

НАСТАВНО-НАУЧНОМ ВЕЋУ УНИВЕРЗИТЕТА У БЕОГРАДУ - БИОЛОШКОГ ФАКУЛТЕТА

На IX редовној седници Наставно-научног већа Универзитета у Београду – Биолошког факултета, одржаној 15. 07. 2026. године, на основу молбе ментора, др Душана Кецкаревића, ванредног професора Универзитета у Београду – Биолошког факултета, одређена је Комисија за преглед и оцену докторске дисертације **Дејана Ј. Шоргића**, докторанда Универзитета у Београду – Биолошког факултета, запосленог у Заводу за судску медицину Ниш, под насловом: **„Предиктивно моделовање у форензичкој генетици: класификација етничке припадности и препознавање инцеста на основу аутозомних *STR* профила“**, у саставу: др Милица Кецкаревић-Марковић, ванредни професор, Универзитет у Београду – Биолошки факултет; др Михаило Јелић, ванредни професор, Универзитет у Београду – Биолошки факултет; др Маја Живковић, научни саветник, Универзитет у Београду – Институт за нуклеарне науке „Винча“, Институт од националног значаја за Републику Србију и др Александра Стефановић, научни сарадник, Универзитет у Нишу – Природно-математички факултет.

Комисија је прегледала урађену докторску дисертацију кандидата и Наставно-научном већу Универзитета у Београду – Биолошког факултета подноси следећи

ИЗВЕШТАЈ

Општи подаци о докторској дисертацији

Веће научних области природних наука Универзитета у Београду, на седници одржаној 26. 03. 2026. године, дало је сагласност на одлуку Наставно-научног већа Универзитета у Београду – Биолошког факултета о прихватању теме докторске дисертације Дејана Ј. Шоргића под насловом „Предиктивно моделовање у форензичкој генетици: класификација етничке припадности и препознавање инцеста на основу аутозомних *STR* профила“ и именовало др Душана Кецкаревића, ванредног професора Универзитета у Београду – Биолошког факултета, за ментора докторске дисертације.

Докторска дисертација Дејана Ј. Шоргића под насловом „**Предиктивно моделовање у форензичкој генетици: класификација етничке припадности и препознавање инцеста на основу аутозомних STR профила**“ представља оригинално истраживање усмерено на развој и евалуацију предиктивних модела заснованих на алгоритмима машинског учења и подацима добијеним анализом аутозомних STR маркера. Истраживање обухвата две тематске целине: развој модела за предикцију порекла на основу популационо специфичних образаца учесталости алела и развој модела за предикцију инцестуозних односа.

Рачунарски део истраживања реализован је применом скрипти написаних у програмском језику *Python*, које су извршаване у *Google Colab* окружењу. Истраживање је засновано на јавно доступним подацима о учесталостима алела аутозомних STR маркера, генерисању синтетичких STR профила, симулацији различитих образаца наслеђивања и примени алгоритама машинског учења. У оквиру прве истраживачке целине анализирани су STR профили генерисани на основу учесталости алела четири референтне популационе групе, док су у другој целини генерисани и анализирани STR профили родитеља и потомака у различитим сценаријима сродства.

Докторска дисертација има укупно 104 стране, од којих 91 страна чине нумерисани део дисертације, док уводне ненумерисане стране обухватају насловне и пратеће елементе. Дисертација је уређена у складу са упутством за техничко обликовање докторских дисертација Универзитета у Београду и садржи насловне стране на српском и енглеском језику, податке о ментору и члановима Комисије за преглед и оцену докторске дисертације, захвалницу, сажетке на српском и енглеском језику са кључним речима, податке о научној области и ужој научној области, листу скраћеница, садржај, текст докторске дисертације организован по поглављима, списак коришћене литературе, биографију аутора, изјаву о ауторству, изјаву о истоветности штампане и електронске верзије докторског рада и изјаву о коришћењу.

Текст докторске дисертације организован је у следеће целине: **Увод** (стр. 1–23), **Предмет научног истраживања** (стр. 24), **Хипотезе и циљеви истраживања** (стр. 25–26), **Материјал и методе** (стр. 27–32), **Резултати** (стр. 33–53), **Дискусија** (стр. 54–70), **Закључци** (стр. 71–72) и **Литература** (стр. 73–86). Након основног текста дисертације

налазе се биографија аутора (стр. 87) и одговарајуће изјаве прописане правилима Универзитета у Београду (стр. 88–91). Докторска дисертација садржи укупно 23 слике и 9 табела. Цитирано је 194 извора.

Анализа докторске дисертације

У докторској дисертацији под насловом „Предиктивно моделовање у форензичкој генетици: класификација етничке припадности и препознавање инцеста на основу аутозомних *STR* профила“ кандидат Дејан Ј. Шоргић испитивао је могућност примене алгоритама машинског учења у анализи стандардних аутозомних *STR* профила. Кандидат је развио, оптимизовао и упоредио више модела машинског учења, испитао утицај броја *STR* локуса, величине базе података и популационе структуре на њихове перформансе и анализирао могућности и ограничења њихове примене у форензичком контексту.

Поглавље **Увод** (стр. 1–23) организовано је у дванаест главних тематских целина, које пружају теоријску и методолошку основу за разумевање предмета истраживања. У првој целини дат је преглед историјског развоја форензичке генетике, од првих система биолошке идентификације до развоја ДНК профилисања, примене *VNTR* маркера, увођења ланчане реакције полимеразе, мултиплекс амплификације и капиларне електрофорезе.

У другој целини приказане су карактеристике аутозомних *STR* маркера. Описани су њихова структура, расподела у геному, механизми настанка варијабилности, мутационе стопе и полиморфност, као и њихов дискриминаторни потенцијал у форензичкој идентификацији. Посебна пажња посвећена је стандардизацији форензичких *STR* панела и развоју система *CODIS* и Европског стандардног сета локуса. Представљен је и значај масивног паралелног секвенцирања, које омогућава разликовање секвенци варијанти *STR* алела исте дужине и тиме повећава њихову информативност.

Трећа целина обухвата основне принципе популационе генетике релевантне за форензичку анализу и предикцију порекла. Кандидат је обрадио значај учесталости алела и генотипова, Харди–Вајнбергове равнотеже, Валундовога ефекта, генетичког дрифта и

миграција. Објашњена је разлика између алела идентичних по пореклу и алела идентичних по стању, као и веза између аутозиготности, сродства и инбридинга. Истакнуто је да се највећи део генетичке варијабилности налази унутар популација и да разлике у њиховим учесталостима алела омогућавају њихову статистичку диференцијацију.

У четвртој целини разматрана је генетичка структура савремених људских популација на континенталном, субконтиненталном и регионалном нивоу. Кандидат указује да популациона структура не подразумева постојање оштрих генетичких граница, већ да се варијабилност испољава у облику генетичких градијената и континуума. Разграничени су појмови етничке припадности и порекла, а анализиране категорије дефинисане су као референтне популационе групе. Категорија *Caucasian* означена је као северноамеричка популација европског порекла (СА-ЕП).

Пета целина посвећена је концепту биогеографског порекла и његовој примени у форензичкој генетици. Описани су методолошки и интерпретативни изазови који произилазе из популационе адмиксије, генетичких континуума и избора референтних популација. Кандидат истиче да резултати предикције не представљају непосредну одредницу културног или етничког идентитета, већ процену генетичке сличности са одабраним референтним групама.

У шестој целини представљени су генетички маркери који се користе за предикцију порекла, укључујући аутозомне маркере информативне за порекло, једнонуклеотидне полиморфизме и унипаренталне маркере Y хромозома и митохондријске ДНК. Описани су и различити статистички и рачунарски приступи популационој класификацији.

Седма целина усмерена је на примену аутозомних *STR* маркера у предикцији порекла. Кандидат даје преглед раних студија доказивања концепта, моделских приступа популационом груписању и Бајесових класификационих поступака, као и примене машинског учења у класификацији популационих група. Истакнуте су предности употребе *STR* профила који су већ доступни у форензичким базама података, али и ограничења која произилазе из релативно малог броја локуса у стандардним форензичким панелима и недовољне информативности појединих маркера за разликовање генетички блиских или адмиксованих популација.

Осма целина обухвата основе консангвинитета, инбридинга и инцеста. Објашњене су појмовне разлике између консангвинитета и инбридинга, значај коефицијента инбридинга, идентичности алела по пореклу и аутозиготности, као и биолошке последице повећане хомозиготности, укључујући инбридинг депресију. Посебно је приказан начин на који се последице инцестних односа могу испољити у аутозомним *STR* профилима кроз повећану хомозиготност, учесталије хетерозиготне подударности између родитеља и детета и повећану укупну сличност профила.

У деветој целини размотрени су специфични форензички изазови у детекцији инцестуозних односа. Кандидат указује на могућност преклапања генетичких сигнала инцестуозног укрштања и повећаног популационог инбридинга, као и на ризик од погрешне интерпретације повећане хомозиготности без разматрања популационе структуре. Приказани су и документовани форензички случајеви у којима анализа ограниченог броја аутозомних *STR* локуса није била довољна за поуздано разликовање блиско сродних могућих очева и алтернативних хипотеза о очинству и сродству, због чега је било потребно проширење панела, укључивање профила мајке или анализа додатних генетичких маркера.

Десета целина посвећена је индексу веродостојности као статистичком оквиру за интерпретацију доказа у форензичкој генетици. Описана је примена *LR* приступа у анализи очинства и других сродничких односа, као и његова ограничења у сложеним инцестуозним сценаријима. Кандидат као један од проблема издваја зависност *LR* вредности од формулисаних хипотеза, употребљених популационих учесталости и доступности профила свих релевантних особа, чиме се образлаже потреба за испитивањем комплементарних предиктивних приступа.

У једанаестој целини представљене су основе вештачке интелигенције и машинског учења, категорије алгоритама и њихове досадашње примене у форензичкој генетици. Описани су принципи рада алгоритама који су тестирани у дисертацији, са посебним освртом на *XGBoost* и *CatBoost*. Размотрене су предности ових модела у анализама великих скупова података, као и питања валидације, објашњивости и стандардизације њихове примене у форензичкој пракси.

У дванаестој целини формулисан је истраживачки проблем, са освртом на ограничења примењиваних приступа и потребу за развојем модела који могу препознати сложене мутилокусне обрасце у генетичким подацима.

Уводно поглавље је свеобухватно и логички организовано и повезује основне концепте форензичке и популационе генетике са савременим методама машинског учења. Кандидат је приказом претходних истраживања указао на методолошке празнине које се односе на употребу стандардних аутозомних *STR* профила у предикцији порекла и инцестуозних односа. На тај начин постављена је одговарајућа теоријска основа за предмет, хипотезе и циљеве докторске дисертације.

У поглављу **Предмет научног истраживања** (стр. 24) предмет дисертације дефинисан је као примена алгоритама машинског учења у анализи аутозомних *STR* профила ради предикције порекла и инцестуозних односа. Истраживање обухвата предикцију порекла на основу популационо специфичних образаца учесталости алела и развој модела за предикцију инцестуозних односа на основу појединачних *STR* профила, као и профила мајке и детета.

У поглављу **Хипотезе и циљеви истраживања** (стр. 25–26) постављене су две основне хипотезе: да аутозомни *STR* маркери садрже довољно популационо информативних података да се применом алгоритама машинског учења може предвидети порекло анализиране особе и да *STR* профили једног родитеља и детета садрже генетичке обрасце који се могу користити за предикцију инцестуозног односа. Као општи циљ дефинисан је развој и евалуација предиктивних модела заснованих на аутозомним *STR* профилима за предикцију порекла и инцестуозних односа применом алгоритама машинског учења. Специфични циљеви подељени су у две истраживачке целине.

У делу који се односи на предикцију порекла циљеви су обухватили поређење више алгоритама машинског учења и унапређење иницијално најуспешнијег модела, испитивање фактора који утичу на његове перформансе — величине базе података, информативности појединачних локуса и комбинација локуса које одговарају комерцијалним комплетима за амплификацију — процену модела у различитим класификационим сценаријима (четири референтне групе, три групе након искључења хиспаноамеричке групе и модел са

обједињеном СА-ЕП и хиспаноамеричком групом), као и примену анализе главних компоненти ради визуелизације генетичке сличности и варијабилности између група.

У делу који се односи на предикцију инцестуозних односа циљеви су обухватили поређење више алгоритама и оптимизацију хиперпараметара најуспешнијег модела, оцену његове ефикасности у различитим популационим и експерименталним условима (применом 29 и 21 локуса, у општој популацији и у популацији Саудијске Арабије са повећаном хомозиготношћу, као и на додатно модификованим учесталостима алела и у присуству једностепених мутација), испитивање могућности предикције искључиво на основу јединствених профила деце, анализу генетичких показатеља консангвинитета (хомозиготности потомака, идентичних хетерозиготних генотипова мајке и детета и укупне и кориговане сличности профила), процену утицаја броја и информативности локуса, као и примену анализе главних компоненти и вишекласне класификације ради разликовања појединачних облика инцеста. Циљеви су јасно дефинисани, међусобно повезани и усклађени са постављеним хипотезама и истраживачким проблемом. Њихова реализација омогућила је систематску оцену предиктивног потенцијала аутозомних *STR* профила у обе испитиване области.

Поглавље **Материјал и методе** (стр. 27–32) подељено је на целине у којима су описани извори података, поступци генерисања синтетичких профила, симулација наслеђивања, припрема база података, избор и оптимизација модела, оцена њихових перформанси и специфични поступци примењени у две истраживачке целине.

За предикцију порекла коришћене су јавно доступне учесталости алела 29 аутозомних *STR* локуса за четири референтне популационе групе: афроамеричку, азијску, СА-ЕП и хиспаноамеричку. Полазни подаци били су засновани на узорку од 1036 испитаника, који је обухватио 342 Афроамериканца, 361 испитаника из СА-ЕП групе, 236 Хиспаноамериканаца и 97 Азијата. На основу ових учесталости алела генерисано је по 90000 синтетичких *STR* профила за сваку групу, односно укупно 360000 профила.

У делу истраживања који се односи на предикцију инцеста коришћене су обједињене учесталости алела наведених популационих група, као и учесталости алела 21 *STR* локуса утврђене у узорку од 523 испитаника из пет региона Саудијске Арабије. Ова популација укључена је због претходно описаног дефицита хетерозиготности, који се доводи у везу са

учесталостћу консангвинитетних бракова. Ради додатног испитивања утицаја популационе структуре формирана је и модификована база, из које су уклоњени алели са учесталостима мањим од 0,003, након чега су преостале учесталости нормализоване.

Синтетички *STR* профили генерисани су помоћу скрипти написаних у програмском језику *Python*, при чему је избор алела на сваком локусу вршен у складу са њиховим учесталостима. За симулацију наслеђивања родитељски профили су насумично упаривани, а профили потомака формирану су избором по једног алела од сваког родитеља на сваком локусу. Генерисани су профили који одговарају нормалном родитељству, нормалном односу брата и сестре, инцестуозном односу отац–ћерка и инцестуозном односу брат–сестра. Ради уравнотежавања класа, у анализама је коришћено по 150000 профила из сваке категорије.

У посебном моделу симулиране су једностепене мутације применом мутационе стопе од 10^{-3} . Кандидат је развио и скрипте за израчунавање стопе хомозиготности профила потомака, учесталости идентичних хетерозиготних генотипова мајке и детета и укупне сличности два *STR* профила.

Базе података подељене су на скупове за обучавање, валидацију и тестирање модела, а ради смањења ризика од прекомерног прилагођавања примењена је десетострука стратификована унакрсна валидација. Најуспешнији алгоритми изабрани су на основу иницијалних перформанси, након чега је извршена оптимизација хиперпараметара методом мрежне претраге. За модел *XGBoost* оптимизовани су стопа учења, дубина стабала, број стабала и параметри узорковања, док су за *CatBoost* оптимизовани број итерација, дубина стабала, стопа учења и регуларизациони параметри.

Перформансе модела оцењене су на основу тачности, прецизности, одзива, F1 резултата, површине испод *ROC* криве, уравнотежене тачности, Метјузовог коефицијента корелације и Коеновог капа коефицијента. Приказане су и матрице конфузије, *ROC* криве и криве прецизности и одзива. Информативност појединачних *STR* локуса процењена је на основу њиховог доприноса одлукама модела, док је анализа главних компоненти примењена ради визуелизације генетичке удаљености испитиваних категорија.

У методолошком делу јасно су дефинисани различити сценарији предикције. За порекло су анализирани модели са четири референтне популационе групе, три групе након искључења хиспаноамеричке групе и три групе у којима су СА-ЕП и хиспаноамеричка група обједињене. За предикцију инцеста анализирани су модели за општу популацију са 29 и 21 локусом, популацију Саудијске Арабије, модификовану популацију Саудијске Арабије, модел заснован искључиво на јединственим профилима деце и модел вишекласне предикције конкретног типа инцестуозног односа. Базе података и скрипте формиране у оквиру студије јавно су доступне у одговарајућим репозиторијумима.

Примењене методе одговарају постављеним циљевима истраживања. Генерисање великих, уравнотежених и контролисаних база синтетичких *STR* профила омогућило је испитивање већег броја класификационих сценарија, док су примењени поступци унакрсне валидације и оцене модела на основу више показатеља допринели поузданијој процени њихових перформанси.

Поглавље **Резултати** (стр. 33–53) подељено је у две главне целине, које одговарају предикцији порекла и инцестуозних односа. Резултати су приказани табеларно и графички, применом матрица конфузије, *ROC* кривих, кривих прецизности и одзива, приказа информативности локуса и *PCA* дијаграма.

У првој целини упоређене су перформансе шест алгоритама машинског учења у предикцији порекла. Најуспешнији иницијални модел био је *XGBoost*, са тачношћу од 93,24% и F1 резултатом од 93,23%. Након оптимизације хиперпараметара, модел заснован на 29 *STR* локуса постигао је тачност од 94,24% при класификацији четири референтне популационе групе. Најуспешнија класификација остварена је за азијску и афроамеричку групу, док је највећи број погрешних класификација забележен између СА-ЕП и хиспаноамеричке групе. Од 9000 тестираних профила по групи, модел је правилно класификовао 8901 азијских, 8841 афроамеричких, 8216 СА-ЕП и 7970 хиспаноамеричких профила. Између СА-ЕП и хиспаноамеричке групе забележено је укупно 1511 погрешних класификација. Вредности површине испод *ROC* криве износиле су 1,00 за афроамеричку и азијску групу, 0,98 за СА-ЕП и 0,95 за хиспаноамеричку групу. Као најинформативнији локуси за предикцију порекла издвојени су *Penta E*, *F13A01* и *SE33*, док су *TPOX* и *CSF1PO* препознати као најмање информативни. Након искључивања најмање информативних

локуса укупна тачност модела смањена је за приближно један проценат, што показује да предикција не зависи искључиво од малог броја појединачних локуса, већ и од кумулативног доприноса већег броја *STR* локуса. Након искључења хиспаноамеричке групе, *XGBoost* модел је при класификацији афроамеричке, азијске и СА-ЕП групе постигао тачност од 99,06%. У моделу у којем су СА-ЕП и хиспаноамерички профили обједињени у једну категорију постигнута је тачност од 98,57%. Ови резултати показују да степен генетичке сличности значајно утиче на ефикасност класификације. *PCA* анализа је показала јасније груписање афроамеричких и азијских профила, док су се СА-ЕП и хиспаноамерички профили преклапали у централном делу дијаграма. Овакав образац у складу је са резултатима матрица конфузије и потврђује да модел успешније разликује генетички удаљеније референтне популационе групе. Испитивањем утицаја величине базе података утврђено је да је при примени 500 профила по групи укупна тачност модела износила 83%, док је повећањем броја профила достигла приближно 94%. Највећи утицај величине базе података забележен је код хиспаноамеричке групе, чија је тачност порасла за 26 процената. Разлика између модела са 10000 и 90000 профила по групи била је мања од једног процента, због чега је 10000 профила по групи препознато као приближна тачка стабилизације перформанси модела. У последњем делу ове истраживачке целине испитане су комбинације локуса које одговарају различитим комерцијалним *STR* комплетима. Комплекти *Investigator 24plex* и *GlobalFiler* постигли су тачност од 88,28% при класификацији све четири групе и 96,30% након искључења хиспаноамеричке групе. Резултати су показали да перформансе модела зависе и од броја и од информативности анализираних локуса, при чему већи број локуса не доводи нужно до веће тачности класификације.

У другој целини поглавља Резултати упоређено је једанаест алгоритама машинског учења у предикцији инцестуозних односа. Међу иницијално тестираним моделима најбоље перформансе постигао је *CatBoost*, са тачношћу од 92,35%, F1 резултатом од 92,31% и вредношћу *ROC AUC* од 97,71%. Након оптимизације хиперпараметара, овај алгоритам је коришћен у даљим анализама. У моделу за општу популацију заснованом на 29 *STR* локуса постигнута је тачност од 96,94%. Модел је правилно класификовао 9694 профила из категорије нормалног очинства и 9695 профила из категорије инцестуозног односа. Вредност *ROC AUC* износила је 99,62%, док су Метјузов коефицијент корелације и Коенов капа коефицијент износили 0,9389, што указује на висок степен сагласности између

стварних и предвиђених категорија. Анализом генетичких показатеља консангвинитета утврђено је да је просечна сличност профила мајке и детета насталог из односа отац–ћерка била за 10,14 процената већа од сличности профила родитеља и детета код нормалног очинства. Сличност профила мајке и детета насталог из односа брата и сестре била је за 7,57 процената већа од сличности профила брата и сестре у нормалном сродству. Разлике између инцестуозних и одговарајућих контролних категорија задржане су и након искључивања локуса са идентичним хетерозиготним генотиповима. Стопа хомозиготности код потомака насталих из инцестуозних односа кретала се између 28% и 48%, док се у контролној групи кретала између 7% и 30%. Учесталост локуса са идентичним хетерозиготним генотиповима мајке и детета износила је 25–27% у инцестуозним категоријама, у поређењу са 5–18,5% у категорији нормалног очинства. Ови резултати показали су да је наведене показатеље потребно тумачити заједно, а не као самосталне показатеље инцестуозног односа. Модел за општу популацију заснован на 21 *STR* локусу постигао је тачност од 95,00%, прецизност од 95,28%, F1 резултат од 94,98% и *ROC AUC* од 99,01%. Смањење броја локуса са 29 на 21 довело је до пада тачности од 1,94 процента. Као најинформативнији локуси издвојени су *SE33*, *D1S1656* и *D18S51*, док су *TH01*, *D3S1358*, *D5S818*, *CSF1PO* и *TPOX* показали најмањи допринос предикцији. Увођење једностепених мутација у профиле потомака није значајно утицало на перформансе модела. У овом сценарију постигнута је тачност од 95,02%, што је приближно једнако резултату модела без симулираних мутација. Модел заснован искључиво на профилима потомака постигао је тачност од 90,37%, прецизност од 92,82%, F1 резултат од 90,09% и *ROC AUC* од 96,34%. Овај резултат показује да профил потомка самостално садржи предиктивни сигнал који може указивати на повећану аутозиготност, али да укључивање профила мајке значајно побољшава тачност анализе. *PCA* анализа показала је делимично просторно раздвајање профила нормалног и инцестуозног сродства, уз њихово преклапање и низак удео варијабилности објашњен првим двема компонентама. При поређењу различитих типова инцеста уочено је знатно преклапање, што је указало на ограничену могућност прецизног одређивања конкретног инцестуозног односа. Модел заснован на учесталостима алела популације Саудијске Арабије постигао је тачност од 94,64%, док је модел са додатно модификованим учесталостима постигао тачност од 94,93%. Добијени резултати показују да повећана популациона хомозиготност утиче на предикцију, али да модел и у таквим условима

задржава релативно високу класификациону способност. У вишекласном моделу испитана је могућност разликовања нормалног очинства, инцеста отац–ћерка и инцеста брат–сестра. *ROC AUC* износио је 98,00% за категорију нормалног очинства и 73,00% за обе категорије инцеста. Између две категорије инцеста забележено је укупно 7496 погрешних класификација. Резултати показују да је бинарна класификација нормалног и инцестуозног односа знатно успешнија од покушаја одређивања конкретног типа инцеста на основу анализираних скупа *STR* локуса.

Резултати су систематично приказани и непосредно одговарају постављеним циљевима истраживања. Применом више показатеља перформанси кандидат је омогућио детаљну оцену предности и ограничења развијених модела, уместо ослањања искључиво на укупну тачност класификације.

Поглавље **Дискусија** (стр. 54–70) организовано је у целине које се односе на предикцију порекла, анализу инцеста, опште предности и изазове примене предиктивних модела у форензичкој генетици, ограничења студије, етичка питања и научни допринос истраживања.

У делу посвећеном предикцији порекла кандидат је резултате упоредио са претходним студијама. Висока тачност класификације афроамеричких и азијских профила протумачена је њиховом већом генетичком удаљеношћу, док су слабије перформансе при разликовању СА-ЕП и хиспаноамеричке групе повезане са сложеном демографском историјом и адмиксованом структуром хиспаноамеричких популација. Кандидат истиче да високе вредности *AUC* не морају нужно подразумевати добру класификацију сваке појединачне групе и да је моделе неопходно процењивати на основу више показатеља. Дискутован је утицај величине тренинг скупа и показано је да се потребна количина података разликује у зависности од генетичке комплексности популације. Идентификовање приближне тачке засићења на 10000 профила по групи има практичан методолошки значај, јер указује да даље повећавање базе података доноси само маргинално побољшање у односу на потребне рачунарске ресурсе. Кандидат је размотрио и значај броја и избора *STR* маркера. Резултати су упоређени са претходним студијама информативности појединачних локуса, а посебно је размотрен значај високо полиморфног маркера *SE33*. Истакнуто је да и локуси са мањим појединачним доприносом побољшавају укупну класификацију кроз

кумулятивни мутилокусни ефекат. Размотрене су могућности примене модела у форензичким истрагама, нарочито у случајевима када *STR* профил непознатог лица не даје подударане у постојећој бази података, а нема довољно преосталог биолошког материјала за додатне анализе.

У делу посвећеном анализи инцеста кандидат је резултате упоредио са стандардним приступима анализи сродства и претходно описаним форензичким случајевима. Посебно је дискутован значај кумулативног генетичког сигнала који обухвата повећану хомозиготност потомка, већи број идентичних хетерозиготних генотипова и повећану сличност профила мајке и детета. Кандидат је показао да алгоритам машинског учења може интегрисати ове обрасце кроз више локуса и остварити високу тачност бинарне класификације. Дискутована је и могућност предикције искључиво на основу профила потомка. Иако је постигнута тачност већа од 90%, кандидат закључује да је укључивање профила мајке методолошки пожељно, јер омогућава анализу дељења алела и повећава тачност модела. Маргиналан утицај симулираних једностепених мутација објашњен је чињеницом да се модел заснива на кумулативном обрасцу већег броја локуса и није условљен променом појединачног алела. Анализирано је и ограничење вишекласне предикције. Значајно генетичко преклапање између потомака насталих у односима отац–ћерка и брат–сестра онемогућило је поуздано одређивање конкретног типа инцеста, иако је бинарна предикција постојања инцестуозног односа била успешна. Кандидат на основу тога закључује да модел има већу вредност као алат за препознавање сигнала блиског инбридинга него као средство за самостално утврђивање прецизне генеалогичке структуре.

У оквиру опште дискусије размотрене су предности и ограничења примене машинског учења у форензичкој генетици. Као предности издвојене су способност препознавања сложених и нелинеарних образаца, анализа великог броја варијабли и могућност интеграције сигнала са више генетичких локуса. Као ограничења истакнути су зависност од квалитета и репрезентативности тренинг података, могућност популационе пристрасности, и потреба за спољашњом валидацијом.

Кандидат је јасно навео ограничења спроведеног истраживања. Модел за предикцију порекла развијен је на четири референтне популационе групе и показује мању ефикасност код генетички блиских и адмиксованих популација. Модел за предикцију инцеста

развијени су на синтетичким профилима и захтевају даљу валидацију на независним и реалним форензичким подацима. Популациона хомозиготност може повећати ризик од лажно позитивних резултата, док велика генетичка разноврсност може умањити уочљивост сигнала инцеста. Као правци будућих истраживања предложени су развој популационо специфичних модела, укључивање *SNP* маркера, микрохаплотипова и других генетичких система, као и примена метода објашњивог машинског учења.

Посебна целина посвећена је етичким и правним аспектима предиктивне форензичке генетике. Размотрени су заштита генетичке приватности, могућност погрешне интерпретације података о пореклу, ризик од стигматизације популационих група и осетљивост информација које указују на могући инцест. Кандидат наглашава да се развијени модели морају примењивати као допунски аналитички алати, уз стручно тумачење и потврду стандардним форензичким поступцима.

Научни допринос дисертације кандидат је дефинисао кроз интеграцију стандардних аутозомних *STR* маркера и алгоритама машинског учења, показивање да *STR* профили садрже употребљив популациони сигнал, развој модела за бинарну предикцију инцестуозних односа, систематско поређење и оптимизацију више алгоритама, развој контролисаног приступа генерисању синтетичких профила и идентификацију фактора који утичу на перформансе модела. Посебан допринос представља показивање да постојећи *STR* профили могу имати додатну аналитичку вредност изван њихове примарне употребе у идентификацији и класичној анализи сродства.

Дискусија је критички постављена и заснована на поређењу добијених резултата са другим релевантним истраживањима. Кандидат не разматра само високе перформансе модела, већ јасно указује и на њихова ограничења, услове применљивости и потребу за независном валидацијом. Тиме показује способност научне интерпретације резултата и разумевање специфичних захтева који се постављају пред методе намењене форензичкој и правној примени.

У поглављу **Закључци** (стр. 71–72), на основу постављених хипотеза и циљева истраживања, а у складу са добијеним резултатима, изведени су закључци који садржински обухватају две целине: предикцију порекла и инцестуозних односа.

У делу који се односи на предикцију порекла, *XGBoost* је издвојен као најуспешнији алгоритам, а модел заснован на аутозомним *STR* профилима постигао је тачност од 94,24% при класификацији четири референтне популационе групе. У алтернативним класификационим сценаријима постигнута је тачност од 99,06% након искључења хиспаноамеричке групе, односно 98,57% када су СА-ЕП и хиспаноамеричка група обједињене. Утврђено је да су на тачност модела утицали величина тренинг скупа, број и информативност локуса и генетичка сличност анализираних популационих група, при чему су као најинформативнији локуси издвојени *Penta E*, *F13A01* и *SE33*. Анализа главних компоненти показала је веће преклапање СА-ЕП и хиспаноамеричких профила, што је било у складу са већим бројем погрешних класификација између ових група.

У делу који се односи на предикцију инцестуозних односа, *CatBoost* је издвојен као најуспешнији алгоритам, а модел заснован на анализи *STR* профила мајке и детета применом 29 локуса постигао је тачност од 96,94%. У различитим популационим и експерименталним условима постигнуте су упоредиве вредности тачности, док увођење једностепених мутација није утицало на тачност модела. Модел заснован искључиво на *STR* профилима деце постигао је тачност од 90,37%. Анализирани генетички показатељи разликовали су се између инцестуозних и контролних категорија: хомозиготност у профилима деце из инцестуозних односа износила је 28–48%, наспрам 7–30% у контролној групи; удео идентичних хетерозиготних генотипова између мајке и детета износио је 25–27%, наспрам 5–18,5%; док је сличност профила мајке и детета била приближно 10 процената већа у случајевима инцеста. Као најинформативнији локуси издвојени су *SE33*, *DIS1656* и *D18S51*. Вишекласна анализа показала је значајно преклапање између два анализирана облика инцеста, због чега је бинарна класификација била погоднија за основни задатак разликовања инцестуозних и неинцестуозних односа.

Изведени закључци логично произлазе из добијених резултата и одговарају постављеним хипотезама и циљевима истраживања.

У поглављу **Литература** (стр. 73–86) наведене су 194 библиографске јединице. Коришћени извори обухватају научне радове, монографије и релевантне учне и институционалне публикације из области форензичке генетике, популационе генетике,

анализе сродства и машинског учења. Литературни извори су релевантни и актуелни, правилно цитирани у тексту и наведени у списку литературе.

Радови и конгресна саопштења из докторске дисертације

Б1. Радови у часописима међународног значаја

1. **Šorgić D**, Stefanović A, Keckarević D, Popović M. XGBoost as a reliable machine learning tool for predicting ancestry using autosomal STR profiles – Proof of method. *Forensic Science International: Genetics*. 2025;76:103183. DOI: 10.1016/j.fsigen.2024.103183. **M21a**
Линк ка публикацији на интернету: <https://doi.org/10.1016/j.fsigen.2024.103183>
2. **Šorgić D**, Stefanović A, Popović M, Keckarević D. From genetic data to kinship clarity: employing machine learning for detecting incestuous relations. *Frontiers in Genetics*. 2025;16:1578581. DOI: 10.3389/fgene.2025.1578581. **M22**
Линк ка публикацији на интернету: <https://doi.org/10.3389/fgene.2025.1578581>

Б2. Конгресна саопштења на скуповима међународног значаја

1. Šorgić, D., Stefanović, A., Keckarević, D., & Popović, M. (2024). Predicting ancestry with the help of autosomal STR profiles. In *Abstract Book of the 30th International Society for Forensic Genetics Congress (ISFG 2024)* (p. 523). **M34**

Провера оригиналности докторске дисертације

На основу Правилника о поступку провере оригиналности докторских дисертација које се бране на Универзитету у Београду и налаза у извештају из програма *iThenticate*, којим је извршена провера оригиналности, утврђено је подударање текста од 3 %. Увидом у Извештај утврђено је подударање са 79 сумарних извора, при чему је подударање са сваким појединачним извором било мање од 1 %. Детаљном анализом поклапања по сегментима утврђено је да се наведена преклапања односе на краће делове различитих реченица, који нису међусобно повезани и не чине смислену целину. У складу са чланом 8. став 2. Правилника о поступку провере оригиналности докторских дисертација које се бране на Универзитету у Београду, ментор др Душан Кецкаревић, ванредни професор Универзитета у Београду – Биолошког факултета, закључио је да извештај указује на оригиналност докторске дисертације кандидата Дејана Ј. Шоргића под насловом „Предиктивно моделовање у форензичкој генетици: класификација етничке припадности и препознавање инцеста на основу аутозомних *STR* профила” и препоручио да се прописани поступак припреме за њену одбрану може наставити.

Мишљење и предлог Комисије

Докторска дисертација кандидата Дејана Ј. Шоргића представља оригиналан научно-истраживачки рад из области форензичке генетике. Истраживањем је показано да стандардни аутозомни *STR* профили, поред њихове примарне примене у идентификацији и анализи сродства, садрже информације које се применом алгоритама машинског учења могу користити и за предикцију порекла и инцестуозних односа. У области предикције порекла кандидат је испитао утицај величине тренинг скупа, броја и информативности *STR* локуса и генетичке структуре референтних популационих група на перформансе предиктивних модела. Посебан оригинални научни допринос дисертације представља развој и евалуација модела машинског учења за предикцију инцестуозних односа на основу аутозомних *STR* профила.

Резултати проистекли из докторске дисертације објављени су у два оригинална научна рада у међународним часописима категорија M21a и M22, у којима је кандидат први аутор, и представљени су у виду саопштења на међународном научном скупу категорије M34. На основу прегледа докторске дисертације, примењених метода, добијених резултата и објављених научних радова, Комисија закључује да су постављени циљеви истраживања испуњени.

Комисија позитивно оцењује докторску дисертацију кандидата Дејана Ј. Шоргића под насловом „Предиктивно моделовање у форензичкој генетици: класификација етничке припадности и препознавање инцеста на основу аутозомних *STR* профила“ и предлаже Наставно-научном већу Универзитета у Београду – Биолошког факултета да прихвати Извештај и омогући кандидату јавну одбрану докторске дисертације.

У Београду, 07. 08. 2026. године

КОМИСИЈА:

др Милица Кецкаревић-Марковић, ванредни професор,
Универзитет у Београду – Биолошки факултет

др Михаило Јелић, ванредни професор,
Универзитет у Београду – Биолошки факултет

др Маја Живковић, научни саветник,
Универзитет у Београду – Институт за нуклеарне науке „Винча“,
Институт од националног значаја за Републику Србију

др Александра Стефановић, научни сарадник,
Универзитет у Нишу – Природно-математички факултет