УНИВЕРЗИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“ - БИТОЛА
ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ И КОМУНИКАЦИСКИ ТЕХНОЛОГИИ - БИТОЛА

студиска програма:
Информатички науки и компјутерско инженерство

МАШИНСКО УЧЕЊЕ ЗА КЛАСИФИКАЦИЈА НА НЕБАЛАНСИРАНИ МЕДИЦИНСКИ МНОЖЕСТВА ПОДАТОЦИ

- докторски проект –

кандидат:
м-р Анита Петреска
бр. индекс: 14/22/III/2022

Ментор:
ред. проф. д-р Благој Ристевски

Битола
септември, 2024 година

СОДРЖИНА

МАШИНСКО УЧЕЊЕ ЗА КЛАСИФИКАЦИЈА НА НЕБАЛАНСИРАНИ МЕДИЦИНСКИ МНОЖЕСТВА ПОДАТОЦИ

Анита Петреска, Универзитет „Св. Климент Охридски“ - Битола
Факултет за нформатички и комуникациски технологии – Битола
https://orcid.org/0009-0006-2098-5769
petreska.anita@uklo.edu.mk

Благој Ристевски, Универзитет „Св. Климент Охридски“ - Битола
Факултет за нформатички и комуникациски технологии – Битола
https://orcid.org/0000-0002-8356-1203
blagoj.ristevski@uklo.edu.mk

Апстракт

Во светски рамки, медицинските состојби со висок ризик претставуваат значителен предизвик за здравствените системи, бидејќи се карактеризираат со сложеност и висока стапка на смртност. Традиционалните методи за класификација на пациенти со висок ризик често се неуспешни во откривањето на ретко застапените случаи во небалансирани множества податоци, резултирајќи во ограничена точност на предвидувањата. Имајќи го ова предвид, целта на истражувањето е да се воведат напредни методи на машинско учење за подобро предвидување на состојбите кај пациентите со висок ризик, користејќи податоци од македонскиот здравствен систем.

Главното истражувачко прашање се однесува на тоа како машинското учење може да се искористи за класификација на високоризичните пациенти од небалансирани множества податоци. Примерокот се состои од реални медицински податоци собрани преку националната платформа „Мој Термин“.

Во анализата ќе се користат техники за балансирање на податоците како SMOTE и Cost-Sensitive Learning, заедно со модели како Random Forest и XGBoost. Очекувањата се дека овие модели ќе постигнат подобра чувствителност при откривање на пациенти со висок ризик. Резултатите се очекува да покажат подобрување на точноста во предвидувањето на компликациите кај високоризичните пациенти.

Очекувањата се дека примената на овие техники ќе резултира со значително подобрување на прецизноста и чувствителноста при класификација на високоризични пациенти и ќе овозможи рана дијагностика и превенција на компликации, што ќе води до подобри клинички резултати.

Ова истражување има потенцијал да предложи препораки за интеграција на вештачката интелигенција во здравствените системи, не само на национално ниво, туку и глобално, со што ќе се овозможи поефикасно управување со здравствените ресурси и автоматизација на клиничките процеси.

Клучни зборови: вештачка интелигенција, машинско учење, небалансирани множества податоци, медицински состојби со висок ризик.

ОБРАЗЛОЖЕНИЕ НА ТЕМАТА НА ДОКТОРСКИОТ ПРОЕКТ

Овој докторски проект се состои од десет глави, секоја од нив детално обработува различни аспекти на истражувањето, почнувајќи од објаснување на значењето на класификацијата на високоризични медицински состојби, па се до етичките импликации и заклучоците за понатамошен развој.

Глава еден опфаќа објаснување на значењето на класификацијата на високоризични медицински состојби, фокусирајќи се на потребата од нови и иновативни решенија преку машинско учење. Се анализираат тековните предизвици во методологијата и како машинското учење може да помогне да се надминат недостатоците при обработката на небалансираните множества податоци.

Глава два опфаќа објаснување на значењето на класификацијата на високоризични медицински состојби, фокусирајќи се на потребата од нови и иновативни решенија преку машинското учење. Се анализираат тековните предизвици во методологијата и како машинското учење може да помогне да се надминат недостатоците при обработката на небалансираните множества податоци.

Третата глава е посветена на подрачјето на истражување и го дефинира неговиот предмет. Фокусот е ставен на употребата на машинско учење во медицинската дијагностика, особено во македонскиот здравствен систем, каде што реалните медицински податоци можат да помогнат во подобрување на клиничките резултати.

Четвртата глава ги истакнува целите на истражувањето, при што главната цел е развојот на модели кои ќе овозможат рана идентификација на пациенти со висок ризик од компликации. Посебен акцент се става на употребата на техники за балансирање на податоци за да се постигне поголема точност и стабилност на предвидувањата.

Петтата глава го разгледува актуелниот контекст на истражувањето, како на глобално така и на локално ниво. Преку анализа на трендовите во примената на машинското учење во здравството, оваа глава ја истакнува важноста на истражувањето за македонскиот здравствен систем, особено во услови на ограничени ресурси.

Шестата глава претставува преглед на релевантната литература и ги разгледува различните стратегии и техники што се користат во справувањето со небалансирани множества податоци во медицината. Оваа анализа го поставува истражувањето во поширок контекст и ги истакнува најдобрите пристапи кои досега се користени во светската пракса.

Во седмата глава се објаснува методологијата на истражувањето, вклучувајќи ги техниките за собирање, обработка и анализа на податоците. Главен фокус е ставен на примената на машински алгоритми како Random Forest и XGBoost, како и на користењето на методи за балансирање на податочните множества за да се подобри прецизноста на предвидувањата.

Во осмата глава се презентираат резултатите од анализата на податоците. Перформансите на различните модели се евалуирани според клучни метрики како AUC-ROC и чувствителност. Резултатите покажуваат значителни подобрувања во идентификацијата на пациенти со висок ризик.

Деветтата глава се фокусира на етичките аспекти и импликациите од употребата на вештачка интелигенција во здравството. Особено внимание се посветува на прашањата поврзани со приватноста и заштитата на податоците, транспарентноста на алгоритмите и потребата од објаснување на одлуките што ги носат овие модели.

Во заклучокот, десетата глава ги обединува главните наоди од истражувањето и потенцира дека моделите за машинско учење значително придонесуваат за подобрување на класификацијата на високоризични медицински состојби. Исто така, се предлагаат насоки за идни истражувања, кои вклучуваат понапредни методи и техники за справување со небалансирани множества податоци.

1. ВОВЕД

Ова истражување е фокусирано на развојот и примената на машинско учење за класификација на високоризични медицински состојби од небалансирани множества податоци. Небалансираните множества податоци претставуваат еден од главните предизвици во медицинската анализа. Тие се одликуваат со нерамномерна застапеност на категориите, што подразбира дека значителен дел од пациентите не развиваат компликации, додека помалку бројната група пациенти се соочува со сериозни здравствени проблеми. Ова претставува предизвик за традиционалните алгоритми за класификација, бидејќи тие честопати покажуваат пристрасност кон мнозинските категории и не успеваат да ги идентификуваат критичните, но помалку застапени случаи ((Ahsan, 2022); (Furizal, 2023)).

Мотивацијата за ова истражување доаѓа од потребата за зголемување на прецизноста на алгоритмите за предвидување во медицинскиот сектор. Раната идентификација на пациенти со висок ризик од компликации е клучна за навремената интервенција и подобрувањето на здравствените резултати (Wehbe, 2023). Со оглед на тоа што традиционалните дијагностички методи често не се доволно ефективни кај пациенти со асимптоматски или комплексни здравствени состојби, машинското учење нуди можност за создавање на автоматизирани и интелигентни системи кои учат од големи множества податоци и можат да предвидат компликации со поголема прецизност (Hamdi, 2022).

Ова истражување е оправдано поради значителниот придонес што може да го даде во областа на медицинската дијагностика и управувањето со ризиците. Преку развој на оптимизирани модели за класификација, истражувањето ќе обезбеди нови решенија за справување со здравствените предизвици, особено во Северна Македонија, каде здравствениот систем се соочува со ограничени ресурси. Со примена на алгоритми за машинско учење како Random Forest, XGBoost и техники како SMOTE (Synthetic Minority Over-sampling Technique) и Cost-Sensitive Learning, ќе се овозможи подобра класификација на пациентите кои се соочуваат со висок ризик од компликации (Ahsan, 2022). На тој начин ќе се овозможи поефективно управување со здравствените ресурси, подобра персонализирана грижа за пациентите и намалување на трошоците за здравствената заштита.

Ограничувањата на ова истражување опфаќаат неколку клучни аспекти. Прво, примената на машинско учење бара големи количини на медицински податоци за да се развијат прецизни модели, а достапноста на такви податоци во македонскиот здравствен систем е ограничена. Второ, постои ризик од „претернираност “ (overfitting) на моделите, што може да предизвика проблеми при нивната примена на нови, невидени податоци. Дополнително, и покрај примената на техники за балансирање на податоците, како што е SMOTE, може да постои одредена пристрасност кон повеќе застапената класа, што може да резултира со намалена точност при идентификација на пациентите со највисок ризик.

Главното истражувачко прашање кое го води ова истражување е: „Како техниките на машинско учење можат да ја подобрат класификацијата на високоризични медицински состојби од небалансираните множества податоци?“ Со ова прашање се истражува способноста на машинските алгоритми да идентификуваат пациентите кои имаат најголем ризик од компликации, користејќи техники за справување со небалансираните податоци. Хипотезата е дека примената на техники како SMOTE, Ensemble Learning и Cost-Sensitive Learning ќе ја зголеми прецизноста и чувствителноста на моделите за предвидување на помалку застапените класи (односно пациентите со висок ризик).

Методологијата на истражувањето вклучува неколку фази. Прво, ќе бидат собрани и обработени реални медицински податоци од националниот здравствен систем преку платформата „Мој Термин“. Овие податоци ќе вклучуваат клинички, демографски и здравствени информации за пациенти со хронични заболувања. Потоа, ќе биде спроведено претпроцесирање на податоците, кое ќе опфати нормализација, справување со недостасувачки вредности и трансформација на категориските податоци во нумерички формат. Во следната фаза, ќе бидат применети техники за балансирање на податочните множества, како што е SMOTE, со цел да се зголеми застапеноста на помалку присутната класа.

2. ВЕШТАЧКА ИНТЕЛИГЕНЦИЈА ЗА РАНА ДИЈАГНОСТИКА ВО ЗДРАВСТВОТО

2.1. Здравствени предизвици и значење на раната дијагностика

Високо ризичните медицински состојби, како што се дијабетот и срцевите заболувања, претставуваат глобален и локален здравствен предизвик. Тие значително влијаат на квалитетот на животот и претставуваат голем товар врз здравствените системи. Според Светската здравствена организација (СЗО), бројот на лица кои се соочуваат со овие состојби драстично се зголемува.

Во Северна Македонија, високоризичните медицински состојби претставуваат значителен товар врз здравствениот систем, со висок процент на морбидитет и морталитет. Овие состојби значително го намалуваат квалитетот на животот и ги зголемуваат трошоците за здравствена грижа (Sun, 2023). Раната дијагностика на овие состојби е клучна за превенција на компликации и подобрување на резултатите кај пациентите. Сепак, традиционалните методи за дијагностика и предвидување на ризик често не се доволно прецизни и ефикасни, особено кај пациенти со повеќе ризик фактори или кај оние со асимптоматски тек на болеста (Rajpurkar P, 2022).

2.2. Вештачка интелигенција и машинско учење во здравството

Системите кои користат вештачка интелигенција (ВИ), благодарение на својата способност да учат од големи количества податоци и да ги анализираат комплексните медицински информации, овозможуваат автоматизирана дијагностика, предвидување на ризик и персонализиран третман (Briganti, 2020). Преку напредни алгоритми, ВИ може да идентификува облици и трендови во медицинските податоци коишто се тешко забележливи за луѓето, со што значително се подобрува прецизноста во дијагностичките процеси.

Една од клучните улоги на ВИ е нејзината примена во анализата на големи медицински множества податоци. ВИ може да обработи огромни количини податоци за кратко време, да ги идентификува клучните ризик фактори и да предложи соодветни третмани врз основа на предвидувањата за текот на болеста (Petreska A. &., 2023). Ова особено се покажува корисно во управувањето со високоризични медицински состојби, каде раната дијагностика и навремената интервенција се од клучно значење за подобрување на здравствените резултати (Rubinger, 2023).

Како подгрупа на ВИ, машинското учење (МУ) се наметна како моќна алатка која овозможува автономно учење од податоците без експлицитно програмирање. МУ користи различни алгоритми кои се способни да откријат скриени обрасци во големите множества податоци и да ги искористат за подобрување на предвидувањата и одлуките во медицината (Jiang, 2021). Особено важна е примената на МУ во обработката на небалансирани медицински податоци, што често се јавува кај високоризичните пациенти со хронични заболувања (Nasr, 2021).

Алгоритми на МУ заедно со техниките на длабоко учење (ДУ), овозможуваат создавање на модели кои прецизно предвидуваат компликации и идентификуваат високо ризични пациенти. Во развиените здравствени системи, овие технологии веќе се користат за откривање на пациенти со зголемен ризик од развој на сериозни здравствени состојби, што значително го подобрува резултатот на третманот (слика бр. 1).

слика бр. 1: интегративен приказ на АИ и МЛ во медицинската примена

Примената на ВИ, МУ и ДУ во здравството не е само технолошка иновација, туку и можност за значително подобрување на квалитетот на грижата за пациентите, оптимизација на ресурсите и намалување на трошоците за здравствената заштита. Ова е особено важно за земји како Северна Македонија, каде што ограничените здравствени ресурси бараат ефикасни и иновативни решенија за подобрување на здравствените услуги.

2.3. Важност за македонскиот здравствен систем

Во контекст на здравствениот систем на Северна Македонија, примената на машинско учење и вештачка интелигенција во управувањето со медицински податоци е сè уште во своите почетни фази. Иако технологијата постои и е достапна, нејзината примена во клинички услови е ограничена. Оваа дисертација претставува прв обид за примена на машинско учење во реални клинички услови за рана дијагностика и предвидување на компликации кај пациенти со високо ризични заболувања. Овој пристап не само што ќе ја подобри ефикасноста на дијагностиката и третманот, туку и ќе го намали товарот врз здравствениот систем преку подобро управување со ресурсите.

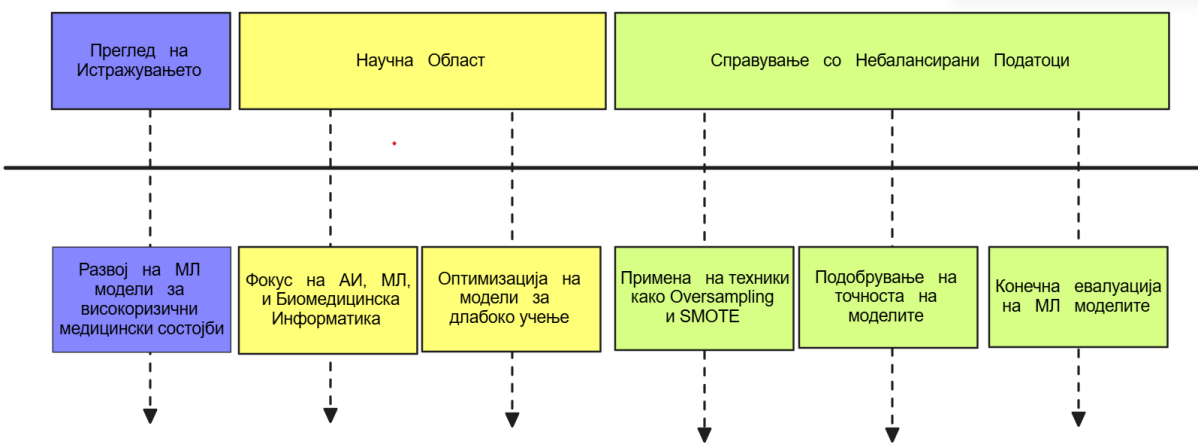
3. ПРЕДМЕТ И ОБЛАСТ НА ИСТРАЖУВАЊЕ

3.1. Преглед на истреажувањето

Главниот предмет на ова истражување е развој и примена на оптимизирани модели за класификација на високоризични медицински состојби, како што се дијабетот и срцевите заболувања, користејќи техники на машинско учење (МУ). Истражувањето се базира на обработка на реални медицински податоци за пациенти во Северна Македонија, земени од националниот здравствен систем преку платформата „Мој Термин“. Овие податоци се комплексни, небалансирани и содржат различни категории на информации, вклучувајќи демографски, клинички и лабораториски параметри (Chen, 2021); (Rawat, S. S., & Mishra, A. K., 2022).

Ова истражување има за цел да ги развие и тестира модели на МУ со цел да се подобри раната идентификација на ризичните пациенти и да се помогне во планирањето на соодветни интервенции. Моделите ќе бидат тренирани да ги анализираат податоците

и да предвидат кои пациенти имаат најголем ризик од развој на компликации. Примената на машинското учење овозможува моделирање на податоци што традиционалните статистички методи не можат ефикасно да го направат, поради комплексноста и небалансираноста на податочните множества (слика бр. 2).



слика бр. 2: развој и примена на модели за прогнозирање на медицински резултати

3.2. Научна област

Научната област на истражувањето е пресекот меѓу податочното рударење, науката за податоци, машинското учење, длабокото учење и биомедицинската информатика. Во последните години, ВИ и МУ се вградени во многу области на медицината, особено во домените каде што треба да се обработуваат големи количини на податоци, како што се електронските здравствени записи и геномските податоци (Savoska, 2023).

Овие технологии овозможуваат автоматизација на сложени процеси како што се дијагностика, класификација на болести и предвидување на здравствени резултати. Во ова истражување, фокусот е ставен на класификација на медицинските податоци и оптимизација на МУ модели. За секој модел ќе се направи хиперпараметарска оптимизација за да се постигне максимална прецизност во предвидувањето на компликации, ќе се користат различни метрики за евалуација на перформансите на моделите со цел да се обезбеди дека моделите даваат точни и применливи резултати (Shu, 2021).

3.3. Небалансирани множества податоци

Во медицинскиот контекст, небалансираните множества податоци претставуваат сериозен проблем, особено кога се класифицираат високоризични медицински состојби. Ова се случува кога бројот на случаи од помаку застапената класа е значително помал во споредба со повеќе застапената (Araf, 2024). Оваа диспропорција води до модели кои често се пристрасни кон повеќе застапената класа, што резултира со ниска точност во предвидувањето на критичните случаи.

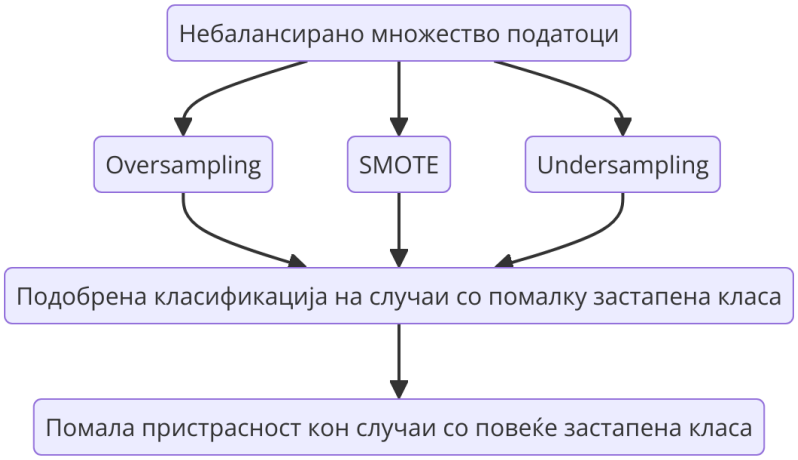
За да се реши овој проблем, применливи се неколку техники и алгоритми кои можат да ги подобрат перформансите на моделите. Ансамбл методите се базираат на комбинирање на повеќе базични модели за да се подобрат предвидувањата. Овие методи имаат посебно добра примена во справување со небалансирани множества бидејќи го намалуваат влијанието на слабите модели и можат поефикасно да се справат со сложеноста на медицинските податоци (Mienye, 2021).

Ансамбл методите како Random Forest, Gradient Boosting Machines (GBM), XGBoost и LightGBM можат да се користат за класификација на високо ризични

медицински состојби. Овие алгоритми се робусни и добро се справуваат со комплексни и хетерогени медицински множества податоци (Rawat, S. S., & Mishra, A. K., 2022). Random Forest е метод кој гради голем број на дрвја за одлуки и ги комбинира за да донесе финална одлука.

Небалансирани множества користи Weighted Random Forest, каде што секоја класа добива различна тежина и може да се додели поголема тежина на помаку застапената класа (пациенти со компликации) за да се осигурате дека моделот ќе ги третира овие пациенти подеднакво важно како и повеќе застапената класа. Ова го подобрува балансот во предвидувањата.

SMOTE (Synthetic Minority Over-sampling Technique) е техника која се користи за зголемување на застапеноста на помаку застапената класа преку синтетичко создавање на нови примероци (Das, 2022). За разлика од класичниот oversampling, каде што се дуплицираат постоечките податоци, SMOTE генерира нови примероци преку интерполација на карактеристиките на постоечките малцински податоци. Во медицинските апликации, каде што податоците за пациентите со ретки компликации се многу малку застапени, SMOTE може да се користи за зголемување на податоците од помаку застапената класа (слика бр. 3).



слика бр. 3: Примена на Oversampling, SMOTE, и Undersampling техники

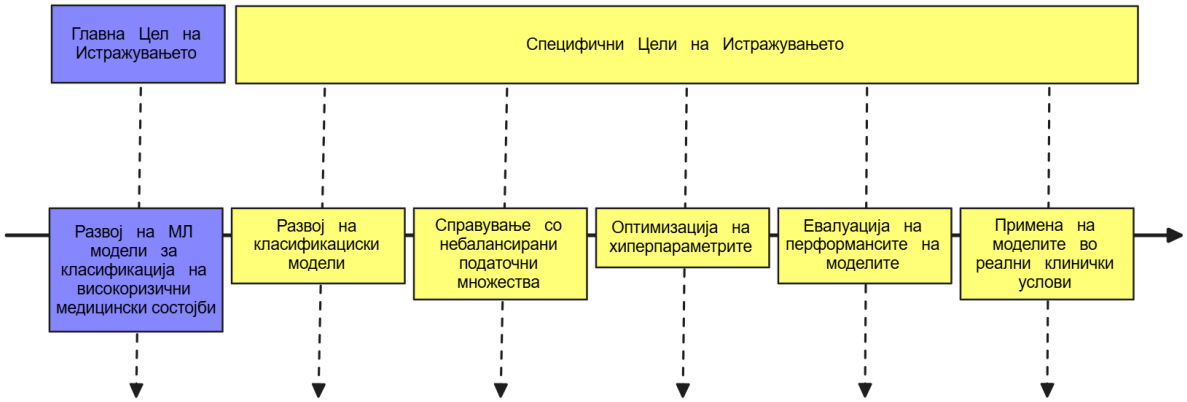
Примената на овие методи значително го подобрува квалитетот на моделите за машинско учење во медицината. Со подобро справување со небалансирани множества податоци, овие модели стануваат попрецизни и праведни, обезбедувајќи подобра идентификација на пациенти со висок ризик од компликации. Ваквите подобрувања во моделите директно придонесуваат кон поефикасно донесување на медицински одлуки, што може да има големо влијание врз навремените и точни интервенции во здравствената заштита.

4. ЦЕЛИ НА ИСТРАЖУВАЊЕТО

4.1. Главна цел на истражувањето

Главната цел на ова истражување е да се развијат и оптимизираат модели за класификација на високо ризични медицински состојби преку примена на техники на машинско учење. Моделите ќе бидат тренирани на реални медицински податоци од пациенти во Северна Македонија, собрани преку платформата „Мој Термин“, со цел да се идентификуваат пациенти со висок ризик од развој на компликации. Овие модели имаат за цел да ја подобрат прецизноста на раната дијагностика и управувањето со

хроничните болести, со што ќе се овозможи подобрување на здравствените резултати и намалување на трошоците за здравствена заштита. Моделите ќе бидат тестирани и оптимизирани со цел да се обезбеди нивната точност во класификација на пациенти и предвидување на компликации (слика бр. 4).



слика бр. 4: патека на имплементација на алгоритми за машинско учење во здравство

4.2. Специфични цели на истражувањето

За да се постигне главната цел, истражувањето е поделено на неколку специфични цели:

- Развој на модели за класификација за рана идентификација

Оваа цел се однесува на развивање на различни модели на машинско учење кои ќе бидат способни да ги класифицираат пациентите врз основа на нивните демографски и клинички податоци. Преку анализа на голем број променливи ќе се обидеме да идентификуваме кои пациенти имаат најголем ризик од развој на компликации (Iparraguirre-Villanueva, 2023).

- Справување со небалансирани множества податоци:

Една од најголемите предизвици во медицинската анализа е небалансираноста на податоците, односно присуството на многу повеќе случаи без компликации отколку случаи со сериозни компликации. Оваа цел ќе се фокусира на примена на техники за балансирање на податоците, како што се Oversampling, SMOTE и Undersampling, со цел да се подобри способноста на моделите да ги предвидат ризичните пациенти. Ќе се анализира како овие техники влијаат на перформансите на моделите и кои пристапи даваат најдобри резултати (Zhou, 2023).

- Оптимизација на хиперпараметрите на моделите:

Оптимизацијата на хиперпараметрите е клучен аспект во развојот на модели на машинско учење. Оваа цел се фокусира на примена на методи за хиперпараметарска оптимизација, како што се Grid Search и Random Search, со цел да се подобри точноста и стабилноста на моделите. Оптимизацијата ќе овозможи модели кои се прецизни и применливи во реални клинички услови (Ali, 2023).

- Евалуација на перформансите на моделите преку различни метрики:

Евалуацијата на моделите е важен аспект за да се осигура дека тие се применливи во реални клинички услови. Оваа цел ќе се фокусира на користење на различни метрики, како што се AUC-ROC, F1-score, точност и чувствителност, за да се оцени ефикасноста на моделите. Целта е да се обезбеди дека моделите даваат прецизни и репродуктивни резултати при предвидување на компликации (Vujović, 2021).

- Примена на моделите во реални клинички услови како:

Се предвидува овие модели да се тестираат во реални клинички услови. Оваа цел има за задача да го истражи потенцијалот на развиените модели за имплементација во здравствениот систем на Северна Македонија, со цел да се подобри раната дијагностика и управувањето со пациенти со хронични болести.

Во истражувањето ќе се комбинираат повеќе модели на машинско учење, вклучувајќи Random Forest, XGBoost и LightGBM, со цел да се подобри предвидувањето на високоризични медицински состојби. Овие модели ќе бидат дополнително оптимизирани со техники за балансирање на податоци, како SMOTE и Cost-Sensitive Learning, со цел да се намали пристрасноста кон повеќе застапената класа и да се подобри точноста и чувствителноста при предвидувањето на ретки, но критични случаи. Комбинирањето на овие модели и техники ќе овозможи постигнување на подобра стабилност и прецизност во класификацијата на високоризични пациенти, што ќе доведе до супериорни резултати во споредба со досегашните пристапи.

5. АКТУЕЛНОСТ НА ИСТРАЖУВАЊЕТО

5.1. Глобални трендови во примена на машинско учење во здравството

Во последните две децении, машинското учење се наметна како моќна алатка во различни научни дисциплини, вклучувајќи ја и медицината. Примената на машинско учење во здравството се зголемува брзо, особено во областите на дијагностика, персонализиран третман и предвидување на резултати од болести. На глобално ниво, напредокот на машинското учење доведе до создавање на системи кои можат да анализираат големи количини на медицински податоци за кратко време, што значително ги подобрува резултатите во клиничката пракса.

Пример за успешна примена на машинско учење е работата на Google Health и неговите партнери, каде што МУ модели се користат за предвидување на развој на хронични заболувања врз основа на податоци од електронски здравствени записи. Во оваа студија, моделите покажаа точност во предвидувањето на ризикот од срцеви заболувања и дијабет (Petreska A. R., 2023). Слични истражувања се спроведени и во други водечки здравствени институции, каде што МУ е користи за предвидување на развој на рак, автоимуни заболувања и компликации кај пациенти со хронични болести (Knevel, 2023).

Во овој контекст, примената на машинско учење во медицината не само што го подобрува здравствениот систем, туку овозможува и индивидуализиран пристап кон пациентите, со што се намалуваат трошоците и времето за дијагноза и третман. Секојдневно се појавуваат нови алгоритми и модели кои нудат иновативни решенија за предизвиците со кои се соочува здравството.

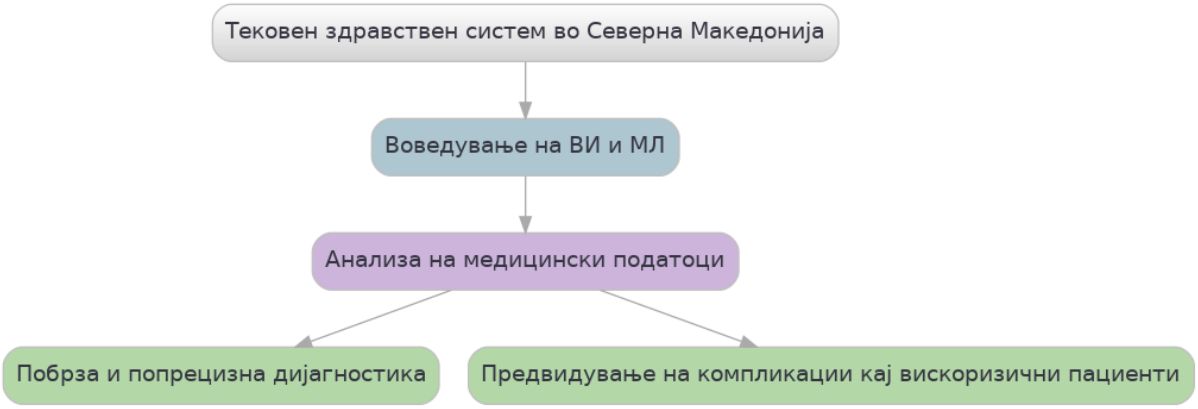
5.2. ВИ и МУ во здравството на Северна Македонија

Во Северна Македонија, здравствениот систем се соочува со бројни предизвици, вклучувајќи недоволна примена на современи технологии и методи за анализа на податоци. Ова истражување претставува прв обид за примена на машинско учење за обработка на медицински податоци од системот „Мој Термин“.

Со ова истражување, Северна Македонија ќе може да го подобри здравствениот систем преку воведување на технологии кои ќе овозможат побрза и попрецизна дијагностика на хроничните болести.

Машинското учење ќе помогне во анализа на сложените медицински податоци и ќе овозможи развој на модели кои ќе се користат за предвидување на компликации кај пациенти со вискоризични заболувања.

Ова истражување ќе ги постави темелите за понатамошен развој на МУ во здравствениот сектор на земјата, со потенцијал да се имплементира како стандард во клиничката пракса (слика бр. 5).



слика бр. 5: трансформација на здравствениот систем со воведување на ВИ и МУ

6. ПРЕГЛЕД НА ЛИТЕРАТУРАТА

Истражувањата за примена на машинско учење во класификацијата на високоризични медицински состојби се соочуваат со сериозни предизвици, како што се небалансираните множества податоци. Овој проблем доведе до развој на различни стратегии за подобрување на прецизноста и чувствителноста на моделите. Истражувањата покажуваат дека моделите како Random Forest, одлуковни дрвја и невронски мрежи се најчесто користени за обработка на медицински податоци, особено при прогноза на хронични заболувања. Истражувањата кои обработуваат податоци од хронична бубрежна болест покажаа дека супервизираното учење е особено ефикасно во класификација на овие состојби, со употреба на демографски и клинички карактеристики од болнички регистри (Sanmarchi, 2023).

Вештачката интелигенција (ВИ), особено моделите на длабоко учење, наоѓа широка примена во предвидување на кардиоваскуларни и невролошки заболувања. Според истражувањата, овие модели, заедно со примената на ВИ, овозможуваат напредни дијагностички придобивки (Rajpurkar, AI in health and medicine., 2022).

Техниките како SMOTE (Synthetic Minority Over-sampling Technique), кои креираат синтетички примероци за помалку застапените класи, со што значително се подобрува точноста и чувствителноста на моделите (Pradipta, 2021). Овие методи овозможуваат подобро справување со ретки, но критични случаи во медицинските апликации.

Студиите за предвидување на дијабетес тип 2 истакнуваат дека примена на ансамбл методи како XGBoost и Random Forest, заедно со техники за зголемување на податоците, овозможува прецизност во клиничките предвидувања. Овие методи покажуваат значајни резултати во раната дијагностика, што ги прави применливи за подобрување на здравствените резултати (Hasan, 2020).

Со ова истражување се придонесува кон полето на медицинската анализа, користејќи комбинација од овие напредни техники за да се подобри точноста на моделите и да се овозможи нивна примена во македонскиот здравствен систем..

7. МЕТОДОЛОГИЈА НА ИСТРАЖУВАЊЕ

Истражувањето е квантитативно по својата природа и вклучува примена на машинско учење за класификација на високоризични медицински состојби користејќи небалансирани множества податоци.

Истражувањето е насочено кон подобрување на предвидувањето на високоризични медицински состојби користејќи машинско учење при обработка на небалансирани податочни множества. Посебно внимание се посветува на споредба на различни техники за балансирање на податоци, како што се SMOTE и Cost-Sensitive Learning, со цел да се утврди кој пристап најмногу ја подобрува точноста и чувствителноста на моделите. Предложениот модел, кој користи комбинација на техники за машинско учење и балансирање на податоци, се очекува да покаже супериорни резултати во споредба со другите модели, особено во детекцијата на ретки и критични медицински случаи.

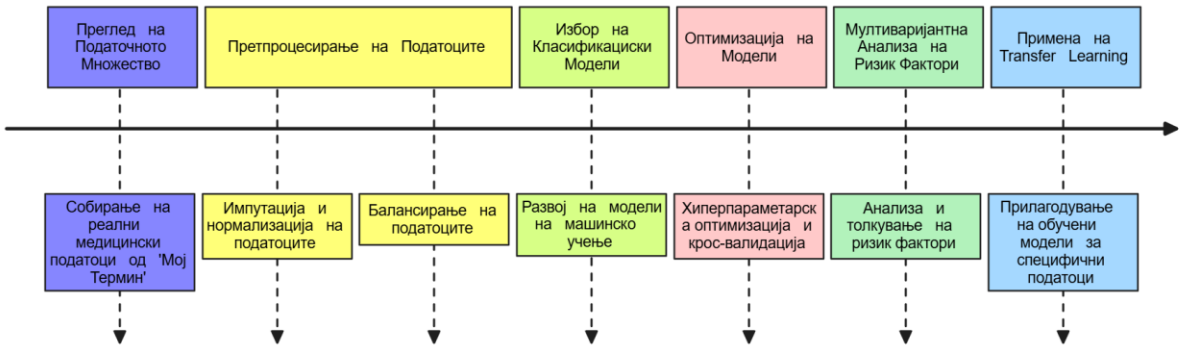
7.1. Преглед на податочното множество

Податочното множество за ова истражување вклучува пациенти од Северна Македонија со дијагностицирани хронични заболувања, како што се дијабетес и кардиоваскуларни заболувања. Податоците ќе бидат извлечени од националната здравствена платформа „Мој Термин“. Примерокот се состои од пациенти, каде што помаку застапената класа (пациенти со компликации) ќе претставува околу 10% од податоците.

Критериуми за вклучување: Пациенти со хронични заболувања, вклучени во платформата „Мој Термин“, со целосни податоци за клиничките параметри и здравствените резултати. Овие податоци се небалансирани, што е еден од најголемите предизвици во анализа на податоците, бидејќи моделите може да покажат пристрасност кон повеќе застапената група, што може да резултира со неточни предвидувања за помаку застапената група (пациенти со компликации).

7.2. Метод за собирање и анализа на податоци

Податоците за истражувањето ќе бидат извлечени од електронските медицински записи, кои ги содржат здравствените информации за пациентите, вклучувајќи клинички резултати, демографски податоци и медицинска историја. Со цел да се обезбеди точност и релевантност на анализата, податоците ќе бидат подложени на повеќе процедури за претпроцесирање и балансирање (слика бр. 6).



слика бр. 6: секвенцијални фази во процесот на машинско учење

Претпроцесирањето на податоците ќе вклучува неколку клучни чекори. Прво, ќе се спроведе импутација на податоците кои недостасуваат, со цел да се пополнат празнините во податочното множество, овозможувајќи поконзистентни и целосни информации. Нормализацијата на податоците ќе се примени за да се осигури дека сите променливи се на ист опсег на вредности, што е клучно за оптималната работа на машинските алгоритми. Дополнително, категориските податоци ќе бидат трансформирани во нумерички формат преку соодветни техники на енкодирање.

За да се справиме со дисбалансот помеѓу повеќе застапената и помаку застапената класа, ќе се применат техники за балансирање на податоците. Една од применетите методи ќе биде SMOTE (Synthetic Minority Over-sampling Technique), која синтетички генерира нови податоци за помаку застапената класа. Покрај тоа, ќе се користат и техники на Cost-Sensitive Learning, кои додаваат поголема тежина на примероците од помаку застапената класа за да се минимизира грешката при предвидувањата на критичните случаи.

Главната анализа на податоците ќе се спроведе преку алгоритми за машинско учење. За класификација на пациентите ќе се користат модели како Random Forest и XGBoost, кои се познати по нивната способност да работат со комплексни и небалансирани множества. Анализата ќе биде поделена во неколку фази. Прво, моделите ќе бидат тренирани со 70% од податоците, додека преостанатите 30% ќе бидат користени за тестирање и валидација на перформансите на моделите. Со овој пристап ќе се обезбеди дека моделите се способни за точна класификација на пациентите и предвидување на високоризичните медицински состојби.

7.3. Методолошки подобрувања и избор на софистицирани алгоритми

Истражувањето ќе користи неколку напредни методи кои се особено применливи за класификација на податоци во здравството, со посебен акцент на справувањето со небалансирани множества и комплексни податочни структури.

Еден од главните пристапи кои ќе се користат за класификација на небалансираните множества податоци е ансамблирањето, преку методи како Random Forest и Gradient Boosting Machines (GBM) (Hu, 2023), (Louk, 2023). Овие модели се познати по нивната робусност и ефикасност во обработка на комплексни множества податоци со многу карактеристики. Конкретно, XGBoost и LightGBM ќе се користат поради нивната докажана способност да ги идентификуваат помалку застапените класи, односно пациенти со сериозни компликации, и да ги минимизираат грешките при класификацијата (Airlangga, 2024). Овие методи создаваат ансамбли од повеќе "слаби" класификатори, кои колективно обезбедуваат посилни предвидувања, со што се подобрува точноста на моделот и се намалува пристрасноста кон повеќе застапената класа.

Покрај директното балансирање на податоците, ќе се применат и Cost-Sensitive Learning техники, каде што тежината на различните класи ќе се прилагоди за да се минимизира грешката при предвидување на критичните случаи. Ова ќе се постигне преку модели како Weighted Random Forest или Cost-Sensitive SVM, кои додаваат поголема казна за погрешни предвидувања на помаку застапената класа (Manokaran, 2023). Со ова, моделите ќе бидат „почувствителни“ на помаку застапената класа, што ќе овозможи подобра детекција на високоризичните пациенти. Ова е клучно во контекст на здравствената примена, каде што неоткривање на компликации може да има сериозни последици.

7.4. Оптимизација на модели

Откако моделите за класификација ќе бидат иницијално обучени, ќе се спроведе процес на хиперпараметарска оптимизација со цел да се подобрат нивните перформанси и стабилност. Овој процес опфаќа систематско истражување на различни конфигурации на хиперпараметрите, како што се бројот на дрвја во Random Forest, длабочината на дрвјата, стапката на учење во XGBoost, и други релевантни параметри зависни од применетите алгоритми. Со примена на методи како Grid Search или Random Search, ќе се идентификуваат најдобрите можни комбинации на параметрите кои овозможуваат оптимално извршување на моделите (Shekar, 2019). Овој пристап ќе овозможи модели

кои се оптимизирани за прецизни предвидувања, со подобра способност за откривање на високо ризични медицински случаи.

Преку овој процес на оптимизација, моделите ќе се подобрат во справувањето со сложеноста на медицинските податоци, ќе се минимизира ризикот од "overfitting" и ќе се обезбеди нивна способност да генерализираат врз основа на податоци што не биле дел од почетната тренинг-околина. Оптимизирањето на хиперпараметрите е критичен чекор за постигнување на највисоко ниво на прецизност, чувствителност и робустност на моделите, особено кога се применуваат во реални клинички услови.

За да се обезбеди стабилноста и генерализацијата на моделите, ќе се примени крос-валидација, со цел да се намали можноста за пристрасност и да се обезбеди пообјективна евалуација на нивната точност. Ќе се користи методот на k-fold cross-validation, каде што податочното множество ќе биде поделено на k делови (folds), и моделот ќе се тренира и тестира на различни комбинации од овие делови (Zhang, 2023). Оваа техника ќе овозможи моделите да бидат тестирани на повеќе подмножества од податоците, со што се минимизира ризикот од преференција кон одредени податочни групи и се обезбедува репродуктивност и објективност на резултатите. Крос-валидацијата е важен елемент за осигурување дека моделите можат да се применат на нови, невидени податоци, што е од клучно значење во клиничката пракса каде што точните предвидувања може да имаат директно влијание врз резултатите кај пациентите.

7.5. Мултиваријантна анализа на ризик фактори

Еден од важните аспекти на ова истражување е идентификација на ризик факторите кои имаат најголемо влијание врз развојот на компликации. Преку примена на мултиваријантна анализа, ќе се идентификуваат клучните променливи кои ќе бидат рангирани според нивната важност, што ќе им помогне на здравствените професионалци да ги насочат своите ресурси кон пациентите кои се во најголем ризик.

Резултатите од анализата ќе бидат претставени преку визуелизации, кои ќе го прикажат односот помеѓу различните фактори и развојот на компликации. Оваа анализа ќе помогне да се разбере како различни фактори заеднички влијаат врз пациентите и кои се најголемите ризици.

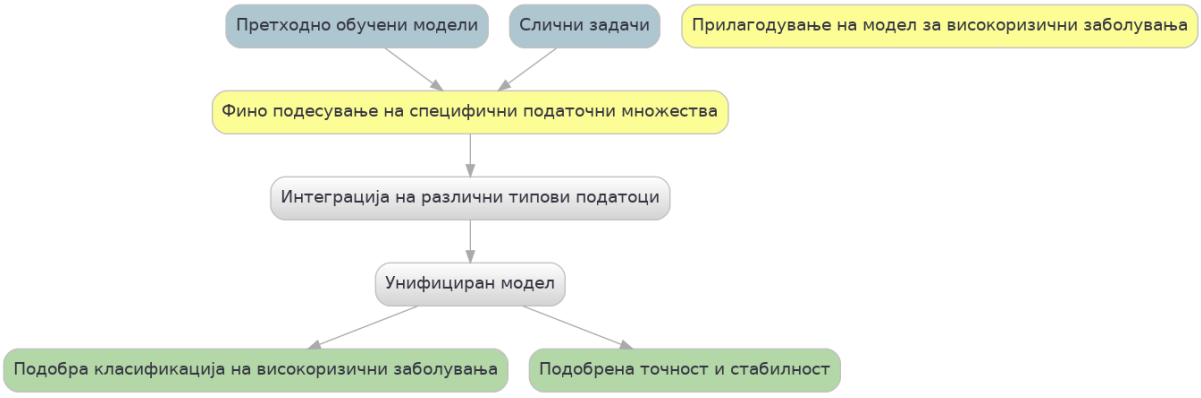
7.6. Примена на Transfer Learning

Во рамките на ова истражување ќе биде применета техниката transfer learning како напредна техника за оптимизација на моделите за класификација на високоризични заболувања (Khan, 2023). Оваа техника овозможува повторно искористување на знаењето стекнато од претходно обучени модели на слични задачи, со што значително се намалува потребата од ресурси и време за тренирање нови модели. На тој начин, transfer learning ќе придонесе за подобрување на точноста и стабилноста на моделите, особено во ситуации каде што податочните множества се ограничени или небалансирани (Alghamdi, 2024).

Една од главните предности на transfer learning е неговата способност да се прилагодува на специфични множества податоци преку подесување на веќе обучени модели. Оваа прилагодливост овозможува моделите да почнат со напредна точка на знаење, што резултира со значително подобрување на перформансите во предвидувањето на високоризични состојби. Во овој контекст, transfer learning ќе биде искористен како средство за надминување на предизвиците што произлегуваат од небалансирани множества податоци.

Оваа техника ќе биде клучна за интеграција на повеќе типови податоци, овозможувајќи создавање на унифицирани модели кои ќе ги користат предностите на различни извори на информации. На овој начин, transfer learning ќе овозможи

поефикасна анализа и подобрување на резултатите, со што ќе се обезбеди поквалитетна класификација и предвидување во областа на здравствените податоци, што е критично за точната дијагностика и управување со високоризични пациенти (слика бр. 7).

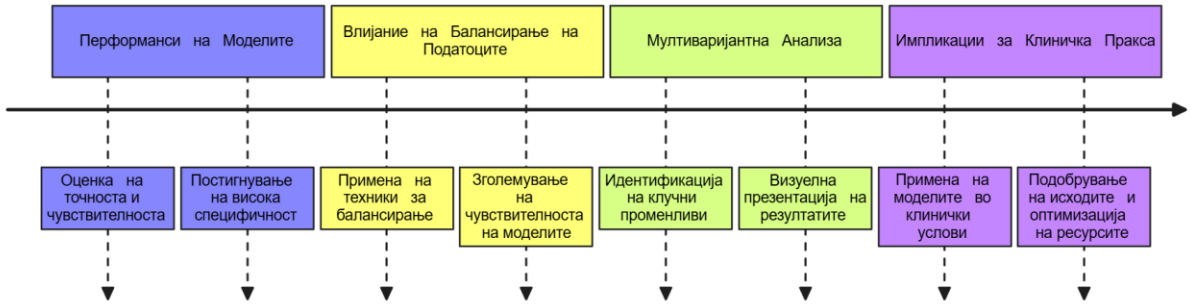


слика бр. 7: оптимизација на модели за машинско учење за класификација на високоризични заболувања

Методолошкиот пристап кој вклучува квантитативна анализа на медицински податоци преку примена на напредни техники за машинско учење ќе придонесе кон подобрување на раната дијагностика на високоризични медицински состојби. Преку ова истражување, очекуваме да демонстрираме дека машинското учење може да ја подобри точноста и чувствителноста во предвидувањата, што ќе резултира со подобра здравствена грижа и управување со ресурсите во македонскиот здравствен систем.

8. АНАЛИЗА, РЕЗУЛТАТИ И ДИСКУСИЈА

Во овој дел, резултатите од спроведената анализа ќе бидат презентирани и дискутирани со цел да се потврдат или отфрлат поставените хипотези, како и да се одговорат истражувачките прашања. Анализата е фокусирана на класификацијата на високоризични медицински состојби користејќи машинско учење, при што главната цел е да се подобри точноста и стабилноста на моделите во услови на небалансирани множества податоци (слика бр. 8).



слика бр. 8: интеграција и примена на алгоритми за машинско учење во клиничка практика

8.1. Анализа на перформансите на моделите

Во рамките на истражувањето, беа применети неколку модели на машинско учење, вклучувајќи Random Forest, XGBoost, и LightGBM, со цел да се процени нивната способност за класификација на високоризични медицински состојби. Перформансите

на моделите беа евалуирани користејќи метрики како AUC-ROC, точност, чувствителност и специфичност. Соодветно, моделите беа тренирани на 70% од достапните податоци, додека преостанатите 30% беа резервирани за тестирање и валидација на резултатите.

Резултатите покажаа дека моделите, особено XGBoost и Random Forest, постигнуваат висока точност и чувствителност, што е критично за раната идентификација на пациенти со висок ризик. AUC-ROC вредностите покажаа дека овие модели се способни да прават точни предвидувања и да ги разликуваат пациентите со и без ризик од компликации. Ова ги потврдува нашите првични хипотези дека примената на напредни алгоритми за машинско учење може да обезбеди значајно подобрување во дијагностиката на високоризични медицински состојби.

8.2. Влијание на техниките за балансирање на податоците

Со оглед на присутната небалансираност на податочните множества, односно малата застапеност на критичните случаи (помаку застапената класа), беа применети техники како SMOTE и Cost-Sensitive Learning. Тие се користеа за да се обезбеди подобра репрезентација на помаку застапената класа во податочното множество, со што се подобри способноста на моделите да идентификуваат пациенти со висок ризик од компликации.

По примената на овие техники, значително се подобри чувствителноста на моделите, особено при предвидување на ретки и критични случаи. SMOTE овозможи зголемување на примероците од помаку застапената класа, додека Cost-Sensitive Learning ги претегна примероците со поголема важност за помаку застапената класа, што го намали нивото на лажни негативни предвидувања. Овие техники, во комбинација со применетите алгоритми за машинско учење, обезбедија подобра урамнотеженост на резултатите и ја зголемија стабилноста на моделите.

8.3. Мултиваријантна анализа на ризик фактори

Мултиваријантната анализа на ризик фактори овозможи идентификација на клучните варијабли кои имаат значајно влијание врз предвидувањата на моделите. Преку мерење на важноста на променливите и анализирање на нивната корелација со резултатите, беше утврдено дека демографските податоци, како возраста и пол, заедно со клиничките резултати, како нивоата на глукоза и крвен притисок, се меѓу најзначајните фактори кои влијаат на предвидувањата за компликации.

Оваа анализа потврди дека моделите на машинско учење можат да идентификуваат специфични комбинации на фактори кои укажуваат на зголемен ризик кај одредени пациенти. Овие наоди се во согласност со постоечката литература во оваа област, која укажува дека демографските фактори, во комбинација со клинички резултати, се клучни за проценка на здравствениот ризик.

8.4. Дискусија на резултатите

Резултатите од ова истражување укажуваат на значајни подобрувања во класификацијата на високоризични медицински состојби, особено кај пациенти со хронични заболувања. Примената на техники за машинско учење, во комбинација со методите за балансирање на податоците, овозможи подобра идентификација на пациенти кои се изложени на најголем ризик. Со оглед на тоа, истражувањето ги потврдува првичните хипотези дека модерните алгоритми на машинско учење, кога се применети со соодветни техники за претпроцесирање на податоци, можат значително да ја подобрат точноста на медицинските предвидувања.

Примената на овие модели во клиничката пракса може да доведе до подобрување на раната дијагностика, навремена интервенција и, како резултат на тоа, намалување на сериозните компликации кај пациентите. Дополнително, оваа технологија може да помогне во оптимизација на здравствените ресурси преку подобра распределба на медицинската грижа кон оние пациенти кои се најмногу изложени на ризик.

8.5 Импликации и насоки за понатамошни истражувања

Ова истражување има важни импликации за здравствениот систем во Северна Македонија, особено во делот на примената на машинско учење за подобрување на клиничките резултати. Технологијата има потенцијал да ја автоматизира и подобри проценката на здравствениот ризик, што може да доведе до поефикасна организација на здравствените услуги и подобрување на грижата за пациентите.

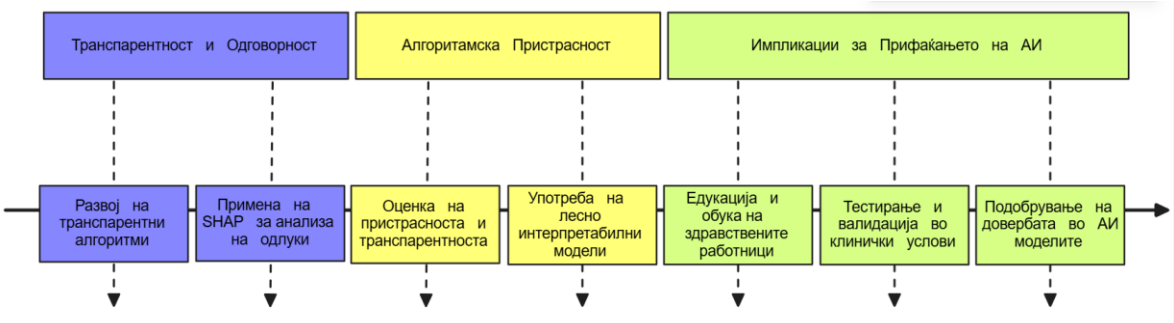
Натамошните истражувања треба да се насочат кон примена на овие модели во други клинички контексти, како и кон интеграција на нови извори на податоци, како што се геномските и сликовните податоци. Дополнително, потребно е да се истражат и други алгоритми и техники за справување со небалансирани множества податоци, со цел да се добијат уште оптимални резултати.

Истражувањето покажува дека моделите на машинско учење претставуваат ефикасна алатка за предвидување на компликации и управување со високоризични пациенти. Преку овие техники, здравствениот систем може да постигне подобри резултати во дијагностиката и третманот на пациентите.

9. ЕТИЧКИ АСПЕКТИ И ИМПЛИКАЦИИ НА ВИ ВО ЗДРАВСТВОТО

9.1. Приватност и заштита на податоци

Приватноста и заштитата на податоците претставуваат еден од најзначајните етички аспекти при примена на машинско учење во здравствениот сектор. Медицинските податоци кои ќе се користат во ова истражување содржат чувствителни информации, вклучувајќи демографски податоци, дијагнози и лабораториски резултати, што бара високо ниво на заштита и усогласеност со законските прописи. За таа цел, податоците ќе бидат целосно анонимизирани преку отстранување на сите лични идентификатори и ќе се применат техники на псевдонимизација за да се овозможи безбедна обработка, без компромитирање на приватноста на пациентите (слика бр. 9). Овие активности ќе се спроведуваат во согласност со Законот за заштита на лични податоци на Северна Македонија и Општата регулатива за заштита на податоци (GDPR) на Европската Унија (Marelli, 2020).



слика бр. 9: интегрирање на етички принципи во алгоритми за АИ

Европската Унија развива AI Act, кој ја дефинира употребата на ВИ во здравството како високо-ризична, барајќи строга усогласеност со принципите на безбедност, транспарентност и одговорност (Madiega, 2021). ISO/IEC 22989:2022

обезбедува насоки за етичка примена на ВИ, фокусирајќи се на транспарентноста и заштитата на податоците при обработка на чувствителни информации, што ќе биде од клучно значење во ова истражување.

Принципите на OECD AI Principles, поставуваат основа за праведна и транспарентна примена на ВИ системи, и со Convention 108+ на Советот на Европа, која дополнително регулира приватноста при користење на автоматизирани системи за обработка на податоци (Fritz). Придржувањето кон овие меѓународни стандарди ќе обезбеди дека податоците се користат одговорно и во согласност со највисоките етички и правни стандарди. Пристапот до податоците ќе биде строго контролиран, овозможувајќи го само на членовите на истражувачкиот тим кои се директно вклучени во проектот.

9.2. Приватност и заштита на податоци

Моделите за машинско учење често се сложени и може да биде тешко да се разбере како точно тие носат одлуки. Ова особено важи за методите на длабоко учење, каде што алгоритмите создаваат нелинеарни модели со повеќе слоеви кои ги обработуваат податоците на начин што не секогаш може лесно да се објасни.

Во рамките на ова истражување, фокусот ќе биде насочен кон транспарентноста и интерпретабилноста на моделите. Еден од етичките предизвици е објаснувањето на одлуките што ги носат алгоритмите. Со цел да се осигура транспарентност, ќе се применат алгоритми кои се лесни за интерпретација, како што се логистичка регресија и Random Forest. Разбирањето на моделите е клучно за изградба на доверба кај здравствените професионалци во одлуките што се базираат на вештачка интелигенција. Алгоритмите кои ќе се користат ќе бидат поддржани со алатки SHAP и LIME, кои ќе обезбедат транспарентност на предвидувањата и ќе им помогнат на лекарите да разберат зошто моделот донел одредена одлука. Оваа транспарентност е критична за етичка и правична примена на технологиите во здравството.

Етичкиот аспект на транспарентноста бара алгоритмите да бидат јасно документирани и резултатите да се презентираат на разбирлив начин за здравствените работници кои ќе ги користат моделите. Покрај тоа, ќе се објават сите методолошки пристапи и ќе се споделат наодите од истражувањето, што ќе овозможи надворешна ревизија и верификација на резултатите.

9.3. Алгоритамска пристрасност

Примената на машинското учење во здравството претставува сериозен предизвик во однос на транспарентноста и одговорноста, особено поради сложеноста на моделите и нивната интерпретабилност. Комплексноста на овие модели може да отежне разјаснување на процесот на донесување одлуки, што е од суштинска важност во клиничките апликации каде што прецизноста и транспарентноста се клучни. Овој проблем е особено акутен кај напредните методи, каде што нелинеарните процеси на обработка на податоци често остануваат нејасни за корисниците, што може да доведе до потешкотии при разбирање и прифаќање на резултатите (Albaroudi, 2024).

Во рамките на ова истражување, ќе биде посветено посебно внимание на обезбедување висока транспарентност и интерпретабилност на моделите. Ќе се користат методи кои овозможуваат појасно разбирање на резултатите и објаснување на основните фактори кои влијаат на одлуките донесени од системот. Овие мерки ќе помогнат да се решат етичките предизвици поврзани со разбирањето на одлуките и ќе обезбедат дека резултатите се презентирани на начин кој е соодветен за клиничките професионалци. Етичкиот аспект на транспарентноста подразбира дека алгоритмите треба да бидат јасно документирани, а методологиите детално објавени, што овозможува надворешна

ревизија и верификација на резултатите. Овој пристап е клучен за да се обезбеди доверливост во процесот и да се осигура дека сите клинички одлуки базирани на моделите можат да бидат лесно разбрани и применети од здравствените работници во реални услови.

9.4. Импликации за прифаќањето на ВИ во здравството

Една од централните етички предизвици во примената на ВИ во здравствениот сектор е степенот на прифаќање од страна на здравствените работници и пациенти. И покрај технолошките напредоци, машинското учење и вештачката интелигенција сè уште претставуваат новина во здравствената пракса, што може да предизвика скептицизам и загриженост, особено во однос на довербата и прецизноста на донесените одлуки. Здравствените работници може да изразат загриженост за вклучувањето на вештачка интелигенција во процесите на клиничко одлучување, особено ако транспарентноста и разбирливоста на алгоритмите не се задоволителни (Singh, 2024).

Во ова истражување, особено внимание ќе се посвети на транспарентноста и разбирливоста на одлуките кои ги носат моделите, со цел да се олесни нивното прифаќање во клиничката пракса. Стратегиите ќе вклучуваат обезбедување на јасни и разбирливи насоки за тоа како алгоритмите функционираат, со фокус на нивната интерпретабилност, што ќе им помогне на здравствените професионалци да стекнат доверба во технологијата. Дополнително, обуките и едукативните програми за здравствените работници ќе бидат клучен дел од процесот, овозможувајќи им полесно да ги интерпретираат резултатите од ВИ моделите и правилно да ги применуваат во клиничкиот контекст.

Со активно вклучување на здравствените работници во развојот и тестирањето на овие модели, ќе се поттикне поголема доверба во технологијата, што ќе доведе до поголема прифатеност на ВИ во секојдневната клиничка пракса. Овој пристап ќе овозможи ефикасна интеграција на системите базирани на вештачка интелигенција во здравствените процедури, на начин кој не само што подобрува точноста на одлуките, туку исто така го поттикнува нивното прифаќање и практичност во употреба.

10. ЗАКЛУЧОК

Ова истражување покажува дека употребата на машинско учење за класификација на високоризични медицински состојби, особено при работа со небалансирани податочни множества, значително го подобрува прецизноста и чувствителноста на предвидувачките модели. Техниките за справување со дисбалансираните податоци, како што се SMOTE (Synthetic Minority Over-sampling Technique), Cost-Sensitive Learning и Ensemble методите (Random Forest и XGBoost), играат клучна улога во зголемување на прецизноста на класификацијата, особено кај помаку застапената класа, што е критично за медицински апликации каде точната детекција на ризичните пациенти има директни импликации врз здравствените резултати.

Преку квантитативна анализа, истражувањето покажа дека алгоритмите како Random Forest и XGBoost се особено ефикасни во работењето со комплексни и небалансирани медицински множества податоци. Хиперпараметарската оптимизација и примената на крос-валидација овозможија моделите да бидат дополнително подобрени, обезбедувајќи не само висока прецизност туку и стабилност при класификацијата. Употребата на мултиваријантна анализа на ризик фактори обезбеди значајни согледувања во врска со тоа кои параметри најмногу придонесуваат кон зголемен ризик од компликации, што дополнително ја потврдува корисноста на овие модели во клиничката пракса.

Техниките за балансирање на податоци, како што е SMOTE, покажаа значајни придобивки во зголемување на застапеноста на помаку застапената класа и намалување на пристрасноста кон повеќе застапената класа. Ова се потврди преку зголемена чувствителност и намалена стапка на лажни негативни предвидувања, што е од клучно значење во медицински апликации каде неуспехот да се идентификува високоризичен пациент може да доведе до сериозни последици по здравјето. Покрај тоа, примената на Cost-Sensitive Learning ја овозможи оптимизацијата на моделите така што тие се повеќе „чувствителни“ на погрешните предвидувања за критичните случаи, со што дополнително се намали бројот на пропуштени високоризични пациенти.

Во однос на идни истражувања, препорачливо е да се истражат понапредни методи за справување со дисбалансирани множества податоци, како што се адаптивни техники за ресемплирање или понапредни алгоритми за длабоко учење, како што е GAN (Generative Adversarial Networks), кои имаат потенцијал да создадат синтетички податоци кои се поавтентични и покорисни за моделите за машинско учење. Исто така, примената на Transfer Learning може да обезбеди поголема флексибилност на моделите, особено во ситуации кога постои недостаток на податоци.

Практичната имплементација на овие модели во клиничката пракса треба да биде следниот чекор, со цел да се процени нивната ефикасност во реални услови. Ова би овозможило поточна проценка на нивното влијание врз здравствените резултати и на потенцијалот за интеграција на машинското учење во здравствениот систем на Северна Македонија. Идните истраживачки напори би можеле да инкорпорираат и геномски податоци и рендгенски слики, со цел да се зголеми точноста на предвидувањата и да се развијат унифицирани модели кои ќе овозможат целосен увид во здравственото состојба на пациентите.

ЛИТЕРАТУРА

Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022, March). Machine-learning-based disease diagnosis: A comprehensive review. In Healthcare (Vol. 10, No. 3, p. 541). MDPI.

Airlangga, G. (2024). Enhancing Medical Diagnostics with Ensemble Machine Learning: A Comparative Study of Gradient Boosting, XGBoost, LightGBM, and Blended Models. Jurasik (Jurnal Riset Sistem Informasi dan Teknik Informatika), 9(2), 1117-1132.

Albaroudi, E., Mansouri, T., & Alameer, A. (2024). A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring. AI, 5(1), 383-404.

Alghamdi, A., Hammad, M., Ugail, H., Abdel-Raheem, A., Muhammad, K., Khalifa, H. S., & Abd El-Latif, A. A. (2024). Detection of myocardial infarction based on novel deep transfer learning methods for urban healthcare in smart cities. Multimedia tools and applications, 1-22.

Ali, Yasser A., et al. "Hyperparameter search for machine learning algorithms for optimizing the computational complexity." Processes 11.2 (2023): 349.

Araf, I., Idri, A., & Chairi, I. (2024). Cost-sensitive learning for imbalanced medical data: a review. Artificial Intelligence Review, 57(4), 80.

Ayyagari, M. R. (2020). Classification of imbalanced datasets using one-class SVM, k-nearest neighbors and CART algorithm. International Journal of Advanced Computer Science and Applications, 11(11).

Briganti, G., & Le Moine, O. (2020). Artificial intelligence in medicine: today and tomorrow. Frontiers in medicine, 7, 509744.

Chen, Z., Duan, J., Kang, L., & Qiu, G. (2021). A hybrid data-level ensemble to enable learning from highly imbalanced dataset. Information Sciences, 554, 157-176.

Chhetri, B., Goyal, L. M., & Mittal, M. (2023). How machine learning is used to study addiction in digital healthcare: A systematic review. International Journal of Information Management Data Insights, 3(2), 100175.

Das, S., Mullick, S. S., & Zelinka, I. (2022). On supervised class-imbalanced learning: An updated perspective and some key challenges. IEEE Transactions on Artificial Intelligence, 3(6), 973-993.

Feng, F., Li, K. C., Shen, J., Zhou, Q., & Yang, X. (2020). Using cost-sensitive learning and feature selection algorithms to improve the performance of imbalanced classification. IEEE Access, 8, 69979-69996.

Fritz, J., & Giardini, T. (2024). Emerging Contours of AI Governance and the Three Layers of Regulatory Heterogeneity. Digital Policy Alert Working Paper 24-001.

Furizal, F., Ma'arif, A., & Rifaldi, D. (2023). Application of machine learning in healthcare and medicine: A review. Journal of Robotics and Control (JRC), 4(5), 621-631.

Hamdi, M., Bourouis, S., Rastislav, K., & Mohmed, F. (2022). Evaluation of neuro images for the diagnosis of Alzheimer's disease using deep learning neural network. Frontiers in Public Health, 10, 834032.

Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.

Hu, J., & Szymczak, S. (2023). A review on longitudinal data analysis with random forest. Briefings in Bioinformatics, 24(2), bbad002.

Iparraguirre-Villanueva, O., Espinola-Linares, K., Flores Castañeda, R. O., & Cabanillas-Carbonell, M. (2023). Application of machine learning models for early detection and accurate classification of type 2 diabetes. Diagnostics, 13(14), 2383.

Jiang, L., Wu, Z., Xu, X., Zhan, Y., Jin, X., Wang, L., & Qiu, Y. (2021). Opportunities and challenges of artificial intelligence in the medical field: current application, emerging problems, and problem-solving strategies. Journal of International Medical Research, 49(3), 03000605211000157.

Khan, T. A., Fatima, A., Shahzad, T., Alissa, K., Ghazal, T. M., Al-Sakhnini, M. M., ... & Ahmed, A. (2023). Secure IoMT for disease prediction empowered with transfer learning in healthcare 5.0, the concept and case study. IEEE Access, 11, 39418-39430.

Knevel, R., & Liao, K. P. (2023). From real-world electronic health record data to real-world results using artificial intelligence. Annals of the Rheumatic Diseases, 82(3), 306-311.

Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2023). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. Journal of ambient intelligence and humanized computing, 14(7), 8459-8486.

Louk, M. H. L., & Tama, B. A. (2023). Dual-IDS: A bagging-based gradient boosting decision tree model for network anomaly intrusion detection system. Expert Systems with Applications, 213, 119030.

Madiega, T. (2021). Artificial intelligence act. European Parliament: European Parliamentary Research Service.

Manokaran, J., & Vairavel, G. (2023). GIWRF-SMOTE: Gini impurity-based weighted random forest with SMOTE for effective malware attack and anomaly detection in IoT-Edge. Smart Science, 11(2), 276-292.

Marelli, L., Lievevrouw, E., & Van Hoyweghen, I. (2020). Fit for purpose? The GDPR and the governance of European digital health. Policy studies, 41(5), 447-467.

Mienye, I. D., & Sun, Y. (2021). Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. Informatics in Medicine Unlocked, 25, 100690.

Mohsen, F., Al-Absi, H. R., Yousri, N. A., El Hajj, N., & Shah, Z. (2023). A scoping review of artificial intelligence-based methods for diabetes risk prediction. NPJ Digital Medicine, 6(1), 197.

Petreska, A., & Ristevski, B. (2023, May). Artificial Intelligence in Intelligent Healthcare Systems–Opportunities and Challenges. In Serbian International Conference on Applied Artificial Intelligence (pp. 123-143). Cham: Springer Nature Switzerland.

Petreska, A., Ristevski, B., Slavkovska, D., Nikolovski, S., Spirov, P., Rendevski, N., & Savoska, S. (2023). Machine Learning Algorithms for Heart Disease Prognosis Using IoMT Devices.

Pradipta, G. A., Wardoyo, R., Musdholifah, A., Sanjaya, I. N. H., & Ismail, M. (2021, November). SMOTE for handling imbalanced data problem: A review. In 2021 sixth international conference on informatics and computing (ICIC) (pp. 1-8). IEEE.

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature medicine, 28(1), 31-38.

Rawat, S. S., & Mishra, A. K. (2022, December). Review of Methods for Handling Class Imbalance in Classification Problems. In International Conference on Data, Engineering and Applications (pp. 3-14). Singapore: Springer Nature Singapore.

Rubinger, L., Gazendam, A., Ekhtiari, S., & Bhandari, M. (2023). Machine learning and artificial intelligence in research and healthcare. Injury, 54, S69-S73.

Sanmarchi, F., Fanconi, C., Golinelli, D., Gori, D., Hernandez-Boussard, T., & Capodici, A. (2023). Predict, diagnose, and treat chronic kidney disease with machine learning: a systematic literature review. Journal of nephrology, 36(4), 1101-1117.

Savoska, S., Ristevski, B., Blazheska-Tabakovska, N., Jolevski, I., Bocevska, A., & Trajkovik, V. (2023). Integration of heterogeneous medical and biological data with electronic personal health records. Medical Informatics, Biometry and Epidemiology (MIBE).

Shekar, B. H., & Dagnew, G. (2019, February). Grid search-based hyperparameter tuning and classification of microarray cancer data. In 2019 second international conference on advanced computational and communication paradigms (ICACCP) (pp. 1-8). IEEE.

Shu, S., Ren, J., & Song, J. (2021). Clinical application of machine learning-based artificial intelligence in the diagnosis, prediction, and classification of cardiovascular diseases. Circulation Journal, 85(9), 1416-1425.

Singh, N., Jain, M., Kamal, M. M., Bodhi, R., & Gupta, B. (2024). Technological paradoxes and artificial intelligence implementation in healthcare. An application of paradox theory. Technological Forecasting and Social Change, 198, 122967.

Singh, N., Jain, M., Kamal, M. M., Bodhi, R., & Gupta, B. (2024). Technological paradoxes and artificial intelligence implementation in healthcare. An application of paradox theory. Technological Forecasting and Social Change, 198, 122967.

Sun, X., Yin, Y., Yang, Q., & Huo, T. (2023). Artificial intelligence in cardiovascular diseases: diagnostic and therapeutic perspectives. European Journal of Medical Research, 28(1), 242.

Vujović, Ž. (2021). Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications, 12(6), 599-606.

Wehbe, R. M., Katsaggelos, A. K., Hammond, K. J., Hong, H., Ahmad, F. S., Ouyang, D., ... & Thomas, J. D. (2023). Deep learning for cardiovascular imaging: A review. JAMA cardiology, 8(11), 1089-1098.

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. Information, 14(1), 54.

Zhang, X., & Liu, C. A. (2023). Model averaging prediction by K-fold cross-validation. Journal of Econometrics, 235(1), 280-301.

Zhou, Q., Qi, Y., Tang, H., & Wu, P. (2023). Machine learning-based processing of unbalanced data sets for computer algorithms. Open Computer Science, 13(1), 20220273.