–ćerka, odnosno 46,83 % za normalan odnos brat–sestra i 58,08 % za scenario brat–sestra. Razlike između odgovarajućih kategorija iznosile su 5,36 i 11,25 procentnih poena.

Tabela 9. Sličnost *STR* profila usled različitih kategorija srodstva

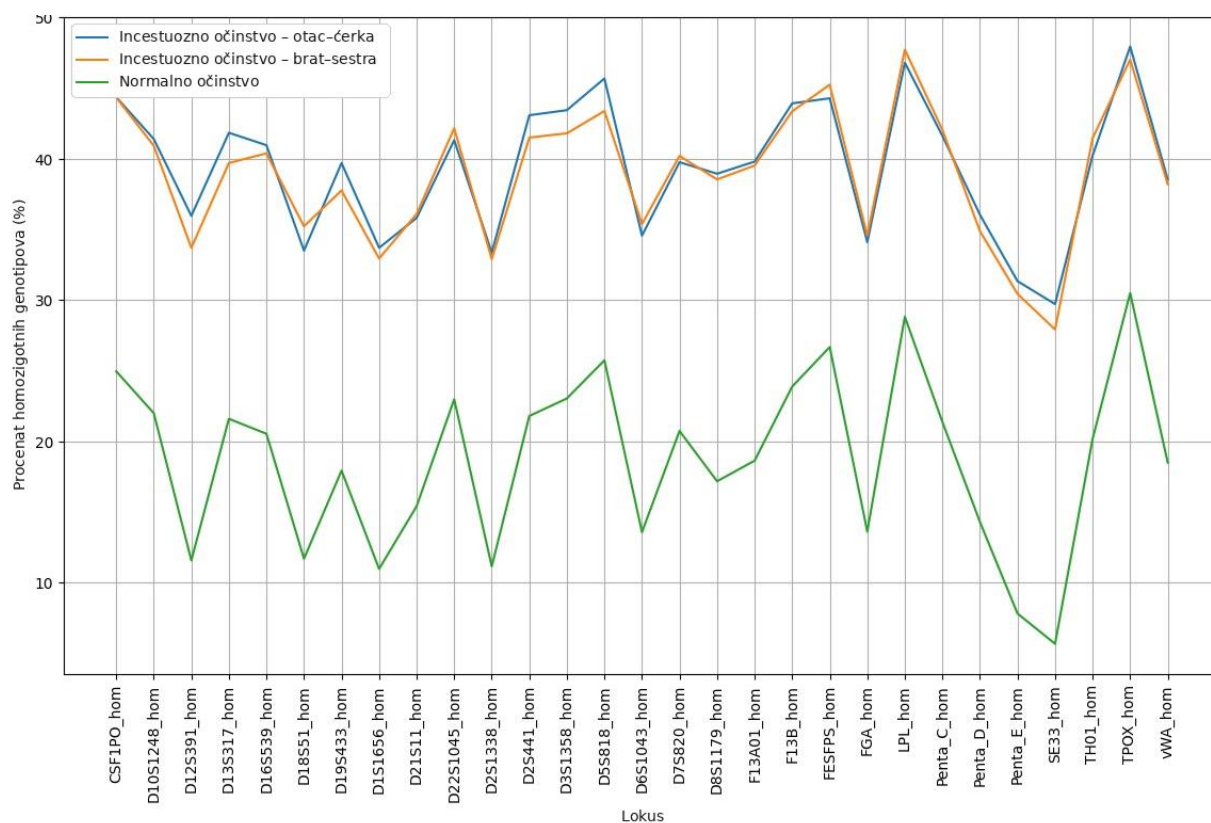
Kategorija srodstva	Ukupna sličnost profila, srednja vrednost $\pm$ SD (%)	Korigovana sličnost profila, srednja vrednost $\pm$ SD (%)
Roditelj–dete (normalno očinstvo)	59,43 $\pm$ 3,58	52,74 $\pm$ 2,25
Majka–dete (dete iz odnosa otac–ćerka)	69,57 $\pm$ 4,50	58,10 $\pm$ 4,00
Brat–sestra (normalno srodstvo)	61,97 $\pm$ 5,81	46,83 $\pm$ 5,59
Majka–dete (dete iz odnosa brata i sestre)	69,54 $\pm$ 4,50	58,08 $\pm$ 3,99

Dalje je ispitan procenat homozigotnosti, čiji su rezultati predstavljeni na Slici 12. Analiza je sprovedena na osnovu 29 lokusa dečjih profila iz kategorija normalnog očinstva i kategorija u kojima su deca nastala usled ukrštanja brat–sestra ili otac–ćerka.

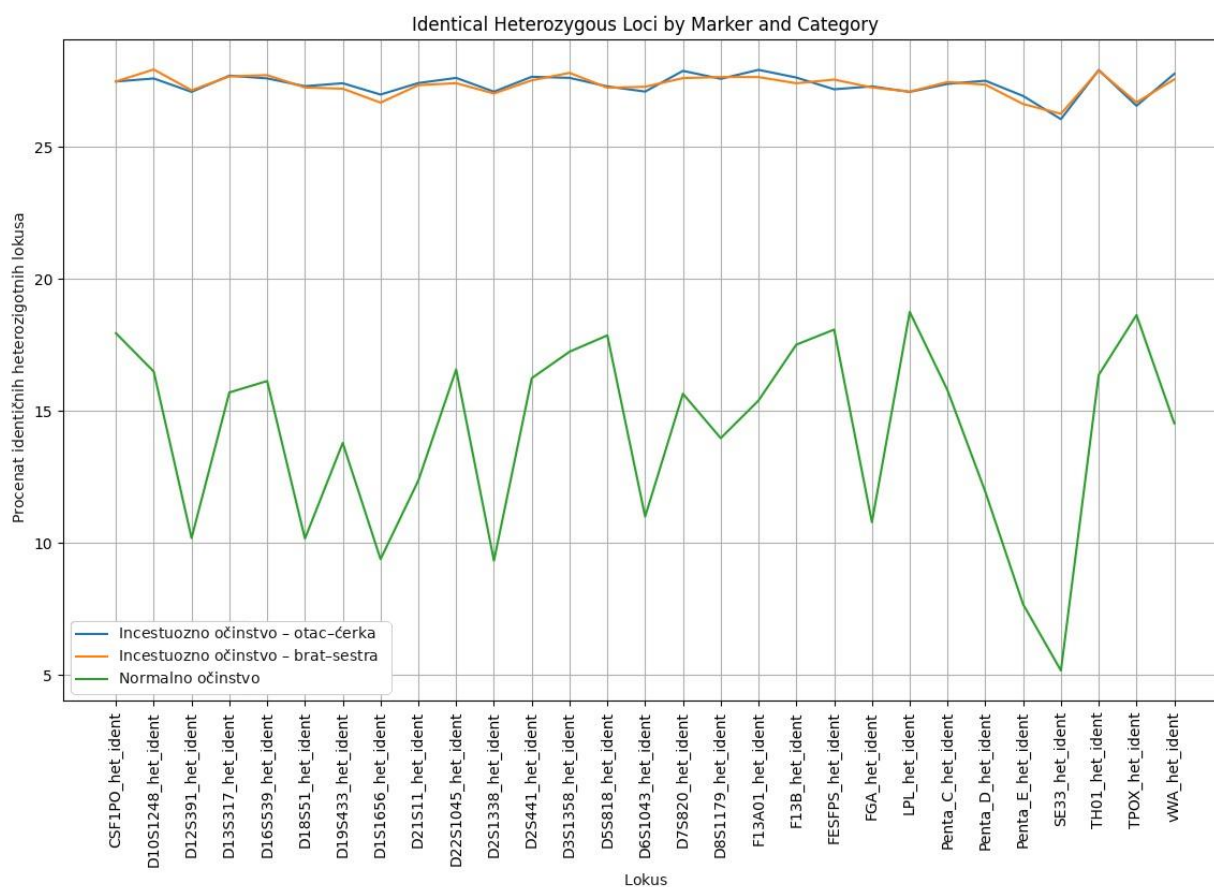
U obe incestuozne grupe primećena je povišena stopa homozigotnosti (28 %–48 %) u odnosu na kontrolnu grupu, u kojoj se vrednosti kretale između 7 % i 30 %. Ovi rezultati sugerišu da su više stope homozigotnosti uzrokovane srodstvom. Veća genetička raznolikost unutar kontrolne grupe smanjuje verovatnoću pojavljivanja identičnih alela na istom lokusu, što objašnjava niže vrednosti homozigotnosti u ovoj grupi.

U slučajevima očinstva kao posledice konsangviniteta, očekuje se veći broj lokusa na kojima dva profila dele identične heterozigotne genotipove (25 %–27 %). Nasuprot tome, u slučajevima normalnog očinstva, vrednosti se kreću u rasponu od 5 % do 18,5 %. Ovi rezultati potvrđuju da srodnički odnosi između roditelja povećavaju učestalost lokusa sa identičnim heterozigotnim genotipovima majke i deteta, što dodatno doprinosi smanjenju genetičke raznolikosti unutar ovih grupa.

Na kraju je ispitana i procentualna zastupljenost lokusa sa identičnim heterozigotnim genotipovima majke i deteta za tri kategorije (Slika 13).



Slika 12. Procentat homozigotnih genotipova po lokusima i kategorijama



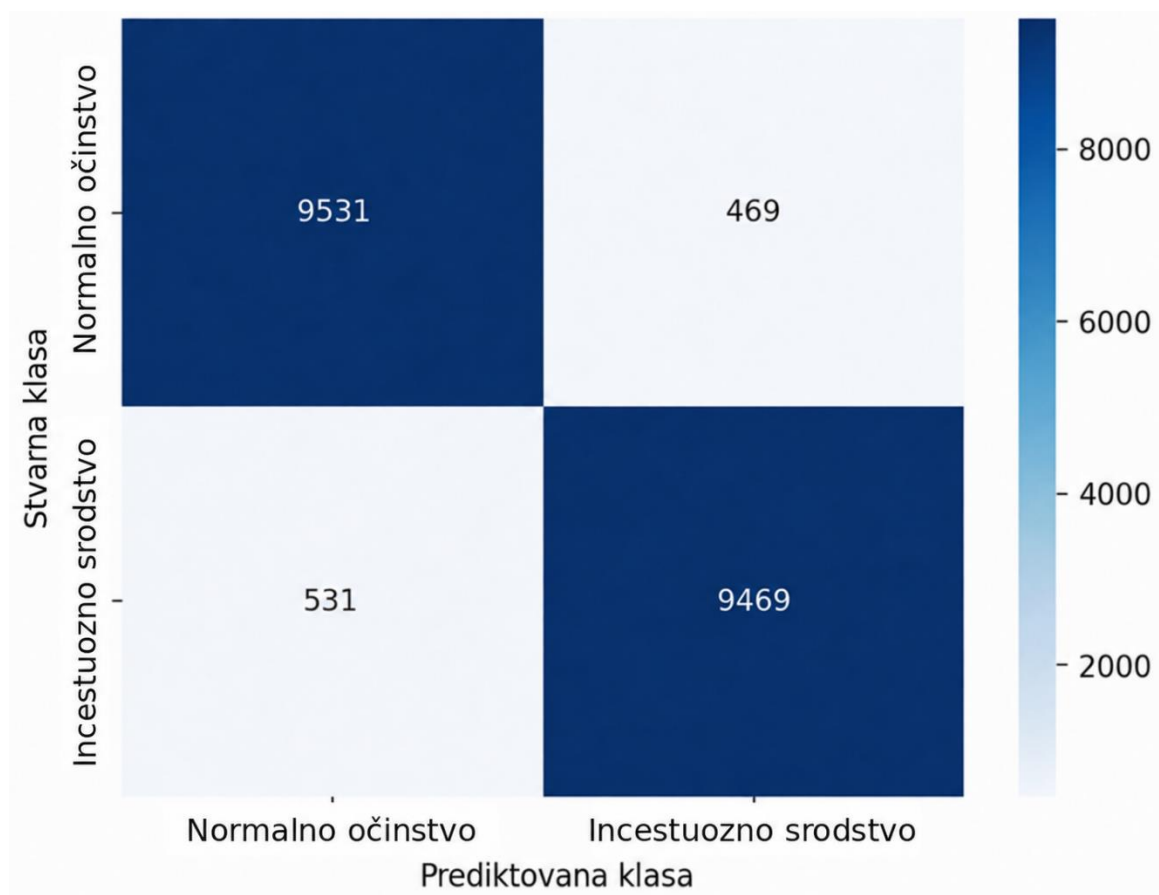
Slika 13. Procentat identičnih heterozigotnih genotipova majke i deteta po lokusima i kategorijama

Rezultati su takođe pokazali da je genetičke indikatore konsangviniteta potrebno sagledati zajedno, a ne tumačiti njihov pojedinačni doprinos kao samostalan pokazatelj incestuoznog odnosa.

3.2.5. Model za opštu populaciju (varijanta sa 21 lokusom)

Analizom 21 lokusa u opštoj populaciji model je postigao ukupnu tačnost od 95,00 %, preciznost od 95,28 %, odziv od 94,69 % i F1 skor od 94,98 %. *ROC AUC* iznosio je 99,01 %, dok su *MCC* i Cohenov kappa koeficijent iznosili približno 0,90.

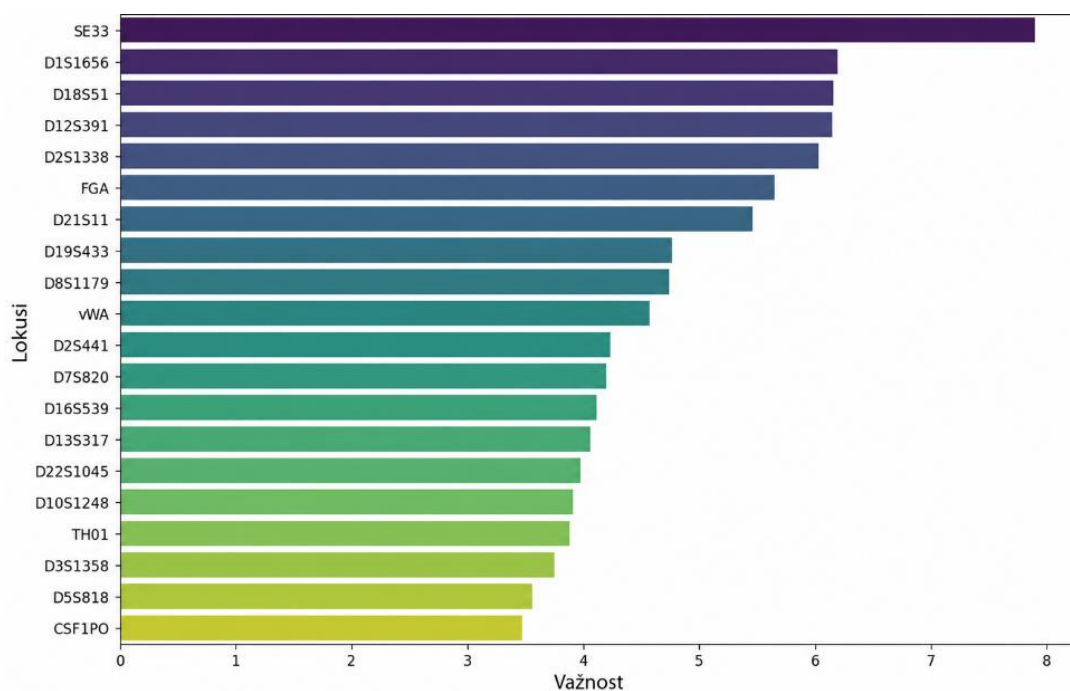
Matrica konfuzije prikazana je na Slici 14. Zapaža se da je analizom manjeg broja markera model načinio više pogrešnih klasifikacija unutar obe kategorije profila. Model je dao 9531 tačnu klasifikaciju u kategoriji normalnog očinstva i 9469 tačnih klasifikacija u kategoriji incestnog srodstva. S druge strane, model je pogrešno klasifikovao 469 profila kao slučajeve incestuoznog srodstva i 531 profil kao slučajeve normalnog srodstva.



Slika 14. Matrica konfuzije modela za opštu populaciju analizom 21 lokusa.

3.2.6. Ocena informativnosti pojedinačnih lokusa za detekciju incesta

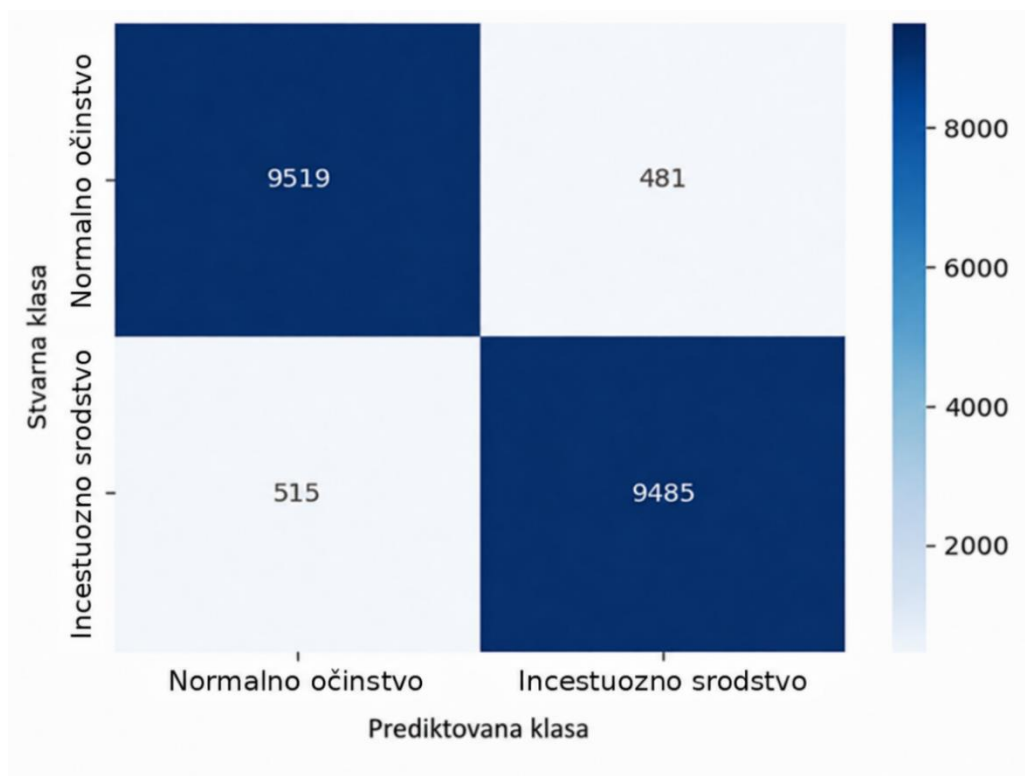
Uvidom u predstavljeni grafikon na Slici 15 zapaža se da su ispitivanjem informativnosti 21 lokusa kao najinformativniji ocenjeni *SE33*, *DIS1656* i *D18S51*, dok su se *TH01*, *D3S1358*, *D5S818*, *CSF1PO* i *TPOX* (nije prikazan na grafikonu) pokazali kao najmanje informativni za predikciju incesta.



Slika 15. 20 najvažnijih lokusa za detekciju incesta.

3.2.7. Model sa jednostepenom mutacijom za opštu populaciju (21 lokus)

Model je pri predikciji incesta na osnovu implementirane jednostepene mutacije u profilima dece postigao tačnost od 95,02 %. Preciznost modela iznosila je 95,17 %, sa odzivom od 94,85 % i F1 skorom od 95,01 %. *ROC AUC* iznosio je 99,01 %, dok je uravnotežena tačnost iznosila 95,02 %. *MCC* i Cohenov kappa koeficijent iznosili su 0,9004. Upoređivanjem sa modelom za opštu populaciju iz tačke 3.2.5, zapaža se da je implementacija jednostepene mutacije imala marginalan uticaj na predikciju. Matrica konfuzije prikazana je na Slici 16.

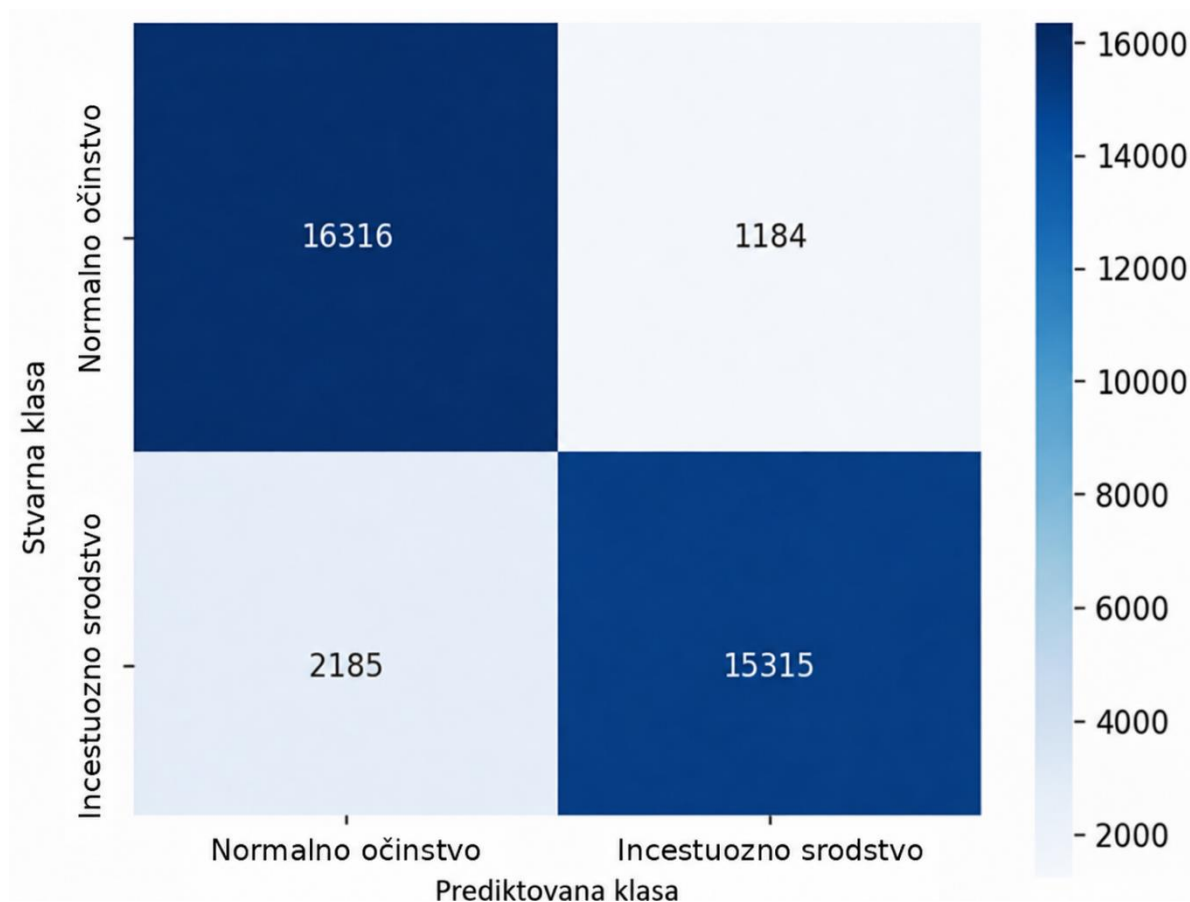


Slika 16. Matrica konfuzije za opštu populaciju sa jednostepenom mutacijom

Model je postigao 9519 tačnih klasifikacija iz kategorije normalnog očinstva i 9485 u kategoriji incestuoznog srodstva. S druge strane, 481 kombinacija profila majka–dete koju je model prepoznao kao incestuozno srodstvo i 515 profila koji su prepoznati kao pripadajući kategoriji normalnog očinstva su pogrešno klasifikovani.

3.2.8. Performanse modela na osnovu jedinstvenih profila dece (21 lokus)

Model za detekciju incesta na osnovu jedinstvenih profila dece analizom 21 lokusa postigao je tačnost od 90,37 %. Matrica konfuzije prikazana je na Slici 17. Preciznost modela iznosila je 92,82 %, uz odziv od 87,51 % i F1 skor od 90,09 %. *ROC AUC* iznosio je 96,34 %, dok je uravnotežena tačnost iznosila 90,37 %. *MCC* je iznosio 0,8088, dok je Cohenov kappa koeficijent iznosio 0,8075.



Slika 17. Matrica konfuzije za model predikcije na osnovu jedinstvenih profila

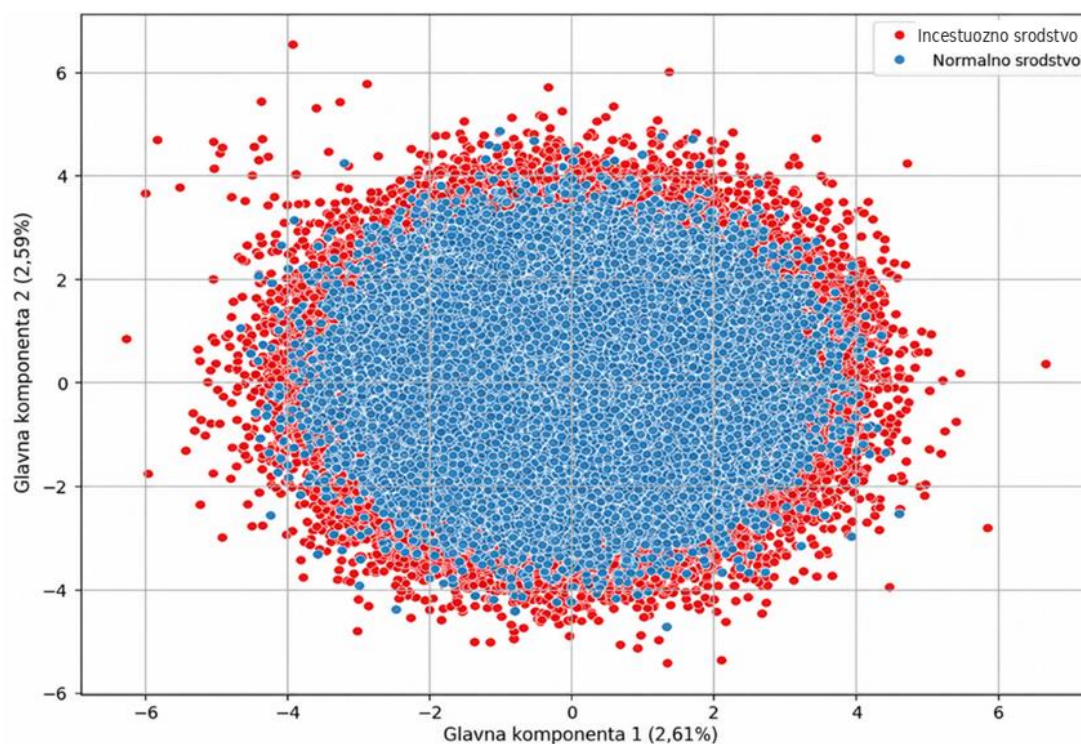
Zapaža se da je model tačno klasifikovao 16316 profila iz kategorije normalnog očinstva i 15315 iz kategorije incestuoznog srodstva. S druge strane, 1184 profila prepoznata su kao pripadajuća kategorija incestuoznog srodstva, a 2185 profila kao pripadajuća kategoriji normalnog očinstva, ali su pogrešno klasifikovani.

3.2.9. PCA analiza

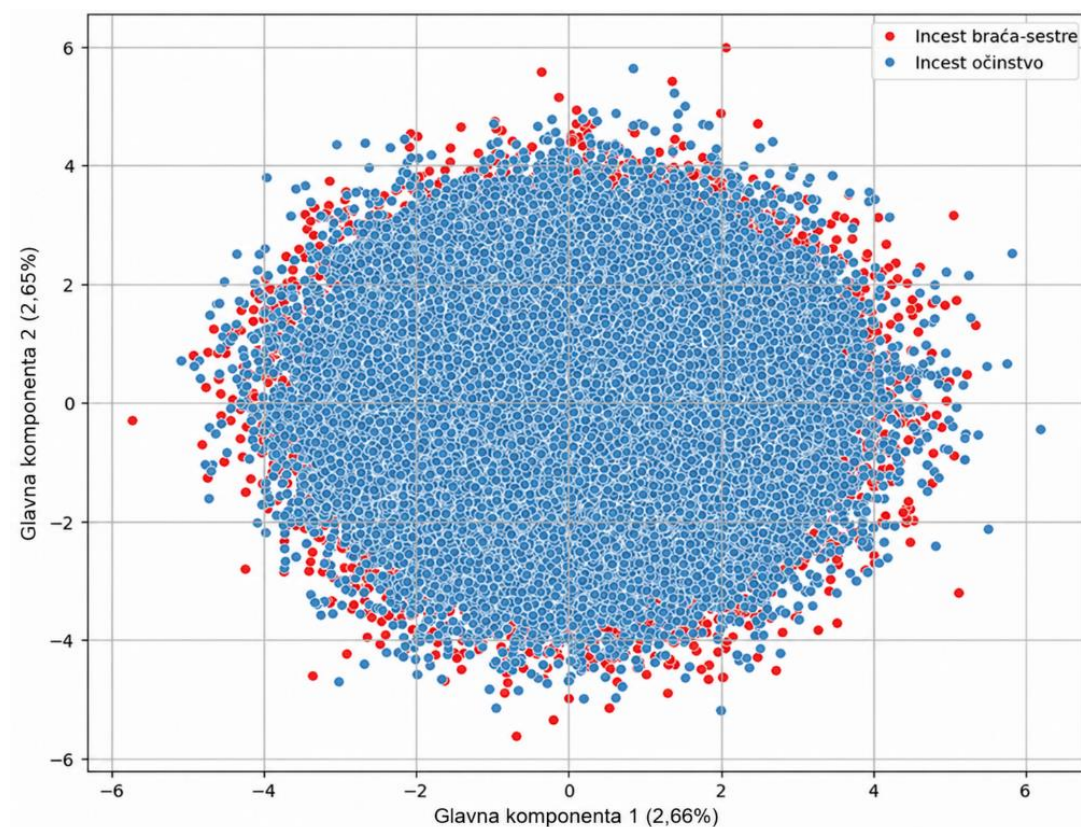
Korišćenjem *PCA* analize cilj je bio da se ispituju razlike između normalnog i incestuoznog srodstva (Slika 18), kao i razlike između različitih kategorija incesta (Slika 19).

Na Slici 18 predstavljen je dijagram disperzije. Profili incestuoznog srodstva prikazani su crvenom bojom, dok su profili usled normalnog srodstva predstavljeni plavom bojom. Na dijagramu se uočava delimična razlika u prostornoj raspodeli profila normalnog i incestuoznog srodstva, ali i njihovo međusobno preklapanje. Prve dve komponente objašnjavaju 2,61 % i 2,59 % varijabilnosti u

podacima, odnosno ukupno 5,20 %. Uočeni obrazac ukazuje da analizirani *STR* lokusi sadrže određenu informaciju relevantnu za razlikovanje ove dve kategorije, ali *PCA* analiza sama po sebi ne pokazuje njihovo potpuno razdvajanje.



Slika 18. *PCA* analiza normalnog i incestuoznog ukrštanja



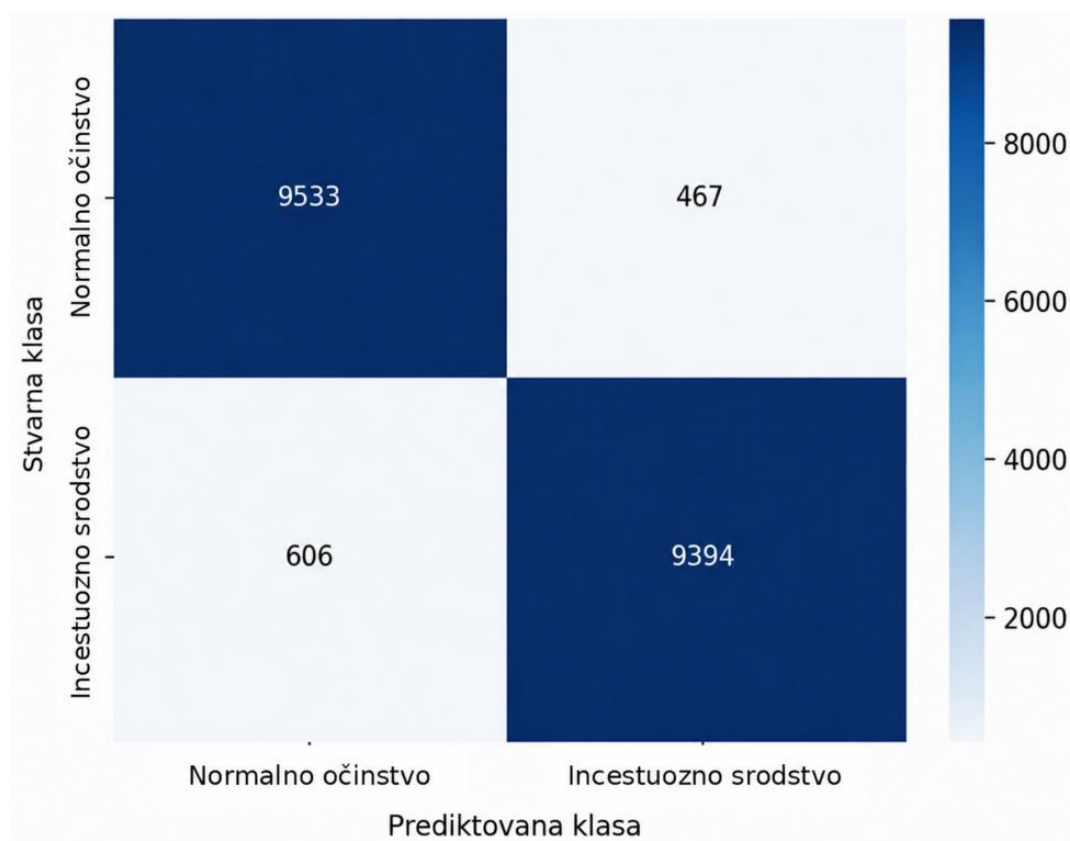
Slika 19. *PCA* analiza ispitanih kategorija incesta

Dijagram na slici 19 prikazuje distribuciju *STR* profila za inceste između braće i sestara (crveno) i očinstvo usled ukrštanja roditelj–dete (plavo) duž prve dve glavne komponente. Slično prethodnom grafikonu disperzije, *PC1* objašnjava 2,66 % i prikazan je na *X*–osi, dok *PC2* objašnjava 2,65 % varijanse i predstavljen je na *Y*–osi.

Grafikon pokazuje da su slučajevi incesta između braće i sestara raspršeni više po periferiji, dok profili incestuoznog očinstva pretežno formiraju gusto centralno jezgro. Ovaj obrazac sugerise blagu, ali primetnu diferencijaciju između dve incestne kategorije, pri čemu incest između braće i sestara pokazuje veću varijabilnost u genetičkim podacima.

3.2.10. Početni model za Saudijsku Arabiju

Početni model, treniran i testiran na skupu podataka Saudijske Arabije analizom 21 lokusa, postigao je tačnost od 94,64 %. Preciznost, odnosno *PPV*, iznosila je 95,26 %, odziv 93,94 %, *F1* skor 94,60 %, a *ROC AUC* 98,85 %. Uravnotežena tačnost iznosila je 94,64 %, dok je *NPV* iznosio 94,02 %. Matrica konfuzije prikazana je na Slici 20.



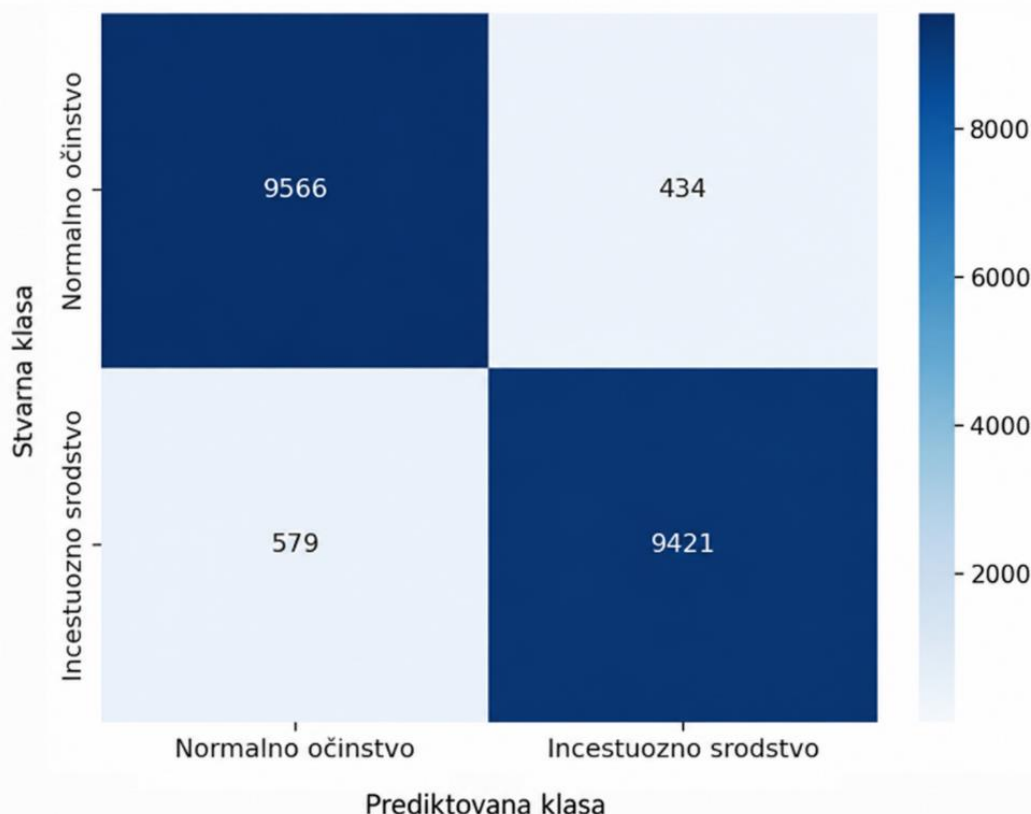
Slika 20. Matrica konfuzije za predikciju incesta u Saudijskoj Arabiji

Ovaj model je pravilno svrstao 9533 profila u kategoriju normalnog očinstva i 9394 profila u kategoriju incestuoznog srodstva. Model je pogrešno klasifikovao 467 profila koji pripadaju kategoriji normalnog očinstva kao incestuozno srodstvo i 606 profila koji pripadaju kategoriji incestuoznog srodstva kao normalno očinstvo.

3.2.11. Modifikovani model za Saudijsku Arabiju

Drugi model, treniran i testiran na modifikovanoj verziji baze podataka za Saudijsku Arabiju, postigao je tačnost od 94,93 %. Preciznost, odnosno *PPV*, iznosila je 95,60 %, odziv 94,21 %, *F1* skor 94,90 %, a *ROC AUC* 98,94 %. Uravnotežena tačnost iznosila je 94,93 %, dok je *NPV* iznosio 94,29 %.

Matrica konfuzije prikazana je na Slici 21. Model je pravilno svrstao 9566 profila u kategoriju normalnog očinstva i 9421 profil u kategoriju incestuoznog srodstva. Model je pogrešno klasifikovao 434 profila iz kategorije normalnog očinstva kao incestuozno srodstvo i 579 profila iz kategorije incestuoznog srodstva kao normalno očinstvo.



Slika 21. Matrica konfuzije za predikciju incesta na modifikovanim učestalostima Saudijske Arabije

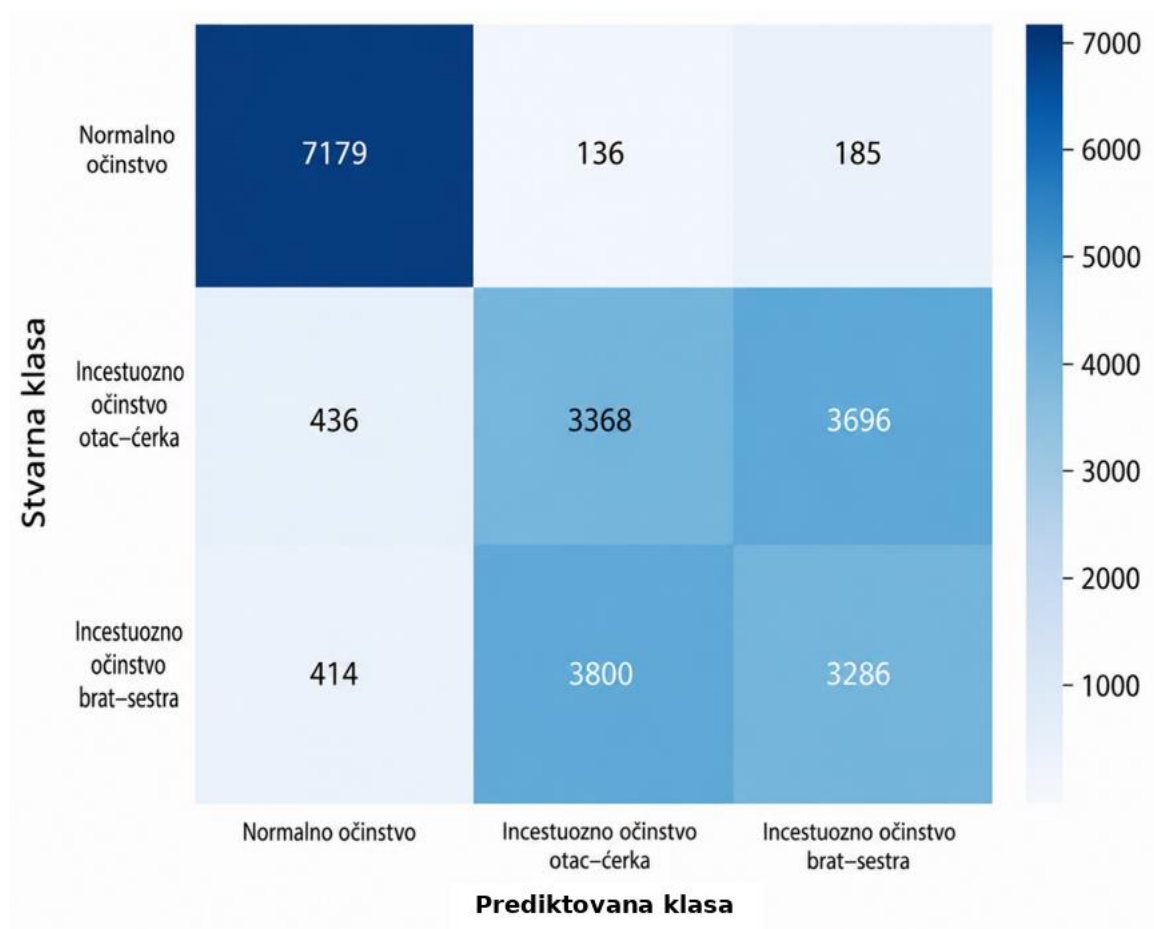
3.2.12. Višeklasna klasifikacija

S obzirom na složenost genetičkih odnosa u slučajevima srodstva, cilj ove analize nije bio samo klasifikacija između normalnog očinstva i incestuoznih veza, već i određivanje specifične vrste incestuoznog odnosa. Na taj način, model je razvijen kao višeklasni klasifikacioni sistem koji može diferencirati tri kategorije: normalno očinstvo, incest kada je otac majke otac deteta i incest kada je brat majke otac deteta.

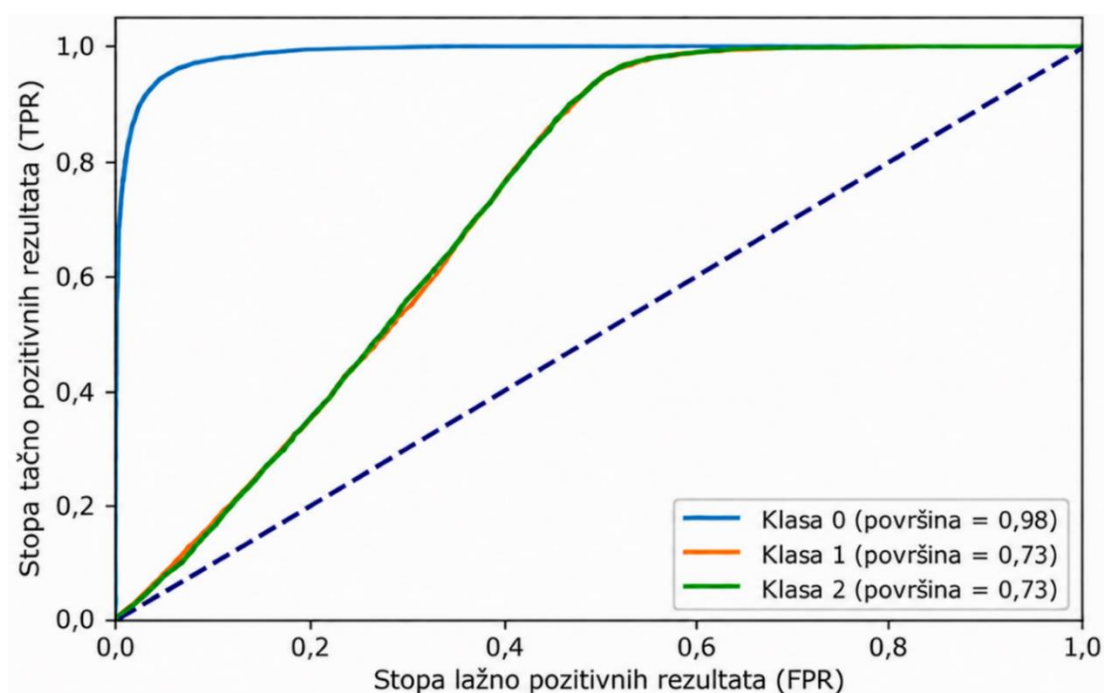
Matrica konfuzije (Slika 22) prikazuje ukupne rezultate klasifikacije za svaku od kategorija. Model je uspešno klasifikovao slučajeve normalnog očinstva 7179 puta. Međutim, uočava se određena konfuzija između dve kategorije incesta, gde je model izvršio ukupno 7496 pogrešnih klasifikacija. Ovo ukazuje na preklapanje karakteristika između ovih kategorija i nemogućnost pouzdanije klasifikacije na osnovu testiranog broja lokusa.

ROC kriva na Slici 23 prikazuje odnos između stope tačno pozitivnih i stope lažno pozitivnih rezultata za svaku kategoriju. *ROC AUC* iznosio je 98,00 % za kategoriju normalnog očinstva, dok je

za obe kategorije incesta iznosio 73,00 %. Ovi rezultati pokazuju da je model znatno bolje razdvajao klasu normalnog očinstva od preostalih klasa nego pojedinačne klase incesta od ostalih klasa.



Slika 22. Matrica konfuzije za višeklasnu predikciju incesta



Slika 23. ROC krive višeklasnog modela za predikciju incesta. Klasa 0 (normalno očinstvo), Klasa 1 (Incest kada je otac majke otac deteta) i Klasa 2 (Incest kada je brat otac deteta).

IV DISKUSIJA

4.1. Predikcija porekla

4.1.1. Tačnost prediktivnog modela pri predikciji porekla

U okviru sprovedene studije izvršena je diferencijacija četiri referentne populacione grupe: afroameričke, SA–EP, azijske i hispanoameričke, primenom autozomnih *STR* profila zasnovanih na učestalostima alela 29 lokusa. Osnovni model, koji je obuhvatao sve četiri populacione grupe, postigao je tačnost od 94,24 % (poglavlje 3.1.3). Isključenje hispanoameričke populacije iz analize dovelo je do povećanja tačnosti na 99,06 % (poglavlje 3.1.5), dok je u modelu u kome su hispanoamerička i SA–EP populacija objedinjene postignuta tačnost od 98,57 % (poglavlje 3.1.6). Rezultati pokazuju da je tačnost modela uslovljena definisanim klasifikacionim scenarijem. Rezultati za pojedinačne klase iz Tabele 3 (poglavlje 3.1.3) pokazuju da je najviša uravnotežena tačnost zabeležena za azijsku populaciju (99,19 %), dok je najniža vrednost dobijena za hispanoameričku populaciju (92,70 %).

Dodatna procena diskriminatorne sposobnosti modela sprovedena je analizom *ROC* krivih (Slika 3, poglavlje 3.1.3). Vrednosti *ROC AUC* od približno 100,00 % za afroameričku i azijsku populaciju ukazuju na dobru separaciju ovih grupa, dok niže vrednosti za SA–EP (98,00 %) i hispanoameričku populaciju (95,00 %) ukazuju na veći stepen preklapanja. Ovaj obrazac potvrđuje i analiza matrice konfuzije (Slika 2, poglavlje 3.1.3), gde je između afroameričke i azijske populacije zabeleženo samo šest pogrešnih klasifikacija od 9000.

Visoka tačnost u diferencijaciji afroameričke i azijske populacije može se objasniti time što ove grupe pripadaju genetički razdvojenim populacionim klasterima, što je potvrđeno u prethodnim studijama populacione strukture (Rosenberg i sar., 2002; The 1000 Genomes Project Consortium, 2015). Sa ovim je saglasna i *PCA* analiza (Slika 7, poglavlje 3.1.7), u kojoj su afroamerički profili formirali klaster na desnoj strani grafikona, dok su se azijski profili grupisali u levom uglu. Njihovo razdvajanje u *PCA* prostoru u skladu je sa visokom uspešnošću klasifikacije između ove dve populacije.

Visoke vrednosti *ROC AUC* ne moraju biti praćene visokom tačnošću klasifikacije za svaku pojedinačnu populaciju. Hispanoamerička grupa pokazuje da relativno visoka *ROC AUC* vrednost može koegzistirati sa većim brojem pogrešnih klasifikacija, zbog čega se uspešnost modela mora procenjivati kombinovanjem više pokazatelja, uključujući matricu konfuzije i tačnost po klasama, a ne samo analizom *ROC* krive.

Model je pri diferencijaciji hispanoameričke i SA–EP populacije dao 1511 međusobno pogrešno klasifikovanih profila. Ovakav rezultat se može objasniti kompleksnom populacionom strukturom hispanoameričke populacije usled istorijskih admiksija (Bryc i sar., 2015; Sofer i sar., 2017). Ovo je u skladu sa nalazima da SA–EP i hispanoamerička populacija pokazuju najniži nivo međusobne genetičke divergencije među analiziranim grupama ($F_{ST} = 0,01$), uz delimično preklapanje u analizama klasterovanja, što otežava njihovo razdvajanje primenom standardnih *STR* markera (Oldt i Kanthaswamy, 2020). Preklapanje je uočeno i *PCA* analizom u ovom istraživanju (Slika 7, poglavlje 3.1.7), gde su se SA–EP i hispanoamerički profili grupisali u centralnom delu dijagrama, bez jasnog međusobnog razdvajanja.

Poboljšanje performansi modela nakon isključenja ili grupisanja hispanoameričke populacije pokazuje da stepen preklapanja učestalosti alela između populacija utiče na tačnost klasifikacije. Ipak, takav pristup, iako metodološki opravdan, pojednostavljuje realne forenzičke scenarije i može dovesti do precenjivanja performansi modela u kontrolisanim uslovima.

Dobijeni rezultati ukazuju da predikcija biogeografskog porekla zasnovana na autozomnim *STR* lokusima daje najpouzdanije rezultate pri diferencijaciji genetički udaljenih populacionih grupa, dok njena uspešnost opada kod populacija sa zajedničkom demografskom istorijom ili izraženom admiksijom.

4.1.2. Uticaj veličine skupa podataka

Kako bi se ispitao uticaj veličine skupa za treniranje na tačnost modela, korišćene su baze podataka različitih veličina (Slika 8, poglavlje 3.1.8). Ukupna tačnost modela rasla je sa povećanjem broja profila po populacionoj grupi, sa 83 % pri 500 profila na 94 % pri 90000 profila. Najizraženije ograničenje uočeno je pri klasifikaciji hispanoameričke populacije, za koju je tačnost pri najmanjem skupu iznosila svega 64 %, dok su afroamerička i azijska populacija i pri 500 profila klasifikovane sa tačnošću većom od 90 % (92 %, odnosno 94 %).

Porast tačnosti nije bio jednako izražen za sve populacije. Između analiza sa 500 i 90000 profila po populacionoj grupi, tačnost za afroameričku grupu povećala se za 6 procentnih poena, za azijsku grupu za 5 procentnih poena, a za SA–EP grupu za 9 procentnih poena, dok je za hispanoameričku grupu zabeleženo povećanje od 26 procentnih poena. Potreba za većim trening skupom stoga je direktno povezana sa genetičkom složenošću populacije: u populacijama sa izraženom admiksijom, kao što je hispanoamerička, veći broj profila omogućava modelu precizniju procenu raspodele alela i genotipova unutar grupe, čime se smanjuje rizik od učenja slučajnih obrazaca na malom uzorku i njihove ekstrapolacije na nove podatke.

Povećanje broja profila iznad 10000 po populaciji dovodi do minimalnog poboljšanja performansi modela, pri čemu je razlika u tačnosti između modela treniranih na 10000 i 90000 profila iznosila oko jednog procentnog poena. Time je tačka zasićenja u ovoj studiji određena na 10000 profila po populacionoj grupi (poglavlje 3.1.8). Ovakav obrazac je u skladu sa nalazima da performanse klasifikacionih modela rastu sa povećanjem veličine uzorka do nivoa stabilizacije, nakon čega dalji rast donosi samo marginalna poboljšanja, pri čemu tačka zasićenja zavisi i od karakteristika samog skupa podataka (Hess i Wei, 2010; Silvey i Liu, 2024).

Ovaj nalaz određuje najmanju veličinu trening skupa pri kojoj model postiže stabilne performanse i pokazuje da dalje uvećanje baze donosi nesrazmerno mali dobitak u odnosu na potrebno vreme obučavanja i računarske resurse. Veličinu skupa treba posmatrati u odnosu na genetičku strukturu populacija, budući da je njen doprinos izraženiji kod genetički heterogenijih i slabije razdvojenih grupa.

4.1.3. Broj testiranih lokusa i njihova informativnost

Rezultati pokazuju jasnu povezanost između broja analiziranih *STR* lokusa i tačnosti modela u predikciji biogeografskog porekla. Najveća tačnost postignuta je analizom 29 lokusa, dok je smanjenje broja lokusa dovelo do postepenog opadanja tačnosti (Tabela 6, poglavlje 3.1.9). Konkretno, analizom 21 lokusa iz komercijalnih setova *Investigator 24plex/GlobalFiler* postignuta je tačnost od 88,28 % za klasifikaciju četiri populacije, odnosno 96,30 % nakon isključenja hispanoameričke populacije. Najniže vrednosti zabeležene su za panele sa 15 i 16 lokusa, što potvrđuje da manji broj lokusa ograničava diskriminatornu moć modela.

Veći broj *STR* lokusa povećava informacioni kapacitet genetičkog profila, jer veći broj lokusa omogućava uočavanje veće genetičke varijabilnosti i bolju rezoluciju između populacija, što može doprineti preciznijoj diferencijaciji populacionih grupa (Dawnay i sar., 2018). Međutim, broj lokusa sam po sebi nije dovoljan uslov za unapređenje modela. Poređenje panela sa sličnim brojem lokusa pokazalo je da *PowerPlex Fusion System*, uprkos većem broju lokusa (22), nije postigao veću tačnost u odnosu na panele sa 21 lokusom. To ukazuje da doprinos dodatnog lokusa zavisi od njegove informativnosti, odnosno od sposobnosti da razdvoji populacije kroz specifične obrasce učestalosti

alela. Ukoliko dodatni lokus ne pokazuje dovoljnu informativnost za razdvajanje populacija, bilo zbog sličnih učestalosti alela među populacijama ili zbog redundantnosti u odnosu na već uključene lokuse, njegovo uključivanje neće značajno unaprediti performanse modela (de la Puente i sar., 2021; Ruiz–Ramírez i sar., 2023).

Analiza pojedinačnih doprinosa lokusa u ovoj doktorskoj tezi potvrđuje da doprinos pojedinačnog lokusa zavisi od njegove informativnosti, a ne samo od ukupnog broja analiziranih lokusa. Kao najinformativniji lokusi izdvojeni su *Penta E*, *F13A01* i *SE33*, dok su *D10S1248*, *D3S1358*, *LPL*, *TPOX* i *CSFIPO* pokazali najniži doprinos modelu. Lokus *SE33* prema ranijim istraživanjima pokazuje visok stepen polimorfizma i diskriminacione moći u odnosu na druge forenzičke *STR* lokuse, što doprinosi njegovoj većoj informativnosti u modelu (Butler i sar., 2009; Bhinder i sar., 2018).

Dodatna analiza, u kojoj su iz modela isključeni najmanje informativni lokusi, pokazala je da je takva modifikacija dovela do smanjenja ukupne tačnosti modela za približno 1 %. Ovaj nalaz ukazuje da i lokusi sa nižom individualnom informativnošću daju doprinos, odnosno da stabilnost modela ne zavisi isključivo od nekoliko visoko informativnih lokusa, već i od njihovog zajedničkog delovanja sa većim brojem drugih lokusa. Takvo tumačenje je u skladu sa zapažanjem da više varijabli sa slabijim pojedinačnim prediktivnim efektom može zajednički doprineti uspešnoj klasifikaciji (Chen i Ishwaran, 2012).

Ovi nalazi su u skladu sa istraživanjima koja pokazuju da standardni forenzički *STR* paneli omogućavaju razlikovanje velikih kontinentalnih populacionih grupa, ali da njihova preciznost ostaje ograničena kada se klasifikacija sprovodi na subkontinentalnom nivou, između geografski i genetički bliskih populacija, kao i kod populacija sa izraženom admiksijom, kod kojih se profili više referentnih grupa međusobno preklapaju (Phillips, 2015; Algee–Hewitt i sar., 2016). Povećanje broja lokusa ili primena markera posebno informativnih za poreklo može dodatno unaprediti performanse modela (Rosenberg i sar., 2002; Phillips, 2015).

4.1.4. Praktična primena predikcije porekla u forenzičkim istragama

Primena predikcije biogeografskog porekla u forenzičkim istragama ilustrovana je u studiji autora Phillips i sar. (2009), koja se odnosi na teroristički napad u Madridu 2004. godine. U tom slučaju standardna *STR* analiza nije omogućila identifikaciju osumnjičenog, dok je primena *AIM–SNP* markera ukazala na severnoafričko poreklo, što je naknadno potvrđeno istragom. Ovaj primer pokazuje da genetička procena porekla može usmeriti istragu kada klasične metode identifikacije ne daju rezultate.

Procene porekla najčešće se zasnivaju na markerima posebno dizajniranim za analizu porekla, poput *AIM–SNP* panela, usled bolje diskriminacione moći u odnosu na standardne forenzičke markere (Phillips i sar., 2013). Nedostatak uniparentalnih markera može se ilustrovati nalazima studije u kojoj je u okviru jedne lokalne engleske genealogije, povezane sa retkim jorkširskim prezimenom, identifikovana Y–hromozomska haplogrupa A1 karakteristična za afričke populacije (King i sar., 2007). Ovo ukazuje da bi oslanjanje isključivo na Y–hromozomske markere moglo dovesti do pogrešne predstave o ukupnom poreklu osobe i, posledično, uticalo na dalji tok istrage. Razlog tome je što uniparentalni markeri odražavaju samo jednu naslednu liniju, dok autozomni markeri omogućavaju procenu doprinosa svih predaka i pružaju kompletniju sliku biogeografskog porekla (Halder i sar., 2008).

Ipak, predikcija porekla zasnovana na autozomnim *STR* markerima ima svoja ograničenja. U poređenju sa panelima specifično dizajniranim za analizu porekla, ona ima nižu preciznost, naročito kod genetički sličnih ili admiksovanih populacija (Halder i sar., 2008; Phillips i sar., 2013). Zbog

toga se takvi rezultati ne mogu tumačiti kao direktni dokazi o populacionom identitetu, već kao indikativne procene sa potencijalnom istražnom vrednošću.

Predikcija porekla zasnovana na *STR* markerima može se primeniti u forenzičkim istragama, posebno kada identitet osumnjičenog nije poznat i kada pretraga DNK baza ne daje rezultate. U takvim okolnostima, preliminarna procena porekla može doprineti sužavanju kruga potencijalnih osumnjičenih i efikasnijem usmeravanju daljeg toka istrage, ali se takvi nalazi moraju tumačiti kao indikativne informacije, a ne kao samostalni dokazni osnov (Dawnay i sar., 2018; Devesse i sar., 2023).

Nalazi ovog istraživanja pokazuju da se postojeći *STR* profili mogu koristiti za dodatne analitičke procene bez potrebe za dodatnim analizama. Prediktivni modeli mogu se primeniti na već dostupne *STR* podatke uz ograničene dodatne resurse, što predstavlja praktičnu prednost ovog pristupa.

4.1.5. Poređenje sa prethodnim studijama

Dosadašnja istraživanja predikcije biogeografskog porekla zasnivala su se na različitim metodološkim pristupima, uključujući analize ograničenog broja *STR* lokusa, Bajesove modele i specijalizovane panele markera informativnih za poreklo.

Jedna od ranijih studija u ovoj oblasti pokazala je da je diskriminativna moć standardnih *STR* profila bila neujednačena između analiziranih grupa: bolji rezultati postignuti su za genetički diferencirane populacije kao što su SA–EP i afro–karijska, dok je uspešnost bila znatno niža za slabije razdvojene grupe, naročito bliskoistočne. To potvrđuje da standardni *STR* lokusi mogu nositi koristan signal za procenu porekla, ali da njegova vrednost ostaje prvenstveno istražna (Lowe i sar., 2001).

Pritchard i sar. (2000) razvili su Bajesov pristup zasnovan na modelovanju populacionog grupisanja, u kome se individue probabilistički svrstavaju u jednu ili više populacija, uz mogućnost procene admiksije. Metod je ilustrovan na simuliranim i stvarnim podacima, uključujući profile Afrikanaca i Evropljana, dok je u simuliranom scenariju mešanja dve populacije pokazano da procena postaje složenija kada je prisutna admiksija ili finija podstruktura populacija.

Graydon i sar. (2009) pokazali su da je diferencijacija genetički udaljenih populacija uspešnija nego razdvajanje srodnijih grupa. U njihovom istraživanju, koje je obuhvatilo Han Kineze, Japance, Korejce, američke belce i indijsku autohtonu grupu Lodha, tačnost je iznosila oko 90 % pri razlikovanju jasno razdvojenih grupa, dok je pri međusobnom razlikovanju istočnoazijskih populacija iznosila približno 70 %. Ovakvi nalazi su u skladu sa rezultatima ove studije, u kojoj je najveći izazov predstavljala diferencijacija SA–EP i hispanoameričke populacije.

Alat PopAffiliator omogućio je klasifikaciju profila u tri široke populacione grupe na osnovu 17 autozomnih *STR* lokusa, uz tačnost od približno 86 %. Ovakav nalaz pokazuje da standardni forenzički *STR* paneli mogu imati istražnu vrednost u gruboj proceni porekla, ali da je njihova primena prvenstveno ograničena na razlikovanje širih, genetički udaljenijih populacionih kategorija (Pereira i sar., 2011). Ograničenja alata PopAffiliator odnosila su se na relativno mali broj lokusa i široko definisane populacione kategorije, što smanjuje preciznost kod mešanja između populacija i kod genetički sličnih populacija.

Razvoj panela markera informativnih za poreklo predstavlja drugi pristup predikciji. Phillips i sar. (2013) razvili su *AIM–STR* panel od 12 lokusa i prilagodili Bajesov klasifikator Snipper za analizu *STR* lokusa. Pri primeni samog *AIM–STR* panela uspešnost klasifikacije razlikovala se između populacionih grupa i kretala se od 51,86 % za Afrikance do 96,82 % za Evropljane. Kombinovanjem

12 *AIM-STR*, 20 identifikacionih *STR* i 34 *AIM-SNP* markera postignuta je uspešnost klasifikacije od 100 % u svih pet analiziranih grupa. Ovi nalazi pokazali su da kombinovanje više markera može unaprediti predikciju porekla, ali i da uspešnost pojedinačnih panela zavisi od analizirane populacione grupe.

Algee-Hewitt i sar. (2016) analizirali su potencijal *CODIS* markera za predikciju biogeografskog porekla na uzorku od 978 osoba iz 53 populacije širom sveta. Na osnovu *STRUCTURE* analiza sa 779 ne-*CODIS* lokusa definisana su četiri velika super-regiona: Afrika, Zapadna Evroazija, Istočna Azija/Pacifik i Amerika. Najveća uspešnost klasifikacije na osnovu 13 *CODIS* lokusa postignuta je za Afriku (91 %) i Ameriku (89 %), dok je bila niža za Zapadnu Evroaziju (52 %) i Istočnu Aziju/Pacifik (64 %). Ovi nalazi pokazuju da standardni forenzički *STR* markeri mogu nositi koristan signal za razlikovanje širokih populacionih grupa, ali i da njihova uspešnost zavisi od stepena genetičke razdvojenosti između analiziranih populacija.

Jedna od novijih studija autora Devesse i sar. (2023) ispitala je primenu *MPS* tehnologije za analizu 26 autozomnih *STR* lokusa na uzorku od 989 osoba iz pet populacionih grupa. U toj studiji, pri konzervativnom kriterijumu procene, pravilno je klasifikovano 84 % uzoraka, što pokazuje da ni savremenije tehnološke platforme ne eliminišu teškoće u diferencijaciji složenijih populacionih obrazaca.

U ovom istraživanju analizirano je do 29 standardnih autozomnih *STR* lokusa za četiri referentne populacione grupe, uz primenu algoritama mašinskog učenja. Ranije studije najčešće su se zasnivale na manjem broju standardnih *STR* lokusa, široko definisanim populacionim kategorijama ili klasičnim populaciono-genetičkim modelima. Pored toga, sistematski su ispitani uticaji veličine skupa za obučavanje, broja markera i strukture populacija na uspešnost modela.

Visoke vrednosti tačnosti postignute su korišćenjem standardnih forenzičkih *STR* panela, bez potrebe za dodatnom genotipizacijom namenski odabranih markera informativnih za poreklo. Pri analizi 21 lokusa iz komercijalnih setova *GlobalFiler* i *Investigator 24plex* postignuta je tačnost od 88,28 % pri klasifikaciji četiri populacije, odnosno 96,30 % nakon isključenja hispanoameričke populacije. Ovakvi rezultati pokazuju da postojeći forenzički *STR* podaci mogu imati prediktivnu vrednost za procenu porekla kada se analiziraju računarskim postupcima.

Poređenje sa prethodnim studijama potvrđuje ranije nalaze o ograničenjima standardnih *STR* markera kod genetički srodnijih i admiksovanih populacija. Rezultati istovremeno pokazuju da se njihov analitički potencijal može proširiti primenom mašinskog učenja i sistematskim unapređenjem modela. Standardni *STR* paneli, iako nisu primarno razvijeni za predikciju biogeografskog porekla, mogu se koristiti kao izvor informacija u forenzičkom kontekstu.

4.1.6. Prednosti i metodološka vrednost sintetičkih *STR* profila

Sintetički profili ne predstavljaju zamenu za realne forenzičke podatke, već metodološki alat koji omogućava razvoj i internu evaluaciju modela u kontrolisanim uslovima (Gonzales i sar., 2023). Omogućavaju testiranje scenarija koji su u realnoj praksi retki, etički osetljivi ili teško dostupni, pre dalje validacije na stvarnim podacima (Mendes i sar., 2025). U ovom istraživanju sintetički *STR* profili omogućili su sistematsko ispitivanje performansi modela, smanjenje pristrasnosti u trening podacima i proširenje analize na scenarije koji nisu dostupni u realnim forenzičkim bazama.

Primena sintetičkih profila u ovom istraživanju omogućila je formiranje uravnoteženih skupova podataka za sve ispitivane populacije. Neujednačene veličine grupa mogu uticati na rezultate populacione analize i uspešnost prediktivnih modela, otežavajući razdvajanje efekata genetičke strukture populacija od efekata neuravnoteženosti podataka (Onogi i sar., 2011; Fortier i sar., 2020;

Koo, 2025). Pristup primenjen u ovoj disertaciji omogućio je da se uticaji veličine skupa za obučavanje modela, broja lokusa i sastava populacionih grupa analiziraju odvojeno i kontrolisano.

Simulacija genetičkih ukrštanja između različitih populacija omogućila je generisanje profila koji odgovaraju različitim scenarijima admiksije. Takvi podaci omogućavaju testiranje modela u uslovima koji su teško dostupni u realnim forenzičkim bazama, a relevantni su za primenu u geografskim regionima sa složenom genetičkom strukturom i izraženim istorijskim mešanjem populacija.

Sintetički profili omogućili su razvoj i internu evaluaciju modela za detekciju incestuoznih odnosa kroz generisanje velikog broja kontrolisanih scenarija srodstva. Na taj način omogućena je sistematska procena uspešnosti modela, uključujući tačnost, osetljivost i specifičnost.

4.1.7. Ograničenja i izazovi primene sintetičkih *STR* profila

Iako primena sintetičkih *STR* profila omogućava kontrolisano ispitivanje performansi modela, ovaj pristup ima metodološka ograničenja koja treba uzeti u obzir pri interpretaciji rezultata.

Jedno od osnovnih ograničenja ove studije odnosi se na pouzdanost ulaznih alelnih učestalosti i reprezentativnost populacionih baza na osnovu kojih su generisani sintetički profili (Aalbers i Gettings, 2026). Sintetički podaci ne mogu nadomestiti nedostatke izvornih podataka, već mogu reprodukovati njihove pristrasnosti i ograničenja (Koul i sar., 2025). Poseban problem predstavlja postupanje sa retkim ili prethodno nezabeleženim alelima, jer njihove procenjene učestalosti mogu znatno uticati na statističku evaluaciju srodstva (Davenport i sar., 2025), dok izbor praga minimalne učestalosti alela može uticati na rezultate procene populacione strukture u genomskim skupovima podataka (Linck i Battey, 2019).

Izbor referentnih populacija predstavlja dodatno ograničenje. Cheung i sar. (2018) navode da izbor populacija utiče na tačnost predikcije biogeografskog porekla i procenu admiksije. Ograničenje se pri tome ne odnosi na tačnost samih učestalosti alela, već na način na koji su populacione kategorije definisane. Referentne grupe su široke i delimično administrativne: grupa označena kao SA–EP obuhvata pretežno osobe evropskog porekla i genetički je relativno homogena, dok hispanoamerička grupa nije genetički jedinstvena, već predstavlja kombinaciju evropske, autohtone američke i afričke komponente u različitim udelima. Model stoga uči obrasce definisane na nivou tako postavljenih kategorija, a ne obrasce stvarne populacione strukture. Rezultati ove teze pokazuju ograničenje u slučaju hispanoameričke populacije, kod koje je zabeležena najniža tačnost klasifikacije i najveći stepen preklapanja sa SA–EP populacijom (poglavlja 3.1.3 i 3.1.7), što je detaljnije razmotreno u poglavlju 4.1.9.

Ograničenje sintetičkih profila odnosi se i na reprezentaciju složene populacione strukture. Iako generisani podaci mogu obuhvatiti širok spektar mogućih genotipova, finiji obrasci genetičke raznolikosti, poput individualne admiksije, subpopulacionih razlika i efekata ograničenih referentnih populacija, teže se prikazuju u potpunosti. Zbog toga sintetički profili mogu verno odražavati dominantne obrasce raspodele alela, ali sa manjom preciznošću obuhvatiti složenije i ređe obrasce porekla, što može uticati na performanse modela u realnim scenarijima (Halder i sar., 2008; Phillips i sar., 2013).

Validacija modela na stvarnim podacima predstavlja dodatni izazov, posebno u heterogenim populacijama. Istočnoazijske i hispanoameričke populacije karakterišu se složenom unutrašnjom strukturom i značajnim varijacijama u učestalostima alela između subpopulacija (Sofer i sar., 2017; Pan i Xu, 2020), što otežava direktno poređenje rezultata dobijenih na sintetičkim i realnim profilima. U odsustvu reprezentativnih baza stvarnih podataka, procena performansi modela ostaje ograničena na kontrolisane uslove simulacije.

Za regionalne modele dodatno ograničenje predstavljaju male razlike između populacija. Studija Novković i sar. (2010), koja je analizirala učestalosti alela 15 *STR* lokusa u populaciji Srbije u poređenju sa drugim populacijama Balkana, pokazala je da su razlike između populacija bivše Jugoslavije uglavnom male, uz statistički značajna odstupanja zabeležena samo za pojedine lokuse u odnosu na crnogorsku populaciju. Ovi nalazi ukazuju na relativnu genetičku homogenost regiona, ali i na to da male razlike u učestalostima alela mogu biti relevantne u forenzičkim analizama. Zbog toga upotreba sintetičkih profila zasnovanih na opštim ili globalnim frekvencijama može biti nedovoljno precizna za regionalne modele predikcije.

Rezultate dobijene na sintetičkim profilima treba tumačiti kao metodološki validne u okviru kontrolisanih uslova, ali samo kao posredan pokazatelj efikasnosti modela u realnim forenzičkim scenarijima. U literaturi o primeni sintetičkih podataka u zdravstvenim istraživanjima naglašava se da njihova spoljašnja validnost mora biti proverena na realnim podacima (Gonzales i sar., 2023; Mendes i sar., 2025). U kontekstu ovog istraživanja, to podrazumeva testiranje modela na stvarnim *STR* profilima koji odražavaju kompleksnost populacione strukture i varijabilnost prisutnu u realnim uzorcima. Sintetički pristup omogućava sistematsko testiranje modela i identifikaciju faktora koji utiču na njihovu tačnost u ranim fazama istraživanja. Dalja istraživanja treba usmeriti ka integraciji sintetičkih i realnih podataka, kao i razvoju regionalno specifičnih modela zasnovanih na pouzdanim populacionim frekvencijama.

4.1.8. Uticaj samoprijavljenog porekla na analizu

Učestalosti alela korišćene u ovoj tezi za generisanje sintetičkih *STR* profila preuzete su iz revidiranog skupa podataka Steffen i sar. (2017), dok su populacione kategorije prvobitno definisane u radu Hill i sar. (2013) na osnovu samoprijavljenog porekla ispitanika (poglavlje 2.1.1). Ovakav pristup predstavlja potencijalni izvor sistematske greške, jer se klasifikacija pojedinaca zasniva na subjektivnoj identifikaciji, koja ne mora nužno odražavati stvarno genetičko nasleđe (Ramos i sar., 2016; Shraga i sar., 2017).

Nepouzdanost samoprijavljenog porekla izražena je u populacijama sa složenom istorijom admiksije. Shraga i sar. (2017) navode da u latinoameričkim i afroameričkim populacijama samoprijavljeno poreklo često ne odgovara stvarnom genetičkom sastavu. Slično tome, Ramos i sar. (2016) pokazuju da samoprijavljeno etničko poreklo objašnjava relativno mali deo varijacije u genetičkom doprinosu različitih populacija, dok je taj udeo još manji kada se posmatra prijavljeno porodično poreklo. Bryc i sar. (2010) pokazuju da osobe koje se identifikuju kao Afroamerikanci mogu imati značajan udeo evropskog genetičkog porekla, što potvrđuje da socijalne kategorije ne predstavljaju precizne genetičke entitete. Royal i sar. (2010) naglašavaju da etničke kategorije imaju ograničenu biološku osnovu, što komplikuje njihovu primenu u genetičkim analizama.

Potencijalne greške u definisanju referentnih populacija prenose se i na generisanje sintetičke profile. Neslaganje između socijalno definisanih kategorija i stvarne genetičke strukture populacija može doprineti preklapanju između pojedinih grupa i time uticati na performanse modela. Deo pogrešnih klasifikacija zato ne mora biti posledica ograničenja samog modela, već i načina na koji su definisane referentne populacije. Model može pravilno prepoznavati genetičke obrasce, ali ostati ograničen neprecizno definisanim referentnim populacijama.

Preporuke u pogledu veličine populacionih baza razlikuju se u zavisnosti od vrste *STR* podataka i polimorfности analiziranih lokusa. Kod manjih uzoraka povećava se nesigurnost pri proceni učestalosti alela, posebno retkih i prethodno nezabeleženih alela (Aalbers i Gettings, 2026). U izvornoj bazi podataka azijska populaciona grupa obuhvatala je 97 ispitanika (Steffen i sar., 2017), što predstavlja metodološko ograničenje. Iako je u rezultatima ove studije azijska populacija pokazala visoku tačnost klasifikacije, ovo ograničenje treba imati u vidu pri tumačenju procenjenih učestalosti alela.

Preciznije definisanje populacionih grupa može se zasnivati na markerima informativnim za poreklo, kao što su *SNP* ili mikrohaplotipovi, koji omogućavaju objektivniju procenu biogeografskog porekla (Bryc i sar., 2015; Oldoni i sar., 2019). U budućim istraživanjima moglo bi se ispitati njihovo kombinovanje sa modelima mašinskog učenja zasnovanim na *STR* markerima radi razvoja hibridnih modela koji koriste oba tipa genetičkih podataka.

4.1.9. Kompleksnost genetičkog nasleđa hispanoameričkih populacija

Rezultati pokazuju da je diferencijacija hispanoameričke populacije bila najmanje uspešna u prediktivnom modelu (poglavlje 3.1.3). Najveći broj pogrešnih klasifikacija zabeležen je upravo između hispanoameričke i SA–EP populacije, što potvrđuje postojanje preklapanja genetičkih obrazaca između ovih grupa. Konkretno, model je dao 1511 pogrešnih klasifikacija između ove dve grupe, pri čemu je hispanoamerička populacija imala i najmanji broj tačno klasifikovanih profila, svega 7970 (Slika 2, poglavlje 3.1.3).

Ovaj nalaz može se objasniti složenom genetičkom strukturom hispanoameričkih populacija, koje predstavljaju tipičan primer populacione admiksije. Sofer i sar. (2017) navode da među različitim latinoameričkim podgrupama postoje značajne razlike u proporcijama evropskog, afričkog i autohtonog američkog genetičkog porekla, pri čemu se posebno izdvajaju razlike između tzv. kontinentalnih i karipskih populacija. Ova heterogenost dovodi do širokog spektra kombinacija alela unutar jedne nominalne populacione kategorije, što otežava njeno jasno genetičko razgraničenje.

Analiza Y–hromozoma u populaciji zapadnog Meksika pokazala je dominantan evropski genetički doprinos (60–64 %), uz značajan udeo autohtonog američkog i afričkog porekla (Rangel–Villalobos i sar., 2008). Sličan polno specifičan obrazac admiksije opisali su Bedoya i sar. (2006), uz dodatno zapažanje da mitohondrijska DNK često pokazuje dominantno autohtono poreklo, što ukazuje na polno specifične obrasce admiksije. Bryc i sar. (2015) potvrđuju postojanje razlika između autozomnih i X–hromozomskih obrazaca nasleđivanja, dok Rothhammer i sar. (2017) opisuju istorijske migracione procese koji su oblikovali genetičku strukturu južnoameričkih populacija.

Nalazi ove disertacije pokazuju da složena genetička struktura hispanoameričke populacije utiče na uspešnost klasifikacije. Povećanje tačnosti nakon isključenja ili objedinjavanja hispanoameričke grupe sa SA–EP pokazuje da je genetičko preklapanje između ovih populacija predstavljalo glavno ograničenje modela.

4.2. Analiza incesta

4.2.1. Značaj analize incesta

Incest se najšire određuje kao seksualna aktivnost unutar porodičnih ili srodničkih odnosa, pri čemu njegovo definisanje i pravno uređenje zavise od društvenih, kulturnih i pravnih okvira konkretne sredine (Choate i Sharan, 2021). U Republici Srbiji, ovo ponašanje je regulisano kroz krivično delo rodoskrvnuća (Republika Srbija, 2005, čl. 197).

Prema izveštaju organizacije Incest Trauma Centar – Beograd (2024), većina žrtava su maloletne osobe, a počinioci su u najvećem broju slučajeva članovi porodice, pri čemu se značajan procenat odnosi na biološke očeve ili očinske figure. Dug vremenski period između izvršenja dela i njegove prijave, kao i nizak procenat pokrenutih krivičnih postupaka, dodatno komplikuju identifikaciju i procesuiranje ovih slučajeva.

DNK analiza koristi se za ispitivanje bioloških odnosa i u forenzičkim istragama seksualnog zlostavljanja. Detekcija incestuoznih veza može biti otežana usled povećanog deljenja alela između

bliskih srodnika (Emre i sar., 2015). Zbog toga standardni skupovi *STR* lokusa ponekad nisu dovoljni za jasno razdvajanje mogućih scenarija srodstva, pa može biti neophodno uključivanje većeg broja lokusa ili primena dodatnih genetičkih markera (Emre i sar., 2015; Garcia i sar., 2022).

Ograničena dostupnost validacionih podataka predstavlja dodatni problem, budući da su potvrđeni slučajevi incesta retki i često nedostupni za sistematska istraživanja (Emre i sar., 2015; Choate i Sharan, 2021). Zbog toga su potrebni metodološki pristupi koji omogućavaju procenu detekcije incestuoznih odnosa u kontrolisanim uslovima.

U ovom istraživanju primenjen je pristup zasnovan na mašinskom učenju nad generisanim *STR* profilima, sa ciljem identifikacije genetičkih obrazaca karakterističnih za incestuozne veze. Ovaj pristup omogućava sistematsku evaluaciju performansi modela u različitim scenarijima srodstva.

4.2.2. Metodološki aspekti genetičke analize incesta

U literaturi su razvijeni različiti pristupi za procenu incestuoznih odnosa, među kojima se izdvajaju indeksi srodstva zasnovani na verovatnoćama. Morris i sar. (1988) uveli su indeks incesta, koji testira hipotezu o bliskom srodstvu potencijalnog oca sa majkom potomka. Ovi indeksi zasnivaju se na proceni verovatnoće genetičkog podudaranja između testiranih osoba, pri čemu veći stepen genetičke povezanosti rezultuje višim vrednostima indeksa.

Ograničen broj lokusa i pojednostavljeni probabilistički modeli smanjuju sposobnost ovih pristupa da obuhvate kompleksne obrasce nasleđivanja kroz čitav genetički profil. Pored toga, validacija ovih metoda često je sprovedena na malim i nedovoljno reprezentativnim skupovima podataka, što otežava procenu njihove stvarne efikasnosti u detekciji incestuoznih odnosa.

Praktični aspekti uzorkovanja dodatno komplikuju analizu. U slučajevima incesta genetički materijal može poticati iz fetalnih tkiva ili drugih specifičnih uzoraka (Emre i sar., 2015). Iz tog razloga, Wenk i sar. (1994) u svom radu koriste termin „potomak“ umesto „dete“, kao širu kategoriju koja obuhvata različite tipove biološkog materijala.

U ovoj doktorskoj disertaciji primenjen je drugačiji metodološki okvir zasnovan na mašinskom učenju i simulaciji *STR* profila. Generisanjem velikog broja kontrolisanih scenarija srodstva omogućena je sistematska evaluacija performansi modela, uključujući izračunavanje tačnosti, osetljivosti i specifičnosti. Za razliku od tradicionalnih metoda zasnovanih na pojedinačnim indeksima, predloženi pristup omogućava istovremenu analizu kombinovanih obrazaca genotipova kroz veći broj lokusa, čime se povećava sposobnost detekcije incestuoznih odnosa u analiziranim genetičkim scenarijima.

4.2.3. Dalji razvoj metodologije za analizu incesta

Dosadašnji pristupi analizi incestuoznih odnosa u velikoj meri zasnivaju se na indeksima srodstva koji procenjuju verovatnoće na nivou pojedinačnih lokusa. Wenk (2008) pokazuje da prisustvo retkih alela može povećati vrednost kombinovanog indeksa incesta, ali istovremeno ukazuje na ograničenu analitičku vrednost ovih metoda. Wenk (2008) pokazuje da analiza 18 *STR* lokusa daje dijagnostičke *CII* vrednosti u približno dva od pet simuliranih incestuoznih slučajeva. Procenjuje da analizom 33 *STR* lokusa može da se postigne dijagnostički prag u 97,5 % slučajeva.

U ovoj tezi analizom 21 *STR* lokusa postignuta je ukupna tačnost detekcije incestuoznih odnosa od 95,00 % (poglavlje 3.2.5). Dok klasični indeksi analiziraju lokuse pojedinačno i oslanjaju se na pretpostavljene distribucije alela, modeli mašinskog učenja istovremeno integrišu informacije sa više lokusa i prepoznaju kompleksne obrasce genotipova, čime koriste ukupnu genetičku informaciju sadržanu u profilu. Model je u ovoj varijanti postigao preciznost od 95,28 %, odziv od

94,69 %, F1 skor od 94,98 % i *ROC AUC* od 99,01 %, što ukazuje na njegovu visoku diskriminatornu sposobnost (Slika 14, poglavlje 3.2.5).

Wenk (2008) ukazuje da efikasnost predikcije incesta nije dovoljno ispitana u populacijama sa povišenim nivoom konsangviniteta, što predstavlja ograničenje postojećih metodologija. U ovoj tezi taj problem je razmotren testiranjem modela na učestalostima alela populacije Saudijske Arabije, za koju je poznato da ima povećan nivo homozigotnosti usled čestih srodničkih brakova (Khubrani i sar., 2019).

Rezultati pokazuju da je model zadržao stabilne performanse i u ovim uslovima, sa odstupanjem tačnosti manjim od 1 % u odnosu na opštu populaciju. To ukazuje da predloženi pristup nije izrazito osetljiv na populacione specifičnosti kao što je povećana homozigotnost. Konkretno, početni model za Saudijsku Arabiju postigao je tačnost od 94,64 %, dok je modifikovani model dao tačnost od 94,93 %, što potvrđuje stabilnost modela u populaciji sa povišenim nivoom konsangviniteta.

Eliminacija retkih alela u modifikovanom modelu nije dovela do smanjenja tačnosti. Iako se retki aleli tradicionalno smatraju visoko informativnim u analizama srodstva, rezultati pokazuju da detekcija incestuoznih odnosa ne zavisi primarno od pojedinačnih alela, već od globalnih obrazaca nasleđivanja, kao što su povećana homozigotnost i specifične kombinacije genotipova između majke i potomka. Blago povećanje tačnosti u modifikovanom modelu, sa 94,64 % na 94,93 %, uz veoma sličan obrazac klasifikacije prikazan na matricama konfuzije (Slike 20 i 21, poglavlja 3.2.10 i 3.2.11), pokazuje da uklanjanje retkih alela nije oslabilo model i da je njegova prediktivna sposobnost zasnovana pre svega na ukupnom obrascu genetičkog nasleđivanja.

Slična ograničenja postoje i kod drugih probablističkih pristupa. Fung (2003) razvija programe za izračunavanje odnosa verovatnoće u testiranju očinstva i srodstva, zasnovane na Bajesovoj teoremi, uslovnoj verovatnoći i analizi rodoslova. Nasuprot tome, metodologija primenjena u ovoj tezi omogućava obučavanje i testiranje modela na velikim skupovima simuliranih podataka i procenu performansi kroz veliki broj kontrolisanih scenarija.

Kombinacija mašinskog učenja i simuliranih *STR* profila omogućila je detekciju incestuoznih odnosa u analiziranim populacionim scenarijima na osnovu obrazaca kroz više lokusa.

4.2.4. Homozigotnost i predikcija incesta na osnovu jedinstvenih profila dece

U profilima potomaka nastalih iz incestuoznih odnosa utvrđena je povećana homozigotnost u poređenju sa potomcima iz nesrodničkih veza. Distribucija broja lokusa sa homozigotnim genotipovima pokazuje jasnu separaciju između ove dve kategorije, što je prikazano u poglavlju 3.2.4 i na Slici 12, i ukazuje da homozigotnost predstavlja snažan diskriminativni parametar u detekciji incestuoznih odnosa.

Ovaj obrazac ima jasno genetičko utemeljenje. U incestuoznim vezama roditelji dele veći broj *IBD* alela, što povećava verovatnoću da potomak nasledi identične alele od oba roditelja (Curik i sar., 2014). Kao posledica toga dolazi do povećanja homozigotnosti i smanjenja ukupne genetičke raznovrsnosti, što predstavlja karakterističan signal konsangviniteta (Erzurumluoglu i sar., 2016).

Model zasnovan isključivo na profilu deteta postigao je tačnost od približno 90 % u razlikovanju incestuoznog i normalnog srodstva. U poglavlju 3.2.8 prikazano je da je tačnost ovog modela iznosila 90,37 %, uz *ROC AUC* od 96,34 %, dok matrica konfuzije (Slika 17) potvrđuje da je i bez uključivanja roditeljskih profila bilo moguće postići visoku uspešnost klasifikacije. To pokazuje da genetički signal incesta može biti detektovan i bez direktnog poređenja sa roditeljskim profilima.

U forenzičkom kontekstu, ovakav pristup može biti posebno koristan u situacijama kada genetički profil potencijalnog oca nije dostupan.

Slični obrasci opisani su u literaturi. Sahoo i sar. (2021) identifikuju duge regione homozigotnosti kao indikator srodničkih brakova, dok Robino i sar. (2006) opisuju slučaj u kojem je veliki broj homozigotnih genotipova u fetalnom *STR* profilu ukazao na moguće incestuozno očinstvo, nakon čega je kasnijom analizom potvrđen incest brata i sestre.

4.2.5. Identični heterozigotni genotipovi i ukupna sličnost *STR* profila

Pored povećane homozigotnosti, incestuozni scenariji odlikovali su se gotovo dvostruko većim udelom lokusa sa identičnim heterozigotnim genotipovima majke i deteta u odnosu na normalno očinstvo. Ovaj nalaz odgovara očekivanom povećanju deljenja alela usled bliskog srodstva roditelja.

Ukupna sličnost *STR* profila takođe je bila veća u oba incestuozna scenarija. U poređenju sa odgovarajućim neincestuoznim odnosima, povećanje je iznosilo od 7,57 do 10,14 procentnih poena. Međutim, pošto su lokusi sa identičnim heterozigotnim genotipovima pri izračunavanju ukupne sličnosti dobijali maksimalnu ocenu 2, bilo je potrebno ispitati u kojoj meri su oni uticali na dobijene razlike. Nakon isključivanja ovih lokusa, korigovana sličnost ostala je veća u oba incestuozna scenarija. Razlika je iznosila 5,36 procentnih poena u poređenju normalnog očinstva sa scenarijem otac–ćerka i 11,25 procentnih poena u poređenju normalnog odnosa brat–sestra sa scenarijem brat–sestra.

Indeks sličnosti korišćen u ovoj tezi nije predstavljao neposrednu procenu udela alela identičnih po poreklu. Tamura i sar. (2000) pokazali su da ukupna genetička sličnost u incestuoznim slučajevima nije uvek odgovarala stvarnom odnosu srodstva, zbog čega ovaj pokazatelj zahteva oprezno tumačenje.

Povećana homozigotnost, veći udeo identičnih heterozigotnih genotipova i povećana ukupna sličnost opisivali su različite, ali međusobno povezane karakteristike *STR* profila u incestuoznim scenarijima. Nijedan od ovih parametara, posmatran pojedinačno, nije predstavljao dokaz incestuoznog odnosa, niti je mogao da zameni formalnu analizu srodstva i druge forenzičke dokaze. Tumačenje ovih parametara zahteva poseban oprez u populacijama sa smanjenim genetičkim diverzitetom ili povećanim nivoom konsangviniteta. U takvim populacijama i nesrodnički parovi mogu imati povišenu genetičku sličnost (Tillmar, 2010; Rohlf i sar., 2012), što bi u prediktivnom modelu moglo povećati rizik od pogrešne klasifikacije. Pragove interpretacije zato treba uskladiti sa populacionim učestalostima alela i karakteristikama analizirane populacije.

4.2.6. Implementacija jednostepene mutacije

Mutacije su izvor varijabilnosti u analizama srodstva, posebno kod autozomnih *STR* markera, gde stopa mutacija iznosi približno 10^{-3} po generaciji i najčešće se manifestuje kao promena alela za jedan ponovak (Butler, 2007). U forenzičkom kontekstu, prisustvo mutacija može dovesti do neslaganja između roditeljskih i potomačkih profila, što otežava interpretaciju rezultata i može uzrokovati pogrešno isključenje bioloških srodnika ukoliko se ne uzme u obzir (Brenner, 2004).

Tradicionalni pristupi zasnovani na verovatnoćama, poput *LR* analize, posebno su osetljivi na genetička nepodudaranja, jer se ona neposredno odražavaju na vrednosti indeksa srodstva. Zbog toga se u savremenim pristupima pri tumačenju srodstva mogućnost mutacije uzima u obzir kroz odgovarajuće mutacione parametre, kako bi se smanjio rizik od pogrešnih zaključaka (Gjertson i sar., 2007; Ge i sar., 2010).

Uticaj mutacija procenjen je uvođenjem jednostepenih mutacija u simulirane *STR* profile dece, čime su stvoreni uslovi bliži realnim forenzičkim scenarijima. Time je procenjena stabilnost modela u prisustvu genetičkih odstupanja.

Implementacija jednostepenih mutacija imala je zanemarljiv numerički uticaj na tačnost modela, koja je iznosila 95,02 %, u poređenju sa 95,00 % u modelu bez mutacija. Poređenje rezultata iz poglavlja 3.2.5 i 3.2.7 pokazalo je da je *ROC AUC* u oba slučaja iznosio 99,01 %, dok su matrice konfuzije prikazane na Slikama 14 i 16 pokazale veoma sličan obrazac klasifikacije. Model je, prema tome, zadržao stabilne performanse u prisustvu simuliranih mutacionih događaja.

Ovakav rezultat može se objasniti time što klasifikacija incestuoznih odnosa ne zavisi od pojedinačnih lokusa, već od globalnih obrazaca genetičke sličnosti, kao što su povećana homozigotnost i obrasci deljenja alela između srodnika. Sporadične mutacije, koje utiču na mali broj lokusa, ne narušavaju ove obrasce u meri koja bi značajno promenila klasifikacioni ishod.

U poređenju sa klasičnim analizama, model zasnovan na mašinskom učenju zadržao je stabilne performanse nakon uvođenja jednostepenih mutacija u simulirane profile. Mutacije u klasičnim analizama mogu dovesti do smanjenja *LR* vrednosti i otežati interpretaciju rezultata, dok su u ovom simulacionom scenariju imale mali uticaj na performanse modela.

4.2.7. Problematika višeklasne klasifikacije

Višeklasna klasifikacija u ovom istraživanju imala je za cilj da razlikuje profile normalnog očinstva od incestuoznih, ali i da unutar incestuoznih profila dodatno klasifikuje tip odnosa (otac–dete ili brat–sestra). Rezultati pokazuju da je model uspešnije razlikovao normalno očinstvo od dve kategorije incesta, pri čemu se najveći broj grešaka javio upravo između te dve kategorije. Ovakav obrazac prikazan je matricom konfuzije za višeklasnu predikciju (Slika 22, poglavlje 3.2.12), gde je model tačno klasifikovao 7179 slučajeva normalnog očinstva, ali je istovremeno zabeleženo ukupno 7496 pogrešnih klasifikacija između dve kategorije incesta.

To pokazuje da genetički signali koji ukazuju na prisustvo bliskog srodstva predstavljaju dominantan obrazac u podacima, dok su razlike između pojedinačnih incestuoznih odnosa znatno suptilnije.

I odnosi otac–dete i brat–sestra predstavljaju bliske srodničke odnose sa visokim stepenom deljenja genetičkog materijala, zbog čega njihovo razlikovanje može biti ograničeno kada se analiza zasniva isključivo na standardnim autozomnim *STR* markerima (Ge i Budowle, 2020; Li i sar., 2025). Ove kategorije se preklapaju u genetičkom prostoru, što ograničava njihovu separaciju na osnovu standardnih *STR* markera.

To potvrđuje i *PCA* analiza, koja pokazuje delimičnu razliku u prostornoj raspodeli profila normalnog i incestuoznog srodstva, ali značajno preklapanje između različitih tipova incestuoznih odnosa. Distribucija ukazuje da varijabilnost u podacima postoji, ali nije dovoljno izražena za pouzdanu višeklasnu klasifikaciju. U poglavlju 3.2.9 ovaj obrazac je prikazan na Slikama 18 i 19: na Slici 18 uočava se delimična razlika u raspodeli između normalnog i incestuoznog srodstva, dok Slika 19 pokazuje znatno izraženije preklapanje između različitih kategorija incesta. Prve dve glavne komponente u tom slučaju objašnjavaju svega 2,66 % i 2,65 % ukupne varijabilnosti, što dodatno ukazuje na ograničenu separaciju ovih kategorija.

ROC AUC vrednosti od 98,00 % za kategoriju normalnog očinstva i 73,00 % za obe kategorije incesta (Slika 23 u poglavlju 3.2.12) model uspešnije razlikovao normalno očinstvo od dva oblika incesta nego što je međusobno razlikovao ta dva oblika incesta, kao i da genetička informacija

sadržana u analiziranim *STR* markerima nije dovoljna za potpuno razdvajanje pojedinačnih kategorija incestuoznih odnosa bez uključivanja dodatnih izvora informacija.

Tokom razvoja metodologije razmatrane su i preliminarne simulacije zasnovane na globalnim učestalostima alela, koje su ukazivale na uspešnost višeklasne klasifikacije. Naknadna analiza je, međutim, pokazala da takav pristup može dovesti do nerealno niske početne sličnosti roditeljskih profila i veštačkog povećanja genetičke raznovrsnosti u simuliranim podacima, čime se performanse modela mogu preceniti. Ovi nalazi zato nisu uključeni u zvanično poglavlje rezultata, ali pokazuju da izbor referentnih populacionih učestalosti utiče na rezultate simulacionih studija.

4.2.8. Binarna klasifikacija incesta

Zbog ograničenja uočenih u višeklasnoj predikciji, u ovom istraživanju primenjen je binarni model koji klasifikuje profile kao „normalno očinstvo“ ili „incestuozno srodstvo“, pri čemu su sve kategorije incesta objedinjene u jednu klasu. Ovakav pristup izabran je u skladu sa primarnim ciljem ovog istraživanja, da se utvrdi da li postoji blisko srodstvo između roditelja. Precizno razgraničenje tipa incestuoznog odnosa predstavlja zaseban korak koji može zahtevati dodatne genetičke markere i analize (Emre i sar., 2015; Garcia i sar., 2022).

Rezultati pokazuju da model zasnovan isključivo na poređenju dva *STR* profila (roditelj–potomak), bez dodatnih karakteristika, postiže visoku tačnost klasifikacije. U scenarijima sa 29 lokusa ostvarena je tačnost od 96,94 %, uz gotovo uravnoteženu i preciznu klasifikaciju obe kategorije. U poglavlju 3.2.3 prikazano je da je model tačno klasifikovao 9694 profila normalnog očinstva i 9695 profila incestuoznog srodstva, dok je pogrešno klasifikovao 306 profila kao incestuozno i 305 kao normalno srodstvo, što potvrđuje uravnoteženost modela (Slika 9). *ROC* kriva i kriva preciznosti i odziva pokazuju diskriminatornu sposobnost modela i nisku stopu lažno pozitivnih rezultata. *ROC AUC* iznosio je 99,62 %, a stopa lažno pozitivnih rezultata 3,06 % (Slike 10 i 11; poglavlje 3.2.3).

Smanjenje broja lokusa na 21 dovelo je do umerenog smanjenja tačnosti modela, pri čemu je postignuta tačnost od 95 %. To ima važnu praktičnu implikaciju, jer ukazuje da se predloženi model može primeniti na standardne *STR* panele koji se rutinski koriste u laboratorijama. Rezultati iz poglavlja 3.2.5 pokazuju da je i pri ovom broju markera *ROC AUC* ostao vrlo visok i iznosio 99,01 %, što potvrđuje da smanjenje broja lokusa nije narušilo sposobnost modela da razlikuje normalno i incestuozno srodstvo (Slika 14).

Efikasnost binarne klasifikacije proizlazi iz činjenice da incestuozni odnosi generišu izražene genetičke obrasce, poput povećane homozigotnosti i većeg stepena deljenja alela između roditelja. Ovi obrasci predstavljaju dominantan signal u podacima i omogućavaju detekciju incestuoznih veza, nezavisno od njihove specifične kategorije. U modelu sa 29 lokusa specifičnost je iznosila 96,94 %, a stopa lažno negativnih rezultata 3,05 % (poglavlje 3.2.3).

Rezultati ove teze pokazali su da informacija sadržana u analiziranim standardnim autozomnim *STR* markerima nije bila dovoljna za pouzdano razlikovanje pojedinačnih tipova incestuoznih odnosa. Prošireni genetički paneli, poput *SNP* markera visoke gustine i sekvenciranja nove generacije, mogu se ispitati radi unapređenja razdvajanja ovih odnosa. Studija Cosenza i sar. (2025) pokazuje da integracija velikog broja *SNP* markera i metoda mašinskog učenja omogućava precizniju analizu složenih srodničkih odnosa.

Čak i primenom naprednijih tehnologija, u situacijama sa više potencijalnih srodnika dodatno profilisanje ostaje neophodno. Binarna klasifikacija može se koristiti za inicijalnu identifikaciju potencijalnih incestuoznih odnosa i usmeravanje daljih ispitivanja.

Kombinacija standardnih *STR* markera i algoritama mašinskog učenja omogućava inicijalnu detekciju incestuoznih odnosa, nakon koje se mogu primeniti dodatni genetički i analitički postupci.

4.2.9. Informativnost lokusa za predikciju incesta

Analiza doprinosa pojedinačnih lokusa pokazala je da su kao najinformativniji za predikciju incesta identifikovani *SE33*, *DIS1656* i *D18S51*, dok su najmanje informativni bili *TH01*, *D3S1358*, *D5S818*, *CSF1PO* i *TPOX* (Slika 15, poglavlje 3.2.6). Rezultat dobijen u ovoj tezi može se tumačiti u skladu sa opštim principima forenzičke genetike, prema kojima lokusi sa većim brojem alela i višim nivoom heterozigotnosti imaju veću diskriminatornu moć (Novokmet i sar., 2022).

Populacioni parametri dodatno potvrđuju ovu razliku. Lokus *SE33* pokazuje visoku heterozigotnost (0,9353), nisku verovatnoću identiteta (0,0066) i visoku verovatnoću isključenja očinstva (0,8881), što ukazuje na njegovu visoku informativnost. Nasuprot tome, lokus *TPOX*, sa nižom heterozigotnošću (0,6899) i višom verovatnoćom identiteta (0,1355), ima znatno manji diskriminatorni potencijal (Steffen i sar., 2017). Pojedinačni lokusi različito doprinose modelu u zavisnosti od svoje polimorfnosti i učestalosti alela u populaciji.

Analizom 21 lokusa postignuta je tačnost od 95 %, što pokazuje da ovaj broj lokusa može obezbediti visoku uspešnost klasifikacije kada se njihovi efekti sagledavaju objedinjeno.

Modeli za predikciju incesta nisu u istoj meri uslovljeni brojem analiziranih markera kao modeli za predikciju biogeografskog porekla. Dok je za preciznu klasifikaciju populacija često neophodan veći broj markera radi detekcije suptilnih razlika u učestalostima alela, detekcija incestuoznih odnosa oslanja se na izraženije genetičke obrasce, kao što su povećana homozigotnost i veći stepen deljenja alela između srodnika.

Predikcija porekla zahteva detekciju finih populacionih razlika, dok se detekcija incesta zasniva na izraženijim obrascima nasleđivanja unutar porodice. Zbog toga relativno mali broj visoko polimorfni *STR* lokusa može sadržati dovoljno informacija za detekciju incestuoznih odnosa.

4.3. Prediktivni modeli u forenzičkoj genetici: prednosti i izazovi u primeni

U ovom istraživanju modeli mašinskog učenja primenjeni su na *STR* profile radi predikcije biogeografskog porekla i detekcije incestuoznih odnosa. Rezultati inicijalnog testiranja različitih algoritama pokazali su da je za predikciju biogeografskog porekla najbolju početnu tačnost postigao *XGBoost*, dok je u detekciji incesta najbolji rezultat ostvario *CatBoost* (Tabele 1 i 7; poglavlja 3.1.1 i 3.2.1).

Razlike u performansama između binarnih i višeklasnih klasifikacionih zadataka, uočene u ovom radu, u skladu su sa opštim principima mašinskog učenja. Višeklasni problemi su metodološki zahtevniji, jer sa porastom broja klasa raste i broj granica odlučivanja koje model mora da razdvoji, što povećava složenost klasifikacije i verovatnoću preklapanja između klasa (Del Moral i sar., 2022). U forenzičkom kontekstu to je posebno izraženo kod predikcije biogeografskog porekla, gde razlikovanje genetički bliskih populacija zahteva prepoznavanje vrlo suptilnih obrazaca u podacima, a standardni forenzički paneli često imaju ograničenu diskriminatornu moć na finijem populacionom nivou (Alladio i sar., 2022). Nasuprot tome, binarni zadaci, poput detekcije incesta, mogu biti stabilniji kada se zasnivaju na izraženijim obrascima razdvajanja između dve klase.

U ovom istraživanju algoritmi zasnovani na gradijentnom pojačanju pokazali su se pogodnim za analizu *STR* podataka. Ovi algoritmi omogućavaju modelovanje nelinearnih odnosa i interakcija između genetičkih obeležja kroz više lokusa. Za razliku od tradicionalnih statističkih pristupa, koji često polaze od pojedinačnih efekata markera ili unapred zadatih interakcija, ovakvi modeli mogu da

prepoznaju složenije obrasce u genotipovima (Elgart i sar., 2022). Među najuspešnijim modelima u ovoj tezi bili su algoritmi gradijentnog pojačanja. U analizi incesta, pored *CatBoost*-a, visoke performanse postigli su i *XGBoost* i *LightGBM*, dok je u predikciji biogeografskog porekla *XGBoost* bio uspešniji od *SVM*-a, naivnog Bajesovog klasifikatora, stabla odlučivanja i logističke regresije, što ukazuje da su u analiziranim uslovima modeli sposobni da obuhvate složenije odnose između lokusa bili pogodniji za analizu *STR* profila (Tabele 1 i 7; poglavlja 3.1.1 i 3.2.1).

Rezultati prediktivnih modela u forenzičkoj genetici moraju se interpretirati kao probabilističke procene, a ne kao konačni dokazi. Biedermann i sar. (2015) naglašavaju da pogrešno tumačenje verovatnoće može dovesti do ozbiljnih grešaka u forenzičkoj interpretaciji. Prediktivni modeli zato pružaju dopunske informacije, ali ne zamenjuju stručnu analizu i druge dokaze.

Interpretabilnost modela predstavlja ograničenje u primeni mašinskog učenja u forenzici. Za razliku od klasičnih statističkih metoda, složeniji modeli često funkcionišu kao „crne kutije“, zbog čega nije uvek moguće neposredno razumeti koji faktori utiču na određeno predviđanje (Solanke, 2022). Metode objašnjivog mašinskog učenja mogu se koristiti za analizu ovih faktora. *SHAP* kvantifikuje doprinos svake pojedinačne varijable odluci modela (Lundberg i Lee, 2017), *LIME* pruža lokalno objašnjenje konkretnog predviđanja aproksimacijom ponašanja modela u neposrednoj okolini posmatranog slučaja (Ribeiro i sar., 2016), dok kontrafaktualna objašnjenja ukazuju na najmanje izmene ulaznih podataka koje bi dovele do promene ishoda (Wachter i sar., 2018). Primena ovih pristupa povećala bi transparentnost modela i olakšala njihovu primenu u forenzičkom kontekstu (Solanke, 2022; Veldhuis i sar., 2022).

Hiperparametri modela optimizovani su radi unapređenja performansi i stabilnosti izabranih modela. Nakon optimizacije, *XGBoost* i *CatBoost* zadržani su kao centralni modeli za dalju analizu, što pokazuje da su njihove performanse procenjene nakon sistematskog prilagođavanja modela strukturi podataka (poglavlja 3.1.2 i 3.2.2).

U forenzičkoj primeni rezultati modela moraju biti tačni, razumljivi i proverljivi. Prediktivni modeli zasnovani na mašinskom učenju mogu identifikovati obrasce u velikim skupovima podataka i pružiti dodatne smernice za istragu, ali njihove rezultate treba tumačiti zajedno sa drugim forenzičkim dokazima.

4.4. Ograničenja studije i buduće smernice istraživanja

Primenljivost rezultata ove studije ograničena je karakteristikama korišćenih podataka, genetičkih markera i samih modela. Navedena ograničenja određuju opseg u kome se rezultati mogu tumačiti i primenjivati.

Jedno od osnovnih ograničenja odnosi se na broj analiziranih populacija. Model je razvijen i testiran na četiri osnovne populacione grupe, što ograničava procenu njegove primenljivosti na druge populacije. Uključivanje dodatnih referentnih populacija, posebno bliskoistočnih, omogućilo bi ispitivanje modela u većem broju forenzičkih scenarija.

Rezultati su takođe ukazali na ograničenja u diferencijaciji između genetički sličnih populacija, poput evropske i hispanoameričke. Ova poteškoća proizilazi iz izražene populacione admiksije, gde preklapanje učestalosti alela otežava klasifikaciju zasnovanu isključivo na autozomnim *STR* markerima. U takvim slučajevima, upotreba dodatnih genetičkih markera, poput *SNP*, *X-STR* ili *Y-STR* markera, može povećati rezoluciju analize i omogućiti preciznije razdvajanje populacionih grupa.

Slična ograničenja uočena su i u detekciji incestuoznih odnosa. Rezultati dobijeni u ovom domenu moraju se tumačiti oprezno i potvrđivati dodatnom genetičkom i forenzičkom analizom.

Tačnost modela može biti uslovljena populacionom strukturom, posebno u slučajevima gde se analizirani pojedinci razlikuju od referentne populacije na kojoj je model treniran.

Genetička raznovrsnost populacije može uticati na tačnost modela. U izolovanim populacijama, povećan nivo homozigotnosti može dovesti do lažno pozitivnih klasifikacija incestuoznih odnosa, dok u populacijama sa visokom genetičkom raznovrsnošću incestuozni odnosi mogu ostati nedovoljno izraženi u genetičkom profilu i dovesti do lažno negativnih rezultata. Interpretacija rezultata zato mora uzeti u obzir populacione karakteristike analiziranih pojedinaca.

Buduća istraživanja treba da obuhvate razvoj populaciono-specifičnih modela, treniranih na reprezentativnim učestalostima alela ciljnih populacija. Integracija različitih tipova genetičkih markera, uključujući *SNP* i mikrohplotipove, mogla bi unaprediti preciznost kako u predikciji porekla, tako i u detekciji složenih srodničkih odnosa.

Dalji razvoj metodologije treba da obuhvati spoljašnju validaciju modela na realnim forenzičkim podacima i primenu metoda za interpretabilnost modela radi povećanja transparentnosti i procene prihvatljivosti ovih pristupa u forenzičkoj i pravnoj praksi.

Dalje unapređenje modela zahteva spoljašnju validaciju na raznovrsnijim populacijama, realnim forenzičkim podacima i dodatnim tipovima genetičkih markera.

4.5. Etika i odgovorna primena prediktivnih modela

Primena prediktivnih modela u forenzičkoj genetici povezana je sa etičkim i pravnim pitanjima koja proizilaze iz prirode genetičkih podataka i načina njihove interpretacije (Scudder i sar., 2018; Wienroth, 2018). Za razliku od standardnih identifikacionih analiza, prediktivni modeli generišu probabilističke procene koje mogu imati implikacije po privatnost, identitet i društveni položaj pojedinca (Scudder i sar., 2018; Samuel i Prainsack, 2019).

Zaštita privatnosti predstavlja jedan od etičkih izazova. Genetičke analize koje omogućavaju procenu biogeografskog porekla ili srodničkih odnosa mogu otkriti informacije koje prevazilaze svrhu forenzičke istrage (Scudder i sar., 2018; Wienroth, 2018). Takvi podaci mogu uključivati informacije koje se smatraju privatnim i koje, ukoliko budu pogrešno interpretirane ili upotrebljene, mogu doprineti stigmatizaciji ili diskriminaciji, naročito određenih manjinskih grupa (Scudder i sar., 2018; Samuel i Prainsack, 2019).

Detekcija incestuoznih odnosa zahteva dodatni oprez, jer takvi nalazi mogu imati ozbiljne emocionalne, porodične, društvene i pravne posledice (Choate i Sharan, 2021). Rezultate koji ukazuju na potencijalni incest treba interpretirati isključivo kao indikativne i dodatno potvrditi kroz standardne forenzičke procedure (Emre i sar., 2015).

Odgovorna primena prediktivnih modela u forenzičkoj genetici zahteva poznavanje njihovih mogućnosti i ograničenja, kao i temeljnu validaciju pre uvođenja u operativni rad. Prediktivni modeli treba da se koriste kao dopunski alati i ne zamenjuju ekspertizu i profesionalnu procenu forenzičkih stručnjaka (Barash i sar., 2024).

Prediktivni modeli ne treba da budu jedina osnova za pravne zaključke. Rezultate treba da tumače stručna lica, uz jasno navođenje ograničenja modela. Transparentnost i proverljivost primenjenih metoda potrebne su zato što nedostatak objašnjivosti može ugroziti prihvatljivost rezultata u pravnom kontekstu (Barash i sar., 2024).

4.6. Naučni doprinos istraživanja

U disertaciji je ispitana primena algoritama mašinskog učenja na standardne autozomne *STR* profile za predikciju biogeografskog porekla i detekciju incestuoznih odnosa. Naučni doprinos istraživanja ogleda se u sledećem.

Pokazano je da standardni autozomni *STR* markeri, koji su primarno razvijeni za individualnu identifikaciju i analizu srodstva, u analiziranim uslovima sadrže populaciono informativne podatke koji se mogu koristiti za predikciju porekla primenom metoda mašinskog učenja. Rezultati ukazuju na mogućnost sekundarne analize postojećih *STR* profila bez dodatne genotipizacije specijalizovanih markera, kao što su *AIM-SNP* paneli.

Razvijen je i evaluiran model za detekciju incestuoznih odnosa zasnovan na *STR* profilima majke i potomka. Model je identifikovao obrasce povezane sa bliskim srodstvom, uključujući povećanu homozigotnost i veći stepen deljenja alela. Binarni klasifikacioni pristup omogućio je detekciju incestuoznih odnosa, dok model nije pouzdano razlikovao pojedinačne analizirane tipove incesta.

Sistematski je upoređeno više algoritama mašinskog učenja i optimizovani su njihovi hiperparametri, čime su izdvojeni modeli sa najboljim performansama u analiziranim zadacima predikcije porekla i detekcije incestuoznih odnosa.

Primenjen je metodološki pristup zasnovan na generisanju sintetičkih *STR* profila prema poznatim učestalostima alela, čime je omogućeno kreiranje velikih i uravnoteženih skupova podataka. Takav pristup omogućio je evaluaciju performansi modela u različitim scenarijima.

Identifikovani su faktori povezani sa performansama modela, uključujući broj analiziranih lokusa, veličinu trening skupa i stepen genetičkog preklapanja analiziranih populacija. Rezultati ukazuju na to da populaciona admiksija i genetičko preklapanje mogu ograničiti tačnost predikcije biogeografskog porekla, dok se detekcija incestuoznih odnosa zasniva pre svega na obrascima nasleđivanja unutar porodice.

Modeli mašinskog učenja omogućili su zajedničku analizu obrazaca genotipova na više lokusa. Za njihovu eventualnu primenu u forenzičkoj praksi potrebne su validacije na realnim profilima i odgovarajući referentni podaci, a rezultate je potrebno tumačiti zajedno sa drugim dokazima.

ZAKLJUČCI

Na osnovu postavljenih hipoteza i ciljeva istraživanja, a u skladu sa dobijenim rezultatima, izvedeni su sledeći zaključci:

1. Poređenjem algoritama *SVM*, *KNN*, Naive Bayes, Decision Tree, multinomijalne logističke regresije i *XGBoost*, algoritam *XGBoost* postigao je najveću početnu tačnost za predikciju porekla i izabran je kao konačni prediktivni model.
2. Model zasnovan na autozomnim *STR* profilima postigao je tačnost od 94,24 % pri klasifikaciji afroameričke, azijske, SA–EP i hispanoameričke grupe.
3. Tačnost modela zavisila je od broja profila korišćenih za treniranje. Najniža ukupna tačnost dobijena je pri skupu od 500 profila po grupi, pri čemu je tačnost bila viša za afroameričku i azijsku grupu, a niža za hispanoameričku i SA–EP grupu. Razlika u tačnosti između skupova od 10000 i 90000 profila po grupi iznosila je oko jednog procentnog poena. Zato je 10000 profila po populacionoj grupi određeno kao najmanja ispitana veličina trening skupa iznad koje povećanje broja profila nije bitno povećavalo tačnost.
4. Najinformativniji lokusi za predikciju porekla bili su *Penta E*, *F13A01* i *SE33*, a najmanje informativni lokusi *TPOX* i *CSFIPO*. Uklanjanje najmanje informativnih lokusa blago je smanjilo tačnost modela, što ukazuje da su i lokusi sa niskim pojedinačnim vrednostima važnosti doprinosili ukupnoj predikciji modela.
5. Pri poređenju skupova markera koji odgovaraju komercijalnim kompletima za amplifikaciju, najveća tačnost zabeležena je za skupove lokusa obuhvaćene kompletima *Investigator 24plex* i *GlobalFiler*, dok su niže vrednosti zabeležene za *AmpFI STR Identifier* i *NGM Detect*. Pošto su se analizirani skupovi razlikovali i prema broju i prema sastavu lokusa, uticaj broja lokusa nije bilo moguće odvojiti od uticaja njihove pojedinačne informativnosti.
6. U alternativnim klasifikacionim scenarijima model je postigao tačnost od 99,06 % pri diferencijaciji afroameričke, azijske i SA–EP grupe, odnosno 98,57 % kada su SA–EP i hispanoamerička grupa objedinjene u jednu kategoriju. Povećanje tačnosti nakon isključivanja ili objedinjavanja hispanoameričke i SA–EP grupe bilo je u skladu sa njihovim većim genetičkim preklapanjem u odnosu na ostale analizirane grupe.
7. Analiza glavnih komponenti pokazala je preklapanje SA–EP i hispanoameričkih profila, dok su afroamerička i azijska grupa formirale prostorno odvojene klastere. Ovakva raspodela bila je u skladu sa većim brojem pogrešnih klasifikacija između SA–EP i hispanoameričke grupe i manjim brojem pogrešnih klasifikacija između afroameričke i azijske grupe.
8. Poređenjem algoritama Random Forest, *SVM*, Naive Bayes, *KNN*, *CatBoost*, *XGBoost*, *LightGBM*, *AdaBoost*, Gradient Boosting, Extra Trees i *MLP*, *CatBoost* je postigao najveću početnu tačnost u binarnoj klasifikaciji incestuoznih i neincestuoznih odnosa i izabran je kao konačni prediktivni model.
9. Model zasnovan na analizi *STR* profila majke i deteta primenom 29 lokusa postigao je tačnost od 96,94 % pri detekciji incestuoznih odnosa u opštoj populaciji.

10. Model je postigao uporedive vrednosti tačnosti u različitim populacionim i eksperimentalnim uslovima: 95,00 % u opštoj populaciji pri analizi 21 lokusa, 94,64 % u populaciji Saudijske Arabije i 94,93 % u istoj populaciji nakon dodatne modifikacije učestalosti alela.
11. Nakon uvođenja jednostepenih mutacija tačnost modela iznosila je 95,02 %, gotovo isto kao u modelu bez mutacija, u kojem je iznosila 95,00 %. U ovoj studiji mutacije nisu imale uticaj na tačnost modela.
12. Model zasnovan isključivo na jedinstvenim *STR* profilima dece, primenom 21 lokusa, postigao je tačnost od 90,37 %. Ovaj rezultat pokazuje da su profili dece sadržali informacije relevantne za detekciju incestuoznih odnosa, ali je tačnost bila niža nego pri zajedničkoj analizi profila majke i deteta.
13. Analizirani genetički pokazatelji razlikovali su se između incestuoznih i kontrolnih kategorija: homozigotnost u profilima dece iz incestuoznih odnosa iznosila je 28–48 %, naspram 7–30 % u kontrolnoj grupi; udeo identičnih heterozigotnih genotipova između majke i deteta iznosio je 25–27 %, naspram 5–18,5 %; sličnost profila majke i deteta bila je približno 10 procentnih poena veća nego pri neincestuoznom ukrštanju.
14. Kao najinformativniji za detekciju incesta prepoznati su lokusi *SE33*, *DIS1656* i *DI8S51*, a kao najmanje informativni *TH01*, *D3S1358*, *D5S818*, *CSF1PO* i *TPOX*. Poređenje modela sa 21 i 29 lokusa pokazalo je da je uključivanje dodatnih osam lokusa povećalo tačnost sa 95,00 % na 96,94 %, odnosno za 1,94 procentna poena. U analiziranim uslovima povećanje broja lokusa dovelo je do umerenog povećanja tačnosti.
15. Analiza glavnih komponenti pokazala je delimično prostorno razdvajanje neincestuoznih i incestuoznih profila, uz preklapanje profila i nizak udeo varijabilnosti objašnjen prvim dvema komponentama. Između dva analizirana oblika incesta zabeleženo je znatno preklapanje. Višeklasni model postigao je *ROC AUC* od 98,00 % za kategoriju normalnog očinstva i 73,00 % za obe kategorije incesta, ali je između dve kategorije incesta zabeleženo 7496 pogrešnih klasifikacija, zbog čega model nije pouzdano određivao tačan oblik incestuoznog odnosa. Binarna klasifikacija bila je pogodnija za osnovni zadatak razlikovanja incestuoznih i neincestuoznih odnosa.

LITERATURA

- Aalbers, S. E., & Gettings, K. B. (2026). Revisiting guidance on population sampling for highly polymorphic STR loci. *Forensic Science International: Genetics*, 80, Article 103336. <https://doi.org/10.1016/j.fsigen.2025.103336>
- Algee-Hewitt, B. F. B., Edge, M. D., Kim, J., Li, J. Z., & Rosenberg, N. A. (2016). Individual identifiability predicts population identifiability in forensic microsatellite markers. *Current Biology*, 26(7), 935–942. <https://doi.org/10.1016/j.cub.2016.01.065>
- Alketbi, S. K. (2023). The role of DNA in forensic science: A comprehensive review. *International Journal of Science and Research Archive*, 9(2), 814–829. <https://doi.org/10.30574/ijrsra.2023.9.2.0624>
- Alladio, E., Poggiali, B., Cosenza, G., & Pilli, E. (2022). Multivariate statistical approach and machine learning for the evaluation of biogeographical ancestry inference in the forensic field. *Scientific Reports*, 12, Article 8974. <https://doi.org/10.1038/s41598-022-12903-0>
- Almars, A. M., Alwateer, M., Qaraad, M., Amjad, S., Fathi, H., Kelany, A. K., Hussein, N. K., & Elhosseini, M. (2021). Brain cancer prediction based on novel interpretable ensemble gene selection algorithm and classifier. *Diagnostics*, 11(10), Article 1936. <https://doi.org/10.3390/diagnostics11101936>
- Al-Mawlah, Y. H. (2026). Advances in DNA phenotyping: From molecular markers to predictive forensic profiling. *MRS Journal of Multidisciplinary Research and Studies*, 3(1), 1–7. <https://doi.org/10.5281/zenodo.18136416>
- Alvarez, G., Ceballos, F. C., & Quinteiro, C. (2009). The role of inbreeding in the extinction of a European royal dynasty. *PLOS ONE*, 4(4), Article e5174. <https://doi.org/10.1371/journal.pone.0005174>
- Aneli, S., Birolo, G., & Matullo, G. (2022). Twenty years of the Human Genome Diversity Project. *Human Population Genetics and Genomics*, 2(4), Article 0005. <https://doi.org/10.47248/hpgg2202040005>
- Balding, D. J., & Nichols, R. A. (1994). DNA profile match probability calculation: How to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International*, 64(2–3), 125–140. [https://doi.org/10.1016/0379-0738\(94\)90222-4](https://doi.org/10.1016/0379-0738(94)90222-4)
- Bandah-Rozenfeld, D., Voskoboinik, L., Rashkovski, K., Avrahami, K., Schayek, H., & Avlas, O. (2025). The power of STR discrimination: Two brothers with high-profile similarity. *Journal of Forensic Sciences*, 70(4), 1622–1628. <https://doi.org/10.1111/1556-4029.70060>
- Barash, M., McNevin, D., Fedorenko, V., & Giverts, P. (2024). Machine learning applications in forensic DNA profiling: A critical review. *Forensic Science International: Genetics*, 69, Article 102994. <https://doi.org/10.1016/j.fsigen.2023.102994>
- Bedoya, G., Montoya, P., García, J., Soto, I., Bourgeois, S., Carvajal, L., Labuda, D., Alvarez, V., Ospina, J., Hedrick, P. W., & Ruiz-Linares, A. (2006). Admixture dynamics in Hispanics: A shift in the nuclear genetic ancestry of a South American population isolate. *Proceedings of the National Academy of Sciences*, 103(19), 7234–7239. <https://doi.org/10.1073/pnas.0508716103>
- Behar, D. M., van Oven, M., Rosset, S., Metspalu, M., Loogväli, E.-L., Silva, N. M., Kivisild, T., Torroni, A., & Villems, R. (2012). A “Copernican” reassessment of the human mitochondrial DNA tree from its root. *The American Journal of Human Genetics*, 90(4), 675–684. <https://doi.org/10.1016/j.ajhg.2012.03.002>
- Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J.-F., Cann, H., Mallick, S., Reich, D., Sandhu, M.

- S., Skoglund, P., Scally, A., Xue, Y., ... Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367(6484), Article eaay5012. <https://doi.org/10.1126/science.aay5012>
- Bhinder, M. A., Zahoor, M. Y., Sadia, H., Qasim, M., Perveen, R., Anjum, G. M., Iqbal, M., Ullah, N., Shehzad, W., Tariq, M., & Waryah, A. M. (2018). SE33 locus as a reliable genetic marker for forensic DNA analysis systems. *Turkish Journal of Medical Sciences*, 48(3), 611–614. <https://doi.org/10.3906/sag-1801-21>
- Biedermann, A., Bozza, S., & Taroni, F. (2015). Prediction in forensic science: A critical examination of common understandings. *Frontiers in Psychology*, 6, Article 737. <https://doi.org/10.3389/fpsyg.2015.00737>
- Bittles, A. H. (2003). Consanguineous marriage and childhood health. *Developmental Medicine & Child Neurology*, 45(8), 571–576. <https://doi.org/10.1017/S001216220300104X>
- Bossart, J. L., & Pashley Prowell, D. (1998). Genetic estimates of population structure and gene flow: Limitations, lessons and new directions. *Trends in Ecology & Evolution*, 13(5), 202–206. [https://doi.org/10.1016/S0169-5347\(97\)01284-6](https://doi.org/10.1016/S0169-5347(97)01284-6)
- Brenner, C. H. (2004). Multiple mutations, covert mutations and false exclusions in paternity casework. *International Congress Series*, 1261, 112–114. [https://doi.org/10.1016/S0531-5131\(03\)01843-0](https://doi.org/10.1016/S0531-5131(03)01843-0)
- Brookes, C., Bright, J.-A., Harbison, S., & Buckleton, J. (2012). Characterising stutter in forensic STR multiplexes. *Forensic Science International: Genetics*, 6(1), 58–63. <https://doi.org/10.1016/j.fsigen.2011.02.001>
- Bryc, K., Auton, A., Nelson, M. R., Oksenberg, J. R., Hauser, S. L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S. A., & Bustamante, C. D. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences*, 107(2), 786–791. <https://doi.org/10.1073/pnas.0909559107>
- Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D., & Mountain, J. L. (2015). The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *The American Journal of Human Genetics*, 96(1), 37–53. <https://doi.org/10.1016/j.ajhg.2014.11.010>
- Butler, J. M. (2005). *Forensic DNA typing: Biology, technology, and genetics of STR markers* (2nd ed.). Elsevier Academic Press.
- Butler, J. M. (2006). Genetics and genomics of core short tandem repeat loci used in human identity testing. *Journal of Forensic Sciences*, 51(2), 253–265. <https://doi.org/10.1111/j.1556-4029.2006.00046.x>
- Butler, J. M. (2007). Short tandem repeat typing technologies used in human identity testing. *BioTechniques*, 43(4), Sii–Sv. <https://doi.org/10.2144/000112582>
- Butler, J. M. (2012). *Advanced topics in forensic DNA typing: Methodology*. Academic Press. <https://doi.org/10.1016/C2011-0-04189-3>
- Butler, J. M. (2015a). *Advanced topics in forensic DNA typing: Interpretation*. Academic Press. <https://doi.org/10.1016/C2011-0-07649-4>
- Butler, J. M. (2015b). The future of forensic DNA analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1674), Article 20140252. <https://doi.org/10.1098/rstb.2014.0252>
- Butler, J. M., Hill, C. R., Kline, M. C., Duewer, D. L., Sprecher, C. J., McLaren, R. S., Rabbach, D. R., Krenke, B. E., & Storts, D. R. (2009). The single most polymorphic STR locus: SE33

- performance in U.S. populations. *Forensic Science International: Genetics Supplement Series*, 2(1), 23–24. <https://doi.org/10.1016/j.fsigss.2009.08.173>
- Cai, M., Lei, F., Chen, M., Lan, Q., Wu, X., Mao, C., Shi, M., & Zhu, B. (2024). Systematic analyses of AISNPs screening and classification algorithms based on genome-wide data for forensic biogeographic ancestry inference. *Forensic Science International*, 357, Article 111975. <https://doi.org/10.1016/j.forsciint.2024.111975>
- Calafell, F., & Larmuseau, M. H. D. (2017). The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research. *Human Genetics*, 136(5), 559–573. <https://doi.org/10.1007/s00439-016-1740-0>
- Canturk, K. M., Emre, R., Gurkan, C., Komur, I., Muslumanoglu, O., & Dogan, M. (2016). An incest case with three biological brothers as alleged fathers: Even 22 autosomal STR loci analysis would not suffice without the mother. *Medicine, Science and the Law*, 56(3), 210–212. <https://doi.org/10.1177/0025802415605539>
- Cerri, N., Presciuttini, S., Verzeletti, A., Notarangelo, L. D., & De Ferrari, F. (2006). Incest by father or by brother? A case report. *International Congress Series*, 1288, 453–455. <https://doi.org/10.1016/j.ics.2005.08.027>
- Charlesworth, B., & Charlesworth, D. (2017). Population genetics from 1966 to 2016. *Heredity*, 118(1), 2–9. <https://doi.org/10.1038/hdy.2016.55>
- Charlesworth, D., & Willis, J. H. (2009). The genetics of inbreeding depression. *Nature Reviews Genetics*, 10(11), 783–796. <https://doi.org/10.1038/nrg2664>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
- Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, 99(6), 323–329. <https://doi.org/10.1016/j.ygeno.2012.04.003>
- Cheung, E. Y. Y., Gahan, M. E., & McNevin, D. (2017). Prediction of biogeographical ancestry from genotype: A comparison of classifiers. *International Journal of Legal Medicine*, 131(4), 901–912. <https://doi.org/10.1007/s00414-016-1504-3>
- Cheung, E. Y. Y., Gahan, M. E., & McNevin, D. (2018). Predictive DNA analysis for biogeographical ancestry. *Australian Journal of Forensic Sciences*, 50(6), 651–658. <https://doi.org/10.1080/00450618.2017.1422021>
- Choate, P., & Sharan, R. (2021). The need to act: Incest as a crime given low priority—A view with India as an example. *Social Sciences*, 10(4), Article 142. <https://doi.org/10.3390/socsci10040142>
- Corach, D., Penacino, G., Marino, M., Polisecki, E., & Sala, A. (2003). Incestuous offspring detection inference by VNTR homozygosity increment. *International Congress Series*, 1239, 861–864. [https://doi.org/10.1016/S0531-5131\(02\)00393-X](https://doi.org/10.1016/S0531-5131(02)00393-X)
- Cosentino, E., Turrina, S., Ferrian, M., Caratti, S., & De Leo, D. (2015). Sibling incest. *Forensic Science International: Genetics Supplement Series*, 5, e472–e473. <https://doi.org/10.1016/j.fsigss.2015.09.187>
- Cosenza, G., Castellino, L., Morelli, S., Alladio, E., & Pilli, E. (2025). Kinship analysis and machine learning algorithms in forensic contexts: A new NGS panel. *Expert Systems with Applications*, 266, Article 126161. <https://doi.org/10.1016/j.eswa.2024.126161>
- Curik, I., Ferenčaković, M., & Sölkner, J. (2014). Inbreeding and runs of homozygosity: A possible solution to an old problem. *Livestock Science*, 166, 26–34. <https://doi.org/10.1016/j.livsci.2014.05.034>

- Davenport, L., Devesse, L., Satmun, S., Syndercombe Court, D., & Ballard, D. (2025). Considerations for the implementation of massively parallel sequencing into routine kinship analysis. *Genes*, 16(3), Article 238. <https://doi.org/10.3390/genes16030238>
- Dawnay, L., Flamson, R., & Dawnay, N. (2018). Expanded and mega-plex STR panels as a tool for presumptive population assignment. *Forensic Science and Criminology*, 3(1), 1–7. <https://doi.org/10.15761/FSC.1000126>
- Dean, L. (2005). *Blood groups and red cell antigens*. National Center for Biotechnology Information (US). <https://www.ncbi.nlm.nih.gov/books/NBK2261/>
- de la Puente, M., Ruiz-Ramírez, J., Ambroa-Conde, A., Xavier, C., Pardo-Seco, J., Álvarez-Dios, J., Freire-Aradas, A., Mosquera-Miguel, A., Gross, T. E., Cheung, E. Y. Y., Branicki, W., Nothnagel, M., Parson, W., Schneider, P. M., Kayser, M., Carracedo, Á., Lareu, M. V., Phillips, C., & VISAGE Consortium. (2021). Development and evaluation of the ancestry informative marker panel of the VISAGE Basic Tool. *Genes*, 12(8), Article 1284. <https://doi.org/10.3390/genes12081284>
- Del Moral, P., Nowaczyk, S., & Pashami, S. (2022). *Why is multiclass classification hard?* *IEEE Access*, 10, 80448–80462. <https://doi.org/10.1109/ACCESS.2022.3192514>
- De Meeûs, T. (2018). Revisiting FIS, FST, Wahlund effects, and null alleles. *Journal of Heredity*, 109(4), 446–456. <https://doi.org/10.1093/jhered/esx106>
- Devesse, L., Davenport, L., Syndercombe Court, D., & Ballard, D. (2023). Biogeographical ancestry estimation from autosomal short tandem repeats in the sequencing era. *Forensic Genomics*, 3(4), 123–137. <https://doi.org/10.1089/forensic.2023.0013>
- de Wit, A. K. J. G., Wagenaar, C. D., Janssen, N. A. C., Hoegen, B., van de Wetering, J., Hoofs, H., Ariëns, S., Benschop, C. C. G., & Ypma, R. J. F. (2026). Making AI accessible for forensic DNA profile analysis. *Forensic Science International: Genetics*, 81, Article 103345. <https://doi.org/10.1016/j.fsigen.2025.103345>
- Elgart, M., Lyons, G., Romero-Brufau, S., Kurniansyah, N., Brody, J. A., Guo, X., Lin, H. J., Raffield, L., Gao, Y., Chen, H., de Vries, P., Lloyd-Jones, D. M., Lange, L. A., Peloso, G. M., Fornage, M., Rotter, J. I., Rich, S. S., Morrison, A. C., Psaty, B. M., ... Sofer, T. (2022). Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Communications Biology*, 5, Article 856. <https://doi.org/10.1038/s42003-022-03812-z>
- Ellegren, H. (2004). Microsatellites: Simple sequences with complex evolution. *Nature Reviews Genetics*, 5(6), 435–445. <https://doi.org/10.1038/nrg1348>
- Emre, R., Canturk, K. M., Komur, I., Dogan, M., Demirel, H., & Baspınar, B. (2015). Evaluation of incest cases of Turkey in terms of DNA profiling difficulties. *Journal of Forensic and Legal Medicine*, 36, 16–21. <https://doi.org/10.1016/j.jflm.2015.08.005>
- Erzurumluoglu, A. M., Shihab, H. A., Rodriguez, S., Gaunt, T. R., & Day, I. N. M. (2016). Importance of genetic studies in consanguineous populations for the characterization of novel human gene functions. *Annals of Human Genetics*, 80(3), 187–196. <https://doi.org/10.1111/ahg.12150>
- European Network of Forensic Science Institutes. (2023, October 5). *Guideline for DNA database management: Review and recommendations* (DNA-GDL-004, Issue No. 001). <https://enfsi.eu/wp-content/uploads/2023/10/ENFSI-GUIDELINE-FOR-DNA-DATABASE-MANAGEMENT-REVIEW-AND-RECOMMENDATIONS.pdf>
- Evetts, I. W., & Weir, B. S. (1998). *Interpreting DNA evidence: Statistical genetics for forensic scientists*. Sinauer Associates.

- Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to quantitative genetics* (4th ed.). Longman.
- Fan, H., & Chu, J. Y. (2007). A brief review of short tandem repeat mutation. *Genomics, Proteomics & Bioinformatics*, 5(1), 7–14. [https://doi.org/10.1016/S1672-0229\(07\)60009-6](https://doi.org/10.1016/S1672-0229(07)60009-6)
- Fang, H., Hui, Q., Lynch, J., Honerlaw, J., Assimes, T. L., Huang, J., Vujkovic, M., Damrauer, S. M., Pyarajan, S., Gaziano, J. M., DuVall, S. L., O'Donnell, C. J., Cho, K., Chang, K. M., Wilson, P. W. F., Tsao, P. S., VA Million Veteran Program, Sun, Y. V., & Tang, H. (2019). Harmonizing genetic ancestry and self-identified race/ethnicity in genome-wide association studies. *The American Journal of Human Genetics*, 105(4), 763–772. <https://doi.org/10.1016/j.ajhg.2019.08.012>
- Federal Bureau of Investigation. (2024). *National DNA Index System (NDIS) operational procedures manual* (Version 13). U.S. Department of Justice. <https://le.fbi.gov/file-repository/ndis-operational-procedures-manual-version-13-070124.pdf/view>
- Fortier, A. L., Kim, J., & Rosenberg, N. A. (2020). Human-genetic ancestry inference and false positives in forensic familial searching. *G3: Genes, Genomes, Genetics*, 10(8), 2893–2902. <https://doi.org/10.1534/g3.120.401473>
- Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119–139. <https://doi.org/10.1006/jcss.1997.1504>
- Fung, W. K. (2003). User-friendly programs for easy calculations in paternity testing and kinship determinations. *Forensic Science International*, 136(1–3), 22–34. [https://doi.org/10.1016/S0379-0738\(03\)00218-4](https://doi.org/10.1016/S0379-0738(03)00218-4)
- Garcia, F. M., Bessa, B. G. O., dos Santos, E. V. W., Pereira, J. D. P., Alves, L. N. R., Vianna, L. A., Casotti, M. C., Trabach, R. S. R., Stange, V. S., Meira, D. D., & Louro, I. D. (2022). Forensic applications of markers present on the X chromosome. *Genes*, 13(9), Article 1597. <https://doi.org/10.3390/genes13091597>
- Garnier-Géré, P., & Chikhi, L. (2013). Population subdivision, Hardy-Weinberg equilibrium and the Wahlund effect. In *eLS*. John Wiley & Sons, Ltd. <https://doi.org/10.1002/9780470015902.a0005446.pub3>
- Ge, J., & Budowle, B. (2020). How many familial relationship testing results could be wrong? *PLOS Genetics*, 16(8), Article e1008929. <https://doi.org/10.1371/journal.pgen.1008929>
- Ge, J., Budowle, B., & Chakraborty, R. (2010). DNA identification by pedigree likelihood ratio accommodating population substructure and mutations. *Investigative Genetics*, 1, Article 8. <https://doi.org/10.1186/2041-2223-1-8>
- Gefridis, L. A., & Welch, K. E. (2011). Forensic biology: Serology and DNA. In A. Mozayani & C. Noziglia (Eds.), *The forensic laboratory handbook: Procedures and practice* (2nd ed., pp. 15–50). Humana Press. https://doi.org/10.1007/978-1-60761-872-0_2
- Gettings, K. B., Aponte, R. A., Vallone, P. M., & Butler, J. M. (2015). STR allele sequence variation: Current knowledge and future issues. *Forensic Science International: Genetics*, 18, 118–130. <https://doi.org/10.1016/j.fsigen.2015.06.005>
- Gettings, K. B., Borsuk, L. A., Steffen, C. R., Kiesler, K. M., & Vallone, P. M. (2018). Sequence-based U.S. population data for 27 autosomal STR loci. *Forensic Science International: Genetics*, 37, 106–115. <https://doi.org/10.1016/j.fsigen.2018.07.013>
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42. <https://doi.org/10.1007/s10994-006-6226-1>

- Gill, P., Benschop, C., Buckleton, J., Bleka, Ø., & Taylor, D. (2021). A review of probabilistic genotyping systems: EuroForMix, DNASTatistX and STRmix™. *Genes*, 12(10), Article 1559. <https://doi.org/10.3390/genes12101559>
- Gjertson, D. W., Brenner, C. H., Baur, M. P., Carracedo, A., Guidet, F., Luque, J. A., Lessig, R., Mayr, W. R., Pascali, V. L., Prinz, M., Schneider, P. M., & Morling, N. (2007). ISFG: Recommendations on biostatistics in paternity testing. *Forensic Science International: Genetics*, 1(3–4), 223–231. <https://doi.org/10.1016/j.fsigen.2007.06.006>
- Goli, R. C., Chishi, K. G., Ganguly, I., Singh, S., Dixit, S. P., Rath, P., Diwakar, V., Sree, C. C., Limbalkar, O. M., Sukhija, N., & Kanaka, K. K. (2024). Global and local ancestry and its importance: A review. *Current Genomics*, 25(4), 237–260. <https://doi.org/10.2174/0113892029298909240426094055>
- Gonzales, A., Guruswamy, G., & Smith, S. R. (2023). Synthetic data in health care: A narrative review. *PLOS Digital Health*, 2(1), Article e0000082. <https://doi.org/10.1371/journal.pdig.0000082>
- Graydon, M., Cholette, F., & Ng, L.-K. (2009). Inferring ethnicity using 15 autosomal STR loci: Comparisons among populations of similar and distinctly different physical traits. *Forensic Science International: Genetics*, 3(4), 251–254. <https://doi.org/10.1016/j.fsigen.2009.03.002>
- Haas, R. J., McCarty, C. A., & Payseur, B. A. (2013). Genetic ancestry inference using support vector machines, and the active emergence of a unique American population. *European Journal of Human Genetics*, 21(5), 554–562. <https://doi.org/10.1038/ejhg.2012.258>
- Hajiloo, M., Sapkota, Y., Mackey, J. R., Robson, P., Greiner, R., & Damaraju, S. (2013). ETHNOPRED: A novel machine learning method for accurate continental and sub-continental ancestry identification and population stratification correction. *BMC Bioinformatics*, 14, Article 61. <https://doi.org/10.1186/1471-2105-14-61>
- Halder, I., Shriver, M., Thomas, M., Fernandez, J. R., & Frudakis, T. (2008). A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: Utility and applications. *Human Mutation*, 29(5), 648–658. <https://doi.org/10.1002/humu.20695>
- Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. *Journal of Big Data*, 7, Article 94. <https://doi.org/10.1186/s40537-020-00369-8>
- Hawass, Z., Gad, Y. Z., Ismail, S., Khairat, R., Fathalla, D., Hasan, N., Ahmed, A., Elleithy, H., Ball, M., Gaballah, F., Wasef, S., Fateen, M., Amer, H., Gostner, P., Selim, A., Zink, A., & Pusch, C. M. (2010). Ancestry and pathology in King Tutankhamun's family. *JAMA*, 303(7), 638–647. <https://doi.org/10.1001/jama.2010.121>
- Hefner, J. T., Spradley, M. K., & Anderson, B. (2014). Ancestry assessment using random forest modeling. *Journal of Forensic Sciences*, 59(3), 583–589. <https://doi.org/10.1111/1556-4029.12402>
- Hess, K. R., & Wei, C. (2010). Learning curves in classification with microarray data. *Seminars in Oncology*, 37(1), 65–68. <https://doi.org/10.1053/j.seminoncol.2009.12.002>
- Hill, C. R., Duewer, D. L., Kline, M. C., Coble, M. D., & Butler, J. M. (2013). U.S. population data for 29 autosomal STR loci. *Forensic Science International: Genetics*, 7(3), e82–e83. <https://doi.org/10.1016/j.fsigen.2012.12.004>
- Incest Trauma Centar – Beograd. (2024). *Sažetak statističkih podataka Incest Trauma Centra – Beograd koji se tiču prijavljenih slučajeva seksualnog nasilja preživljenog u detinjstvu sa teritorije Srbije u periodu maj 2022 – decembar 2023. godine.* <https://incesttraumacentar.org.rs/wordpress/wp-content/uploads/2024/04/ITC-Statisticki-podaci-2022-2023-1-1.pdf>

- Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985). Individual-specific “fingerprints” of human DNA. *Nature*, 316(6023), 76–79. <https://doi.org/10.1038/316076a0>
- Jobling, M. A., & Gill, P. (2004). Encoded evidence: DNA in forensic analysis. *Nature Reviews Genetics*, 5(10), 739–751. <https://doi.org/10.1038/nrg1455>
- Jobling, M. A., & Tyler-Smith, C. (2003). The human Y chromosome: An evolutionary marker comes of age. *Nature Reviews Genetics*, 4(8), 598–612. <https://doi.org/10.1038/nrg1124>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 3146–3154). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf
- Khubrani, Y. M., Wetton, J. H., & Jobling, M. A. (2019). Analysis of 21 autosomal STRs in Saudi Arabia reveals population structure and the influence of consanguinity. *Forensic Science International: Genetics*, 39, 97–102. <https://doi.org/10.1016/j.fsigen.2018.12.006>
- Kidd, K. K., Speed, W. C., Pakstis, A. J., Furtado, M. R., Fang, R., Madbouly, A., Maiers, M., Middha, M., Friedlaender, F. R., & Kidd, J. R. (2014). Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Science International: Genetics*, 10, 23–32. <https://doi.org/10.1016/j.fsigen.2014.01.002>
- Kimpton, C. P., Gill, P., Walton, A., Urquhart, A., Millican, E. S., & Adams, M. (1993). Automated DNA profiling employing multiplex amplification of short tandem repeat loci. *PCR Methods and Applications*, 3(1), 13–22. <https://doi.org/10.1101/gr.3.1.13>
- King, T. E., Parkin, E. J., Swinfield, G., Cruciani, F., Scozzari, R., Rosa, A., Lim, S. K., Xue, Y., Tyler-Smith, C., & Jobling, M. A. (2007). Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy. *European Journal of Human Genetics*, 15(3), 288–293. <https://doi.org/10.1038/sj.ejhg.5201771>
- Kirin, M., McQuillan, R., Franklin, C. S., Campbell, H., McKeigue, P. M., & Wilson, J. F. (2010). Genomic runs of homozygosity record population history and consanguinity. *PLOS ONE*, 5(11), Article e13996. <https://doi.org/10.1371/journal.pone.0013996>
- Kivisild, T. (2015). Maternal ancestry and population history from whole mitochondrial genomes. *Investigative Genetics*, 6, Article 3. <https://doi.org/10.1186/s13323-015-0022-2>
- Koo, K.-C. (2025). Rapid forensic ancestry inference in selected Northeast Asian populations: A Y-STR based attention-based ensemble framework for initial investigation guidance. *Frontiers in Genetics*, 16, Article 1631529. <https://doi.org/10.3389/fgene.2025.1631529>
- Koul, A., Duran, D., & Hernandez-Boussard, T. (2025). Synthetic data, synthetic trust: Navigating data challenges in the digital revolution. *The Lancet Digital Health*, 7(11), Article 100924. <https://doi.org/10.1016/j.landig.2025.100924>
- Kristensen, T. N., Pedersen, K. S., Vermeulen, C. J., & Loeschcke, V. (2010). Research on inbreeding in the ‘omic’ era. *Trends in Ecology & Evolution*, 25(1), 44–52. <https://doi.org/10.1016/j.tree.2009.06.014>
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. <https://doi.org/10.1007/978-1-4614-6849-3>
- Lazaridis, I., Patterson, N., Mitnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P. H., Schraiber, J. G., Castellano, S., Lipson, M., Berger, B., Economou, C., Bollongino, R., Fu, Q., Bos, K. I., Nordenfelt, S., Li, H., de Filippo, C., Prüfer, K., ... Krause, J. (2014). Ancient human

genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518), 409–413. <https://doi.org/10.1038/nature13673>

Levinson, G., & Gutman, G. A. (1987). Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, 4(3), 203–221. <https://doi.org/10.1093/oxfordjournals.molbev.a040442>

Li, R., Zang, Y., Liu, J., Wu, E., Wu, R., & Sun, H. (2025). Inferring the degree of relatedness and kinship types using an all-in-one marker set. *Genes*, 16(4), Article 455. <https://doi.org/10.3390/genes16040455>

Linck, E. B., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources*, 19(3), 639–647. <https://doi.org/10.1111/1755-0998.12995>

Lipson, M., Cheronet, O., Mallick, S., Rohland, N., Oxenham, M., Pietrusewsky, M., Pryce, T. O., Willis, A., Matsumura, H., Buckley, H., Domett, K., Nguyen, G. H., Trinh, H. H., Kyaw, A. A., Win, T. T., Pradier, B., Broomandkhoshbacht, N., Candilio, F., Changmai, P., ... Reich, D. (2018). Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science*, 361(6397), 92–95. <https://doi.org/10.1126/science.aat3188>

Londin, E. R., Keller, M. A., Maista, C., Smith, G., Mamounas, L. A., Zhang, R., Madore, S. J., Gwinn, K., & Corriveau, R. A. (2010). CoAIMs: A cost-effective panel of ancestry informative markers for determining continental origins. *PLOS ONE*, 5(10), Article e13443. <https://doi.org/10.1371/journal.pone.0013443>

Lowe, A. L., Urquhart, A., Foreman, L. A., & Evett, I. W. (2001). Inferring ethnic origin by means of an STR profile. *Forensic Science International*, 119(1), 17–22. [https://doi.org/10.1016/S0379-0738\(00\)00387-X](https://doi.org/10.1016/S0379-0738(00)00387-X)

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Lynch, M. T., Maloney, K. A., Xu, H., Perry, J. A., Regeneron Genetics Center, Shuldiner, A. R., & Mitchell, B. D. (2023). Associations of genome-wide and regional autozygosity with 96 complex traits in Old Order Amish. *BMC Genomics*, 24, Article 134. <https://doi.org/10.1186/s12864-023-09208-5>

Macan, M., Uvodić, P., & Botica, V. (2003). Paternity testing in case of brother-sister incest. *Croatian Medical Journal*, 44(3), 347–349.

Mayo, O. (2008). A century of Hardy-Weinberg equilibrium. *Twin Research and Human Genetics*, 11(3), 249–256. <https://doi.org/10.1375/twin.11.3.249>

McNevin, D. (2020). Forensic inference of biogeographical ancestry from genotype: The Genetic Ancestry Lab. *WIREs Forensic Science*, 2(2), Article e1356. <https://doi.org/10.1002/wfs2.1356>

McQuillan, R., Eklund, N., Pirastu, N., Kuningas, M., McEvoy, B. P., Esko, T., Corre, T., Davies, G., Kaakinen, M., Lyytikäinen, L.-P., Kristiansson, K., Havulinna, A. S., Gögele, M., Vitart, V., Tenesa, A., Aulchenko, Y. S., Hayward, C., Johansson, Å., Boban, M., ... Wilson, J. F. (2012). Evidence of inbreeding depression on human height. *PLOS Genetics*, 8(7), Article e1002655. <https://doi.org/10.1371/journal.pgen.1002655>

McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., MacLeod, A. K., Farrington, S. M., Rudan, P., Hayward, C., Vitart, V., Rudan, I., Wild, S. H., Dunlop, M. G., Wright, A. F., ... Wilson,

- J. F. (2008). Runs of homozygosity in European populations. *The American Journal of Human Genetics*, 83(3), 359–372. <https://doi.org/10.1016/j.ajhg.2008.08.007>
- Mendes, J. M., Barbar, A., & Refaie, M. (2025). Synthetic data generation: A privacy-preserving approach to accelerate rare disease research. *Frontiers in Digital Health*, 7, Article 1563991. <https://doi.org/10.3389/fdgth.2025.1563991>
- Mengel-From, J., Børsting, C., Sanchez, J. J., Eiberg, H., & Morling, N. (2010). Human eye colour and HERC2, OCA2 and MATP. *Forensic Science International: Genetics*, 4(5), 323–328. <https://doi.org/10.1016/j.fsigen.2009.12.004>
- Moretti, T. R., Moreno, L. I., Smerick, J. B., Pignone, M. L., Hizon, R., Buckleton, J. S., Bright, J.-A., & Onorato, A. J. (2016). Population data on the expanded CODIS core STR loci for eleven populations of significance for forensic DNA analyses in the United States. *Forensic Science International: Genetics*, 25, 175–181. <https://doi.org/10.1016/j.fsigen.2016.07.022>
- Morris, J. W., Garber, R. A., D’Autremont, J., & Brenner, C. H. (1988). The avuncular index and the incest index. In B. Brinkmann & K. Henningsen (Eds.), *Advances in forensic haemogenetics 2* (pp. 607–611). Springer. https://doi.org/10.1007/978-3-642-73330-7_124
- Mullis, K. B., Faloona, F. A., Scharf, S. J., Saiki, R. K., Horn, G. T., & Erlich, H. A. (1986). Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, 51, 263–273. <https://doi.org/10.1101/SQB.1986.051.01.032>
- Myles, S., Stoneking, M., & Timpson, N. (2009). An assessment of the portability of ancestry informative markers between human populations. *BMC Medical Genomics*, 2, Article 45. <https://doi.org/10.1186/1755-8794-2-45>
- National Research Council. (1996). *The evaluation of forensic DNA evidence*. National Academies Press. <https://doi.org/10.17226/5141>
- Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., & Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, 456(7218), 98–101. <https://doi.org/10.1038/nature07331>
- Novković, T., Panić, B., Banjac, A., Kovač Đekić, T., Tomišić-Kosić, I., Vučetić-Dragović, A., Stamenković, G. G., Blagojević, J. V., Marjanović, D., & Pojskić, N. (2010). Genetic polymorphisms of 15 AmpFISTR Identifiler loci in a Serbian population. *Forensic Science International: Genetics*, 4(5), e149–e150. <https://doi.org/10.1016/j.fsigen.2009.08.012>
- Novokmet, N., Peričić Salihović, M., Škaro, V., Projić, P., Šarac, J., Havaš Auguštin, D., Rudan, P., Primorac, D., & Marjanović, D. (2022). Influence of genetic substructuring of statistical forensic parameters on genetic short tandem repeat markers in the populations of Southeastern Europe. *Croatian Medical Journal*, 63(3), 244–256. <https://doi.org/10.3325/cmj.2022.63.244>
- Novroski, N. M. M., King, J. L., Churchill, J. D., Seah, L. H., & Budowle, B. (2016). Characterization of genetic sequence variation of 58 STR loci in four major population groups. *Forensic Science International: Genetics*, 25, 214–226. <https://doi.org/10.1016/j.fsigen.2016.09.007>
- O’Connor, K. L., & Tillmar, A. O. (2012). Effect of linkage between vWA and D12S391 in kinship analysis. *Forensic Science International: Genetics*, 6(6), 840–844. <https://doi.org/10.1016/j.fsigen.2012.03.008>
- Oldoni, F., Kidd, K. K., & Podini, D. (2019). Microhaplotypes in forensic genetics. *Forensic Science International: Genetics*, 38, 54–69. <https://doi.org/10.1016/j.fsigen.2018.09.009>
- Oldt, R. F., & Kanthaswamy, S. (2020). Expanded CODIS STR allele frequencies: Evidence for the irrelevance of race-based DNA databases. *Legal Medicine*, 42, Article 101642. <https://doi.org/10.1016/j.legalmed.2019.101642>

- Onogi, A., Nurimoto, M., & Morita, M. (2011). Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods. *BMC Bioinformatics*, 12, Article 263. <https://doi.org/10.1186/1471-2105-12-263>
- Ostrer, H., & Skorecki, K. (2013). The population genetics of the Jewish people. *Human Genetics*, 132(2), 119–127. <https://doi.org/10.1007/s00439-012-1235-6>
- Pan, Z., & Xu, S. (2020). Population genomics of East Asian ethnic groups. *Hereditas*, 157, Article 49. <https://doi.org/10.1186/s41065-020-00162-w>
- Pemberton, T. J., Absher, D., Feldman, M. W., Myers, R. M., Rosenberg, N. A., & Li, J. Z. (2012). Genomic patterns of homozygosity in worldwide human populations. *The American Journal of Human Genetics*, 91(2), 275–292. <https://doi.org/10.1016/j.ajhg.2012.06.014>
- Pereira, L., Alshamali, F., Andreassen, R., Ballard, R., Chantratita, W., Cho, N. S., Coudray, C., Dugoujon, J. M., Espinoza, M., González-Andrade, F., Hadi, S., Immel, U. D., Marian, C., Gonzalez-Martin, A., Mertens, G., Parson, W., Perone, C., Prieto, L., Takeshita, H., ... Fonseca, N. A. (2011). PopAffiliator: Online calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile. *International Journal of Legal Medicine*, 125(5), 629–636. <https://doi.org/10.1007/s00414-010-0472-2>
- Phillips, C. (2015). Forensic genetic analysis of bio-geographical ancestry. *Forensic Science International: Genetics*, 18, 49–65. <https://doi.org/10.1016/j.fsigen.2015.05.012>
- Phillips, C., Fernandez-Formoso, L., Gelabert-Besada, M., Garcia-Magariños, M., Santos, C., Fondevila, M., Carracedo, Á., & Lareu, M. V. (2013). Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing. *Electrophoresis*, 34(8), 1151–1162. <https://doi.org/10.1002/elps.201200621>
- Phillips, C., Prieto, L., Fondevila, M., Salas, A., Gómez-Tato, A., Álvarez-Dios, J., Alonso, A., Blanco-Verea, A., Brión, M., Montesino, M., Carracedo, Á., & Lareu, M. V. (2009). Ancestry analysis in the 11-M Madrid bomb attack investigation. *PLOS ONE*, 4(8), Article e6583. <https://doi.org/10.1371/journal.pone.0006583>
- Phillips, C., Salas, A., Sánchez, J. J., Fondevila, M., Gómez-Tato, A., Álvarez-Dios, J., Calaza, M., Casares de Cal, M., Ballard, D., Lareu, M. V., Carracedo, Á., & SNPforID Consortium. (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Science International: Genetics*, 1(3–4), 273–280. <https://doi.org/10.1016/j.fsigen.2007.06.008>
- Polychronopoulos, V. (2012, December 11). *Applying machine learning methods to a mitochondrial DNA dataset to guess ethnicity* [Course project report]. University of California, Santa Cruz. https://users.soe.ucsc.edu/~vassilis/projects/CMPS242_ProjectReport_Polychronopoulos.pdf
- Powell, J. E., Visscher, P. M., & Goddard, M. E. (2010). Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics*, 11(11), 800–805. <https://doi.org/10.1038/nrg2865>
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945–959. <https://doi.org/10.1093/genetics/155.2.945>
- Ralph, P., & Coop, G. (2013). The geography of recent genetic ancestry across Europe. *PLOS Biology*, 11(5), Article e1001555. <https://doi.org/10.1371/journal.pbio.1001555>
- Ramos, B. R., D’Elia, M. P. B., Amador, M. A. T., Santos, N. P. C., Santos, S. E. B., da Cruz Castelli, E., Witkin, S. S., Miot, H. A., Miot, L. D. B., & da Silva, M. G. (2016). Neither self-reported ethnicity nor declared family origin are reliable indicators of genomic ancestry. *Genetica*, 144(3), 259–265. <https://doi.org/10.1007/s10709-016-9894-1>

- Rangel-Villalobos, H., Muñoz-Valle, J. F., González-Martín, A., Gorostiza, A., Magaña, M. T., & Páez-Riberos, L. A. (2008). Genetic admixture, relatedness, and structure patterns among Mexican populations revealed by the Y-chromosome. *American Journal of Physical Anthropology*, 135(4), 448–461. <https://doi.org/10.1002/ajpa.20765>
- Rebbeck, T. R., & Sankar, P. (2005). Ethnicity, ancestry, and race in molecular epidemiologic research. *Cancer Epidemiology, Biomarkers & Prevention*, 14(11), 2467–2471. <https://doi.org/10.1158/1055-9965.EPI-05-0649>
- Ren, H., An, C., Fu, W., Wu, J., Yao, W., Yu, J., & Liang, P. (2023). Prediction of local tumor progression after microwave ablation for early-stage hepatocellular carcinoma with machine learning. *Journal of Cancer Research and Therapeutics*, 19(4), 978–987. https://doi.org/10.4103/jcrt.jcrt_319_23
- Republika Srbija. (2005). Krivični zakonik (Službeni glasnik RS, br. 85/2005, 88/2005 – ispr., 107/2005 – ispr., 72/2009, 111/2009, 121/2012, 104/2013, 108/2014, 94/2016, 35/2019, 94/2024 i 63/2026). <https://www.pravno-informacioni-sistem.rs/SlGlasnikPortal/eli/rep/sgrs/skupstina/zakon/2005/85/6/reg>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939778>
- Robino, C., Barilaro, M. R., Gino, S., Chiarle, R., Palestro, G., & Torre, C. (2006). Incestuous paternity detected by STR-typing of chorionic villi isolated from archival formalin-fixed paraffin-embedded abortion material using laser microdissection. *Journal of Forensic Sciences*, 51(1), 90–92. <https://doi.org/10.1111/j.1556-4029.2005.00013.x>
- Rohlf, R. V., Fullerton, S. M., & Weir, B. S. (2012). Familial identification: Population structure and relationship distinguishability. *PLOS Genetics*, 8(2), Article e1002469. <https://doi.org/10.1371/journal.pgen.1002469>
- Rosenberg, N. A., Li, L. M., Ward, R., & Pritchard, J. K. (2003). Informativeness of genetic markers for inference of ancestry. *The American Journal of Human Genetics*, 73(6), 1402–1422. <https://doi.org/10.1086/380416>
- Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. *Science*, 298(5602), 2381–2385. <https://doi.org/10.1126/science.1078311>
- Rothhammer, F., Fehren-Schmitz, L., Puddu, G., & Capriles, J. (2017). Mitochondrial DNA haplogroup variation of contemporary mixed South Americans reveals prehistoric displacements linked to archaeologically-derived culture history. *American Journal of Human Biology*, 29(6), Article e23029. <https://doi.org/10.1002/ajhb.23029>
- Royal, C. D., Novembre, J., Fullerton, S. M., Goldstein, D. B., Long, J. C., Bamshad, M. J., & Clark, A. G. (2010). Inferring genetic ancestry: Opportunities, challenges, and implications. *The American Journal of Human Genetics*, 86(5), 661–673. <https://doi.org/10.1016/j.ajhg.2010.03.011>
- Ruiz-Linares, A., Adhikari, K., Acuña-Alonso, V., Quinto-Sánchez, M., Jaramillo, C., Arias, W., Fuentes, M., Pizarro, M., Everardo, P., de Ávila, F., Gómez-Valdés, J., León-Mimila, P., Hunemeier, T., Ramallo, V., Silva de Cerqueira, C. C., Burley, M.-W., Konca, E., de Oliveira, M. Z., Veronez, M. R., ... Gonzalez-José, R. (2014). Admixture in Latin America: Geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLOS Genetics*, 10(9), Article e1004572. <https://doi.org/10.1371/journal.pgen.1004572>
- Ruiz-Ramírez, J., de la Puente, M., Xavier, C., Ambroa-Conde, A., Álvarez-Dios, J., Freire-Aradas, A., Mosquera-Miguel, A., Ralf, A., Amory, C., Katsara, M. A., Khellaf, T., Nothnagel, M., Cheung,

- E. Y. Y., Gross, T. E., Schneider, P. M., Uacyisrael, J., Oliveira, S., Klautau-Guimarães, M. D. N., Carvalho-Gontijo, C., ... Phillips, C. (2023). Development and evaluations of the ancestry informative markers of the VISAGE Enhanced Tool for Appearance and Ancestry. *Forensic Science International: Genetics*, 64, Article 102853. <https://doi.org/10.1016/j.fsigen.2023.102853>
- Sahoo, S. A., Zaidi, A. A., Anagol, S., & Mathieson, I. (2021). Long runs of homozygosity are correlated with marriage preferences across global population samples. *Human Biology*, 93(3), 201–216. <https://doi.org/10.1353/hub.2021.0011>
- Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., & Arnheim, N. (1985). Enzymatic amplification of β -globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732), 1350–1354. <https://doi.org/10.1126/science.2999980>
- Samuel, G., & Prainsack, B. (2019). Civil society stakeholder views on forensic DNA phenotyping: Balancing risks and benefits. *Forensic Science International: Genetics*, 43, Article 102157. <https://doi.org/10.1016/j.fsigen.2019.102157>
- Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3), Article 160. <https://doi.org/10.1007/s42979-021-00592-x>
- Schneider, P. M. (2009). Expansion of the European Standard Set of DNA database loci: The current situation. *Profiles in DNA*, 12(1), 6–7. <https://www.promega.com/-/media/files/resources/profiles-in-dna/1201/expansion-of-the-european-standard-set.pdf>
- Schrider, D. R., & Kern, A. D. (2018). Supervised machine learning for population genetics: A new paradigm. *Trends in Genetics*, 34(4), 301–312. <https://doi.org/10.1016/j.tig.2017.12.005>
- Scudder, N., McNevin, D., Kelty, S. F., Walsh, S. J., & Robertson, J. (2018). Forensic DNA phenotyping: Developing a model privacy impact assessment. *Forensic Science International: Genetics*, 34, 222–230. <https://doi.org/10.1016/j.fsigen.2018.03.005>
- Sessa, F., Esposito, M., Cocimano, G., Sablone, S., Karaboue, M. A. A., Chisari, M., Albano, D. G., & Salerno, M. (2024). Artificial intelligence and forensic genetics: Current applications and future perspectives. *Applied Sciences*, 14(5), Article 2113. <https://doi.org/10.3390/app14052113>
- Setser, C. H., Planz, J. V., Barber, R. C., Phillips, N. R., Chakraborty, R., & Cross, D. S. (2020). Differentiation of Hispanic biogeographic ancestry with 80 ancestry informative markers. *Scientific Reports*, 10, Article 7745. <https://doi.org/10.1038/s41598-020-64245-4>
- Shraga, R., Yarnall, S., Elango, S., Manoharan, A., Rodriguez, S. A., Bristow, S. L., Kumar, N., Niknazar, M., Hoffman, D., Ghadir, S., Vassena, R., Chen, S. H., Hershlag, A., Grifo, J., & Puig, O. (2017). Evaluating genetic ancestry and self-reported ethnicity in the context of carrier screening. *BMC Genetics*, 18, Article 99. <https://doi.org/10.1186/s12863-017-0570-y>
- Shriver, M. D., & Kittles, R. A. (2004). Genetic ancestry and the search for personalized genetic histories. *Nature Reviews Genetics*, 5(8), 611–618. <https://doi.org/10.1038/nrg1405>
- Shukla, R., & Singh, T. R. (2024). AlzGenPred: CatBoost-based gene classifier for predicting Alzheimer's disease using high-throughput sequencing data. *Scientific Reports*, 14, Article 30294. <https://doi.org/10.1038/s41598-024-82208-x>
- Silvey, S., & Liu, J. (2024). Sample size requirements for popular classification algorithms in tabular clinical data: Empirical study. *Journal of Medical Internet Research*, 26, Article e60231. <https://doi.org/10.2196/60231>
- Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139(1), 457–462. <https://doi.org/10.1093/genetics/139.1.457>

- Slooten, K.-J., & Egeland, T. (2014). Exclusion probabilities and likelihood ratios with applications to kinship problems. *International Journal of Legal Medicine*, 128(3), 415–425. <https://doi.org/10.1007/s00414-013-0938-0>
- Sofer, T., Wong, Q., Hartwig, F. P., Taylor, K., Warren, H. R., Evangelou, E., Cabrera, C. P., Levy, D., Kramer, H., Lange, L. A., Horta, B. L., COGENT-BP Consortium, Kerr, K. F., Reiner, A. P., & Franceschini, N. (2017). Genome-wide association study of blood pressure traits by Hispanic/Latino background: The Hispanic Community Health Study/Study of Latinos. *Scientific Reports*, 7, Article 10348. <https://doi.org/10.1038/s41598-017-09019-1>
- Solanke, A. A. (2022). Explainable digital forensics AI: Towards mitigating distrust in AI-based digital forensics analysis using interpretable models. *Forensic Science International: Digital Investigation*, 42, Article 301403. <https://doi.org/10.1016/j.fsidi.2022.301403>
- Song, M., Zhao, C., Wang, Z., & Hou, Y. (2019). Applying machine learning algorithms to a real forensic case to predict Y-SNP haplogroup based on Y-STR haplotype. *Forensic Science International: Genetics Supplement Series*, 7(1), 637–638. <https://doi.org/10.1016/j.fsigs.2019.10.120>
- Steffen, C. R., Coble, M. D., Gettings, K. B., & Vallone, P. M. (2017). Corrigendum to “U.S. population data for 29 autosomal STR loci” [Forensic Sci. Int. Genet. 7 (2013) e82–e83]. *Forensic Science International: Genetics*, 31, e36–e40. <https://doi.org/10.1016/j.fsigen.2017.08.011>
- Tadmouri, G. O., Nair, P., Obeid, T., Al Ali, M. T., Al Khaja, N., & Hamamy, H. A. (2009). Consanguinity and reproductive health among Arabs. *Reproductive Health*, 6, Article 17. <https://doi.org/10.1186/1742-4755-6-17>
- Tamura, A., Tsuji, H., Miyazaki, T., Iwata, M., Nishio, H., Hashimoto, T., Kamiyama, K., & Suzuki, K. (2000). Sibling incest and formulation of paternity probability: Case report. *Legal Medicine*, 2(4), 189–196. [https://doi.org/10.1016/S1344-6223\(00\)80039-5](https://doi.org/10.1016/S1344-6223(00)80039-5)
- Tan, M., Tan, Y., Jiang, H., Xue, J., Wu, Q., Zheng, Y., Liu, G., Xiao, Y., Lv, M., Liao, M., Zhang, L., Qu, S., & Liang, W. (2025). Explainable artificial intelligence in forensic DNA analysis: Alleles identification in challenging electropherograms using supervised machine learning methods. *Forensic Science International: Genetics*, 78, Article 103289. <https://doi.org/10.1016/j.fsigen.2025.103289>
- The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74. <https://doi.org/10.1038/nature15393>
- Tillmar, A. (2010). *Populations and statistics in forensic genetics* (Linköping University Medical Dissertations No. 1175) [Doctoral dissertation, Linköping University]. Linköping University Electronic Press. <https://liu.diva-portal.org/smash/get/diva2:309703/FULLTEXT01.pdf>
- Tillmar, A. O., Kling, D., Butler, J. M., Parson, W., Prinz, M., Schneider, P. M., Egeland, T., & Gusmão, L. (2017). DNA Commission of the International Society for Forensic Genetics (ISFG): Guidelines on the use of X-STRs in kinship analysis. *Forensic Science International: Genetics*, 29, 269–275. <https://doi.org/10.1016/j.fsigen.2017.05.005>
- Trikalinos, T. A., Salanti, G., Khoury, M. J., & Ioannidis, J. P. A. (2006). Impact of violations and deviations in Hardy-Weinberg equilibrium on postulated gene-disease associations. *American Journal of Epidemiology*, 163(4), 300–309. <https://doi.org/10.1093/aje/kwj046>
- Tvedebrink, T. (2022). Review of the forensic applicability of biostatistical methods for inferring ancestry from autosomal genetic markers. *Genes*, 13(1), Article 141. <https://doi.org/10.3390/genes13010141>
- Veldhuis, M. S., Ariëns, S., Ypma, R. J. F., Abeel, T. E. P. M. F., & Benschop, C. C. G. (2022). Explainable artificial intelligence in forensics: Realistic explanations for number of contributor

- predictions of DNA profiles. *Forensic Science International: Genetics*, 56, Article 102632. <https://doi.org/10.1016/j.fsigen.2021.102632>
- Volgin, L., Taylor, D., Bright, J. A., & Lin, M. H. (2021). Validation of a neural network approach for STR typing to replace human reading. *Forensic Science International: Genetics*, 55, Article 102591. <https://doi.org/10.1016/j.fsigen.2021.102591>
- Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–887. <https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf>
- Wenk, R. E. (2008). Incest indices from microsatellite genotypes of mother-child pairs. *Transfusion*, 48(2), 341–348. <https://doi.org/10.1111/j.1537-2995.2007.01528.x>
- Wenk, R. E., Chiafari, F. A., & Houtz, T. (1994). Incest diagnosis by comparison of alleles of mother and offspring at highly heterozygous loci. *Transfusion*, 34(2), 172–175. <https://doi.org/10.1046/j.1537-2995.1994.34294143949.x>
- Wienroth, M. (2018). Governing anticipatory technology practices: Forensic DNA phenotyping and the forensic genetics community in Europe. *New Genetics and Society*, 37(2), 137–152. <https://doi.org/10.1080/14636778.2018.1469975>
- Willems, T., Gymrek, M., Highnam, G., The 1000 Genomes Project Consortium, Mittelman, D., & Erlich, Y. (2014). The landscape of human STR variation. *Genome Research*, 24(11), 1894–1904. <https://doi.org/10.1101/gr.177774.114>
- Yaghoubi, E., Yaghoubi, E., Khamees, A., & Vakili, A. H. (2024). A systematic review and meta-analysis of artificial neural network, machine learning, deep learning, and ensemble learning approaches in field of geotechnical engineering. *Neural Computing and Applications*, 36(21), 12655–12699. <https://doi.org/10.1007/s00521-024-09893-7>
- Yang, F.-C., Tseng, B., Lin, C.-Y., Yu, Y.-J., Linacre, A., & Lee, J. C.-I. (2021). Population inference based on mitochondrial DNA control region data by the nearest neighbors algorithm. *International Journal of Legal Medicine*, 135(4), 1191–1199. <https://doi.org/10.1007/s00414-021-02520-3>
- Yuan, V., Price, E. M., Del Gobbo, G., Mostafavi, S., Cox, B., Binder, A. M., Michels, K. B., Marsit, C., & Robinson, W. P. (2019). Accurate ethnicity prediction from placental DNA methylation data. *Epigenetics & Chromatin*, 12, Article 51. <https://doi.org/10.1186/s13072-019-0296-3>
- Zhu, J., Zou, H., Rosset, S., & Hastie, T. (2009). Multi-class AdaBoost. *Statistics and Its Interface*, 2(3), 349–360. <https://doi.org/10.4310/SII.2009.v2.n3.a8>

Biografija autora

Dejan Šorgić rođen je u Zadru, u Republici Hrvatskoj, a osnovnu i srednju školu završio je u Nišu. Osnovne akademske studije završio je 2014. godine na Departmanu za biologiju i ekologiju Prirodno–matematičkog fakulteta Univerziteta u Nišu. Zvanje diplomiranog biologa stekao je odbranom diplomskog rada pod nazivom „Ispitivanje akutne toksičnosti biljne vrste *Artemisia absinthium* na miševima soja BALB/c“.

Specijalističke strukovne studije iz oblasti forenzike završio je 2016. godine na Fakultetu za fizičku hemiju Univerziteta u Beogradu, odbranom specijalističkog rada pod nazivom „Primena EPR dozimetrije za kvantifikaciju izloženosti visokim dozama radioaktivnog zračenja“.

Od 2016. godine zaposlen je u Zavodu za sudsku medicinu Niš, gde se bavi poslovima iz oblasti forenzičke genetike. Član je Međunarodnog društva za forenzičku genetiku (ISFG) i Komisije za medicinska sredstva Instituta za standardizaciju Srbije. Suosnivač je bioinformatičkog startapa Peptolyte. Autor je ili koautor šest naučnih radova, od kojih su dva kategorisana kao M21a. Govori tri strana jezika.

Изјава о ауторству

Потписани—а Дејан Ј. Шоргић

број индекса Б3049/2015

Изјављујем

да је докторска дисертација под насловом

Предиктивно моделовање у форензичкој генетици: класификација етничке припадности и препознавање инцеста на основу аутозомних STR профила

1. резултат сопственог истраживачког рада,
2. да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
3. да су резултати коректно наведени и
4. да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, _____

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Дејан Ј. Шоргић

Број индекса Б3049/2015

Студијски програм Биологија

Наслов рада Предиктивно моделовање у форензичкој генетици: класификација етничке припадности и препознавање инцеста на основу аутозомних STR профила

Ментор проф др Душан Кецакевић,

Потписани/а Дејан Ј. Шоргић

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду**.

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанда

У Београду, _____

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Предиктивно моделовање у форензичкој генетици: класификација етничке припадности и препознавање инцеста на основу аутозомних STR профила

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство
2. Ауторство – некомерцијално
- ☒ 3. Ауторство – некомерцијално – без прераде
4. Ауторство – некомерцијално – делити под истим условима
5. Ауторство – без прераде
6. Ауторство – делити под истим условима

(Молимо да заокружите само једну од шест понуђених лиценци, кратак опис лиценци дат је на полеђини листа).

Потпис докторанда

У Београду, _____

1. Ауторство – Дозвољавање умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.
2. Ауторство – некомерцијално. Дозвољавање умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.
3. Ауторство – некомерцијално – без прераде. Дозвољавање умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.
4. Ауторство – некомерцијално – делити под истим условима. Дозвољавање умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.
5. Ауторство – без прераде. Дозвољавање умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.
6. Ауторство – делити под истим условима. Дозвољавање умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.