



Универзитет „Св. Кирил и Методиј“ – Скопје
Факултет за електротехника и информациски технологии



- Докторска дисертација -

Тема:

Идентификација на расна пристрасност во податочни множества со помош на
машинско учење

Ментор:

Проф. д-р Марија Календар

Докторанд:

м-р Бојана Величковска

Скопје, 2024

Посветено на моите родители.

Содржина

Вовед	1
2 Преглед на истражувањето	6
2.1 Цели на истражувањето.....	7
2.2 Образложение на работните хипотези и тези.....	8
2.3 Научни методи кои ќе се применат	8
2.4 Очекуван научен придонес	9
2.5 Примена на резултатите од истражувањето	9
3 Мотивација и преглед на литература	11
3.1 Мотивација	12
3.2 Преглед на литература.....	21
4 Методологија.....	27
4.1 Избор на податочно множество.....	28
4.2 Етички аспекти	45
4.3 Претпроцесирање на податоци и избор на карактеристики	45
4.4 Алгоритми и оптимизација на параметри	54
4.5 Валидација на моделот, евалуациски метрики, и интерпретабилност на моделот.....	58
5 Резултати.....	65
5.1 Пристрасност која произлегува од витални знаци, медицинска опрема и пракса.....	65
5.2 Разрешавање на потенцијални збунувачки променливи.....	78
5.3 Екстерна валидација	81
5.4 Трансферабилност: Полова пристрасност	82
6 Дискусија	85
6.1 Ограничувања на истражувањето.....	87
6.2 Генерализираност на пристапот	88
7 Заклучок.....	93
Референци	95

Листа на слики

Слика 3-1 Видови на ПРИСТРАСНОСТ ВО МЕДИЦИНА	12
Слика 3-2 ФИЛТРИРАЊЕ (ДЕГРАДАЦИЈА) НА СЛИКА, СПОРЕД [70], ЗА ДЕТЕКЦИЈА НА РАСА. ПРВИОТ РЕД ИЛУСТРИРА ВИСОКОПРОПУСЕН ФИЛТЕР (ОЗНАЧЕНИ СО HPF), ДОДЕКА ВТОРИОТ РЕД ИЛУСТРИРА НИСКОПРОПУСЕН ФИЛТЕР (ОЗНАЧЕНИ СО LPF).....	23
Слика 4-1 БАРАЊА ЗА КВАЛИТЕТ НА ПОДАТОЦИТЕ.....	29
Слика 4-2 ОРГАНИЗАЦИЈА НА СЛЕДЕЊЕ НА ПАЦИЕНТИ ВО eICU-CRD. СЛИКА ПОЗАЈМЕНА ОД [90].....	33
Слика 4-3 ПРЕГЛЕД НА СТРУКТУРАТА НА MIMIC-III ПОДАТОЧНОТО МНОЖЕСТВО. СЛИКАТА ПОЗАЈМЕНА ОД [94].....	35
Слика 4-4 ИНСТАЛИРАЊЕ НА MIMIC-III ПОДАТОЧНОТО МНОЖЕСТВО И ПОМОШНИТЕ КОНЦЕПТИ.....	35
Слика 4-5 ПРЕСМЕТКА НА OASIS [107]. СЕКОЈА МЕРКА НА ПРОМЕНЛИВА ДОБИВА СООДВЕТНА ВРЕДНОСТ (ДАДЕНА ВО ЗАДЕБЕЛЕНО), ТАКА ШТО СЕ ЗЕМА НАЈЛОША ВРЕДНОСТ ВО ПРВ ДЕН, А ФИНАЛНАТА ВРЕДНОСТ НА OASIS Е СУМА ОД СИТЕ ИНДИВИДУАЛНИ ВРЕДНОСТИ. LOS = LENGTH OF STAY (ВРЕМЕ НА ПРЕСТОЈ), GCS = GLASGOW COMMA SCALE (ГЛАЗГОВ КОМАТОЗНА СКАЛА), MAP = MEAN ARTERIAL PRESSURE (СРЕДЕН АРТЕРИСКИ ПРИТИСОК).	38
Слика 4-6 РЕПРЕЗЕНТАЦИЈА НА ЕТНИЧКО-РАСНИ ГРУПИ ВО eICU-CRD И MIMIC-III МНОЖЕСТВАТА	40
Слика 4-7 ТИПОВИ НА ОБЗЕРВАЦИОНИ СТУДИИ	47
Слика 4-8 ДИЈАГРАМ ЗА СЕЛЕКЦИЈА НА ПАЦИЕНТИ: eICU-CRD	49
Слика 4-9 ДИЈАГРАМ ЗА СЕЛЕКЦИЈА НА ПАЦИЕНТИ: MIMIC-III	50
Слика 4-10 ПРОЦЕС НА СПАРУВАЊЕ НА ПАЦИЕНТИ	52
Слика 4-11 СИГМОИДНА ФУНКЦИЈА	55
Слика 4-12 ОД ДРВА НА ОДЛУКА ДО XGBOOST	55
Слика 4-13 XGBOOST ПАРАМЕТАРСКА МРЕЖА ЗА RANDOMIZEDSEARCHCV	58
Слика 4-14 СТРАТИФИКУВАНА К-КРАТНА ВКРСТЕНА ВАЛИДАЦИЈА.....	59
Слика 4-15 МАТРИЦА НА ГРЕШКА ЗА БИНАРНА КЛАСИФИКАЦИЈА И ДОПОЛНИТЕЛНИ ИЗВЕДЕНИ МЕТРИКИ.....	60
Слика 4-16 AUC-ROC КРИВА (ROC – ВЕРОЈАТНОСНА КРИВА, AUC – СТЕПЕН НА РАЗДВОЈУВАЊЕ МЕЃУ КЛАСИТЕ)	61
Слика 4-17 ПРАГОВИ НА AUC-ROC КРИВА. СЕКОЈ ПАР ФИГУРИ ЈА ПОКАЖУВА МАТРИЦАТА НА ГРЕШКИ И КАКО ТИЕ ВРЕДНОСТИ СЕ ПРЕНЕСУВААТ НА AUC-ROC КРИВАТА.....	61
Слика 4-18 SHAP ГРАФИК: ПРИМЕР [162].....	64
Слика 5-1 РЕЗУЛТАТИ ОД ПАЦИЕНТИ ОД БЕЛА И ЦРНА РАСА ЗА eICU-CRD СО СИТЕ ВИТАЛНИ ЗНАЦИ (ЛЕВО) И СО ОСНОВНИ ВИТАЛНИ ЗНАЦИ (ДЕСНО)	66
Слика 5-2 АНАЛИЗА НА КАРАКТЕРИСТИКИ ЗА КОХОРТА ОД ПАЦИЕНТИ ОД БЕЛА И ЦРНА РАСА СО СИТЕ ВИТАЛНИ ЗНАЦИ (eICU-CRD).....	67
Слика 5-3 АНАЛИЗА НА КАРАКТЕРИСТИКИ ЗА КОХОРТА ОД ХИСПАНЦИ И ПАЦИЕНТИ ОД БЕЛА РАСА (ЛЕВО) И КОХОРТА ОД ХИСПАНЦИ И ПАЦИЕНТИ ОД ЦРНА РАСА (ДЕСНО) СО СИТЕ ВИТАЛНИ ЗНАЦИ (eICU-CRD)	67
Слика 5-4 РЕЗУЛТАТИ ОД ПАЦИЕНТИ ОД БЕЛА И ЦРНА РАСА ЗА eICU-CRD СО НЕИНВАЗИВЕН КИСЛОРОД (ЛЕВО) И СО ИНВАЗИВЕН КИСЛОРОД (ДЕСНО).....	69
Слика 5-5 РЕЗУЛТАТИ ОД ХИСПАНЦИ И ПАЦИЕНТИ ОД БЕЛА РАСА ЗА eICU-CRD СО НЕИНВАЗИВЕН КИСЛОРОД (ЛЕВО) И СО ИНВАЗИВЕН КИСЛОРОД (ДЕСНО)	69
Слика 5-6 РЕЗУЛТАТИ ОД ХИСПАНЦИ И ПАЦИЕНТИ ОД ЦРНА РАСА ЗА eICU-CRD СО НЕИНВАЗИВЕН КИСЛОРОД (ЛЕВО) И СО ИНВАЗИВЕН КИСЛОРОД (ДЕСНО)	70
Слика 5-7 АНАЛИЗА НА КАРАКТЕРИСТИКИ ЗА КОХОРТА ОД ХИСПАНЦИ И ПАЦИЕНТИ ОД ЦРНА РАСА ЗА eICU-CRD СО НЕИНВАЗИВЕН КИСЛОРОД (ЛЕВО) И СО ИНВАЗИВЕН КИСЛОРОД (ДЕСНО).....	70
Слика 5-8 РЕЗУЛТАТИ ОД ПАЦИЕНТИ ОД БЕЛА И ЦРНА РАСА ЗА eICU-CRD СО НЕИНВАЗИВЕН КРВЕН ПРИТИСОК (ЛЕВО) И СО ИНВАЗИВЕН КРВЕН ПРИТИСОК (ДЕСНО).....	71
Слика 5-9 РЕЗУЛТАТИ ОД ХИСПАНЦИ И ПАЦИЕНТИ ОД БЕЛА РАСА ЗА eICU-CRD СО НЕИНВАЗИВЕН КРВЕН ПРИТИСОК (ЛЕВО) И СО ИНВАЗИВЕН КРВЕН ПРИТИСОК (ДЕСНО).....	71
Слика 5-10 РЕЗУЛТАТИ ОД ХИСПАНЦИ И ПАЦИЕНТИ ОД ЦРНА РАСА ЗА eICU-CRD СО НЕИНВАЗИВЕН КРВЕН ПРИТИСОК (ЛЕВО) И СО ИНВАЗИВЕН КРВЕН ПРИТИСОК (ДЕСНО).....	72
Слика 5-11 АНАЛИЗА НА КАРАКТЕРИСТИКИ ЗА НЕИНВАЗИВЕН КРВЕН ПРИТИСОК ЗА ДВЕТЕ КОХОРТИ: ХИСПАНЦИ И ПАЦИЕНТИ ОД БЕЛА РАСА (ЛЕВО) И ХИСПАНЦИ И ПАЦИЕНТИ ОД ЦРНА РАСА (ДЕСНО)	72
Слика 5-12 РЕЗУЛТАТИ ОД eICU-CRD КОХОРТА ОД ПАЦИЕНТИ ОД ЦРНА И БЕЛА РАСА СО САМО НЕИНВАЗИВНИ МЕРЕЊА	73
Слика 5-13 АНАЛИЗА НА КАРАКТЕРИСТИКИ ОД eICU-CRD КОХОРТА ОД ПАЦИЕНТИ ОД ЦРНА И БЕЛА РАСА СО САМО НЕИНВАЗИВНИ МЕРЕЊА	73
Слика 5-14 РЕЗУЛТАТИ ОД НЕИНВАЗИВНИ МЕРЕЊА ЗА ДВЕ КОХОРТИ: ХИСПАНЦИ И ПАЦИЕНТИ ОД БЕЛА РАСА (ЛЕВО) И ХИСПАНЦИ И ПАЦИЕНТИ ОД ЦРНА РАСА (ДЕСНО)	74
Слика 5-15 АНАЛИЗА НА КАРАКТЕРИСТИКИ ОД НЕИНВАЗИВНИ МЕРЕЊА ЗА ДВЕ КОХОРТИ: ХИСПАНЦИ И ПАЦИЕНТИ ОД БЕЛА РАСА (ЛЕВО) И ХИСПАНЦИ И ПАЦИЕНТИ ОД ЦРНА РАСА (ДЕСНО)	74
Слика 5-16 РЕЗУЛТАТИ ЗА КОХОРТАТА ОД ПАЦИЕНТИ ОД ЦРНА И БЕЛА РАСА (eICU-CRD) ОД ВИТАЛНИ ЗНАЦИ И ФРЕКВЕНЦИИ НА МЕРЕЊА КОМБИНИРАНО	76
Слика 5-17 РЕЗУЛТАТИ ЗА КОХОРТАТА ОД ПАЦИЕНТИ ОД ЦРНА И БЕЛА РАСА (eICU-CRD) ОД ФРЕКВЕНЦИИ НА МЕРЕЊА САМО	76
Слика 5-18 АНАЛИЗА НА КАРАКТЕРИСТИКИ ЗА КОХОРТА ОД ПАЦИЕНТИ ОД ЦРНА И БЕЛА РАСА (eICU-CRD) ОД ФРЕКВЕНЦИИ НА МЕРЕЊА САМО.....	77
Слика 5-19 РЕЗУЛТАТИ ОД ЕКСТЕРНА ВАЛИДАЦИЈА ЗА НЕИНВАЗИВНИ МЕРЕЊА: ОД eICU-CRD ДО MIMIC-III (ЛЕВО) И ОД MIMIC-III ДО eICU-CRD (ДЕСНО)	82
Слика 5-20 ПОЛОВА ПРИСТРАСНОСТ ЗА eICU-CRD СО СИТЕ ВИТАЛНИ ЗНАЦИ	83

Слика 5-21 Анализа на карактеристики за полова пристрасност за eICU-CRD со сите витални знаци.....	83
Слика 5-22 Полова пристрасност за eICU-CRD по исклучување на мерења на крвен притисок	84
Слика 6-1 Предлог архитектура за детекција на пристрасност во податоци	88

Листа на табели

ТАБЕЛА 3-1 ПРИКАЗ НА ТРУДОВИ КОИ ПОКАЖУВААТ ПРИСТРАСНОСТ ВО МЕДИЦИНСКА ПРАКСА	14
ТАБЕЛА 3-2 ПРИКАЗ НА ТРУДОВИ КОИ ПОКАЖУВААТ ПРИСТРАСНОСТ ВО МЕДИЦИНСКА ТЕХНОЛОГИЈА.....	15
ТАБЕЛА 3-3 ПРИКАЗ НА ТРУДОВИ КОИ ПОКАЖУВААТ ПРИСТРАСНОСТ ВО МЕДИЦИНСКИ ПОДАТОЧНИ МНОЖЕСТВА.....	17
ТАБЕЛА 3-4 ПРИКАЗ НА ТРУДОВИ КОИ ПОКАЖУВААТ ПРИСТРАСНОСТ ВО МЕДИЦИНСКИ АПЛИКАЦИИ БАЗИРАНИ НА ВИ	19
ТАБЕЛА 3-5 ПРИКАЗ НА ТРУДОВИ ПОВРЗАНИ СО ТЕМАТА НА ИСТРАЖУВАЊЕ.....	23
ТАБЕЛА 3-6 АНАЛИЗА НА ПРЕДНОСТИ, СЛАБОСТИ, МОЖНОСТИ И ЗАКАНИ НА РЕЛЕВАНТНИТЕ ТРУДОВИ.....	24
ТАБЕЛА 3-7 АНАЛИЗА НА КОРЕЛАЦИЈА И ДИСАСОЦИЈАЦИЈА ПОМЕЃУ РЕЛЕВАНТНИТЕ ТРУДОВИ И ОВАА ДОКТОРСКА ДИСЕРТАЦИЈА	26
ТАБЕЛА 4-1 ПРЕГЛЕД НА МЕДИЦИНСКИТЕ ПОДАТОЧНИ МНОЖЕСТВА	31
ТАБЕЛА 4-2 СПОРЕДБА НА ИЗБРАНИТЕ ПОДАТОЧНИ МНОЖЕСТВА (eICU-CRD И MIMIC-III)	36
ТАБЕЛА 4-3 ПРЕСМЕТКА НА APACHE-IV [106]. ПОДАТОЦИ КОИ СЕ КОРИСТАТ ЗА КРЕИРАЊЕ НА МОДЕЛОТ ЗА ПРЕДВИДУВАЊЕ И ДЕТАЛИТЕ ЗА ИСТИТЕ, СИТЕ ЗЕМЕНИ ВО ПРВ ДЕН ОД ПРЕСТОЈ. APS = ACUTE PHYSIOLOGY SCORE (РЕЗУЛТАТ ЗА ФИЗИОЛОШКА СОСТОЈБА), APACHE = ACUTE PHYSIOLOGY AND CHRONIC HEALTH EVALUATION (КОЕФИЦИЕНТ ЗА СЕРИОЗНОСТ НА СОСТОЈБА НА ПАЦИЕНТ), GCS = GLASGOW COMMA SCALE (ГЛАЗГОВ КОМАТОЗНА СКАЛА), PAO ₂ = ARTERIAL OXYGEN PARTIAL PRESSURE (АРТЕРИСКИ ПАРЦИЈАЛЕН ПРИТИСКОК НА КИСЛОРОД), FIO ₂ = FRACTIONAL INSPIRED OXYGEN (ФРАКЦИОНА КОНЦЕНТРАЦИЈА НА ИНСПИРИРАН КИСЛОРОД), P(A-a)O ₂ = ALVEOLAR-ARTERIAL GRADIENT (АЛВЕОЛАРНО-АРТЕРИСКИ ГРАДИЕНТ, Т.Е. А-А ГРАДИЕНТ).....	37
ТАБЕЛА 4-4 ПРИЕМНА ДИЈАГНОЗА ВО ОРИГИНАЛЕН ФОРМАТ ЗА ПАЦИЕНТИ ВО eICU-CRD, ЗА СИТЕ ПАЦИЕНТИ И СПОРЕД ЕТНИЧКО-РАСНА ПРИПАДНОСТ	41
ТАБЕЛА 4-5 ПРИЕМНА ДИЈАГНОЗА ВО ОРИГИНАЛЕН ФОРМАТ ЗА ПАЦИЕНТИ ВО MIMIC-III, ЗА СИТЕ ПАЦИЕНТИ И СПОРЕД ЕТНИЧКО-РАСНА ПРИПАДНОСТ	42
ТАБЕЛА 4-6 ПОЛОВИ РАЗЛИКИ НА НИВО НА ПАЦИЕНТИ И ПРИЕМИ ВО СООДВЕТНИТЕ ЕТНИЧКО-РАСНИ ГРУПИ (Р - УНИКАТНИ ПАЦИЕНТИ, S - УНИКАТНИ ПРЕСТОИ ВО ЕИИ).....	43
ТАБЕЛА 4-7 СТАРОСНИ РАЗЛИКИ НА НИВО НА ПАЦИЕНТИ И ПРИЕМИ ВО СООДВЕТНИТЕ ЕТНИЧКО-РАСНИ ГРУПИ (Р - УНИКАТНИ ПАЦИЕНТИ, S - УНИКАТНИ ПРЕСТОИ ВО ЕИИ).....	43
ТАБЕЛА 4-8 КОЕФИЦИЕНТ ЗА ОЦЕНУВАЊЕ НА СЕРИОЗНОСТ НА ЗАБОЛУВАЊЕ, ЗА СИТЕ ПАЦИЕНТИ И ПО ЕТНИЧКО-РАСНА ГРУПА	44
ТАБЕЛА 4-9 СТАТИСТИКА ЗА КОЕФИЦИЕНТ ЗА СЕРИОЗНОСТ НА БОЛЕСТ ПО ПРИЕМИ ВО ЕИИ, ЗА СИТЕ ПАЦИЕНТИ И ПО ЕТНИЧКО-РАСНА ГРУПА	45
ТАБЕЛА 4-10 КАРАКТЕРИСТИКИ ОД ИНТЕРЕС, ЕДИНИЦА МЕРКА И ОПСЕГ НА ВАЛИДНИ ВРЕДНОСТИ.....	48
ТАБЕЛА 4-11 ПРЕГЛЕД НА ПРАЗНИ ПОЛИЊА ПО ВИТАЛЕН ЗНАК ВО MIMIC-III.....	51
ТАБЕЛА 4-12 КАРАКТЕРИСТИКИ НА ПАЦИЕНТИТЕ ВО ДОБИЕНИТЕ КОХОРТИ	53
ТАБЕЛА 5-1 РЕЗУЛТАТИ ДОБИЕНИ ОД СИТЕ ВИТАЛНИ ЗНАЦИ И ОД ОСНОВНИ ВИТАЛНИ ЗНАЦИ (ОТЧУКУВАЊЕ НА СРЦЕ И РЕСПИРАТОРНА СТАПКА) ЗА СИТЕ ТРИ КОХОРТИ.....	65
ТАБЕЛА 5-2 РЕЗУЛТАТИ ОД eICU-CRD ЗА СИТЕ КОХОРТИ СО СИТЕ КАРАКТЕРИСТИКИ ЗА ОРИГИНАЛНАТА КОХОРТА И НОВА КОХОРТА РЕАЛИЗИРАНА СО ВОЗРАСТ И ПОЛ ЗА ПРОВЕРКА НИВНО ВЛИЈАНИЕ КАКО ЗБУНУВАЧКИ ПРОМЕНЛИВИ	78
ТАБЕЛА 5-3 АНАЛИЗА НА ЗБУНУВАЧКИ ПРОМЕНЛИВИ: СПАРУВАЊЕ НА ПАЦИЕНТИ ВО КОХОРТИ СО ЕДНАКВА ДИЈАГНОЗА ПРИ ПРИЕМ, ПОЛ И ВОЗРАСТ	80
ТАБЕЛА 5-4 АНАЛИЗА НА КОМОРБИДИТЕТИ НА ПЕТ ИЗБРАНИ ДИЈАГНОЗИ (ДАДЕНИ ВО ОРИГИНАЛНА ФОРМА) ЗА КОХОРТА ОД ПАЦИЕНТИ ОД ЦРНА И БЕЛА РАСА	81
ТАБЕЛА 5-5 РЕЗУЛТАТИ ОД ЕКСТЕРНА ВАЛИДАЦИЈА НА КОХОРТА ОД ПАЦИЕНТИ ОД ЦРНА И БЕЛА РАСА ЗА ТРИ КОМБИНАЦИИ НА КАРАКТЕРИСТИКИ	81

Апстракт

Преку Хипократовата заклетва, која ги дефинира обврските и однесувањето на медицинските работници, лекарите ветуваат да препишуваат корисни третмани и да се воздржат од предизвикување на штета или повреда, најдобро можно според нивните способности и судење. Ова индиректно кажува дека пациентите треба секогаш да добиваат квалитетна здравствена заштита без разлика на лични карактеристики; сепак, одредени популации на пациенти се маргинализирани поради постојни, имплицитни предрасуди базирани на нивниот пол и сексуалност, нивната боја на кожа, нивниот социоекономски статус и други демографски карактеристики. Пристрасните практики може штетно да влијаат на пристапот и квалитетот на здравствената заштита што една личност ја добива и последователно негативно да влијаат и на квалитетот на живот. Со цел да се подобри квалитетот на здравствената заштита за сите пациенти подеднакво, без оглед на социјална категоризација, неопходно е да се идентификува присуството на различни видови на пристрасност во здравствениот систем, да се разберат изворите на таа пристрасност и да се развијат стратегии за ублажување на истите. Постојат бројни извештаи за расна пристрасност во медицинската пракса, поточно пристрасност во медицинската технологија, однесувањето на медицинските работници, податочните множества собирани од пациенти во медицинските центри и медицинските апликации базирани на вештачка интелигенција. Сепак, многу помалку се знае за потенцијалната расна пристрасност во рутински собирани, суштински информации што се користат во клиничкото одлучување, како што се виталните знаци. Оваа дисертација главно се фокусира на проверката дали виталните знаци имаат вградени етничко-расни информации што може да се научат преку алгоритми за вештачка интелигенција, со дополнителна анализа за идентификација на степенот на постоење на познати, имплицитни предрасуди (медицинска технологија, медицинска пракса, итн.) во податоците што се користат за истражувањето.

Истражувањето беше дизајнирано како ретроспективна кохортна студија на критично болни пациенти од две јавно достапни податочни множества, примени во единици за интензивна нега (ЕИН), на возраст од 18 и повеќе години, живи во првите 24 часа по приемот, со валидна самопријавена раса, со најмалку две мерења за отчукување на срцето, заситеност со кислород, респираторна стапка и крвен притисок. Добиени беа три кохорти од пациенти: пациенти од црна и бела раса, Хиспанци и пациенти од бела раса, и Хиспанци и пациенти од црна раса; секоја кохорта беше оформена преку спојување на пациенти од двете етничко-расни групи врз основа на возраста, полот, дијагнозата при прием и коефициентите за евалуација на сериозност на болеста. За да се предвиди самопријавената етничко-расна припадност на пациентите беа користени традиционални алгоритми за машинско учење: XGBoost и логистичка регресија. На добиените модели беше направена и внатрешна и надворешна валидација. Резултатите покажуваат дека моделите тренирани само на шест рутински мерени витални знаци можат да ја предвидат самопријавената раса на пациентите со AUC од 0,74 ($\pm$ 0,022) помеѓу пациенти од бела и црна раса. Понатаму, инструментите што се користат за мерење на заситеност со кислород се значителен извор на самопријавени расни информации (AUC од 0,72 $\pm$ 0,026), а истото се забележува и кај мерењата на крвниот притисок (AUC од 0,64 $\pm$ 0,029). Практиките за грижа на пациенти се покажаа како непристрасни во поглед на расни информации (AUC од 0,58 $\pm$ 0,031). Но, и при контролирање на познати фактори на пристрасност, експериментите покажаа дека самопријавената раса сè уште може да се научи од витални знаци, а потеклото на

пристрасноста не може целосно да се објасни со добиените анализи. Бидејќи резултатите покажуваат како виталните знаци кријат расни информации што алгоритмите за вештачка интелигенција може да ги научат, оваа студија открива значителен ризик за правично клиничко одлучување. Истовремено, имајќи ја во предвид основната улога на виталните знаци, ова истражување дополнително ги нагласува предизвиците што треба да се надминат за добивање на квалитетни мерки за ублажување на пристрасноста.

Клучни зборови: податоци од центри за интензивна нега, критична нега, eICU-CRD, MIMIC-III, кохортна студија, витални знаци, предвидување на етничко-расна припадност, етничко-расна пристрасност, машинско учење

Abstract

The Hippocratic oath, which dictates the obligations of conduct by medical professionals, has physicians pledging to prescribe beneficial treatments, and refrain from causing harm or hurt, to the best of their abilities and judgement. Indirectly, this speaks that people should always receive quality healthcare regardless of personal characteristics; however, certain patient populations are marginalised due to existing, implicit biases based on their gender and sexuality, their skin colour, their socioeconomic status, and other demographics. These biased practices can have detrimental effects on the access and quality of healthcare a person receives and consequently negatively influence their quality of life. In order to improve healthcare quality for all patients equally, regardless of social categorisation, it is necessary to identify the presence of different types of bias in the healthcare system, understand the sources of that bias, and develop mitigation strategies. There have been numerous reports of racial bias in medical practice, namely biases in medical technology, behaviour of medical professionals, data collected from patients, and AI-based medical applications. However, far less is known about the potential racial bias in routinely-collected and essential information in clinical decision-making, namely vital signs. This dissertation mainly focusses on analysing whether vital signs embed ethno-racial information that can be learned by artificial intelligence algorithms, with additional analysis extending into identifying the degree of known implicit biases (medical technology, medical practice, etc.) embedded in the data used for the research.

The research was designed as a retrospective cohort study of critically ill patients from two publicly available datasets, admitted into intensive care units (ICUs), aged 18 and over, alive during the first 24 hours after admission, with recorded self-reported race as well as at least two measurements of heart rate, oxygen saturation, respiratory rate, and blood pressure. Three cohorts were created: Black and White patients, Hispanic and White patients, and Black and Hispanic patients; each cohort was created by matching pairs of racial subgroups based on age, gender, admission diagnosis, and illness severity scores. Traditional machine learning algorithms, XGBoost and Logistic Regression, were used to predict self-reported race using vital signs as an input. Both internal and external validation was performed. The results show that models derived from only six, routinely-measured vital signs can predict patients' self-reported race with an AUC of 0.74 (± 0.022) between White and Black patients. Furthermore, technologies used to measure oxygen saturation are a significant source of self-reported racial information (AUC of 0.72 ± 0.026), in addition to blood pressure measurements (AUC of 0.64 ± 0.029). Care delivery practices do not carry significant bias regarding racial information (AUC of 0.58 ± 0.031). However, even when controlling for these known factors, experiments showed that self-reported race can still be learned from vital signs, and the origin of the bias cannot be currently explained. As results show how vital signs embed racial information that can be learned by artificial intelligence algorithms, this study uncovers a significant risk to equitable clinical decision-making. Moreover, this only emphasises challenges in obtaining quality mitigating measures, considering the fundamental role of vital signs.

Keywords: ICU data, critical care, eICU-CRD, MIMIC-III, cohort study, vital signs, ethno-race prediction, ethno-racial bias, machine learning

Вовед

Како што технологијата напредува и нејзините секојдневни апликации се зголемуваат, податоците и науката за податоци станаа движечка сила во светот, толку суштинска, што последователно компјутерите станаа неопходна алатка за собирање, обработка и анализа на податоците. Науката за податоци ги опфаќа сите сознанија кои можат да се извлечат од податоците со помош на класична анализа. Еден чекор понатаму е користењето на компјутери и програмски јазици за предавање на податоците како влез во алгоритми за тие да ги екстраполираат важните информации со цел да се обучи функционален модел кој може да решава секојдневни проблеми на начин сличен на луѓето. Концептот е дефиниран како вештачка интелигенција (ВИ), а нејзината употреба е обемна и широко распространета, во индустриски, образовни и многу други цели.

Бидејќи новите пристапи и апликации на ВИ во различни области се покажаа како успешни, влијанието на ВИ дополнително се зајакнува и примената се проширува во попредизвикувачки, но и ризични професии, каде што нејзиниот потенцијал е неспоредлив, неистражен и неограничен, со што ВИ стана составен дел од модерната здравствена заштита, [1]. Главната цел на ВИ во медицината е поддршка за медицинските работници во клинички услови и при тековни истражувања, преку обработка на големи количини медицински податоци и информации за пациентите со цел да се добие увид за подобрување на здравствените исходи и искуствата на пациентите, но и да се постигне она што медицината треба да го овозможи на прво место, а тоа е најдобра можна грижа на оние пациенти на кои им е најпотребна. Водечките примени на ВИ во медицината во денешно време вклучуваат поддршка при клинички одлуки (анг. clinical decision support), [2], кои влијаат на третманите, лековите, менталното здравје и дополнителните потреби на пациентите, како и медицинска дијагностика во слика (анг. medical imaging), каде што слики добиени при скенирање на пациенти се анализираат од страна на алгоритми за ВИ со цел да се добијат наоди кои можеби не биле забележани од медицинскиот персонал, [3]. Притоа, и двата типа на системи активно се користат во медицински услови за да се донесат одлуки што може да значат разлика помеѓу живот и смрт, па затоа нивниот квалитет мора да биде неоспорен.

КОВИД-19 пандемијата изврши екстреман притисок врз здравствените системи на глобално ниво и ги изнесе на виделина прикриените слабости, [4], и во најмоќните земји во светот, [5], [6], но и во земјите во развој, [7], покажувајќи ги бројните недостатоци во медицинскиот сектор; а, резултатот беше неверојатен број на загубени животи, со прекумерна смртност, [8]. Овие слабости се дел и од медицинската опрема и апликациите на ВИ во медицината, на различни начини: едно, ВИ не влијаеше на текот на КОВИД-19 пандемијата бидејќи нејзината функционалност беше нарушена и од премногу и од недоволна количина податоци, [9], додека институциите се соочуваа со проблем на неснабденост со медицинска опрема, и второ, се појавија бројни извештаи за случаи на сериозна пристрасност во медицинската технологија и алгоритмите базирани на ВИ, [10]. Двете остро го демонстрираа влијанието што длабоко вкоренетите проблеми во медицинската пракса и системите за медицинска ВИ може да го имаат и ја истакнаа неопходноста да се адресираат и решат тие прашања пред ВИ да може да стане доверлив основен дел од дигиталните здравствени системи кои ја обликуваат и поддржуваат модерната медицина. Имено, овие пристрасни пристапи и методи обезбедуваат послаб квалитет на здравствена заштита кај одредени популации на пациенти, што води до неточни дијагнози, доцнење на третмани или целосен недостиг од третман.

Со отворање на прашањето за пристрасност, континуиран прилив на извештаи и истражувања во новопопуларизираната тема открија екстремни недостатоци на примена на ВИ во медицината и доведоа до сомневање во точноста на активно-користени апликации. Потребата да се намалат овие конфликти се совпаѓа со концептот на доверлива ВИ (анг. trustworthy AI), [11], кој има за цел да го насочи вниманието кон клучни барања за намалување на ризикот при развој на системи базирани на ВИ и обезбедување на доверливост на системите и нивно прифаќање од страна на корисници на светско ниво. Следствено, доверливата ВИ станува се попопуларна и го привлекува вниманието на огромен број организации, меѓу кои се и истражувачки центри, универзитети,

бизниси, па дури и влади. Дополнително, се формираат и независни експертски групи за да работат на креирање на етички насоки и создавање доверливост, како што е случајот со Европската комисија, [12]. Бидејќи, доверливата ВИ се здоби со голема привлечност, загриженоста за пристрасност расте. Без разлика дали пристрасноста е експлицитна или имплицитна, дали потекнува од нерепрезентативни или неточни податоци во кои се вградени дискриминаторски општествени однесувања или од претпоставки базирани на предрасуди при развој на моделите, самата пристрасност влијае на моделите на ВИ бидејќи тие имаат способност во процесот на тренирање да се оформат и приспособат врз основа на добиените податоци, па така можат да ја пропагираат нееднаквоста кон одредени групи или поединци врз основа на етничко-расна припадност (анг. *race-ethnicity*), пол, социјална класа или други карактеристики. Без разлика на дискриминаторскиот фактор на кој е базирана пристрасноста, последиците врз маргинализираните популации можат да бидат екстремни. Додека сите наведени демографски фактори се подеднакво важни и имаат различни последици врз групи на пациенти, главниот фокус во оваа докторска дисертација е етничко-расната пристрасност (анг. *ethno-racial bias*).

Татковците-основачи (анг. *Founding Fathers*) на Соединетите Американски Држави (САД) ја потпишале Декларацијата за независност, во која е напишано: „Ние сметаме дека овие вистини се очигледни, дека сите луѓе се создадени еднакви, дека тие се обдарени од нивниот Творец со одредени неотуѓиви права, меѓу кои се Животот, Слободата и потрагата по среќа“¹. И, сепак, „демократијата“ изградена на овие зборови, кои дефинираат восхитувачки принципи, има богата историја на непочитување – често насилно – на одредени популации. И покрај тоа што се направени огромни чекори напред од мрачното минато на киднапирања, експлоатација и ропство на цели раси засновани на вкоренети верувања за супериорност на други, и покрај многуте залагања направени за еднаквост во изминатите неколку години, светот сè уште е воден од расната поделба воспоставена од закоравените, стари практики и научниот расизам (анг. *scientific racism*) кој датира од пред многу векови. Овие верувања и практики опстојуваат иако современиот научен консензус посочува дека не постои биолошка основа за расна категоризација и дека расата е всушност општествено-политичка конструкција. Сепак, расата продолжува да игра клучна улога во медицинското учење и практика, преку ставовите на медицинските лица и имплицитните предрасуди во нивното однесување, стереотипни верувања за болести, клиничката номенклатура и клиничките алгоритми, алатки и упатства за третман, кои придонесуваат до и ги зајакнуваат тековните расни разлики.

Доказите за етничко-расна пристрасност во медицината се многубројни. Студии покажуваат дека главните движечки сили на етничко-расни разлики во здравството се пристрасноста и дискриминацијата кај медицинските лица и институции, [13]. При споредба со однесувањето кон пациенти од бела раса, третманот кон малцинските етничко-расни групи открива предрасудни ставови кај медицинските лица, кои како даватели на здравствена грижа сметаат дека пациентите од црна раса помалку се придржуваат кон пропишаниот медицински третман, [14], додека Хиспанците ги поврзуваат со ризично и неусогласено однесување, а и дополнително, пациентите од малцински етничко-расни групи генерално ги сметаат за помалку интелигентни и повеќе склони кон насилство, [15], [16]. Ваквите негативни ставови кон пациенти со потемна боја на кожа се во корелација со понеквалитетни интеракции пациент-лекар и супстандарден квалитет на нега, а се особено забележливи во управување на болка кај пациентите, каде пациенти од црна раса се систематски под-третираны во споредба со пациенти од бела раса, [17], додека кај Азијците и Хиспанците постои помала веројатност тие да добијат проценка на степенот на болката и да им бидат препишани адекватни лекаства, [18].

¹ “We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness.”

Со оглед на тоа што овие разлики потекнуваат од однесувањето на медицинскиот персонал, особено лекарите, очекувано пациентите го имаат забележано и пријавено овој некоректен третман базиран на нивната етничко-расна припадност при нивниот пристап до здравствена заштита. Имено, кај пациенти од црна раса постои поголема веројатност да пријават негативни искуства при интеракција со медицински лица во споредба со припадници на бела раса, [19], а припадниците на црната раса и Хиспанците сметаат дека е особен предизвик да најдат лекари кои ќе ги третираат со достоинство и почит, [20]. Уште полошо, во материјалите за медицинска обука се користат расни стереотипи, кои следствено имаат улога во создавањето на грешно-информирани нови генерации на медицински персонал, [21]. Имено, материјали од предавања прикажуваат етничко-расни разлики во болестите, што придонесува студентите да поврзуваат одредени заболувања со одредени етничко-расни припадности, како на пример, предавачите често ги учат студентите дека рекурентни инфекции на бели дробови кај пациенти од бела раса се индикација за цистична фиброза што може да резултира со пропуштени дијагнози на цистична фиброза кај пациенти од црна раса, [22], медицински помагала и ресурси на слики често ја прикажуваат Лајмската болест на посветла боја на кожа и затоа болеста кај пациенти од црна раса често се дијагностицира дури по нејзина прогресија во стадиум на артритис, [23], учебници најчесто користат слики од потемна боја на кожа за да илустрираат сексуално преносливи болести, [24], како и медицински текстови кои содржат сведоштва за предрасуди во пристапот за управување на болка кај пациенти врз основа на бојата на кожа, [18]. Овие проблеми индиректно влијаат и создаваат професионалци кои не се свесни за здравствените разлики кои ненамерно ги пропагираат при креирање на нови клинички алатки, алгоритми и упатства кои се потпираат на расата како фактор. Еден пример каде етничко-расната припадност прави разлика во функционирањето на клиничките алатки е во нефрологијата, каде на пациенти од црна раса се доделуваат различни мерки за бубрежни функции во споредба со пациенти од други етничко-расни групи, [25]. Друг пример е често користен калкулатор за предвидување на успех при вагинално породување по претходно извршен царски рез, кај кој за припадниците на црна раса и Хиспанците постои фактор за корекција на пресметката што предвидувањето на успешно породување кај овие групи пациенти го намалува за над 50%, [25]. Слично на тоа, фактор на корекција постои и кај пулмоналните тестови, кај кои како главен ризик за неонатална жолтица се смета припадност во источноазиска група народи, а за пациенти со потекло од Азија и Пацифичките острови се користи различен праг за индексот на телесна маса при препораки за тестирање за дијабетес, [25].

Сите претходни примери, кои се само дел од постоечките пристрасни ставови и големата злоупотреба на етничко-расната припадност во медицината, ја покажуваат недоследноста од користењето на демографски информации во клиничката пракса и како таа доведува до недоволно или преголемо лекување на расистички третирани поединци предизвикувајќи одложувања или грешки во дијагноза и клиничка нега. Понатаму, овие примери илустрираат колку голем предизвик е да се спречи ширењето на етничко-расната пристрасност во алгоритмите за ВИ кои учествуваат или го водат клиничкото одлучување, и до кој степен истиот сè уште не е успешно адресиран од истражувачкиот свет, што се забележува земајќи ги во предвид бројните извештаи за пристрасност во активно-користени медицински апликации засновани на ВИ.

Постоењето и ширењето на пристрасност во медицината, а дури и во алгоритмите за ВИ, може да се должи на една или повеќе од неколку можни причини: првин, бидејќи постои очигледна пристрасност кај здравствените работници, таа пристрасност ненамерно се пренесува на моделите; второ, присуство на прикриена пристрасност во податоците што се користат за обука на моделите и/или недоволна застапеност на различни популации во податоците; трето, ненамерна пристрасност како последица на дизајнот, имплементацијата и толкувањето на алгоритмот за ВИ, итн. Сите наведени проблеми јасно се илустрирани преку конкретен пример за широко користен медицински алгоритам заснован на ВИ, кој имал за цел да идентификува пациенти кои ќе имаат корист од препорака за посложени здравствени процедури, и по долгогодишно користење на истиот, беше откриено дека алгоритмот е расно пристрасен и фаворизирал препорака на пациенти од бела раса

за посложена нега во споредба со пациенти од црна раса и покрај тоа што пациентите од црна раса биле значително поболни, [26]. Студијата во прашање откри дека пристрасноста се должи на дизајнот на алгоритмот, со што на алгоритмот му биле дадени трошоците направени за здравствена нега како информација за предвидување на потребите за здравствена заштита, и бидејќи пациентите од црна раса се соочуваат со поголеми бариери во пристап до здравствена заштита и не купуваат здравствени услуги исто колку и пациентите од бела раса, ова на моделот му сигнализирало дека кај пациентите од црна раса болеста е помалку сериозна. Алгоритмот се користел за над 200 милиони пациенти годишно, во периодот помеѓу 2013 и 2015 година, а според проценка на истражувачите, во случај на непристрасен алгоритам бројот на пациенти од црна раса кои би добиле дополнителна нега би се зголемил скоро три пати. Друг пример се и непретставителни дерматолошки податочни множества кај кои се јавува мала или воопшто не постои застапеност на пациенти со потемна боја на кожата, и во случаи на сериозни медицински состојби како што е ракот на кожа, [27]. Понатаму, речиси 50% од дерматолозите сметаат дека се несоодветно обучени за дијагностицирање на кожни болести кај пациенти со потемна боја на кожа, [28]. Со податочни множества во кои доминира светол тен на кожа и дерматолози кои не можат правилно да ги означат кожните заболувања кај луѓето со потемна боја на кожата, очекувано е моделите добиени за дијагноза на кожни болести да имаат намалена способност за детекција на кожни состојби кај потемните тенови на кожата.

Внимателно дизајнираните алгоритми можат да ја ублажат пристрасноста и да помогнат во намалувањето на етничко-расните разлики во здравствената грижа и заштита, а бидејќи пристрасноста продира во секој аспект на здравствената заштита – персоналот, технологијата, алатките и алгоритмите – неопходно е да се провери дали пристрасноста постои во податоците и информациите кои при креирање на алгоритмите се сметаат и користат како непристрасни. Затоа, ова истражување го анализира присуството на пристрасност во медицински информации на пациенти и мерења на витални знаци. За таа цел, студијата се обидува да разбере дали алгоритмите за ВИ се способни да ја идентификуваат самопријавената (анг. self-reported) етничко-расна припадност на пациентите притоа користејќи само статистички информации од виталните знаци измерени кај возрасни пациенти во првите 24 часа по приемот во единици за интензивна нега (ЕИН).

Истражувањето на докторската дисертација вклучува болнички приеми на критично болни пациенти од две јавно достапни податочни множества, собрани од повеќе медицински центри низ САД, за кои има валидни демографски информации – поконкретно, пол, возраст и етничко-расна припадност – како и документирана дијагноза при прием и коефициент за сериозност на болеста (анг. severity score), но и најмалку две мерења на витални знаци за време на нивниот престој во ЕИН. Пациентите кои ги исполнуваа овие критериуми беа споени во бинарни кохорти врз основа на нивните демографски информации со цел да се исклучи влијанието на т.н. интервенирани или збунувачки варијабли (анг. confounders), а добиените бинарни кохорти од пациенти оформени од сите приеми во ЕИН се користат за да се провери дали самопријавената етничко-расна припадност на пациентите може да се одреди користејќи само мерења од витални знаци како влезни карактеристики за алгоритмите. Првично се разгледуваат два алгоритми за машинско учење (МУ): логистичка регресија (со цел да се воспостави основен резултат) и XGBoost (како популарен и успешен пристап во решавање на слични проблеми). За споредба тренирани беа и невронски мрежи. Перформансите на моделите беа оценети со помош на внатрешна и надворешна валидација, при што внатрешната валидација е стратификувана петкратна вкрстена валидација (анг. stratified five-fold cross-validation), а надворешната валидација се реализира користејќи ги податоците од двете податочни множества наизменично. Однесувањето на моделот со најдобри перформанси беше анализирано со помош на алатки за интерпретирање на ВИ со цел да се утврдат причините зад предвидувањата на моделот за сите направени тестирања. Дополнително, направена е и анализа на етничко-расна пристрасност во технологија, т.е. инструментите за мерење на витални знаци и во медицинската пракса, иако двете веќе се имаат покажано како пристрасни во литературата и претходните истражувања, главно со цел да се провери дали избраните податочни множества во

себе кријат имплицитна пристрасност, и доколку истата постои, каква е природата на таа пристрасност.

Придонесот на истражувањето ќе опфати: едно, детална анализа на истражувања кои презентираат различни видови на пристрасност во медицинската пракса и импликациите кои може да се извлечат од нив; второ, споредба на јавно достапните податочни множества за витални знаци и достапност на етничко-расни информации; трето, нов предлог пристап за идентификување на етничко-расна пристрасност во медицинските податоци; четврто, пристап за идентификување на изворите на етничко-расна пристрасност во податоците, доколку таа пристрасност навистина постои; и конечно, воопштен пристап за идентификација на различни типови на пристрасност во табеларни податоци, силните и слабите страни на пристапот и потенцијал за применливост на пристапот на пошироко ниво.

Остатокот од оваа дисертација е организиран во неколку поглавја. Поголавје 2 дава преглед на истражувањето, односно ги сумира истражувачките цели, главната и придружната хипотеза и очекуваниот научен придонес. Поголавје 3 го дефинира концептот на пристрасност и ги прикажува познатите случаи на пристрасност во медицината, што пропратно ја дефинира мотивацијата за истражувањето, и тука истовремено се претставени најсовремените истражувања тесно поврзани со темата на дисертацијата. Поголавје 4 ги презентира релевантните медицински податочни множества и процесот на селекција на податочните множества кои се користат во истражувањето, како и осврт на методологијата со теоретски аспект на истражувањето. Поголавјето 5 ги прикажува резултатите од двете главни истражувачки прашања, имено го дефинира пристапот и резултатите од идентификацијата на етничко-расна пристрасност кај виталните знаци на пациенти и инструментите за мерење на виталните знаци, како и пристрасност во медицинската пракса. Во поглавје 6 се дадени интерпретацијата и дискусијата за добиените резултати, ограничувањата на студијата, како и генерализиран пристап за детекција на пристрасност во податоци. Поголавјето 7 дава заклучок од истражувањето и нуди увид во некои идеи за идна работа.

2 Преглед на истражувањето

Во ова истражување се разгледуваат можностите за детекција на етничка и расна пристрасност во медицински податоци на пациенти собирани за време на престој во ЕИН. Притоа, главен осврт е даден на употреба на виталните знаци на пациенти за детекција на нивната етничко-расна припадност на пациенти со помош на алгоритми базирани на ВИ. Пристапот се состои во примена на статистичка анализа на избраните податочни множества и примена на алгоритми од ВИ на витални знаци на пациенти од ЕИН преземени од две дистинктивни податочни множества. Добиениот пристап има за цел да анализира и провери постоење на расна пристрасност во медицински податочни множества (генерално), во алатките кои се користат за мерење на витални знаци (споредба на различни алатки кои се користат за иста цел) и во медицинска пракса (на пр. во однесувањето на медицинските лица).

Примарната цел на истражувањето е примена на алгоритми од класично МУ за класификација на етничка и расна група на пациенти врз база на статистички информации од виталните знаци на пациенти мерени за време на првите 24 часа од нивниот престој во ЕИН. Со цел да се реализира ова, рамката на ова истражување ќе се состои од неколку фази: фаза на претпроцесирање на податоците (во која ќе бидат селектирани сите валидни мерења на витални знаци од пациенти забележани во првите 24 часа по прием во ЕИН), фаза на регрупирање на пациентите т.е. процесот на креирање на кохорти (во која ќе се овозможи креирање на кохорти од пациенти базирано на идентични или слични вредности за дијагноза при прием, годините, полот и коефициентите за евалуација на сериозност на болеста), тренирање и валидација на моделите од МУ (фаза во која различни алгоритми ќе бидат тестирани и валидирани на кохортата пациенти или на подгрупа од пациенти од кохоратата, со цел да се одреди прецизноста на алгоритмите во предвидување на самопријавена етничко-расна припадност) и анализа на однесувањето на моделите (каде селектирани алгоритми ќе бидат темелно анализирани со цел да се добие претстава за тоа кои од влезните карактеристики на моделот се незаменливи за квалитетот на перформансите на истиот).

Пропратно, како секундарна цел на истражувањето е и употреба на алгоритмите од МУ за детекција на етничка и расна пристрасност во уредите (инструментите) кои се користат за мерење на витални знаци кај пациентите и во начинот на кој се одвиваат мерењата на виталните знаци кај пациентите (т.е. проверка за потенцијални грешки во медицинска практика). За да се постигне ова, дисекција на точноста на медицинската опрема алгоритмите базирани на МУ ќе ја реализираат така што ќе бидат употребени како алатка за утврдување на варијабилноста во мерењата што се добиваат од различни уреди кои првично биле наменети да служат буквално за иста цел (на пр. проверка на разликите во нивото на кислородна заситеност измерено инвазивно од артериска анализа на гас во крвта, т.е. артериски крвен тест, и неинвазивно со употреба на пулсниот оксиметар). Истражувањето на недоследностите во медицинската пракса ќе вклучува извлекување на карактеристики од достапните информации за да се пресмета фреквенцијата со која се мерат виталните знаци, по што овие вредности ќе бидат дадени на моделите за МУ како влезни параметри. И во двата случаи ќе се направи проверка на значајните карактеристики на моделот, односно ќе се анализира кои параметри се најзначајни за моделот.

На крајот, за понатамошно зајакнување на резултатите од студијата, целиот процес ќе се примени на две различни, јавно достапни податочни множества. Покрај тоа, за да се надополнат резултатите добиени од алгоритмите на МУ, оваа докторска дисертација исто така ќе ги тестира перформансите на моделите на длабоко учење (ДУ) во идентификувањето на самопријавената етничка и расна припадност на пациентите.

Анализата на моделот и значајните карактеристики ќе се користи не само за да се разберат перформансите на моделот, туку и во дефинирање на заклучоци за недостатоците на медицинската пракса, а со тоа и ќе се обезбедат забелешки и детали потребни за да се предложат потенцијални методи што ќе ја намалат расната пристрасност во медицината. Понатаму, дисертацијата ќе опфати

краток преглед на применливоста на целосниот пристап при откривање и намалување на различни форми на пристрасност во медицинското опкружување.

2.1 Цели на истражувањето

Акумулативната цел на докторската дисертација е да се даде придонес во детекција на самопријавена етничко-расна припадност во витални знаци на пациенти, последователно идентификувајќи етничко-расна пристрасност која постои во вредностите на виталните знаци на пациенти во критична состојба, во медицинската опрема која се употребува за мерење, како и во медицинската пракса преку анализа на фреквенција на направените мерења. Ова овозможува истражување на присуство на расна пристрасност во медицинските информации на пациентите со користење на алгоритми на МУ, со примарен осврт на анализа на податоци кои во никој случај не треба да носат етничко-расна пристрасност. Понатаму, пристапот ќе овозможи увид во изворите на етничката и расната пристрасност, како и неопходните мерки на претпазливост и чекорите потребни за намалување и целосно елиминирање на пристрасноста.

Во насока на постигнување на замислената цел, се предложуваат неколку потцели на истражувањето:

- Систематски преглед на литературата што јасно ја илустрира мотивацијата за ова истражување, вклучувајќи ја и дефиницијата за пристрасност и различните видови на пристрасност што постојат, обемот на пристрасност која е присутна во медицинската пракса и однесувањето на здравствените работници, пристрасноста што постои во медицинската технологија и опрема, пристрасност создадена од процесот на собирање на медицински податоци и пристрасност присутна во широко користени, медицински апликации базирани на ВИ. Секоја од наведените видови на пристрасност ќе биде придружена со резиме во кое ќе бидат претставени импликациите на соодветната пристрасност.
- Компаративна анализа на различни медицински податочни множества, со најголем фокус на базите на податоци (податочните множества) кои ќе бидат избрани за целите на оваа студија.
- Статистичка анализа на избраните податочни множества, што ќе обезбеди увид во процесот на создавање на медицински податочни множества и самата медицинска пракса, со што ќе се обезбеди и проценка на постоење на овие два вида пристрасност во избраните податочни множества.
- Градење, тренирање и валидација на алгоритмите на МУ за идентификување на етничко-расна пристрасност кај витални знаци на пациенти.
- Градење, тренирање и валидација на алгоритмите на МУ за идентификување на етничко-расна пристрасност во медицинската опрема која се користи за мерење на витални знаци.
- Градење, тренирање и валидација на алгоритмите на МУ за идентификување на етничко-расна пристрасност во рутинска медицинска пракса.
- Детална анализа на резултатите, со фокус на квалитетот на перформансите на моделите на МУ, со темелна анализа на карактеристиките значајни за моделот со најдобри перформанси. Резултатите добиени од различни комбинации на влезни карактеристики за моделот и графици за значење на влезните карактеристики ќе се користат за изведување на заклучоци околу изворите на етничка и расна пристрасност.
- Дефинирање на конечен, динамичен, пренослив пристап за откривање на етничко-расна пристрасност во медицинските податоци.
- Дефинирање на чекори за реализација на потребни прилагодувања со цел примена на предложениот пристап при идентификација на различни типови пристрасност во слично структурирани бази на податоци.
- Апликации и придобивки од предложениот пристап во реална примена.

2.2 Образложение на работните хипотези и тези

Во оваа докторска дисертација, главната хипотеза вели дека предложениот пристап ќе овозможи идентификација на етничка и расна пристрасност во клинички податоци на пациенти, поконкретно во витални знаци, и последователно ќе овозможи подлабоко разбирање на изворите на пристрасност присутни во податоците, а со тоа и ќе даде идеја за неопходните чекори со цел да се подобри пристапот до неа и искуството на пациентите, со што ќе се овозможи побезбедна, висококвалитетна здравствена заштита, како и создавање непристрасни популациони групи од пациенти при истражување, непристрасни медицински податочни множества и непристрасни алгоритми базирани на ВИ.

Главната хипотеза се заснова на неколку хипотези за поддршка, наведени во продолжение:

- Хипотеза 1: Предложениот модел ќе овозможи детекција на етничко-расна пристрасност во медицински податоци.
- Хипотеза 2: Предложениот модел ќе овозможи детекција на етничко-расна пристрасност во виталните знаци на пациенти.
- Хипотеза 3: Предложениот модел ќе даде објаснување на изворите на расната пристрасност.
- Хипотеза 4: Предложениот модел ќе има динамичен и пренослив пристап која ќе може лесно да се модифицира за детекција на различни видови на пристрасност во различни податочни множества.

2.3 Научни методи кои ќе се применат

Истражувањето ќе се спроведе користејќи неколку специфични методски постапки и техники на истражување: анализа, класификација, компарација, синтеза, индукција, дедукција и експериментирање.

Прво, анализата ќе се фокусира на разбирање на влијанието и импликациите на различните типови на пристрасност присутни во медицинската пракса, технологијата и податочните множества, како и медицинските апликации базирани на ВИ, со изразен фокус на последиците во случаите со појава на расна пристрасност. Второ, ќе бидат прикажани претходните истражувања во областа на идентификација на етничко-расна пристрасност, а пропратно и детална анализа за ефикасноста на постоечките модели, нивните сличности и разлики со ова истражување, како и нивните силни страни, слабости, можности и закани. Конечно, ќе бидат анализирани и споредени неколку медицински податочни множества, со што ќе се воведат потенцијалните аспекти и елементи на податоците кои содржат предиспозиција за воведување или зголемување на пристрасноста. Врз основа на добиените резултати од анализата, од сите дискутирани податочни множества, ќе бидат избрани оние кои се најпогодни за целите на докторската дисертација. Откако ќе се изберат базите на податоци, значајните карактеристики (карактеристиките од интерес) ќе бидат избрани и преработени за да одговорат на барањата на студијата. Дополнително, ќе се користи споредба на пациентите со цел да се создадат групи на пациенти врз основа на релевантната сличност на демографските и клиничките информации на пациентите (на пр. возраст, пол, дијагноза и/или коефициентите за евалуација на сериозност на болеста). Откако ќе се завршат овие чекори, ќе се синтетизира пристап за откривање на пристрасност во медицинските податоци на пациентите. Врз база на предложениот пристап со избраните податоци ќе се извршат експерименти со цел да се утврди присуството на етничка и расна пристрасност во медицинските информации на пациентите. Експерименталната фаза ќе ги спореди перформансите на различни МУ алгоритми, по што ќе се разгледа можност за примена на длабоко учење (ДУ) и невронски мрежи. На крајот, користејќи достапни методи за интерпретабилност (анг. interpretability) на МУ моделот, ќе се спроведе истрага за карактеристиките важни за моделот, што ќе ги разоткрие изворите на пристрасност од кои што учат моделите. Со цел да се зајакне кредибилитетот на истражувањето, целиот процес ќе се повтори на најмалку на две посебни бази на податоци. За да се формира конечна проценка на перформансите на предложениот пристап, ќе се користат методи на споредба и евалуација (статистички и други),

како и метрики за евалуација на перформансите на МУ моделот. Како алатки за извлекување на заклучоци од резултатите, ќе се користат методите на индукција и дедукција.

Со анализа на заклучоците добиени од перформансите на моделите, како и со користење на индукција и дедукција, ќе се дискутираат чекорите за ублажување на пристрасноста. Дополнително, предложениот пристап ќе биде анализиран за негова применливост во откривањето на пристрасност во медицинските податоци, како и неговата потенцијална примена во пристапи со поширок опсег, како што би била употребата на предложениот пристап во решавањето на предизвиците од различни видови на пристрасност во слични услови и податочни параметри.

2.4 Очекуван научен придонес

Според целите на истражувањето, очекуваниот научен придонес се очекува да биде синтеза на ефикасен систем за откривање на расна пристрасност во медицинските информации на пациентите, поконкретно виталните знаци на пациентите. Се очекува предложениот пристап да даде јасна слика за изворите кои предизвикуваат расна пристрасност (ако навистина постои пристрасност), што пак ќе овозможи разбирање на неопходните чекори за ублажување на пристрасноста. Затоа, индиректно предложениот метод ќе влијае на намалувањето на конфликтите и проблемите кои произлегуваат од пристрасната практика, технологијата, податочните множества и алгоритмите, со што би се обезбедила подобра точност на активно-користени апликации базирани на ВИ. Во рамките на истражувањето се очекуваат следните придонеси:

- Класификација, анализа и импликации на различни форми и извори на пристрасност во медицински услови;
- Компаративна анализа на повеќе јавно достапни податочни множества од витални знаци, но и етничката и расната транспарентност на собраните податоци;
- Компаративна анализа на етничките и расните разлики што може да се забележат во податочните множества избрани за студијата (доколку постојат такви разлики);
- Нов пристап за откривање на самопријавената етничко-расна припадност од виталните знаци на пациентите, со што се откриваат потенцијалните извори на пристрасност во медицинската пракса, медицинската опрема и медицинските податоци (дури и во случаи на податоци коишто треба да бидат целосно непристрасни);
- Анализа на перформансите на системот, неговиот потенцијал за примена во широк опсег и предложена структура за понатамошна повторна употреба и надградба на системот.

2.5 Примена на резултатите од истражувањето

Пристапот предложен во ова истражување ќе се примени за брзо и точно откривање на етничко-расна пристрасност во медицински информации на пациенти и како надополнување на алгоритмот, пристапот ќе ги истакне и изворите од кои потекнува таа пристрасност. Згора на тоа, кодот што го имплементира пристапот ќе биде динамичен, па лесно ќе може да се пренесе (што ќе биде илустрирано со вклучување на повеќе податочни множества) во различни структури и организации на бази на податоци, на различни карактеристики и различни типови на пристрасност со мали измени. Ова ќе овозможи спроведување и употреба на пристапот во бројни здравствени установи и истражувачки центри кои работат на откривање, разбирање и ублажување на расната пристрасност (и други форми на пристрасност) за медицински цели.

Самата употреба на предложениот пристап ќе резултира во откривање на сомнителни аспекти поврзани со постоење на пристрасност, и затоа, индиректно ќе ги истакнува недостатоците (поврзани со маргинализирани популациони групи на пациенти) кои постојат во здравствениот систем. Покрај тоа, пристапот предложен во оваа докторска дисертација ќе биде тестиран не само на етничка и расна пристрасност, туку и на различни видови пристрасност, што може да вклучуваат пол, возраст, социоекономски статус, итн. Ова дополнително ќе илустрира дека предложениот пристап е флексибилен и може да биде подложен на корекција, што значи дека во случаи кога на

системот се даваат дополнителни информации, истиот пристап по минимални промени треба успешно да се прилагоди на нив.

Како заклучок, извршената анализа и предложениот пристап може да послужат како повик за будење за здравствените работници и да ја истакнат потребата од ублажување на пристрасни практики, а во исто време ќе послужат како темел за понатамошни истражувања и подобрувања.

3 Мотивација и преглед на литература

Благодарение на брзиот развој на ВИ и ефикасноста што истата ја нуди, денес постојат многу индустрии кои целосно се потпираат на нејзината примена во решавање на секојдневни задачи. Имено, ВИ е важен дел од широк спектар на активно-користени апликации со примена која опфаќа сè од социјални медиуми и персонализирани системи за препораки, па сè до паметни домови, паметни автомобили, безбедносни системи, итн. Притоа, придобивките што ги овозможува ВИ се многубројни. За жал, со позитивните аспекти на ВИ доаѓаат и слабостите. На пример, социјалните медиуми и системите за препораки користат значителна количина на лични информации за да влијаат и ги обликуваат мислењата на корисниците до алармантен степен, паметните домови и паметните автомобили се потпираат на уреди кои имаат безбедносни пропусти и можат да бидат хакирани, итн.

Имајќи ги предвид овие недостатоци, нормално е да се постави прашањето: „Согледувајќи дека ВИ има толку висока маргина на грешка, што се случува кога истата ВИ технологија се користи во чувствителни области како што е медицината?“ Постојат голем број на активно-користени медицински апликации кои се базирани на ВИ, [1]. Имено, ВИ е широко користена во медицински услови со примена во сè, од грижа за пациенти и одржување на системи за медицинска евиденција, па сè до работа со наплатни системи. Со тоа, употребата на ВИ во медицината директно, а понекогаш и индиректно, ги дефинира пристапите кои се користат за лекување на пациентите, а со тоа и исходот од нивните престои. Една од многуте примени на ВИ во медицината во денешно време е идентификација и дијагноза на различни, тешки медицински состојби, [29]. Како подгранка на ВИ во дијагностички капацитет е медицинската дијагностика во слика, каде што алгоритмите на ВИ се учат да препознаваат сложени концепти во слика и, врз основа на наученото, да обезбедат здравствени проценки на состојбата на пациентите, што како дијагностичка алатка има постигнато голем успех во изминатите години, [3]. Други апликации на ВИ во медицината вклучуваат, но не се ограничени на: персонализирана медицина, [30], електронска здравствена евиденција, [31], клиничко испитување и истражување, [32], откривање и производство на лекови, [33], итн.

Сепак, технологијата е склona кон различни форми на неисправност, па се очекува дека кај активно-користените апликации може да се појават различни проблеми. Од сите потенцијални проблеми постојат такви кои се случуваат како ненамерен и неочекуван нуспроизвод на ВИ, што ги прави посериозни бидејќи се т.н. „тивки“ или „скриени“ проблеми. И покрај тоа што има случаи каде нуспроизводите имаат позитивна страна и придонесуваат до подобрување на одреден систем, најчесто тоа не е така и конотацијата на добиениот нуспроизвод е негативна. Од таа причина, употребата на ВИ во медицината и здравството е полна со потенцијал за клинички, социјални и етички конфликти. Имено, пациентите може да бидат оштетени поради преовладувачки грешки во моделите базирани на ВИ, кои пак настануваат под влијание на пристрасни нееднакости во здравствениот систем, дополнително зајакнати од недостаток на транспарентност во процесот на селекција на пациенти при креирање медицински податочни множества, како и евидентен недостаток на транспарентност во развојот на медицински апликации базирани на ВИ.

Имајќи ги предвид овие ризици, од огромен интерес е да се разберат предрасудите кои постојат во медицинската пракса и да се адресираат потенцијалните проблеми и стапици при користење на пристрасни податочни множества и пристрасни апликации базирани на ВИ. Исто така, потребно е поттикнување на развој на свеста за сите проблеми кои се појавуваат при примена на ВИ во медицински услови. Следствено, ова би значело подобрување на здравствениот систем чија цел е „помагање на сите пациенти подеднакво“ и кој во моментот очигледно не функционира баш така.

3.1 Мотивација

Според македонскиот речник, зборот пристрасност (анг. bias²) се дефинира како:

- нешто што се става на нечија страна;
- уважува нешто пред другите;
- што дава предност на нешто;
- што е засновано на необјективност.

Според тоа, значењето на зборот пристрасност во медицинската пракса може да се објасни како склоност да се покаже уважување кон одредени пациенти или групи пациенти на сметка на други маргинализирани групи, врз база на необјективни гледишта, така што последователно може да придонесе до негативно влијание врз маргинализираната група. Од тука, може да се заклучи дека присуството на пристрасност во медицинската пракса претставува дискриминација кон селективна група пациенти, што може да резултира со значително послаби здравствени услуги и здравствена нега за тие групи пациенти, што пак може да доведе до долгорочни компликации во здравствената состојба на пациентите, дополнителни нарушувања и во најлош случај дури и смртни случаи кои можеле да бидат спречени.

Во неодамнешна презентација на Marzyeh Ghassemi за примена на ВИ во здравството, тема на дискусија бил фактот дека пристрасноста е веќе широко присутна во клиничката пракса, [34]. Пристрасноста присутна во здравствениот систем може: прво, да постои свесно и да се манифестира преку предрасуди како што се расизмот и сексизмот; второ, да биде вгнездена во несвесни, но зацементирани мислења засновани на научени стереотипи; или, трето, може да се случи по грешка при работа со мал, униформен дел од населението. Во која било од трите наведени можности, може да се појават различни форми на пристрасност: етничко-расна пристрасност, пристрасност врз основа на пол, род и/или сексуален идентитет, социоекономска пристрасност, образовна пристрасност, како и пристрасност што произлегува од географската локација, прекумерна тежина или возраст. Поделбата на пристрасност во медицината е илустрирана и на **Слика 3-1**. Некои форми на пристрасност се поддоминантни во медицината од другите, но секоја од нив има повеќекратни последици што може негативно да се одразат на одредени групи на пациенти.



Слика 3-1 Видови на пристрасност во медицина

² a tendency (either known or unknown) to prefer one thing over another that prevents objectivity; a systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others.

Пристрасноста поврзана со ВИ во медицината може да се расчлени на четири различни типа: податочна, алгоритамска, технолошка и човечка, [35]. Во суштина, сите овие аспекти се директно или индиректно поврзани со човеково влијание во медицината. Имено, пристрасните медицински одлуки и ставови, како и пристрасното однесување на медицинските лица се секојдневна појава, а за жал, исто така, се пропагираат и се предаваат на студентите по медицина, [36].

3.1.1 Пристрасност во медицинска пракса

Првиот вид на пристрасност што ја истраживме е човечкиот аспект на пристрасност, т.е. пристрасноста која постои во медицинската пракса. Според клучните наоди презентирани во Американскиот национален извештај за здравствена заштита и диспаритети, [37], спроведен за 2019 година, забележано е дека и покрај напорите за отстранување на пристрасноста, ваквите некоректни ставови во медицината опстојуваат, а дел од нив дури и се влошиле. Притоа, влошувањето е најизразено кај посиромашните и неосигурени популации. Извештајот [37], исто така, илустрира разлики во медицинската грижа врз основа на локација на живеење. Дополнително, посочени се и етничко-расни разлики, т.е. било откриено дека припадниците на бела раса добиваат значително подобра нега во споредба со другите етничко-расни групи: припадници на црна раса, американски Индијанци, домородци од Алјаска, Хиспанците, Азијците и домородните Хавајци. Во [38] авторите дискутираат за тоа како луѓето припадници на одредени раси (пациентите со потемна боја на кожата) во САД се соочуваат со значајни разлики во пристапот до здравствените услуги и во квалитетот на добиената нега, што последователно влијае на здравствената состојба на овие пациенти и исходите од грижа. Според [39], луѓето со потемна боја на кожа се соочуваат со повеќе бариери при пристапот до нега, што за нив значи и намален пристап до превентивни услуги и третмани. Авторите на [39] наведуваат дека дури и кога пристапот до нега е обезбеден, кај пациентите со потемна боја на кожа има поголем број на незадоволителни интеракции со давателите на здравствените услуги. Расна пристрасност постои и во кардиохирургијата, каде студија спроведена во 2016 година покажува дека постојат диспаритети при избор на пациенти што ќе бидат препорачани за бајпас операција, [40]. Поточно истражувањето тврди дека причината за намалената препорака на пациенти од црна раса се должи на тоа што некои лекари имаат тенденција да препорачуваат пациенти од бела раса наместо пациенти од црна раса, бидејќи сметаат дека припадниците на црната раса нема да се придржуваат до потребната физичка активност потребна по операцијата.

Пријавени се и бројни случаи на родова пристрасност и пристрасност врз база на сексуален идентитет. Студијата [41] посочува дека лекарите почесто ја отфрлаат хроничната болка кај жените отколку кај мажите, на тој начин што оваа разлика ја објаснуваат како дистинкција помеѓу „храбри мажи“ во споредба со „емоционални жени“. Ваквите предрасуди може да доведат до замолчување на пациентите при појава и решавање на важни здравствени проблеми. Незадоволството од однесувањето на медицинските лица при прегледи е честа појава и кај трансродовите лица, кои не се чувствуваат удобно при барање на соодветна здравствена нега поради тоа што очекуваат некоректно однесување и предрасуди, [42].

Постојат и други видови на пристрасност кои се присутни во однесувањето, ставовите и мислењата на медицинските лица. Имено, според истражувањето во [43] лекарите сметаат дека работата со постари пациенти и нивните семејства претставува предизвик, а постарите пациенти ги опишуваат како напорни, навредливи и со желба сами да го управуваат сопственото лекување. Исто така, пациентите со посебни потреби имаат ограничен пристап до одредени делови на здравствените центри. Авторите на [44] дискутираат за тоа како над 80% од медицинските лица повеќе би сакале да работат со луѓе без попреченост. Луѓето со голема телесна тежина, исто така, во голем број случаи се соочуваат со лош третман од страна на здравствените работници, а дополнително често нивните симптоми се занемаруваат поради или се препишуваат на зголемената телесна тежина, [45]. Кај луѓето со понизок социоекономски статус постои поголема веројатност за доцнење во

тестирањето и лекувањето, што создава проблеми со одржување на редовни и квалитетни превентивни третмани, [46].

Севкупно, претставените типови на пристрасност резултираат со намалување на пристапот и квалитетот на здравствената заштита за одредени групи пациенти врз основа на предрасудни мислења и верувања длабоко вкоренети во однесувањето на медицинскиот персонал. Последователно, овие пациенти се изложени на различни сериозни ризици кои директно влијаат на нивното здравје, како што се: задоцнет или непостоечки третман, неправилна дијагноза, занемарување на сериозни состојби и компликации кај пациентите, итн.

Истражувањето од оваа секција е прикажано во **Табела 3-1**, каде по колони е наведен бројот на референцата од текстот, типот на пристрасност кој постои во соодветното истражување, опис на содржината на разгледаниот труд, изворот на пристрасност, како и импликациите од наведеното.

Табела 3-1 Приказ на трудови кои покажуваат пристрасност во медицинска пракса

Реф.	Тип на пристрасност	Опис на трудот	Извор на пристрасност	Импликации
[37]	социоекономска, локациска, етничко-расна	влошена грижа за посиромашни популации; лоша грижа за одредени етничко-расни групи	медицинска пракса	слаба ефикасност на стратегии за намалување на пристрасност
[38]	етничко-расна	луѓе со потемна боја на кожа се соочуваат со диспаритети во пристап и квалитет на нега	медицинска пракса	намалена превентивна грижа и влошени исходи кај пациенти со потемна боја на кожа
[39]	етничко-расна	луѓе со потемна боја на кожа се соочуваат со бариери при пристап до медицинска грижа	медицинска пракса	намалена превентивна грижа и влошени исходи кај пациенти со потемна боја на кожа
[40]	етничко-расна	препораки за бајпас операција базирани на боја на кожа	медицинска пракса	на пациенти од црна раса им била одбиена можност за операција за бајпас
[41]	полова	немарност за хронична болка кај жените	медицинска пракса	може да доведе до замолчување на пациенти за важни здравствени проблеми
[42]	полова	трансродовите пациенти не се чувствуваат удобно да побараат соодветна здравствена нега	медицинска пракса	трансродовите пациенти може да избегнуваат посета на лекари освен во итни случаи
[43]	старосна	медицинските лица ја сметаат работата со постари лица како предизвикувачка	медицинска пракса	некоректен третман кон постари пациенти
[44]	способносна	преку 80% од медицинските лица преферираат да не работат со луѓе со попреченост	медицинска пракса	поголема желба за работа само со работоспособни пациенти може да создаде проблем за пристап до нега кај пациентите со посебни потреби
[45]	тежинска	кај пациенти со зголемена тежина симптомите често се препишуваат на тежината	медицинска пракса	одредени медицински состојби може да се превидат
[46]	социоекономска	кај посиромашни луѓе може да се појавува поголемо доцнење во тестирања и третмани	медицинска пракса	повисоки шанси за влошување на медицински состојби и компликации поради време на чекање

3.1.2 Пристрасност во медицинска технологија

Истражувања за пристрасност во медицината се проширени и на квалитативна анализа на медицинската технологија која активно се користи за проценка на состојбата на пациентите. Вроден во самата медицинска технологија, поради фактот што тестирањето на уредот не опфатило разновидна популација, овој тип на пристрасност долго време бил занемаруван, додека самите апарати кои ја инкорпорираат оваа пристрасност постојано се користеле и имале директно влијание на мерења од пациентите користени за проверка на здравствена состојба.

Првата регистрирана појава на пристрасност во медицинската технологија датира од 1968 година, [47]. Имено, според упатствата дадени од производителите на рендген машините, било дефинирано дека има потреба за поголема изложеност на зрачење за пациенти од црната раса, како последица на лажното верување дека луѓето со потемна боја на кожата имаат погусты коски и подебела кожа. Ова понатаму било апсорбирано во медицинската пракса, па прераснало во секојдневност и резултирало со тоа што рендген техничарите со години рутински и непотребно ги изложувале пациентите со потемна боја на кожа на повисоки дози на зрачење во споредба со оние на припадници од бела раса. Во [48] авторите заклучуваат дека кај бесконтактните инфрацрвени термометри, кои што ја мерат температурата преку кожата на челото со помош на инфрацрвена технологија, постои 26 % помала веројатност да истите успешно детектираат треска кај припадници на црна раса во споредба со употреба на оралните термометри. Друга студија, [49], работена за време на пандемијата КОВИД-19, покажува дека при употреба на пулс оксиметрите постои три пати помала веројатност тие да детектираат ниски нивоа на кислород кај пациенти со потемна боја на кожа (особено припадници на црна раса), што резултирало со одложување на потребниот третман и придонело до поголем ризик кај пациентите со потемна боја на кожа.

Всушност, пристрасноста во медицинската опрема, особено кај опрема тестирана на лимитирана група пациенти која сепак се користи секојдневно на разновидни популации, може да има непростливи последици за несоодветно опслужените пациенти. Проблемите со опремата доведуваат до казнување на одредени групи на пациенти, што може да придонесе за одложување на дијагнозата или администрацијата на третманот, и последователно да резултира со зголемен број на смртни случаи.

Истражувањето од оваа секција е прикажано во **Табела 3-2**, каде по колони е наведен бројот на референцата од текстот, типот на пристрасност кој постои во соодветното истражување, опис на содржината на разгледаниот труд, изворот на пристрасност, како и импликациите од наведеното.

Табела 3-2 Приказ на трудови кои покажуваат пристрасност во медицинска технологија

Реф.	Тип на пристрасност	Опис на трудот	Извор на пристрасност	Импликации
[47]	етничко-расна	повисок степен на радијација за пациенти од црна раса	медицинска опрема	усвоено и рутински применувано во медицинска пракса
[48]	етничко-расна	челни термометри со 26% помала веројатност за детекција на треска кај пациенти од црна раса во споредба со орални термометри	медицинска опрема	пропуштени високи температури може да влијаат на одложување на третмани, а со тоа и зголемена смртна стапка кај пациенти од црна раса
[49]	етничко-расна	пулс оксиметри имаат три пати помала шанса за детекција на ниски нивоа на кислород кај пациенти од црна раса	медицинска опрема	придонесува до одложување на третмани и може да доведе до зголемена смртна стапка кај пациенти од црна раса

3.1.3 Пристрасност во медицински податочни множества

Без никаква рестрикција, системските човечки предрасуди, стигматичните мислења и пристрасност во медицинската опрема можат лесно да се пренесат во медицинските податочни множества и во медицинските алгоритми базирани на ВИ. На тој начин би се зголемило присуството на пристрасност во широк спектар на апликации развиени со цел да се подобрат здравствената нега и искуството на болните пациенти.

Пристрасните податочни множества се или полни со пристрасни маркери или имаат недоволно застапеност на одредени групи на пациенти (на пр. доминантна раса, пол, социоекономски статус, и сл.), што произлегува од една или повеќе во листата причини:

- систематска дискриминација што произлегува од несправеден третман на пациентите од страна на медицинскиот персонал поради етничко-расни, полови, социоекономски или дополнителни видови на пристрасност;
- пристрасност во процесот на собирање на податоците;
- недостаток на разновидност и интердисциплинарност при креирање на медицинска опрема и тестирање на квалитетот на креираната технологијата;
- или едноставно, недостаток на детални квалитативни испитувања на остварените технолошки, клинички и научни истражувања.

Собирањето податоци од медицинските установи без отстранување (или барем ублажување) на пристрасни мислења, практики и третмани на медицинскиот персонал може ефективно да доведе до пристрасност во медицинските податочни множества. Имајќи ги во предвид веќе презентирани некоректности во однесувањето на медицинските работници, разбирливо е зошто постојат многу пријавени случаи на пристрасност во медицински податочни множества. Имено, во податочните множества пристрасноста може да потекнува од имплицитна пристрасност на медицинските лица, или од имплицитна пристрасност за време на процесот на селекција на податоци и од недоволна застапеност на различни групи пациенти.

Податоците треба да бидат репрезентативни за разновидноста на популацијата, во спротивно тие може само да го засилат недостатокот на генерализација и да ги истакнат различните форми на пристрасност, [35], [50]. Во [51] авторите ја истражуваат медицинската историја на 18.259 пациенти преку анализа на лекарски белешки собирани во урбан академски медицински центар. Студијата го анализира присуството на негативни дескриптори во деталите за пациентите и нивното однесување, како што се на пример несогласен или настојчив. Резултатите од истражувањето покажуваат присуство на расна пристрасност во електронската здравствена евиденција. Според [51] кај пациентите од црна раса има 2,5 пати поголема веројатност за постоење на негативни дескриптори во медицински белешки во споредба со пациентите од бела раса.

Податочните множества составени од медицински слики се исто така богати со различни форми на пристрасност. Авторите во [52] се осврнуваат на важноста на половата рамнотежа во податочни множества од медицински слики, на тој начин што истражувањето го прикажува намалувањето на перформансите кај модели за ВИ при работа со небалансирани податочни множества, покажувајќи значително влошени перформанси за пациенти чиј пол е недоволно застапен. Авторите го истражуваат и степенот на влијание што небалансираните податочни множества го имаат врз перформансите на моделот. Тие заклучуваат дека кога се работи со однос на нерамнотежа помеѓу класите кој изнесува 25 % / 75 %, перформансите на моделот се значително послаби за малцинската класа во споредба со мнозинската класа. Од друга страна, авторите не ја забележуваат таа разлика во балансираните податочни множества.

Податочните множества составени од медицински слики се анализираат за небалансираност во поглед на етничко-расна припадност на пациенти. При анализа на податочно множество составено на рендгенски снимки од бели дробови на 53.000 пушачи кои имале рак на белите дробови, [53], се покажало дека само 4 % од пациентите вклучени во креирање на податочното

множество се пациенти од црна раса. Друго податочно множество познато по пристрасност на податоците е International Skin Imaging Collaboration, [54]. Ова податочно множество е со отворен пристап и претставува едно од најупотребуваните множества за проверка на лезии на кожата со цел дијагноза на меланом (најсериозната форма на рак на кожата). И покрај достапноста и широката употреба на множеството, собраните податоци претежно се од пациенти со посветла боја на кожа, [55]. Друга форма на пристрасност е географската пристрасност, детално анализирана во [56]. Студијата [56] е спроведена во 2020 година и покажува дека во 71 % од студиите спроведени во САД каде што била забележана и географската локација, се користат податоци од само три сојузни држави: Калифорнија, Масачусетс и Њујорк. Дополнително, истражувањето открива дека постоечките податочни множества користат информации само од 16 сојузни држави, додека за останатите 34 не постојат достапни податочни множества.

Севкупно, презентираниите случаи во кои пристрасност е забележана во медицински податочни множества покажуваат дека податоците кои се отворени и со тоа достапни за истражувачите може да бидат значително ограничени и фокусирани на група на пациенти од одредена географска област и истовремено да припаѓаат на една доминантна раса или пол. Ова влијае на рамномерна застапеност на различни групи на пациенти. На тој начин, податочните множества во себе носат ограничено знаење што не дозволува темелно разбирање на медицинските состојби и суптилните промени кои може да се појават кај различни популации на пациенти.

Истражувањето од оваа секција е сумаризирано во **Табела 3-3**, каде по колони е наведен бројот на референцата од текстот, типот на пристрасност кој постои во соодветното истражување, опис на содржината на разгледаниот труд, изворот на пристрасност, како и импликациите од наведеното.

Табела 3-3 Приказ на трудови кои покажуваат пристрасност во медицински податочни множества

Реф.	Тип на пристрасност	Опис на трудот	Извор на пристрасност	Импликации
[51]	етничко-расна	медицински белешки од 18.259 пациенти покажуваат повеќе негативни дескриптори за одредени етничко-расни групи	медицинска пракса проследува во медицински податочни множества	2,5 повеќе негативни дескриптори кај пациенти од црна наспроти пациенти од бела раса што води во апликации кои ја учат таа разлика и работат врз база на пристрасноста
[52]	полова	намалување на перформанси на модели во случаи на недоволна репрезентација на одредени популации	медицински податочни множества проследуваат во медицински апликации	работа со небалансирани податоци прави значителен пропуст за помалку застапената група и може да резултира со апликации кои имаат лоши перформанси за одредени групи на пациенти
[53]	етничко-расна	само 4% од избрани пациенти во податочно множество за дијагноза на рак се припадници на црна раса	медицински податочни множества	недоволна застапеност која води до апликации кои даваат послаби резултати за одредени етничко-расни групи
[54]	етничко-расна	податочно множество за детекција на меланом собирано од луѓе со светла боја на кожа	медицински податочни множества	слаба репрезентација на одредени групи може да доведе до апликации кои не го детектираат ракот кај пациенти со потемна боја на кожа
[56]	локациска	71% од студии со вклучена географска локација содржат информации од само три сојузни држави	медицински податочни множества	недоволна географска покриеност може да доведе до модели кои нема да работат точно за пациенти од други географски локации

3.1.4 Пристрасност во медицински апликации базирани на ВИ

Кога алгоритмите за ВИ користат пристрасни податочни множества за време на процесот на тренирање на моделот, тие имаат ограничен преглед на проблемот кој треба да го решаваат и добиваат подобро разбирање за проблемот само од перспектива на доминантно присутната група. Затоа моделите можат да ја научат пристрасноста што ја вклучуваат податоците и да имаат помала точност, т.е. послаби перформанси, за одредени групи на пациенти. Следствено, обучените алгоритми ги зајакнуваат нееднаквостите кои постојат во здравствените системи за многу, различни примени: од алгоритми за откривање на рак кои се помалку ефикасни за припадници на црна раса, [57], до алгоритми за процена на ризик за срцев удар кои ја потценуваат потребната нега за пациентите од црна раса [58]. Постојат извештаи за пристрасност во алгоритмите што се користат во мајчина здравствена заштита. Имено, широко користен алгоритам за предвидување на успешно вагинално породување по претходно направен царски рез (Vaginal Birth after Caesarean, т.е. VBAC алгоритмот) во САД придонел за повисоки стапки на царски рез кај жените со потемна боја на кожа бидејќи предвидувал пониски успешни стапки за бремени жени со потемна боја на кожа, [59].

Студијата [26] презентира докази за расна пристрасност во активно-користен алгоритам, кој носел одлуки за повеќе од 200 милиони луѓе во САД на годишно ниво. Потеклото на пристрасноста доаѓа од употреба на здравствените трошоци како информација за детекција на здравствените потреби на пациентите. Од таа причина, бидејќи помалку пари се трошат на пациенти од црна раса кои имаат исто ниво на потреба како и пациенти од бела раса, алгоритмот го научил ова несовпаѓање на финансии како обележје за проблемот и затоа моделот доделувал исто ниво на процена на ризик за пациенти од црна и бела раса, и покрај тоа што пациентите од црна раса биле во значително полоша здравствена состојба во споредба со оние од бела. Според проценката направена од авторите на [26], бројот на пациенти од црна раса идентификувани за дополнителна нега бил намален за 2,5 пати во споредба со она што истиот требало да биде.

Истражувањето [60] ги евалуира перформансите на современи алгоритми во откривањето на абнормалности (на пр. пневмонија, јазли на белите дробови, лезии, фрактури, итн.) во рендгенски снимки на градниот кош. Добиените резултати покажуваат дека кај младите жени има највисока стапка на погрешна дијагноза, следни се пациентите од црна раса, а потоа следуваат пациентите што имаат јавно здравствено осигурување поради ниски примања. Грешните дијагнози се уште повеќе изразени кај оние пациенти кај кои постои спој на повеќе од наведените критериуми, т.е. пациенти од црна раса со јавно осигурување и ниски примања имаат највисока стапка на погрешна дијагноза.

Истрага за алатка базирана на ВИ за рано откривање на сепса, активно-користена од повеќе од 170 болници, покажала неспособност на моделот да ја предвиди оваа животозагрозувачка болест кај 67% од пациентите кои ја развиле, [61]. Од друга страна, моделот генерирал голем број на лажни предупредувања за можна појава на сепса кај илјадници пациенти кои никогаш воопшто не ја развиле болеста.

Друг алгоритам, [62], предложил екстремни намалувања на домашната нега за пациентите со посебни потреби, што предизвикало екстремно нарушување на животот на пациентите и резултирало со зголемена хоспитализација. Друга студија која анализира присуство на социоекономска пристрасност имала за цел да го процени степенот до кој квалитетот на податоците во електронските здравствени досиеја е поврзан со социоекономскиот статус, [63]. Моделите за МУ испитувани во [63] имаат за цел да го предвидат влошувањето на состојбата кај деца со астма. Резултатите покажуваат полоши перформанси на моделот за предвидување кај пациенти со понизок социоекономски статус. Модел заснован на ВИ за дијагноза на Алцхајмеровата болест преку аудио податоци, развиен во Канада, покажал слаби перформанси за пациенти со одредени акценти, бидејќи процесот на тренирање на моделот вклучувал примероци од говор со само еден акцент, што ја направило апликацијата неупотреблива за сите останати во земјата, [64].

Севкупно, медицинските апликации засновани на ВИ, кои активно се користат во медицински услови, се имаат покажано како пристрасни врз различни групи на пациенти. Како последица на нив, постојат случаи во кои сериозно бил нарушен квалитетот на грижата што требало да биде понудена на пациентите. Многу пријавени случаи придонеле до загрозување на животите на пациентите поради тоа што ја пропуштиле дијагнозата на болеста во ситуации опасни по живот. Други случаи покажуваат непотребен стрес предизвикан поради неточни дијагнози за постоење на заболување кои подоцна се покажани како непостоечки. Дополнително, недостатоците на алгоритмите се поизразени кај пациенти кои имаат намален пристап до здравствена заштита, поради што се создаваат сериозни потешкотии, т.е. пациентите со ниски приходи и ограничен пристап до медицинска нега не можат да си дозволат мислење од друг лекар, што ја прави погрешната дијагноза во овие случаи на некој начин и казнена акција кон пациентот.

Истражувањето од оваа секција е прикажано во **Табела 3-4**, каде по колони е наведен бројот на референцата од текстот, типот на пристрасност кој постои во соодветното истражување, опис на содржината на разгледаниот труд, изворот на пристрасност, како и импликациите од наведеното.

Табела 3-4 Приказ на трудови кои покажуваат пристрасност во медицински апликации базирани на ВИ

Реф.	Тип на пристрасност	Опис на трудот	Извор на пристрасност	Импликации
[57]	етничко-расна	алгоритам за детекција на рак со слаби перформанси за пациенти од црна раса	медицински апликации	зголемена смртна стапка кај пациенти од црна раса што можела да се избегне
[59]	етничко-расна	VBAC алгоритам предвидува поголем степен за царски рез кај жени со потемна боја на кожа	медицински апликации	зголемен број на потенцијално непотребни процедури кај жени со потемна боја на кожа
[26]	етничко-расна	ист ризик фактор доделен и кај пациенти од црна и бела и покрај тоа што тие од црна се во полоша здравствена состојба	медицински апликации	бројот на пациенти од кои не биле препратени за дополнителна нега е 2,5 помал од тоа што требал да биде
[60]	социоекономска, полова, етничко-расна	рендгенски снимки од граден кош покажуваат најголем степен на погрешна дијагноза се појавува кај млади жени, проследени со пациенти од црна раса, и на крај пациенти со јавно здравствено осигурување	медицински апликации	занемарување на сериозни заболувања
[62]	способносна	намалувања на домашна нега на хендикепирани пациенти	медицински апликации	нарушување на животот на пациентите и последователно зголемена хоспитализација
[63]	социоекономска	анализа на степен до кој квалитетот на податоците на електронски здравствени досиеја зависи од социоекономскиот статус на пациентите	медицински податочни множества проследуваат во медицински апликации	влошени перформанси на модели кај пациенти со понизок социоекономски статус
[64]	лингвистичка	дијагностичка алатка за Алцхајмерова болест не работи за пациенти со различен акцент	медицински податочни множества проследуваат во медицински апликации	апликации неупотребливи за голема популација на пациенти

3.1.5 Сумарен преглед на пристрасност и значајни заклучоци

Од изобилството наведени примери, очигледно е дека пристрасноста е присутна во медицинската пракса и технологија. Дополнително, таа пристрасност може лесно да се пренесе од медицинската пракса во медицински податочни множества, и на крајот, до апликации базирани на ВИ. Накратко, податоците за кои се претпоставува дека правично ги рефлектираат различните групи на население во многу наврати не успеале да го направат токму тоа, алгоритмите засновани на ВИ за кои се претпоставува дека имаат еднакви перформанси кај различни групи на население се покажале пристрасни во многу случаи, што е сосема разбирливо кога ќе се земе во предвид дека алгоритмите учат од податочни множества богати со најразновидни влијанија од медицинските лица кои за жал практикуваат примена на општествени предрасуди кон пациенти од различна етничка група, пол, возраст, физички изглед, социоекономски статус, итн. Единствениот начин да се избегне проследувањето на пристрасност во секој дел од развој на модели базирани на ВИ е преземање на соодветни превентивни мерки.

Од согледаните факти и нивната сериозна природа, се отвара прашањето за ублажување (па дури и целосно разрешување) на пристрасноста. Примарниот извор на пристрасност е човечкиот аспект во медицинската пракса. Од таа причина, најважниот чекор е спротивставување на пристрасните практики преку наметнување задолжителна обука на целиот медицински персонал со цел да се зацврсти фундаменталното разбирање за пристрасноста во здравствениот систем и можните последици. Постоене на неисправна и пристрасна медицинска опрема во 21 век е уште еден огромен проблем. Исто, потребна е дивергентност во тестирањето на нова медицинска опрема пред таа да биде пуштена во масовно производство со цел да се потврди дека опремата не може да нанесе никаква ненамерна штета на одредени групи пациенти.

Создавањето податочни множества во кои има присуство на една раса, пол, социоекономски статус и слично, е неетички и таквите податочни множества не треба да бидат достапни за секого ширум светот. Работа со мултидисциплинарни тимови во процес на создавање на медицински податочни множества може да биде корисно за намалување, па дури и потенцијално искоренување на културни или академски слепи точки, што би овозможило создавање на правични податочни множества.

Секоја следна фаза во развојот на модели базирани на ВИ може да подлегне на пристрасност. Со цел да се обезбеди одговорен развој на алгоритам, мора да се преземат чекори во пред-обработка, процесирање и пост-обработка на податоците. Имено, кога се работи на модел, развојните тимови треба да бараат максимизирање на точноста на моделот и истовремено минимизирање на влијанието на разновидни пристрасни маркери. Солиден пристап вклучува и темелна анализа на перформансите на моделот на различни групи на население, со што ќе може да се откријат проблеми со апликациите базирани на ВИ во раните фази на продукција. На тој начин се овозможува наново тренирање и оптимизација на моделите, а со тоа и добивање на модели кои работат еднакво за сите.

Накратко, обезбедувањето исправна медицинска опрема која работи со споредливи перформанси кај најразлични групи од пациенти, како и собирање на широко претставителни податоци прецизно лабелирани е исклучително важно, бидејќи со пристрасни податоци тешко може да се спречи пренос на пристрасност во апликации за ВИ. Понатаму, при креирање на апликации базирани на ВИ мора да се почитуваат одредени правила. Тимовите за развој на апликацијата треба да бидат способни успешно да се соочат со различни аспекти на проблемот, поради што е неопходна огромна палета на познавања. На крајот, дури и откако ќе се преземат сите мерки на претпазливост, моделот мора ригорозно да се тестира за преостанати форми на пристрасност пред да се примени во пракса.

Откако ќе се исполнат сите барања за правичност во фазата на развој, споделувањето на деталите за тоа како е развиен моделот е од суштинско значење, поради неколку причини: прво, тоа ѝ овозможува на истражувачката заедница подобро да ги разбере чекорите што треба да се преземат

за да се развијат непристрасни апликации; второ, ќе се дефинира упатство за тоа како треба да се користи моделот; трето, непристрасни тимови би имале пристап да спроведат дополнителни проценки на пристрасност; четврто, транспарентноста на апликацијата ќе осигура пациентите да имаат доверба во работниот процес на истата; и многу повеќе.

На кратко, овој дел од истражувањето покажува дека постојат фундаментални прашања што треба да се разгледаат и коригираат, и тоа итно, бидејќи тие влијаат на животите на пациенти ширум светот. И, уште побитно, таа промена треба да дојде од сите нас. Ова е главната мотивација врз база на која се заснова истражувањето во оваа докторска дисертација.

3.2 Преглед на литература

Истражувањето презентирano во секција досега го илустрира постоењето на различни видови на пристрасност во медицината. Меѓутоа, од досега спроведените истражувања, најголем дел се фокусираат на анализа на пристрасност во медицинската пракса и однесувањето на медицинските лица, како и на медицински податочни множества и алгоритми базирани на ВИ. Притоа, истражувањата кои се фокусираат на анализа на податочните множества внимаваат на статистичките аспекти на податочните множества, односно прават проверка на застапеност на класи (на пр. род, раса, итн.). Понатаму, истражувањата кои ги анализираат алгоритмите базирани на ВИ се фокусираат на проверка на постоење на пристрасност откако алгоритмот ќе биде целосно истрениран и подготвен за во продукција, а во најлош можен случај дури откако моделот потфрлил при работата за одредена популација пациенти по негово пуштање во продукција.

Сумарно, досегашните истражувања ги анализираат перформансите кои ги постигнуваат методите за МУ при решавање на одреден проблем за одредена раса, етничка група, пол. Но, во рамките на оваа докторска дисертација, тема на истражување е детекција на етничко-расна пристрасност со користење на алгоритми од МУ. Дистинкцијата во споредба со досегашните истражувања е во тоа што наместо користење само на анализи за етничко-расна пристрасност во оценување на податочни множества и на перформанси на модели од МУ, ова истражување се фокусира на примена на алгоритмите во детекција на етничко-расна пристрасност во медицински податочни множества.

Употреба на МУ за детекција на етничко-расна пристрасност во медицински податоци е тема во подем. Постојат истражувања кои покажуваат како алгоритмите базирани на ВИ можат да идентификуваат разни демографски информации за пациенти користејќи медицински слики. Во [65], покажано е дека моделите можат да го предвидат полот и да направат јасна дистинкција помеѓу возрасни и педијатриски пациенти врз база на рендгенски снимки од граден кош. Објавено е и истражување во кое е постигната солидна точност во одредување на возраст на пациенти во слики од различни делови на телото, [66]. Во офталмологија, слики на мрежница на око се користат за предвидување на пол, возраст и срцеви маркери на пациенти, [67], [68], [69].

Досегашни истражувања за детекција на расна пристрасност со помош на методи за МУ се имаат фокусирано на два различни типа на медицински податочни множества и тоа: работа со медицински слики, [70], и работа со текстуални податоци од клинички белешки, [71].

Во [70] авторите користат колекција од приватни и јавно достапни податочни множества во детекција на расна и полова пристрасност во слики со помош на алгоритми на МУ. Најпрвин, во истражувањето авторите применуваат ВИ на рендгенски слики од граден кош за детекција на самопријавена раса во примероци од три големи податочни множества (две јавно достапни и едно приватно податочно множество). Од податоците во јавно достапните множества, 60% ги користат за тренирање на моделот, а 10% ги користат за валидација на моделот, додека преостанатите 30% од податоците се употребуваат за интерна валидација на резултатите (така што овие податоци се невидени од моделот). Дополнително, направена е и екстерна валидација, што значи дека по тренирање на моделите на едно податочно множество, моделите биле тестирани на друго податочно

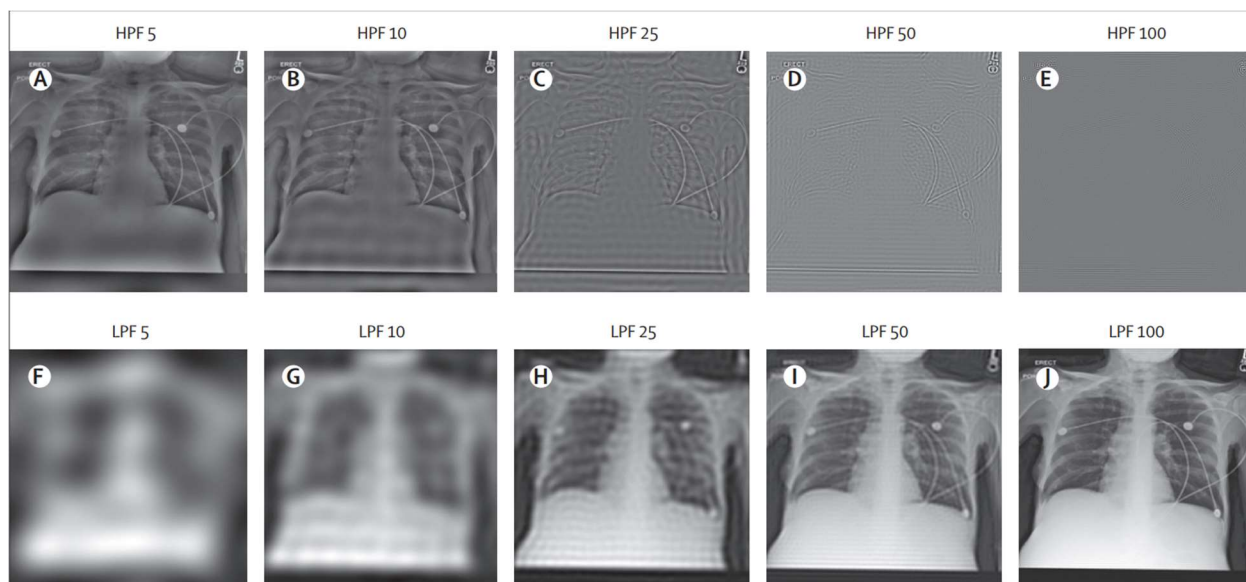
множество. И во случај на двата типа на валидација, алгоритмот покажува дистинкција помеѓу класите (т.е. расите) во опсег од 0.94 до 0.99. Тренирани се и модели за детекција на раса на податочни множества составени од слики од други делови на телото (на пр. екстремитети) кои биле направени со друг вид на технологија (на пр. компјутерска томографија), при што добиените резултати биле релативно слични.

По установување дека моделите на ДУ можат да направат успешна детекција на раса без разлика на тоа со каква технологија се креирани податочните множества и кој дел од телото истите го разгледуваат, целта на авторите на [70] била да разберат како моделите за ДУ ја постигнуваат прецизноста во нивните предвидувања. За таа цел, анализирани се неколку натпреварувачки хипотези:

- разлики во физички карактеристики помеѓу пациенти од различни расни групи (како телесен хабитус или густина на градите),
- разлика во дистрибуција на болести кај пациенти од различни расни групи за да се елиминираат случаи каде една расна група има поголем број на одредени заболувања како потенцијална информација која би можел да ја научи моделот и врз основа на која би ги правел предвидувањата (на пр. ако кај пациенти од црна раса има поголемо присуство на одредена болест, како срцева слабост, моделот може да ја научи оваа информација и да ја користи при предвидување на раса на пациентот, што не би требало да се случува бидејќи пациенти од бела раса може да го имаат истото заболување), и
- разлики специфични за локацијата или ткивото кое се разгледува (на пр. постојат докази дека припадници на црна раса имаат поголема минерална густина на коски и помала стапка на опаѓање на густина низ годините во споредба со припадници на белата раса, [72]).

Дополнително, авторите тестираат дали информациите за раса опстануваат и во случај на високо деградирани слики. За таа цел, врз сликите е направено нископропусно и високопропусно филтрирање со различни дијаметри на филтерот. Резултатите од филтрирањето кај случајно избран примерок се прикажани на **Слика 3-2**, каде првиот ред ги покажува промените од примена на високопропусен филтер (со дијаметар 5, 10, 25, 50 и 100, соодветно), додека вториот ред ги покажува промените во сликата при употреба на нископропусен филтер (со дијаметар 5, 10, 25, 50 и 100, соодветно). Но, и покрај значајна деградација кај сликите која може да се забележи во сите случаи, моделите сепак успешно ја детектираат расата на пациентите.

Авторите на [70] покажуваат дека моделите на ДУ се способни да предвидат раса низ различни клинички средини (покажано преку употреба на податочни множества од различни клинички центри), модалитети на медицинска слика (покажано со користење слики добиени со употреба на различни технологии) и различни популации на пациенти. На тој начин, добиен е заклучок дека моделите не се потпираат на разлики во начинот на кој сликите се земани кај пациенти со различен расен идентитет. Исто така, авторите посочуваат дека врз база на нивното согледување не постојат специфични карактеристики во сликата кои се одговорни за препознавање на расата, ниту во случаи на разгледување на различни локации од сликите, ниту пак при пренос на сликите во фреквентен домен, а ни во самата анатомија на пациентите. Едно ограничување кое го потенцираат авторите на студијата е недоволна репрезентација на повеќе етнички групи, што можеби не илустрира целосна слика за проблемот. Сепак, потенцирано е дека се разгледани познати и постоечки разлики помеѓу расите, но дека и покрај бројни пристапи кон прашањето, причината за успешноста на алгоритмот останува непозната, што е мала контрадикција на бројни истражувања кои посочуваат дека форензички антрополози користат историски и културолошки познавања кои придаваат значење на физички карактеристики кои се појавуваат помеѓу различни популации, [73], [74], како и студии кои прикажуваат разлики во коскената геометрија помеѓу разни етнички и расни групи, [75].



Слика 3-2 Филтрирање (деградација) на слика, според [70], за детекција на раса. Првиот ред илустрира високопропусен филтер (означени со HPF), додека вториот ред илустрира нископропусен филтер (означени со LPF).

Во [71], авторите го истражуваат нивото на имплицитни информации за раса достапни за алгоритмите базирани на ВИ и импликациите од можна детекција на етничко-расни разлики во клинички белешки. Притоа, истражувањето се фокусира на три потцелини:

- се прави проверка на тоа дали моделите базирани на ВИ може да идентификуваат самопријавена раса на пациенти користејќи редактирани клинички белешки од кои биле отстранети експлицитни индикатори на раса,
- се потврдува дека медицински лица, експерти, не можат точно да ја одредат расата на пациентите од редактираните белешки, и
- се дава преглед на потенцијалната штета која што може да биде нанесена од овој тип на информациите во клинички белешки преку симулација.

Истражувањата поврзани со дисциплината на истражување се прикажани во **Табела 3-5**, каде по колони е наведен бројот на референцата од текстот, типот на пристрасност кој постои во соодветното истражување, опис на содржината на разгледаниот труд, изворот на пристрасност, како и импликациите од наведеното.

Табела 3-5 Приказ на трудови поврзани со темата на истражување

Реф.	Тип на пристрасност	Опис на трудот	Извор на пристрасност	Импликации
[65]	полова, старосна	предвидување на полот и јасна дистинкција помеѓу возрасни и педијатриски пациенти врз база на рендгенски снимки од граден кош	биолошки фактори ³	алгоритмите разбираат биолошки фактори подобро од медицински лица

³ под биолошки фактори се подразбираат познати медицински сознанија, како губење на коскена маса со стареење, постоење на маркери за сериозни заболувања (како висок притисок, дијабетес, артритис, итн.) во ретина, итн.

[66]	старосна	солидна точност во одредување на возраст на пациенти во снимки од различни делови на телото	биолошки фактори	алгоритмите несвесно може да носат одлуки врз основа на старост на пациенти, што може да доведе до пропусти во детекција на заболувања врз база на статистички веројатности
[67] [68] [69]	полова, старосна	предвидување на пол, возраст и срцеви маркери на пациенти од слики на мрежница на око	биолошки фактори	алгоритмите може индиректно заклучоците за одредени дијагнози да ги базираат врз основа на несакани параметри (на пр. постоење на маркери за повисок крвен притисок би покажувал на една раса, што моделот може да го научи и користи во предвидување)
[70]	полова, етничко-расна	успешна детекција на етничко-расна припадност во снимки од различни делови на телото земени со различни технологии	биолошки фактори	алгоритмите разбираат биолошки фактори подобро од медицински лица
[71]	етничко-расна	детекција на етничко-расни разлики во клинички белешки	медицинска пракса проследува во медицински податочни множества	некоректна репрезентација на одредени етничко-расни групи, што може да доведе до креирање на алгоритми кои одлуките ги носат врз основа на всадени пристрасни однесувања

За истражувањата релевантни за темата на оваа докторска дисертација во **Табела 3-6**, се прикажани предностите и слабости, можностите и заканите на релевантните трудови. Од табелата може да се заклучи дека предностите на студиите кои анализираат расна пристрасност се состои во работа со различни податочни множества, додека слаби страни се ограничувањата на различни популации на пациенти (најчесто поради недостаток на достапни податоци) и работа со самопријавена раса како параметар за анализа (кој не вклучува детални информации за генетско потекло). Понатаму, анализите од овие студии ги поткрепуваат заклучоците прикажани во **Секција 3.1.5**, а можностите на овие анализи се фокусираат на развој на свест за пристрасност во медицината, т.е. коректно однесување и еднаква репрезентација на различни групи на пациенти (без разлика на пол, раса, старост, идентитет, итн.), како и ригорозно тестирање на моделите од ВИ кои се пуштаат во активна употреба во медицински околина.

Табела 3-6 Анализа на предности, слабости, можности и закани на релевантните трудови

Реф.	Предности	Слабости	Можности	Закани
[70]	1. Анализа на различни податочни множества 2. Анализа на слики од различни делови од телото 3. Анализа на различни технологии за медицински слики 4. Анализа на влијание на деградација на слика	1. Истражувањето се базира на самопријавена раса, што е повеќе општествен, отколку биолошки концепт 2. Мал број на пациенти од одредени етнички и расни групи резултира со анализа на само три доминантни групи (Азијци и пациенти од црна и бела раса)	1. Развој на свест за потребата од ревизии за перформанси на моделот врз база на раса, пол, возраст 2. Развој на свест за вклучување на информации за раса, пол, итн. во медицински податочни множества 3. Намалување или целосно спречување на развојот на модели со расна пристрасност	1. Алгоритмите може да научат да ги вршат предвидувањата врз база на раса и покрај тоа што самите креатори на алгоритмите не го сакаат ова, т.е. можно вложување на расна пристрасност во модели базирани на ВИ

[71]	1. Анализа на пристрасност во медицински белешки од два засебни извори 2. Фокус на диспаратети кај пациенти од црна раса во споредба со пациенти од бела раса 3. Споредба на резултати постигнати од моделите и способноста на медицински лица да предвидат раса од истите белешки 4. Анализа на разликите во темите дискутирани кај различни раси на пациенти (на пр. различни заболувања)	1. Истражувањето се базира на самопријавена раса, што е повеќе општествен, отколку биолошки концепт 2. Студијата е спроведена само помеѓу пациенти од бела и црна раса 3. Предвидувањата од медицинските лица се базирани само на тестирање, без никаков преглед на белешките кои моделите ги користат за тренирање	1. Развој на свест во креирање податочни множества со непристрасни дистрибуции и од непристрасни податоци 2. Обука на медицински лица за намалување на пристрасност	1. Алгоритмите може да научат да ги вршат предвидувањата врз база на раса и покрај тоа што самите креатори на алгоритмите не го сакаат ова
------	--	---	--	--

3.2.1 Претходни истражувања наспроти нас

Според нашето познавање, двете претходно анализирани студии се единствените истражувања кои се релевантни на темата на оваа докторска дисертација. Имено, сличностите на претходните студии и истражувањето спроведено во дисертацијата се состојат во употреба на медицински податоци за предвидување на етничка и расна припадност на пациенти. Меѓутоа, претходни студии користат различни видови на податочни множества (слика и текст), и според спроведените анализи не постојат досегашни студии кои истражувањето го спроведуваат на табеларни податоци, како што се на пример витални знаци на пациенти мерени во ЕИИ. Од таа причина, ние се одлучивме да примениме модели од МУ врз табеларни податоци и да се обидеме да ја детектираме самопријавената раса на пациенти од ЕИИ, што ја прави оваа дисертација првото истражување кое применува алгоритми на ВИ на табеларни податоци за детекција на етничко-расна пристрасност кај пациенти во критична состојба.

Притоа, во **Табела 3-7** се дадени детали за корелација и дисасоцијација на истражувањата од релевантните трудови и истражувањето кое е спроведено во оваа дисертација. Главната разлика помеѓу пристапите обработени во релевантните трудови и ова истражување се состои во работа со различни податочни типови. Од друга страна, корелационите аспекти се побројни. Имено, првиот проблем со кој се соочивме во рамките на истражувањето е бројот на достапни пациенти по различни етничко-расни групи, што само по себе е огромен предизвик во тренирање на алгоритми базирани на ВИ. Истиот тренд, каде пациенти од бела раса се генерално опширно репрезентирани и други етничко-расни припадности се занемарени, може да биде забележан и во претходни истражувања без разлика на типот на податоци кои се користат (слика, текст, итн.). Следно, во рамките на претходните истражувања од огромно значење е проверка на моделот на различни податочни множества, интерно и екстерно. Истовремено, истражувањата се фокусираат на идентификација на самопријавена раса и етничка припадност на пациентите, но и на анализа на изворите на пристрасност, што е елемент и на нашето истражување.

Дополнително, претходни истражувања ја прикажуваат важноста за проверка на перформансите на моделот за различни податочни множества, интерно и екстерно. Следствено, проблемот се анализира од различни перспективи, т.е., во [70] авторите анализираат различни делови од телото и различни техники за процесирање на податоците. Понатаму, претходни истражувања се фокусираат на детекција на самопријавена етничка и расна припадност на пациенти, но и анализа на изворите на пристрасност како значајни елементи на студиите.

Табела 3-7 Анализа на корелација и дисасоцијација помеѓу релевантните трудови и оваа докторска дисертација

Реф.	Корелација	Дисасоцијација
[70]	1. Ограничување на бројот на етничко-расни групи поради достапност на пациенти 2. Проверка на моделот на повеќе податочни множества 3. Анализа на извори на пристрасност 4. Анализа на познати биолошки фактори како носители на расна пристрасност 5. Анализа на различни технологии како носители на расна пристрасност	1. Процесирање на слика
[71]	1. Ограничување на бројот на етничко-расни групи поради достапност на пациенти 2. Проверка на моделот на повеќе податочни множества 3. Анализа на извори на пристрасност 4. Анализа на однесување на медицинските лица како извор на расна пристрасност	1. Работа со текстуални податоци 2. Споредба на само две етничко-расни групи (пациенти од бела и црна раса)

4 Методологија

Оваа секција ги опишува теоретскиот и практичниот аспект на методологијата која се користи за проверка дали во рутински мерења на витални знаци може да се идентификува самопријавена етничка и расна припадност, почнувајќи од селекција на податочни множества и процесирање на податоците, преку градење на кохорти, па се до апликација на алгоритмите, валидација и интерпретабилност.

МУ е поле на ВИ, кое се фокусира на тренирање на алгоритми користејќи податоци, и сè што процесот дополнително опфаќа, со цел да се креира модел кој може да идентификува шаблони и соодноси во податоците и да ги користи тие информации за да реализира предвидувања на истата тематика така што ќе се примени на непознати податоци. Гледајќи дека алгоритмите се градат за да учат од податоците без експлицитно програмирање или потреба од присуство на експерти во областа на примена, МУ е добро прилагодено за решавање на сложени проблеми од многу разновидни дисциплини. Алгоритмите на МУ се развиваат да се справат со сè покомплексни клинички предизвици, [76], [77], [78], со цел да се добијат повеќекратни придобивки во медицинската пракса и да обезбеди автоматско донесување на одлуки со помош на алгоритам за оптимизирање на клинички работни процеси, дијагностички интервенции и подобрување на персонализирана нега.

Поради широката примена на овие видови на апликации и очекувањата за нивна примена во случаи кога животи на пациенти се изложени на ризик, потребно е истите да бидат робусни, доверливи и непристрасни. Имено, една од главните причини за овие барања е тоа што алгоритмите најверојатно ќе бидат интегрирани во клиничка пракса и може значително да влијае на животите на пациентите. Со тоа, интегритетот на целиот процес на развој на МУ модели е особено значаен, така што постои потреба да се елиминираат ризици низ целиот животен процес на алгоритмите за МУ на тој начин што ќе бидат следени група правила кои ги осигуруваат квалитетот на податоците, перформансите на моделот, правичноста, безбедноста и транспарентноста на истиот, [79], [76].

Најпрво, квалитетот на развиените модели на МУ строго зависи од податоците кои се користат во процесот, [80]; поточно, квалитетот и конзистентноста на податоците може директно да влијае и да ја наруши функционалноста на моделот во процесот на развој, дополнително придонесувајќи непоправлива штета за корисниците на моделите на МУ кои што биле тренирани на податоци со слаб квалитет, [81]. Второ, квалитетот на развиените модели на МУ може да биде засегнат од правичноста на развојниот процес, што треба да се стреми да „кодира програма која темелно ги наведува сите значајни фактори кои ни овозможуваат да препознаваме објекти од секоја можна перспектива и во сите нивни можни визуелни конфигурации“⁴, [82], па последователно, потребно е да се воспостават тестирање и верификација на системите за МУ пред истите да се користат во продукција, обезбедувајќи диверзитет во тимот кој го гради системот, овозможувајќи транспарентност на процесот на развој и многу повеќе, како техники за обезбедување интегритет и елиминирање на пристрасни одлуки од страна на апликациите на МУ.

Пристрасноста во медицинската пракса е сложена појава, [83], а разликите во здравствената заштита што произлегуваат од дискриминација се широко распространети. Постоечките нееднакости може да се зацврстат или дури и да се зголемат од страна на алгоритми развиени за клиничко донесување одлуки поради пристрасност во податоците што се користат за развој на моделите, пристрасност воведена за време на развојот на моделот или пристрасност за време на употреба на моделот и негово следење откако ќе биде ставен во продукција. Секое од овие може да резултира со донесување одлуки што може да биде дискриминирачко и штетно за социјално загрозените групи на население кои не се соодветно застапени во податоците.

⁴ “hand code a program that exhaustively enumerates all the relevant factors that allow us to recognize objects from every possible perspective or in all their potential visual configurations”

Варијација на третманите во однос на етничко-расна припадност, која не може да се објасни преку фактори карактеристични за пациент или болест, е детално опишана во неколку студии, кои се придружени со значителни докази за несвесна пристрасност во ставовите, очекувањата и однесувањето на медицинските лица, [84], [85], [86]. Присуството на овој тип на пристрасност во медицинската пракса дополнително се засилува ако дискриминаторските ставови и однесувања за возврат се моделираат како карактеристични за болеста или пак како алгоритми за поддршка на одлуки кои се имплементирани во медицинска средина или им се предаваат на студентите. Овој феномен е опишан од Брукс во субјективната студија што несвесната пристрасност ја дефинира како „тивка наставна програма“⁵, наведувајќи дека „помеѓу двајца пациенти со болка кои чекаат во собата за испитување на одделот за итни случаи, белегот е поверојатно да добие лекови, а црнецот поверојатно е да биде отпуштен со белешка која документира „однесување за барање наркотици“.“⁶, [36]. Понатаму, една неодамнешна студија ја илустрира расната пристрасност во описите во електронските здравствени досиеја на пациентите, покажувајќи дека пациентите од црна раса имаат 2,5 пати поголема веројатност да имаат еден или повеќе негативни дескриптори во нивните досиеја во споредба со пациенти од бела раса, [51]. Истражувањето Obermeyer et al. јасно ги илустрира случаите кога пристрасноста е вградена во податоците и последователно пропагирана за време на развојот на моделот, каде што „со даден резултат на ризик, пациенти од црна раса се значително поболни од пациенти од бела раса, што е потврдено со знаци на неконтролирани болести.“⁷, [26]. Алгоритмот научил да ги предвидува трошоците за нега, ставајќи ги пациентите од црна раса во иста ризична категорија како подгрупа на пациенти од бела раса, и покрај тоа што пациентите од црна раса имаат значително полоши симптоми. За да се зголеми сериозноста на проблемот, бројот на пациенти од црна раса кои требало да бидат упатени на покомплексна нега е намален за половина врз основа на препораката на алгоритмот.

Иако има свест за пристрасноста која постои во медицинска пракса, медицинска технологија, клинички резултати, податочни множества и алгоритми на МУ, помалку се знае за потенцијалната расна пристрасност во рутински собрани информации, имено виталните знаци, кои се од суштинско значење во клиничкото одлучување и кои се користат за развој на алгоритми за помош на процесот на донесување одлуки. Ако чувствителните атрибути како етничко-расна припадност лесно се учат од основните клинички податоци, тогаш е загрижувачки доколку овие атрибути станат главен дел од клиничкото одлучување и оптимизацијата на третманот на пациенти.

4.1 Избор на податочното множество

Изборот на податочното множество може значајно да влијае на целиот процес и на квалитетот на добиените резултати, [80]. Затоа, потребно е да се задоволат одредени стандарди. На *Слика 4-1* се прикажани дел од најчесто цитираните карактеристики потребни за квалитет на податоци според истражување публикувано во 1995, [87]. Притоа, овие стандарди и денес се едни од главните барања кои треба да бидат исполнети од страна на податоците со цел да истите се сметаат за квалитетни. Податоците мора да бидат веродостојни, валидни, точни, конзистентни, целосни и ажурирани; притоа, со исклучување на било кое од наведените барања квалитетот на податоците се компромитира. Како што може да се забележи од *Слика 4-1*, важно е добиените податоци да потекнуваат од сигурен извор, да носат изобилство валидни информации и точно да ја одразуваат реалноста на процесот за кој ќе се користат. Понатаму, податоците треба да бидат ажурирани за прецизно да ги прикажуваат тековните трендови во полето на истражување и мора да бидат

⁵ “silent curriculum”

⁶ “among two patients in pain waiting in an emergency department examination room, the white one is more likely to get medications and the Black one is more likely to be discharged with a note documenting “narcotic-seeking behavior”.”

⁷ “at a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses.”

конзистентни, т.е. да се под влијание само на строго дефинирани и прифатливи правила. Идентификување на одредени аспекти кои се карактеристични за податоци со лош квалитет може да се постигне преку истражувачка анализа на податоци (ИАП), [88], на пр. ИАП овозможува откривање на: информации што недостасуваат преку испитување на бројна и процентуална застапеност на полиња без вредност (анг. missing values) во податоците, отстапувања во податоците (анг. data outliers) кои делуваат на способноста на моделот да го генерализира проблемот, нерамнотежа на податоците што се случува кога целните променливи не се подеднакво претставени, итн.



Слика 4-1 Барања за квалитет на податоците

Бидејќи изборот на податоци е од клучна важност за развојот на апликациите базирани на ВИ, првиот чекор во методологијата се фокусира на избор на податочни множества кои одговараат на дисертацијата. Во **Табела 4-1** е даден преглед на достапните податочни множества кои се значајни за студијата, каде по колони се дадени името и референцата на податочното множество, земјата на потекло и временскиот опсег опфатен од податоците, вкупниот број на пациенти, етничките и расни групи репрезентирани во податоците, мерените витални знаци, типот на податочното множество врз база на какви пациенти опфаќа (прво, популациони, англ. population-based, каде податоците се собирали од болни и здраво индивидуи надвор од болничка средина; второ, од болничка средина, англ. hospital-based, податоците се собирали од сите центри во болничка средина; и трето, од ЕИН, англ. ICU-based, каде податоците чуваат информации само од ЕИН во болниците), како и годината кога множеството било споделено.

Податочното множество UK Biobank содржи детални здравствени информации вклучувајќи демографски карактеристики, фактори на начин на живот (на пр. физичка активност, консумирање алкохол, пушење, итн.), медицинска историја и употреба на лекови, кои биле собрани преку самостојно пополнети прашалници од страна на испитаниците и кратко компјутерски-потпомогнато интервју, пропратено со мерења на физички и функционални способности (меѓу кои се собирање на крв, урина и плунка). Дополнително, UK Biobank содржи обемни генетски информации (податоци за генотипирање со над 800.000 генетски варијанти). Измерените витални знаци вклучуваат крвен притисок, отчукување на срцето, респираторна стапка и телесна температура. Податоците биле

собирани од 22 центри за проценка ширум Обединетото Кралство, покривајќи различни регии за да се обезбеди социоекономска и етничка хетерогеност и урбано-рурален микс. Меѓутоа, бидејќи податоците биле собирани периодично, за време на првична проценка на пациентите и последователни посети, не постојат континуирани мерења на виталните знаци. Дополнително ограничување претставува фактот дека вклучените пациенти биле на возраст од 40-69 години. eICU-CRD е мултицентрично податочно множество, кое содржи демографски и клинички информации на пациенти во критична состојба примени во ЕИИ во над 200 болници во САД. Податоците вклучуваат континуирани мерења на виталните знаци на пациентите во текот на целиот нивен престој од приемот во ЕИИ, заедно со информации за дијагнозата и коефициентот за сериозност на болеста. Податоците опфаќаат над 200.000 приеми на пациенти, со родова и етничка хетерогеност, покривајќи родова и етничка хетерогеност. NHANES (The National Health and Nutrition Examination Survey) податочното множество содржи податоци од интервјуа спроведени во домаќинства и од стандардизирани физички прегледи спроведени во мобилни центри за испитување, со цел да се добијат проценки на национално ниво репрезентативни за неинституционализираното население во САД. Податоците содржат континуирано мерени витални знаци, вклучувајќи го и крвниот притисок (мерен со монитор за крвен притисок кој се носи 24 часа), отчукување на срцето и телесната температура. Сепак, податоците се популациони, односно не ги земаат во предвид пациентите во најкритична состојба. Податочното множество eMERGE (Electronic Medical Records and Genomics) поврзува генетски податоци со електронски здравствени досиеја собирани од клинички центри ширум САД, при што обезбедената медицинска евиденција вклучува резултати од лабораториски тестови, телесна маса и индекс на телесна маса (анг. body mass index), физиолошки тестови (на пр. крвен притисок), како и претходна здравствена историја. Податоците не се ограничени само на ЕИИ, па следствено на тоа, за учесниците во истражувањето нема континуирано мерени витални знаци. Податочното множество ICNARC (The Intensive Care National Audit & Research Centre) содржи податоци од над 2,8 милиони пациенти примени во ЕИИ за возрасни во болниците на Националната здравствена служба во Англија, Велс и Северна Ирска. Податочното множество е колекција од демографски информации и клинички податоци (прием, третман, исходи) на пациенти примени во единици за критична нега за широк спектар на медицински и хируршки состојби и истото е собрано во 2019 година. MIMIC-III (Medical Information Mart for Intensive Care) е голема база на податоци собирана во еден клинички центар која содржи деидентификувани податоци за пациенти примени во медицинскиот центар Beth Israel Deaconess Medical Center (BIDMC). Податоците содржат демографски информации за пациентите, клинички мерења и интервенции, детали за наплата, клинички лабораториски тестови, но и други медицински информации. Притоа, множеството е собирано помеѓу 2001 и 2012 година и содржи нешто над 53.000 приеми на пациенти. Податочното множество MIMIC-IV претставува продолжение на MIMIC-III и се состои од податоци собрани во BIDMC по 2012 година. Двете множества се компатибилни, при што главната значајна разлика е во тоа што MIMIC-IV содржи информации за пациентите примени или во оддели за итни случаи или во ЕИИ помеѓу 2008-2019 година. Податочното множество CERNER HealthFacts (CERNER HF) е голема, деидентификувана база на електронска здравствена евиденција, која содржи демографски и клинички информации за над 50 милиони пациенти дијагностицирани со срцева слабост во болниците низ САД, собрани за време на речиси 300 милиони средби со пациентите. Множеството вклучува демографски информации, средби со медицински лица, дијагнози, рецепти, процедури, лабораториски тестови, локација на секоја пружена услуга (на пр. ЕИИ, клиника, итн.). Но и покрај тоа, постојат случаи каде не се земани континуирани мерења, а пропратно едно од главните ограничувања е тоа што базата на податоци се состои само од пациенти со срцева слабост.

Од наведените податочни множества, кај сите постојат одредени предности и ограничувања. Главното ограничување е тоа што сите податочни множества се собирани на ниво на една земја, со што самите податоци креираат можности за селективна пристрасност. Понатаму, некои податочни множества се собрани во рок од само една година (eMERGE, ICNARC). UK Biobank и NHANES се популациони податочни множества (со информации собрани претежно од здрави пациенти),

eMERGE и CERNER HF се од болничка средина (и ги вклучуваат сите болнички единици, не само ЕИИ), додека останатите се податоци собирани во ЕИИ. И покрај тоа што кај CERNER HF постои голема популациона покриеност, податоците се собирани во болничка средина и притоа не се правени мерења за кислородна заситеност, додека податочните множества NHANES и eMERGE покриваат само мала популација на пациенти. И покрај тоа што UK Biobank и eMERGE имаат генетски информации (и со тоа потенцијално и попрецизен преглед на етничка и расна припадност), ниту едно од податочните множества не е собирано во ЕИИ.

Табела 4-1 Преглед на медицинските податочни множества

Име [Реф]	Собрани во	Примероци	Етничко-расна група	Витални знаци	Тип	Објавени
UK Biobank [89]	Обединето Кралство, 2006-2010	> 500,000	Бела раса, црна раса, Азијци, други	Крвен притисок, отчукување на срцето, температура, респираторна стапка	популационен	2012
eICU-CRD [90]	САД, 2014-2015	> 200,000	Бела раса, црна раса, Хиспанци, Азијци, други	Крвен притисок, отчукување на срцето, температура, респираторна стапка, кислородна заситеност	од ЕИИ	2019
NHANES [91]	САД, 2015-2016	> 9,000	Бела раса (без Хиспанци), црна раса (без Хиспанци), Мексиканци, Хиспанци, други/мешани раси	Крвен притисок, отчукување на срцето, температура, респираторна стапка	популационен	2019
eMERGE [92]	САД, 2015	> 24,000	Европејци, Афроамериканци, Азиски Американци, други	Крвен притисок, отчукување на срцето, температура, респираторна стапка, кислородна заситеност	болничка средина	2019
ICNARC [93]	Обединето Кралство, 2019	> 2.8 милиони	Бела раса, црна раса, Азијци, други	Крвен притисок, отчукување на срцето, температура, респираторна стапка, кислородна заситеност	од ЕИИ	2019
MIMIC-III [94]	САД, 2001-2012	> 53,000	Бела раса, црна раса, Хиспанци, Азијци, други/мешани раси	Крвен притисок, отчукување на срцето, температура, респираторна стапка, кислородна заситеност	од ЕИИ	2019
CERNER Health Facts [95]	САД, 2000-2019	> 300 милиони	Бела раса, црна раса, Хиспанци, Азијци, други/мешани раси	Крвен притисок, отчукување на срцето, температура, респираторна стапка	болничка средина	2019
MIMIC-IV [96]	САД, 2008-2019	> 100,000	Бела раса, црна раса, Хиспанци, Азијци, други/мешани раси	Крвен притисок, отчукување на срцето, температура, респираторна стапка, кислородна заситеност	од ЕИИ	2021

Иницијално, податочните множества врз кои ќе е базирано истражувањето треба да исполнат два услови: прво, податоците треба да имаат информација за самопријавената етничко-расна припадност на пациентите, и второ, треба да постојат континуирани мерења за виталните

знаци на пациентите. Гледајќи дека повеќето од разгледаните податочни множества ги исполнуваат овие критериуми, беше потребно да се направи проширување на барањата, што значеше потреба од доволен број на пациенти за солидна анализа, како и разновидност во застапените етнички и расни групи. Дополнително, приоритет се дава пациенти во критична состојба, за кои постојат континуирани мерења на витални знаци од нивниот престој во ЕИИ, во споредба со истражувања спроведувани на ниво на болници или пак популациони истражувања, најмногу поради сериозноста на здравствената состојба на пациенти во ЕИИ и непосредната опасност од смрт која е пропратна карактеристика кај овие пациенти. Меѓутоа, примената на истиот пристап во анализа на болнички базирани и популациони податочни множества, би овозможило етничко-расна анализа на поздрави пациенти и би допринело до разбирање на биолошки маркери надвор од социјални концепти доколку анализата е придружена од генетски маркери за пациентите, што би резултирало со резултати кои не се под влијание на заболувањата на пациентите, сериозноста на нивните состојби или степенот на итност кој бил доделен на нивните состојби од пристрасен медицински персонал.

За целите на оваа докторска дисертација, фокусот останува на критично болни пациенти, со што врз база на опишаните барања, од податочните множества собирани во ЕИИ, изборот беше направен за eICU-CRD и MIMIC-III, бидејќи двете множества покриваат разнобразни популации на пациенти и ги исполнуваат критериумите потребни за реализација на темата на истражувањето. Важно е да се напомене дека демографските податоци кои се главниот интерес на оваа студија, поточно етничко-расната припадност, се самопријавени од страна на пациентите. Имено, откако биле примени во ЕИИ, пациентите ја пријавиле уникатна етничка или расна припадност, главно базирана на физички и социјални идентификатори како што се гледаат во соодветното општество, па последователно, оваа студија етничко-расната пристрасност ја гледа како социјален, а не биолошки концепт. Понатака, базата на податоци е организирана така што пациентите можеле да пријават само една етничко-расна припадност. Дополнително, двете податочни множества немаат информации за историјатот на пациентите, што значи дека нема никаква биолошка основа за додефинирање на мултиетничка или мултирасна припадност.

Двете податочни множества, eICU-CRD и MIMIC-III, се дел од збирката на податоци PhysioNet, [97], и можат да бидат пристапени по завршување на онлајн курс за обука креиран од MIT⁸ и потпишување на договор за употреба на податоците, кој наложува одговорно ракување со истите и почитување на принципи на заедничко истражување. Курсот ги истакнува етичките принципи и директивите за заштита на човечки субјекти во клинички истражувања, со посебен фокус на три основни принципи при работа со човечки субјекти: почитување на личности, добродетелство и правда, [98]. Главната цел на курсот е да се воведат правилата што се следат при креирањето на податочните множества и дополнително да се дефинираат правилата што треба да се следат при користење на податоците. Дополнително, објаснета е и потребата за заштита на приватноста на пациентите, т.е. барањата што истражувачите треба да ги исполнат при креирање, собирање, користење и споделување на резултатите од множествата во оваа збирка на податоци. Овие барања го следат HIPAA (Health Insurance Portability and Accountability Act) законот, [99]. Врз основа на побарувањата на HIPAA, двете податочни множества се деидентификувани, што всушност е процес во кој се отстрануваат сите директни идентификатори од податоците на пациентите (на пр. име и презиме, телефонски и матичен број⁹, адреса, итн.). Идентификаторите кои постојат во податоците се исфрлаат за да идентитетите на пациентите останат скриени, што овозможува ублажување на ризикот од прекршување на приватноста на поединците вклучени во податочните множества. Понатаму, ова ја подржува секундарната употреба на податоците за научно истражување и други напори. Меѓутоа, податоците содржат и таканаречени индиректни идентификатори кои на крајот остануваат во податочните множества бидејќи тие информации се потребни за одредени видови студии (на пр., проучување на трендовите во податоците). Во случај

⁸ Massachusetts Institute of Technology, <https://www.mit.edu/>

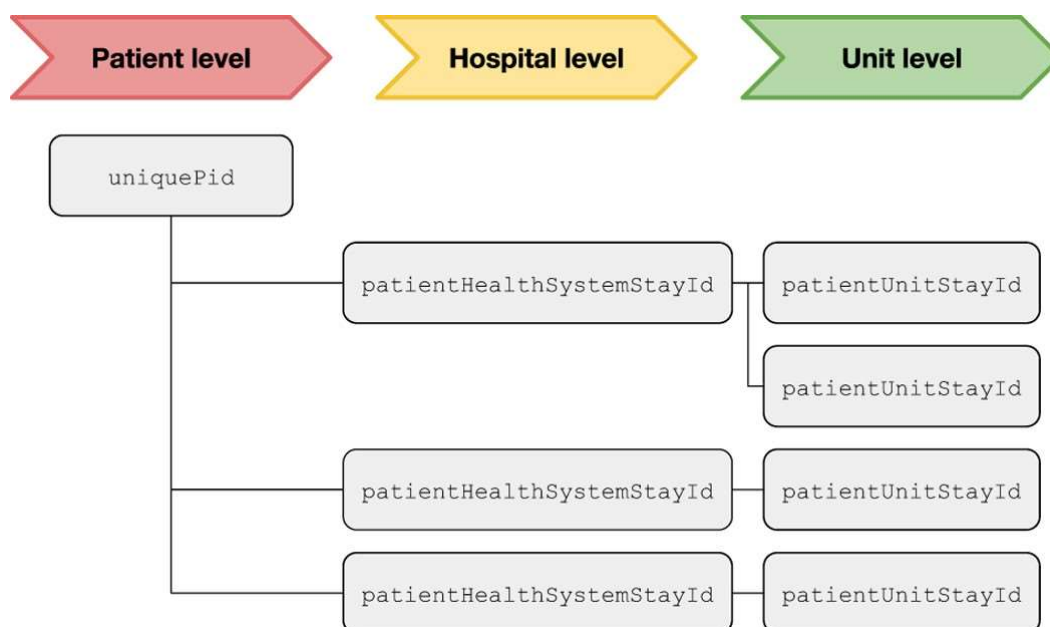
⁹ social security number, чиј пандан не постои кај нас, т.е. најблиското нешто е матичниот број

на постојни индиректни идентификатори (на пр. пол, раса, возраст, итн.), од особена важност е да се оцени ризикот за повторна идентификација на пациентите. Постојат бројни пријавени случаи за обиди за повторна идентификација направени врз основа на објавени медицински податоци, [100]. Од таа причина, за пристап до збирката од податочни множества PhysioNet, корисникот мора да потпише изјава дека нема да ги споделува податоците, ќе ги користи само за да ги реализира целите на истражувањето и нема да се направи обид за каква било повторна идентификација на пациентите.

4.1.1 eICU-CRD

eICU-CRD (eICU Collaborative Research Database) множеството содржи преку 200.000 приеми во ЕИИ за 139.000 различни пациенти, собирани во 335 ЕИИ во 208 болници низ САД. Базата на податоци содржи демографски информации за пациентите, вклучува инвазивни и неинвазивни мерења на витални знаци, детали за дијагнози и третмани, документација за грижата за пациентот, коефициент за евалуација на сериозноста на болеста, и многу други информации. Детална документација за податочното множество може да се најде на [101], а до помошните концепти (дадени од креаторите на базата на податоци) може да се пристапи на GitHub, [102].

Базата на податоци е организирана во свездена шема, е високо денормализирана, па како таква е централизирана околу пациентите (на тој начин што секој пациент е придружен со уникатен идентификатор) и нивните престои во ЕИИ (дополнителен единствен идентификатор по престој), кои делуваат како референтни клучеви за сите останати табели во базата на податоци (каде се чуваат дополнителни информации за пациентот и соодветниот престој). Како што може да се забележи на **Слика 4-2**, постојат три нивоа кои дефинираат единственост на инстанците (или пациентите): ниво на пациенти каде уникатниот идентификатор на пациентот (uniquePid) ги дефинира сите пациенти примени во ЕИИ (и ги поврзува сите потребни информации за пациентот низ сите табели во базата на податоци), ниво на болница (каде што секоја уникатна хоспитализација е означена со patientHealthSystemStayId) и на ниво на ЕИИ (patientUnitStayId) дефинирана на тој начин што за време на еден единствен престој во болницата може да постојат неколку различни вредности за повеќе засебни престои во различни оддели на болницата.



Слика 4-2 Организација на следење на пациенти во eICU-CRD. Слика позајмена од [90].

4.1.2 MIMIC-III

MIMIC-III (Medical Information Mart for Intensive Care) е база на податоци собирана во еден клинички центар, која содржи информации за прием на над 50.000 возрасни пациенти во ЕИН во болница за специјализирана нега (анг. tertiary care) со седиште во Бостон помеѓу 2001 и 2012 година. Множеството содржи детали за демографски информации, инвазивни и неинвазивни мерења на витални знаци, лекови, лабораториски мерења, набљудувања и белешки од медицински лица, кодови за дијагноза и неколку коефициенти за евалуација на сериозноста на болеста и многу повеќе. Детална документација во врска со базата на податоци може да се најде на [103], а до помошните концепти може да се пристапи на GitHub, [104].

Базата на податоци е организирана така што секој пациент има уникатен идентификатор (SUBJECT_ID), а секоја уникатна хоспитализација и единствен прием во ЕИН имаат соодветни идентификатори (HADM_ID за прием во болница и ICUSTAY_ID за прием во ЕИН). Овие три идентификатори ги поврзуваат дополнителните информации за пациентите за време на нивниот престој во соодветните ЕИН. На *Слика 4-3* е прикажана организациската структура на податоците, информациите за пациентите кои се зачувани во базата на податоци MIMIC-III, но и подобрувањата направени во самата база врз основа на повратни забелешки од корисниците на истата. Имено, MIMIC базата на податоци е една од ретките подложна на верзонирање, а секоја нова верзија е компатибилна со претходните верзии (без да се земат предвид нови дополнувања на податоците); па така, постојат:

- MIMIC-II содржела податоци собрани помеѓу 2001-2008 година, измерени со CareVue монитори. Базата на податоци повеќе не е јавно достапна. Сепак, податоците може да се добијат со селектирање на податоците од CareVue мониторите во множеството MIMIC-III;
- MIMIC-III содржи податоци за приеми од 2001-2012 година, кои се собирани и од Metavision и од CareVue мониторите; и,
- MIMIC-IV содржи податоци за приеми од 2008-2019 година, собирани само од Metavision мониторите.

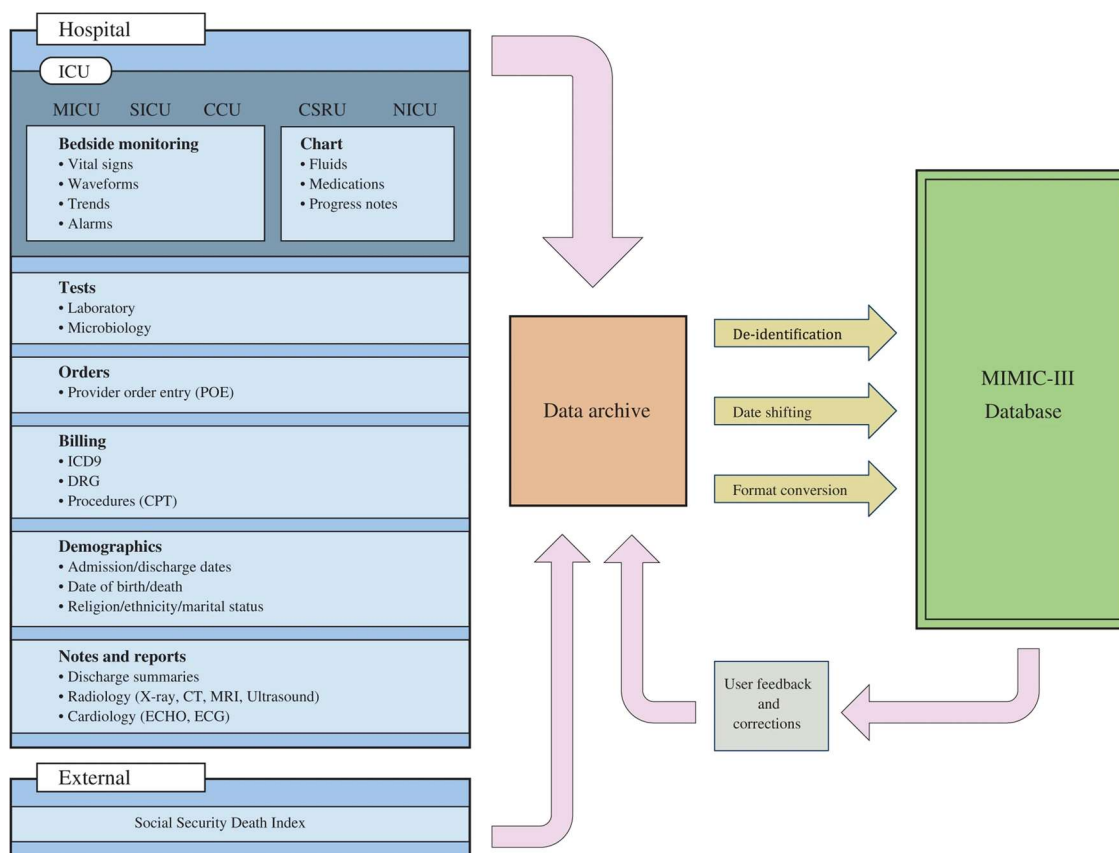
4.1.3 Поставување и инсталација на помошни концепти (eICU-CRD и MIMIC-III)

Двете податочни множества се организирани како релациони бази на податоци, така што се воспоставени неколку механизми за пристап до податоците: со пристап во облак (на BigQuery, AWS, или Google Storage) или инсталација на множеството локално, што се сведува на едноставно преземање на податоците. Меѓутоа, за да може да се работи со помошните концепти кои се понудени од креаторите на податочните множества, потребно е истите да бидат инсталирани локално. Локалното поставување на помошните концепти може да се реализира со инсталација преку системот за управување со релациони бази на податоци PostgreSQL, [105]. Пред да се отпочне со инсталацијата на множествата на PostgreSQL системот, потребно е да:

- Преземање на eICU-CRD и MIMIC-III податочните множества (откако пристапот до нив е обезбеден);
- Екстрахирање на базите од податоци како .csv фајлови на локална машина (и истите се потребни за да се постават како патека до податоците откако ќе почне градење на помошните концепти преку PostgreSQL), и
- Преземање на PostgreSQL системот, како и достапните скрипти во папките на соодветните складишта (eICU-CRD и MIMIC-III).

Откако сите побарувања ќе бидат задоволени, може да се отпочне сесија преку терминалот на PostgreSQL системот, по што потребно е да се извршат низа наредби за да се воспостават базите на податоци и сите концепти овозможени за нив. Инструкциите кои го опишуваат процесот за MIMIC-III податочното множество се прикажани на *Слика 4-4* (сите наредби резултираат со излез на терминалот, меѓутоа поради обемноста на податочното множество и самите излезни информации

се обемни, па од таа причина не се вклучени во приказот). Не постојат клучни разлики помеѓу поставките потребни за двете податочни множества (eICU-CRD и MIMIC-III), т.е. најголемите разлики се состојат во промената на патеките во зависност од локацијата на која ќе биде инсталирано податочното множество и структурата на самата база на податоци.



Слика 4-3 Преглед на структурата на MIMIC-III податочното множество. Сликата позајмена од [94].

```
Server [localhost]:
Database [postgres]:
Port [5432]:
Username [postgres]:
Password for user postgres:
psql (14.6)
postgres=# DROP DATABASE IF EXISTS mimic;
postgres=# CREATE DATABASE mimic OWNER postgres;
postgres=# \c mimic;
mimic=# CREATE SCHEMA mimiciii;
mimic=# set search_path to mimiciii;
mimic=# \i D:/Bojana/Personal/Studies/PhD/PhD/TheThing/data/mimic/mimic-code/mimic-iii/buildmimic/postgres/postgres_create_tables.sql
mimic=# \set ON_ERROR_STOP 1
mimic=# \set mimic_data_dir
'D:/Bojana/Personal/Studies/PhD/PhD/TheThing/data/mimic/mimic/mimic'
mimic=# \i D:/Bojana/Personal/Studies/PhD/PhD/TheThing/data/mimic/mimic-code/mimic-iii/buildmimic/postgres/postgres_load_data.sql
mimic=# select icustay_id, intime, outtime from icustays limit 10;
mimic=# \i D:/Bojana/Personal/Studies/PhD/PhD/TheThing/data/mimic/mimic-code/mimic-iii/concepts_postgres/postgres-make-concepts.sql
```

Слика 4-4 Инсталирање на MIMIC-III податочното множество и помошните концепти

4.1.4 Детали за податочните множества (eICU-CRD и MIMIC-III)

Основните информации за двете податочни множества се дадени во **Табела 4-2**, каде се истакнати значајните сличности и разлики. Двете податочни множества се собирали во САД, но додека eICU-CRD множеството било собирано во 335 ЕИН во 208 болници во опсег од две години и содржи преку 200.000 приеми, MIMIC-III податочното множество е собирано од една болница во нешто повеќе од 10 години и содржи над 50.000 приеми. Со тоа, главната разлика помеѓу двете множества се крие во работа со податоци собирали во еден центар и податоци собирали во повеќе центри. Следствено, eICU-CRD множеството опфаќа повеќе локации ширум САД, додека MIMIC-III покрива поголем временски опсег.

Табела 4-2 Споредба на избраните податочни множества (eICU-CRD и MIMIC-III)

Детали	eICU-CRD	MIMIC-III
Извор	Приеми во ЕИН во повеќе центри	Приеми во ЕИН во еден центар
Земја и период на земање на податоците	САД 2014 – 2015	САД 2001 – 2012
ЕИН и болници	335 ЕИН во 208 болници	5 ЕИН во 1 болница
Приеми во ЕИН	200.859	53.423
Опсег на години	18 и постари	15 и постари
Етничко-расна припадност на пациентите	пациенти од бела раса (77,31%) пациенти од црна раса (10,61%) Хиспанци (3,72%) Азијци (1,63%) домородни Американци (0,85%)	пациенти од бела раса (66,01%) пациенти од црна раса (17,44%) Хиспанци (2,7%) Азијци (1,78%) домородни Американци (< 0,1%)
Мерени витални знаци	Крвен притисок (систолен, дијастолен, среден), отчукување на срце, респираторна стапка, кислородна заситеност, температура	Крвен притисок (систолен, дијастолен, среден), отчукување на срце, респираторна стапка, кислородна заситеност, температура
Коефициент за сериозност на болест	APACHE IV	OASIS
Ограничувања	Ограничена на пациенти од САД, потенцијална селективна пристрасност	Ограничена на еден центар, неможност да се генерализира за други средини

Двете податочни множества ги содржат истите витални знаци (кои се од интерес на истражувањето): инвазивни и неинвазивни мерења на крвен притисок (систолен, дијастолен, среден крвен притисок), отчукување на срцето, респираторна стапка, инвазивни и неинвазивни мерења на кислородна сатурација, и температура. Неинвазивните мерења на крвниот притисок се земаат преку притисокомер со манжетна за рака (анг. arm cuff) и за да се направи мерење потребно е присуство на медицински персонал, па затоа по пациент има мал број на мерења. Инвазивните мерења на крвниот притисок се земаат преку артериска линија (анг. arterial blood line), која се вметнува во артерија на пациентот како средство за следење; сепак, артериска линија се инсталира само кај дел од пациенти (на пр. високоризични хируршки, критично болни, итн.) и тоа врз основа на одлуките на лекарите. Неинвазивната кислородна заситеност се мери со пулсен оксиметар, додека инвазивната кислородна заситеност се мери преку лабораториска анализа (артериски крвен тест).

Дополнително, кај двете податочни множества има коефициенти за проценка на морталитет (коефициент за оценување на сериозноста на болеста), кој покажува на сериозноста на состојбата на пациентот, така што eICU-CRD го користи коефициентот APACHE (Acute Physiology and Chronic Health Evaluation) IV, [106], додека MIMIC-III го користи коефициентот OASIS (Oxford Acute

Severity of Illness Score), [107]. Иако и двата коефициенти се валидни при проценка на сериозност на болест и резултираат слични вредности во предвидувањето на смртноста за време на болнички престој и други негативни исходи кај критично болни пациенти, сепак постојат разлики во процесот на пресметка на коефициентите. Имено, OASIS има за цел да го поедностави пресметувањето на ризикот користејќи само неколку физиолошки променливи, додека APACHE IV е покомплексен систем за бодување и зема предвид широк опсег на клинички податоци. За подетален преглед, кај OASIS во пресметката се користат 10 рутински собирани информации за пациентите при престој во ЕИН, како што може да се види на **Слика 4-5**, а бидејќи самиот принцип на пресметки не бара податоци од лабораториски тестови, пресметката е експедитивна. Спротивно на тоа, APACHE IV има за цел сеопфатна проценка со земање предвид на широк опсег на клинички податоци направени во првиот ден од престој во ЕИН, користејќи 142 променливи, т.е. резултатот бара детали за променливи што ја опфаќаат демографијата на пациентот, претходна медицинска историја, физиолошките мерења (витални знаци, лабораториски вредности, други параметри). Податоците што се користат за пресметка на APACHE IV се прикажани во **Табела 4-3**.

Табела 4-3 Пресметка на APACHE-IV [106]. Податоци кои се користат за креирање на моделот за предвидување и деталите за истите, сите земени во прв ден од престој. APS = acute physiology score (резултат за физиолошка состојба), APACHE = Acute Physiology and Chronic Health Evaluation (коефициент за сериозност на состојба на пациент), GCS = Glasgow Comma Scale (Глазгов Коматозна Скала), PaO₂ = arterial oxygen partial pressure (артериски парцијален притисок на кислород), FiO₂ = fractional inspired oxygen (фракциона концентрација на инспириран кислород), P(A-a)O₂ = alveolar-arterial gradient (алвеоларно-артериски градиент, т.е. A-a градиент).

Променливи	Детали
Старост	Континуирано мерена, и дополнителни 5 точки за сплајн функција
APS променливи	Најабнормални вредности за тежина во прв ден на APACHE; сума од тежини еднаква на APS, кој се движи од 0 до 252. Додадени пет термини за сплајн. Променливите вклучуваат пулс, среден крвен притисок, температура, респираторна стапка, однос PaO ₂ /FiO ₂ (или P(A-a)O ₂ за интубирани пациенти со FiO ₂ ≥ 0,5), хематокрит, број на бели крвни зрнца, креатинин, урина, азот од уреа во крвта, натриум, албумин, билирубин, гликоза, абнормалности на киселинска база и невролошки абнормалности врз основа на GCS
Хронични здравствени променливи	СИДА, цироза, хепатална инсуфициенција, имunosупресија, лимфом, леукемија или миелом, метастатски тумор. Не се користи за пациенти под елективна хирургија
Приемна дијагноза во ЕИН	116 категории, повеќе детали во апендикс на [66]
Извор на прием во ЕИН	Кат, сала за итни случаи, соба за оздравување/операциона сала, единица за повлекување на пациенти, директен прием, друга ЕИН, друга болница, друг извор на прием
Должина на престој пред прием во ЕИН	Квадратен корен плус 4 точки од сплајн функција
Итна операција	Да/Не
Пристап до GCS	Да/Не
Тромболитичка терапија	За пациенти со акутен миокарден инфаркт (Да/Не)
Скалиран GCS	15 минус измерената вредност за коефициентот GCS
PaO ₂ /FiO ₂ сооднос	
Механичка вентилација	Да/Не

5 <0.17		3 0.17-4.94		Pre-ICU LOS 0 4.95-24.00 Hours		2 24.01-311.80		1 >311.80					
				Age 0 <24 Years		3 24-53		6 54-77		9 78-89		7 >90	
10 3 - 7		4 8 - 13		3 14		GCS 0 15							
				4 <33		Heart Rate 0 33-88 min ⁻¹		1 89-106		3 107-125		6 >125	
4 <20.65		3 20.65-50.99		2 51-61.32		MAP 0 61.33-143.44 mmHg		3 >143.44					
			10 <6		1 6-12		Respiratory Rate 0 13-22 min ⁻¹		1 23-30		6 31-44		9 >44
3 <33.22		4 33.22-35.93		2 35.94-36.39		Temperature 0 36.40-36.88 °C		2 36.89-39.88		6 >39.88			
10 <671		5 671-1426.99		1 1427-2543.99		Urine Output 0 2544-6896 Cc/day		8 >6896					
						Ventilated 0 NO		9 YES					
				6 NO		Elective Surgery 0 YES							

Слика 4-5 Пресметка на OASIS [107]. Секоја мерка на променлива добива соодветна вредност (дадена во задебелено), така што се зема најлоша вредност во прв ден, а финалната вредност на OASIS е сума од сите индивидуални вредности. LOS = length of stay (време на престој), GCS = Glasgow Comma Scale (Глазгов Коматозна Скала), MAP = mean arterial pressure (среден артериски притисок).

Бидејќи податоците мора да се придржуваат до потребни барања (главно веродостојност, валидност, точност, конзистентност, целосност и ажурност) потребно беше да се анализираат избраните податоци, пред да се продолжи со дефинирање на остатокот од пристапот. Затоа, беа анализирани: популацијата на пациенти за двете податочни множества со цел да се разбере дистрибуцијата на лабелите, т.е. дистрибуцијата на етничко-расната пристрасност пред тренирање на моделите. Дополнително, на резултатите може да влијаат и дополнителни параметри, како што се дијагнозата при прием, полот, годините и коефициентот за сериозност на болеста, па од таа причина беа разгледани промените што се случуваат кај овие параметри за различни етничко-расни групи.

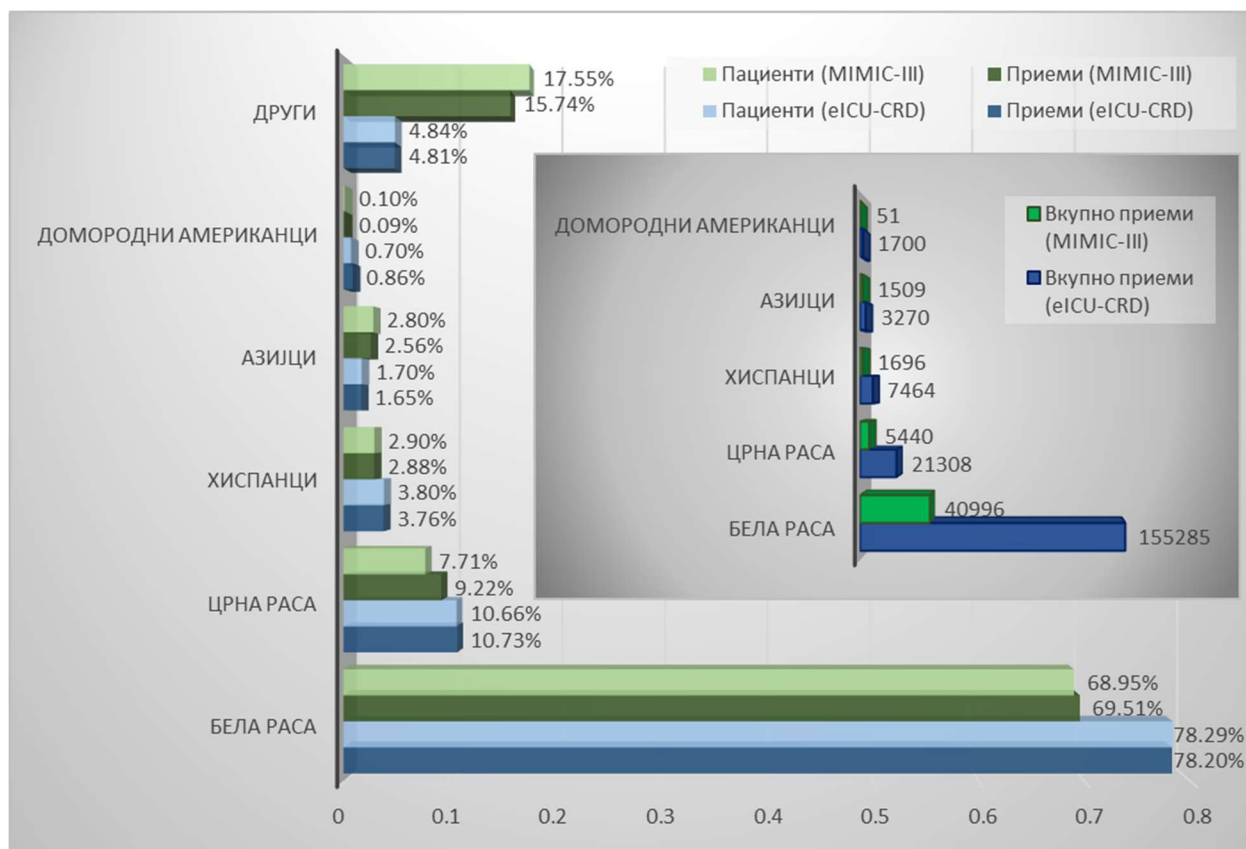
Во врска со популацијата, двете податочни множества опфаќаат неколку етнички и расни групи: кај eICU-CRD пациентите се сместени во една од пет етничко-расни групи (означени како: Caucasian, African American, Hispanic, Asian, и Native American) или пак не е позната етничко-расната група, додека кај MIMIC-III пациентите се сместени во повеќе етничко-расни групи (при што главната организација на подгрупите може да се структурира во: White, Black/African American, Hispanic or Latino, и Asian). Меѓутоа, во MIMIC-III множеството кај дел од пациентите постои дополнителна спецификација во врска со нивната раса, односно во множеството има 40.996 пациенти од бела раса, меѓутоа, постојат и 329 пациенти за кои освен информацијата дека се од бела раса, постои и додатна спецификација за етничко-расната припадност (на пр. White – Russian, White – Other European, White – Brazilian, White – Eastern European), што е случај со сите останати етничко-расни групи во множеството. Меѓутоа, и покрај ова, најдоминантните етничко-расни групи во двете податочни множества се исти, па од таа причина еквиваленција на ниво на етничко-расна припадност меѓу двете множества е возможна.

Во податочното множество eICU-CRD се користи терминологијата “Caucasian” како опис за етничка припадност, што претставува некоректна употреба гледајќи дека голем дел од припадници на бела раса во САД, каде податочното множество е собирано, всушност не потекнуваат од Кавкаскиот регион помеѓу Европа и Азија, што во целост ја негира единствената примена за која терминот оригинално бил користен. Понатаму, терминот е застарен начин за дефинирање на ограничувачки, субјективни антрополошки категории, преостанати од расистички верувања за супериорност на белата раса во споредба со други (тогаш нарекувани деген раси) што датираат од 18 век, [108]. Бидејќи терминот е застарен и потекнува од расистички гледишта за расна припадност, во рамките на студијата терминот ќе биде заменет и на негово место за означување на етничката група ќе се користи прифатената нотација White.

Фокусот на студијата ќе биде наменет на најзастапените категории на етничко-расни припадности, па од таа причина за нив ќе биде даден соодветен превод. Имено, во понатамошна употреба: White ќе биде преведувано како бела раса, Black/African American како црна раса, Hispanic or Latino како Хиспанци, Asian како Азијци, и Native American како домородни Американци. Меѓутоа, иницијалните етничко-расни термини, како што се оригинално дефинирани во податочните множества, беа презентирани и тука за илустрација на содржината на самите податочни множества.

Дистрибуцијата на најзастапените етничко-расни припадности за двете множества е дадена во **Слика 4-6**; двете податочни множества се составени предоминантно од пациенти од бела раса со 68,95% застапеност кај MIMIC-III и 78,29% кај eICU-CRD, проследени со пациенти од црна раса, Хиспанци и Азијци. Преку ознаката „Други“ прикажани се пациенти од други етнички групи (каде бројот на приеми е значително помал во споредба со најзастапените популации на пациенти, на пр. во случајот со домородни Американци кај eICU-CRD податочното множество, застапеноста на пациентите од оваа група во податоците е помала од 1%), или пациенти кои не ја специфицирале нивната етничко-расна припадност или воопшто не биле прашани за истата. Исто така, на сликата за секое множество се прикажани процентуалната застапеност на етничко-расни групи и по пациенти и по приеми; процентите кои илустрираат информација за пациенти во предвид ги земаат уникатните идентификатори за секој пациент и пресметката за етничко-расна припадност ја прават врз основа на таа вредност, додека пресметките за приеми се реализирани преку уникатен идентификатор за прием на одреден пациент во ЕИН и секој прием се зема како засебен податок за одредена етничко-расна група и во случај на повеќе приеми на истиот пациент. Како што може да се забележи од сликата, не постојат значителни вредносни разлики помеѓу уникатни пациенти и уникатни престои во ЕИН (што покажува дека повторните примања на пациентите во ЕИН, иако присутни, не се истакнуваат особено).

Уште една важна карактеристика на **Слика 4-6** е тоа што илустрира еден значаен недостаток за двете податочни множества, во тоа што забележаните и собраните информации доаѓаат најмногу од пациенти од бела раса, со што постои огромна незастапеност за други етничко-расни групи. Ова е значајно илустративен пример, бидејќи покажува огромен проблем кој реално постои кај медицинските податочни множества, а тоа е дека множествата сами по себе не ги претставуваат сите етничко-расни групи подеднакво, што дополнително прераснува во предизвик за алгоритмите на ВИ кои треба да научат и да ги објаснат информациите согледани од овие податоци. Оваа дистинкција е уште поизразена кога ќе се обрне внимание на вгnezдениот график во сликата (кој содржи точен број на приеми во ЕИН за двете податочни множества), каде може да се забележи дека бројот на приеми на пациенти во ЕИН е значително помал за Хиспанци, Азијци и домородни Американци, особено во случајот на MIMIC-III податочното множество.



Слика 4-6 Репрезентација на етничко-расни групи во eICU-CRD и MIMIC-III множествата

Друг значаен аспект беа и приемните детали и дијагнози бидејќи при прием во ЕИН пациентите можат да бидат во различна здравствена состојба, базирана на критериуми при прием на пациентите, но и да имаат различни дијагнози и коефициенти за сериозност на болеста, [109]. Од сите документираны заболувања, најзастапените 10 за eICU-CRD се прикажани во **Табела 4-4**, додека најзастапените 11 за MIMIC-III множеството се прикажани во **Табела 4-5**. Главната причина за избор на 11 најзастапени дијагнози при прием при анализа на MIMIC-III множеството (во споредба со изборот на само 10 кај eICU-CRD) лежи во тоа што BIDMC здравствениот центар врши прием и на новороденчиња, па со самото тоа ова е најзастапената категорија на дијагноза при прием кај MIMIC-III. Бидејќи самата студија ќе се фокусира само на возрасни пациенти, новороденчињата не се релевантни за потребите на истражувањето и од таа причина во табелата се обележани со црвена боја, и покрај тоа што не можеше да не бидат споменати гледајќи дека се најзастапена дијагноза при прием за MIMIC-III податочното множество.

Двете табели се организирани на тој начин што го прикажуваат бројот на пациенти примени во ЕИН со специфична дијагноза и процентуалната застапеност на примените пациенти со истата дијагноза во споредба со сите приеми во податочното множество. Дополнително, за секоја етничко-расна група пресметани се две информации (означени со R и D), каде R ја означува процентуалната застапеност на пациенти кои ја имаат соодветната дијагноза при прием во споредба со други пациенти во истата етничко-расна група, додека D ја прикажува процентуалната застапеност на пациенти од одредена етничко-расна група кои ја имаат соодветната дијагноза при прием во споредба со сите пациенти кои биле примени со истата дијагноза. Резултатите во табелите беа добиени преку спојување во лево (анг. left join) на дијагнози на прием меѓу сите пациенти и пациентите припадници на етничко-расните групи од интерес. Во **Табела 4-5** може да се забележат полиња со празно место, кои покажуваат дека кај соодветната етничко-расна група не постојат

приеми на пациенти со соодветна дијагноза. Полињата обоени со зелена боја ги покажуваат најзастапените дијагнози детектирани кај пациенти од одредена етничко-расна група за секое од податочните множества, додека полињата во сина боја покажуваат која етничко-расна група има популација од пациенти со одредена дијагноза при прием над 3% (за eICU-CRD) и 1,5% (за MIMIC-III), при што, двете се структурирани во релација со 10 најзастапени дијагнози при прием.

Табела 4-4 Приемна дијагноза во оригинален формат за пациенти во eICU-CRD, за сите пациенти и според етничко-расна припадност

Дијагноза при прием (оригинална форма)	Сите пациенти		Од бела раса		Од црна раса		Хиспанци		Азијци		Домородни Американци	
	Број	P (%)	R (%)	D (%)	R (%)	D (%)	R (%)	D (%)	R (%)	D (%)	R (%)	D (%)
Sepsis, pulmonary	8862	4,98	5,19	79,95	3,85	8,62	4,16	3,2	5,08	1,73	5,05	0,74
Infarction, acute myocardial (MI)	7228	4,06	4,17	78,67	3,11	8,52	3,21	3,03	3,69	1,54	1,38	0,25
CVA, cerebrovascular accident/stroke	6647	3,74	3,73	76,55	3,66	10,92	4,27	4,39	4,38	1,99	2,52	0,5
CHF, congestive heart failure	6617	3,72	3,49	71,95	5,12	15,35	4,64	4,79	4,22	1,92	1,99	0,39
Sepsis, renal/UTI (including bladder)	5273	2,96	3	77,55	2,57	9,67	3,37	4,36	2,99	1,71	3,75	0,93
Rhythm disturbance (atrial, supraventricular)	4827	2,71	2,94	83,01	1,8	7,38	2,91	4,12	1,36	0,85	1,83	0,5
Diabetic ketoacidosis	4825	2,71	2,42	68,5	4,37	17,95	3,43	4,85	1,3	0,81	5,35	1,45
Cardiac arrest (with or without respiratory arrest)	4580	2,58	2,45	72,9	3,29	14,26	2,5	3,73	2,39	1,57	2,83	0,81
CABG alone, coronary artery bypass grafting	4543	2,55	2,68	80,34	1,43	6,25	2,25	3,39	3,62	2,4	1,15	0,33
Emphysema/bronchitis	4494	2,53	2,74	83,13	2,19	9,68	1,51	2,29	1,03	0,69	1,3	0,38

Од двете табели може да се забележи дека помеѓу приемните дијагнози кај двете податочни множества има препокривање – двете податочни множества имаат висока застапеност на пациенти примени со сепса, конгестивна срцева слабост, коронарна артериска болест и коронарен артериски байпас графт. MIMIC-III има голема застапеност на новороденчиња, кои ќе бидат исклучени од истражувањето. Сепса е една од најзастапените дијагнози за сите пациенти и на ниво на етничко расна група. Во случај на eICU-CRD од сите пациенти примени со сепса 79,95% се од бела раса, 8,62% се од црна раса, 3,2% се Хиспанци, 1,73% се Азијци, а само 0,74% се домородни Американци, додека останатите 5,76% се пациенти за кои нема документирана етничко-расна припадност. Слична дистрибуција може да се забележи и кај MIMIC-III каде 72,97% од пациентите со сепса се од бела раса, 11,57% се од црна раса, 2,36% се Хиспанци, а 1,6% се Азијци. Дополнително, интересно е да се забележи дека и кај двете податочни множества, повеќето од примените пациенти од црна раса страдаат од конгестивна срцева слабост, што е во корелација со постоечки медицински истражувања кои покажуваат дека пациенти од црна раса имаат највисоки инцидент и превалент за срцева слабост, како и најлоши клинички исходи во споредба со останати етничко-расни групи, [110], [111], [112], а истиот тренд може да се забележи и кај тинејџери, [113].

Табела 4-5 Приемна дијагноза во оригинален формат за пациенти во MIMIC-III, за сите пациенти и според етничко-расна припадност

Дијагноза при прием (оригинална форма)	Сите пациенти		Од бела раса		Од црна раса		Хиспанци		Азијци		Домородни Американци	
	Број	P (%)	R (%)	D (%)	R (%)	D (%)	R (%)	D (%)	R (%)	D (%)	R (%)	D (%)
New-born	7823	13,27	11,6	60,78	15,9	11,06	21,76	4,72	47,45	9,14	52,94	0,35
Pneumonia	1566	2,66	2,85	74,46	3,02	10,47	2,48	2,68	2,52	2,43		
Sepsis	1184	2,01	2,11	72,97	2,52	11,57	1,65	2,36	1,26	1,6		
CHF, congestive heart failure	928	1,57	1,5	66,16	3,24	18,97	1,12	2,05	0,46	0,75	5,88	0,32
Coronary artery disease	840	1,42	1,29	62,86	0,24	1,55	0,88	1,79	0,4	0,71	1,96	0,12
Chest pain	778	1,32	1,4	73,91	1,34	9,38	1,12	2,44	0,73	1,41		
Intracranial haemorrhage	713	1,21	1,3	74,61	0,64	4,91	0,88	2,1	1	2,1		
Altered mental status	712	1,21	1,23	70,51	1,97	15,03	1,24	2,95	0,53	1,12		
Gastrointestinal bleed	686	1,16	1,21	72,3	1,2	9,48	0,83	2,04	0,8	1,75		
Coronary artery disease/Coronary artery bypass graft/SDA	583	0,99	1,1	77,53	0,28	2,57	0,94	2,74	1	2,57		
Upper GI bleed	580	0,98	0,97	68,79	1,07	10	1,36	3,97	0,86	2,24		

Освен деталите за приемна дијагноза, интересно е да се направи анализа и во разликите помеѓу пол и возраста на пациентите помеѓу етничко-расни група, со цел да се види дали постојат видливи разлики во пол и возраст при работа со уникатни пациенти или уникатни приеми во ЕИН за двете податочни множества. Резултатите добиени за разликите според пол се прикажани во **Табела 4-6**, додека старосните разлики се дадени во **Табела 4-7**. Притоа, двете табели се организирани на тој начин што ги прикажуваат разликите според пол и возраст за уникатни пациенти (означени со P) и за уникатен престој во ЕИН (означени со S) за сите етничко-расни групи во множествата заедно и за секоја група индивидуално. Од **Табела 4-6** може да се забележи дека за двете податочни множества, кај мажите постои поголем прием во споредба со женската популација, и овој тренд постои кај секоја етничко-расна припадност, освен во случај за пациентите од црна раса кај MIMIC-III податочното множество каде може да се забележи спротивното.

Старосните разлики, прикажани во **Табела 4-7**, беа анализирани во рамки на четири старосни групи, каде првата група се состои од пациенти под 30 години, втората група ги опфаќа пациентите во опсег од 30 до 55 години, третата група се пациенти во опсег од 55 до 70 години и четвртата група ги вклучува пациентите над 70 години. Како што може да се забележи од податоците во eICU-CRD, бројот на приеми рамномерно расте со порастот на староста на популацијата од интерес; имено, кај постари пациенти има почест прием во ЕИН. Набљудувајќи го трендот кај сите пациенти генерално, постојат 7.925 пациенти под 30, па 35.020 во втората група, и 46.319 и 49.329 пациенти во третата и четвртата старосна група соодветно. Сличен пораст на броеви може да се забележи и кај пациенти од бела раса, Хиспанци и Азијци, исто така. Меѓутоа, овој тренд не може да се забележи кај пациенти од црна раса и домородни Американци, каде прием на средновековни пациенти е во пораст, со приближно преполовување на бројот на пациенти помеѓу третата и четвртата старосна група. Ова е дополнително илустрирано со вредностите добиени од MIMIC-III

податочното множество. Истиот пораст на број на приеми со зголемување на старосна група може јасно да се забележи за сите пациенти и кај пациентите од бела раса, поспецифично; но, ова не важи за другите групи, па така спротивните трендови постојат кај Хиспанци и пациенти од црна раса.

Табела 4-6 Полови разлики на ниво на пациенти и приеми во соодветните етничко-расни групи (P - уникатни пациенти, S - уникатни престои во ЕИН)

Пол	Сите пациенти		Од бела раса		Од црна раса		Хиспанци		Азијци		Домородни Американци	
	P	S	P	S	P	S	P	S	P	S	P	S
eICU-CRD												
Машки	75177	108379	58348	84207	7534	10896	2796	4022	1224	1711	490	854
Женски	64039	92303	49362	71065	7133	10407	2435	3440	1114	1559	474	846
MIMIC-III												
Машки	26121	32950	18121	23187	1671	2427	813	1027	732	855	25	27
Женски	20399	26026	13953	17809	1915	3013	536	669	572	654	20	24

Табела 4-7 Старосни разлики на ниво на пациенти и приеми во соодветните етничко-расни групи (P - уникатни пациенти, S - уникатни престои во ЕИН)

Старост	Сите пациенти		Од бела раса		Од црна раса		Хиспанци		Азијци		Домородни Американци	
	P	S	P	S	P	S	P	S	P	S	P	S
eICU-CRD												
a ≤ 30	7925	10901	5223	7228	1218	1693	447	604	135	162	120	188
30 < a ≤ 55	35020	50323	24771	35448	5195	7542	1546	2251	620	821	414	759
55 < a ≤ 70	46319	67950	36248	53191	5028	7465	1422	2046	719	1039	266	488
70 < a ≤ 90	49329	70611	40960	58700	3119	4462	1770	2497	843	1211	146	233
MIMIC-III												
a < 30	9954	10546	6014	6331	1053	1189	490	509	755	801	29	31
30 < a ≤ 55	9801	12979	6459	8534	937	1548	450	609	165	210	6	8
55 < a ≤ 70	11485	15453	8187	11025	824	1463	233	330	150	195	5	5
70 < a ≤ 90	13132	17181	9699	12831	666	1081	150	218	199	260	5	7

Најголем дел од процесот на анализа на информации на пациенти во истражувањето се фокусира на витални знаци и на коефициенти за сериозност на болест како мерка за разбирање на одреден проблем. Со тоа, важно е да се забележи дали постојат некои значајни контрадикции во вредностите за виталните знаци помеѓу различни етничко-расни групи, а гледајќи дека тие витални знаци влијаат на вредностите на коефициентот за сериозност на болест, анализата се фокусира на статистика на промените кај APACHE IV коефициентот за eICU-CRD и OASIS коефициентот за MIMIC-III. Како што секој засебен прием може да биде поврзан со различни вредности за APACHE IV и OASIS коефициентите (и во случај на ист пациент), статистиката прикажана во секцијата е

направена на ниво на уникатен престој на пациент во ЕИН, а резултатите се дадени во **Табела 4-8**. Вредностите за APACHE IV се поделени во опсези од 50 вредности, додека кај OASIS коефициентот вредностите се поделени во опсези од 25 вредности. Понатаму, за секоја етничко-расна група има две пресметани вредности (R и S), каде R ја означува процентуалната застапеност на пациенти кои припаѓаат во соодветниот опсег за коефициентот во споредба со сите останати пациенти во истата етничко-расна група, додека пак S ја означува процентуалната застапеност на пациенти од соодветната етничко-расна група кои припаѓаат во соодветниот ранг за коефициентот од сите примени пациенти во соодветниот опсег за вредноста на коефициентот. Табелата исто така содржи информација за бројот на приеми во секоја етничко-расна група од целосната кохорта, но и бројот на приеми во секој опсег за вредноста на коефициентот.

Според податоците во **Табела 4-8**, кај eICU-CRD сите етничко-расни групи имаат највисок број на приеми на пациенти со вредности за APACHE IV коефициентот под 50. Од друга страна, кај MIMIC-III, сите етничко-расни групи (со исклучок на Азијци) имаат најголем број на приеми на пациенти за кои вредностите на OASIS коефициентот се во опсег од 25 до 50. Од сите примени домородни Американци, нешто повеќе од 8% се во третиот вредносен опсег на APACHE IV коефициентот, т.е. со вредности за коефициентот од 100 до 150, а следна етничко-расна група по застапеност во истата категорија се пациентите од црна раса со 6,58%. Понатаму, овие две етничко-расни групи имаат поголема процентуална застапеност на приеми во нивната група и за четвртиот вредносен опсег за коефициентот во споредба со останатите етничко-расни групи. Поточно, и кај домородните Американци и кај пациентите од црна раса има над 6% од регистрирани приеми (во нивната етничко-расна група) во третиот вредносен опсег и над 0,6% од регистрирани приеми во четвртиот вредносен опсег за коефициентот на сериозност на болест, што не е случај со останатите етничко-расни групи или на ниво на сите приеми разгледани за податочното множество. Дополнително, трите колони означени со зелена боја ги посочуваат етничко-расните групи каде целокупната процентуална застапеност на приеми расте со порастот на коефициентот за сериозност на болест: ова може да се забележи кај приеми на Хиспанци и пациенти од црна раса кај eICU-CRD множеството, и кај пациентите од бела раса за MIMIC-III податочното множество.

Табела 4-8 Коефициент за оценување на сериозност на заболување, за сите пациенти и по етничко-расна група

Коефициент	Сите групи	Од бела раса		Од црна раса		Хиспанци		Азијци		Домородни Американци	
	Број на пациенти	R (%)	S (%)	R (%)	S (%)	R (%)	S (%)	R (%)	S (%)	R (%)	S (%)
eICU-CRD											
≤ 50	73.870	49,08%	75,93%	52,07%	11,93%	49,47%	3,89%	50,43%	1,44%	48,21%	0,68%
50 < s ≤ 100	65.720	45,13%	78,47%	40,75%	10,49%	44,25%	3,91%	43,22%	1,39%	42,80%	0,67%
100 < s ≤ 150	8.268	5,38%	74,35%	6,58%	13,46%	5,74%	4,03%	5,93%	1,51%	8,12%	1,02%
s > 150	674	0,41%	70,62%	0,6%	15,28%	0,54%	4,6%	0,42%	1,34%	0,87%	1,34%
Број на пациенти		114.282		16.922		5.803		2.108		1.035	
MIMIC-III											
≤ 25	20.664	32,43%	67,18%	37,35%	10,21%	41,37%	3,51%	57,95%	4,25%	46,3%	0,12%
25 < s ≤ 50	39.514	65,33%	70,77%	60,79%	8,69%	57,44%	2,55%	40,53%	1,55%	53,7%	0,07%
50 < s ≤ 75	1.354	2,24%	70,9%	1,86%	7,75%	1,19%	1,55%	1,52%	1,7%	0%	0%
Број на пациенти		42.809		5.646		1.755		1.515		54	

Средната вредност и стандардната девијација, како и медијаната и првиот и третиот квартил за коефициентите за сериозност на болеста при приеми на пациентите за целата кохорта и по секоја етничко-расна група се дадени во **Табела 4-9**, каде што може да се забележи дека не постојат особени разлики помеѓу вредностите на коефициентот и покрај очигледните разлики прикажани во **Табела 4-8**, освен во случај на Азијци и домородни Американци кај MIMIC-III податочното множество, што може да се должи на фактот дека за овие две групи застапеноста на приеми во однос на сите приеми изнесува 2,93% and 0,1% соодветно. Разлики се забележливи и кај Хиспанците кои врз глобалната слика влијаат со 3,39% пристапи. Слични, но не толку изразени, карактеристики може да се забележат и кај eICU-CRD множеството со Азијците и домородните Американци.

Табела 4-9 Статистика за коефициент за сериозност на болест по приеми во ЕИН, за сите пациенти и по етничко-расна група

Коефициент	Сите групи	Од бела раса	Од црна раса	Хиспанци	Азијци	Домородни Американци
eICU-CRD						
mean $\pm$ std	54,79 $\pm$ 26,13	54,99 $\pm$ 25,69	54,21 $\pm$ 27,96	55,06 $\pm$ 26,82	54,48 $\pm$ 26,73	57,83 $\pm$ 29,3
median [first, third]	51 [37, 68]	51 [37, 68]	49 [35, 68]	51 [36, 69]	50 [36, 68]	51 [36, 74]
MIMIC-III						
mean $\pm$ std	29,75 $\pm$ 9,83	30,01 $\pm$ 9,71	29,03 $\pm$ 9,82	27,67 $\pm$ 9,66	24,11 $\pm$ 10,88	27,39 $\pm$ 10,16
median [first, third]	30 [23, 36]	30 [24, 36]	29 [22, 36]	27 [21, 34]	24 [15, 32]	26 [17, 34]

Направените анализи покажуваат дека во податочните множества постојат непополнети податоци, кои треба да се регулираат. Дополнително, постојат значајни разлики во етнички и расни припадности за двете податочни множества, но и разлики во возрасна и полова репрезентација, како и разлики во дијагноза при прием и во коефициентите за сериозност на болеста за различни популации од пациенти. Гледајќи дека е скоро невозможно да се најдат податочни множества каде овие параметри ќе бидат слични, овие разлики треба да се адресираат при претпроцесирање на податоците и да бидат истражени откако моделите ќе бидат истренирани. Затоа, овие мали разлики се прифатлив баланс (анг. trade-off) за ова истражување. Оваа анализа дополнително го зацврстува фактот дека избраните податочни множества се адекватни за тематиката на оваа студија.

4.2 Етички аспекти

Податоците употребени за истражувањето се етички собирани и јавно достапни преку [97]. Методологијата на истражувањето може да биде следена и рекреирана, а добиените резултати треба да се споредливи со резултатите прикажани во оваа дисертација. Дополнително, експерименталниот дел од студијата, кој ќе биде презентираан, ги следи правилата за истражување на Strengthening the Reporting of Observational Studies in Epidemiology (STROBE), [114].

4.3 Претпроцесирање на податоци и избор на карактеристики

Откако ќе бидат детектирани одредени проблеми во квалитетот на податоците, може да се применат стратегии за ублажување на нивното влијание како дополнителен чекор во процесот на предпроцесирање на податоците, како дел од подготовка на податоците како влезни параметри во алгоритмот. Решавање на проблемот на полиња без вредност е од особено значење бидејќи голем број од алгоритмите на МУ не може да функционираат без целосност кај податоците; процесот се состои или од бришење (отстранување на редиците или колоните каде што има полиња без вредност при што како недостаток може да се избришат значајни податоци) или од замена на полињата без вредност (така што вредноста што недостасува се препокрива со произволна вредност, средна

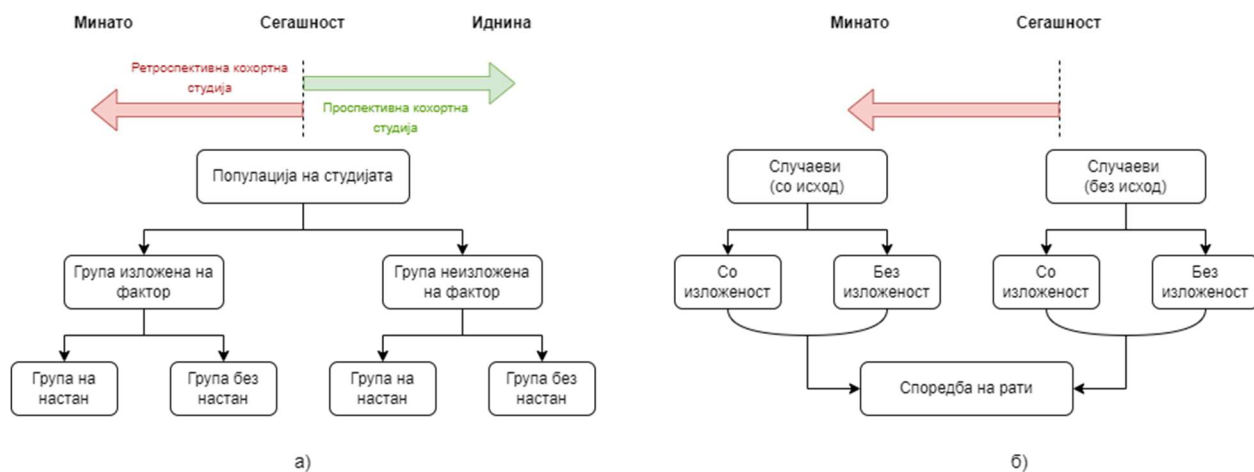
вредност, мода или медијана, или преку интерполација и други техники за обработка при што како недостаток може процесот да резултира со неточни податоци ако бројот на вредности што недостасуваат е голем или ако карактеристиката има голема варијанса). Кога станува збор за отстапувања во податоците, без разлика дали тие се должат на човечка грешка, технологијата за мерење на податоците или манипулација и семплирање на податоците, важно е истите да се детектираат и да се отстранат со цел да се добие робустен модел, кој нема да ги детектира овие вредности и да даде главен фокус на истите, [115]; отстапувањата во податоците може или да се избришат од податоците или да бидат заменети со одредена статистичка вредност, или со теоретски максимум (анг. capping) или теоретски минимум (анг. flooring).

Следна значајна карактеристика која го одредува квалитетот на податоците е рамнотежата на целните променливи во податочното множество, при што податоците се сметаат за неурамнотежени доколку во пропорцијата на класите се гледа значителна разлика, така што постојат мнозински класи кои заземаат огромен дел од податоците и малцински класи кои опфаќаат помал дел од податоците. Небалансираност кај податоците може да биде од блага до екстремна во зависност од пропорцијата која ја зафаќа малцинската класа, така што како блага небалансираност се смета онаа кај која малцинската класа зафаќа 20-40% од податоците, умерена небалансираност е кога малцинската класа зафаќа 1-20% од податоците, а како екстремна небалансираност се смета онаа каде застапеноста на малцинската класа во податоците изнесува под 1%. Небалансираност на податоците ја попречува способноста на алгоритмот да ги идентификува случаите од интерес во доста често, [116], бидејќи повеќето алгоритми се направени за да ја оптимизираат точноста (анг. accuracy) што е доверлива метрика кај небалансирани податоци, што резултира со слаби перформанси, како последица на една од две можни причини: или не постојат доволно податоци за една од класите за да се реализира тренирањето на моделот (па тогаш моделот учи само од мнозински застапената класа) или постојат голем број на специјални случаи во малцинската класа (и постои дополнително групирање во самата класа, што го попречува процесот на идентификација), [117]. Доколку тренирањето на оригиналната дистрибуција на податоците не доведе до задоволителни резултати, најчесто решавањето на проблемот на небалансираност се сведува на натсемплирање (креирање на нови примероци во малцинската класа преку дуплирање на постоечки или синтеза на нови) или потсемплирање (бришење на примероци од малцинската класа) на податоците.

Иако натсемплирањето е валидна техника, но кога процесот вклучува правење на егзактни копии од постојните примероци во малцинската класа се појавува недостаток во кој моделите ги учат дупликатите и не знаат да се снајдат со нови податоци, т.н. пренаучување (анг. overfitting), а друга форма на натсемплирање е користење на други техники преку кои новите примероци се генерираат синтетички. Една таква метода за генерирање податоци е SMOTE (Synthetic Minority Oversampling Technique), [118], која генерира нови примероци со вметнување на нов податок помеѓу примероците од оригиналните податоци користејќи ја стратегијата KNN (K-Nearest Neighbours), [119]; втора метода за иста намена е ADASYN (Adaptive Synthetic Sampling Approach), [120], со ист пристап како и кај методата SMOTE, но новите примероци не се создаваат помеѓу блиски точки од податоците кои припаѓаат на една класа и наместо тоа методата користи отстапувања во податоците за генерирање на новите информации. Според [121] натсемплирањето во формата во која тековно се користи е несигурно за учење. Згора на тоа, и SMOTE и ADASYN имаат недостатоци, т.е. SMOTE имплицитно претпоставува дека распределбата на класите е доволно хомогена во соседството околу примероците од малцинската класа, и следствено, процесот на генерирање може да скока помеѓу класи и да користи инстанци од повеќе класи за да генерира податоци за малцинската класа, поради што SMOTE не е применлив за сите сценарија, [122], додека ADASYN може да користи соседства на еден примерок ако отстапувањата на податоците кај малцинските инстанци се ретко распределени, [123].

Потсемплирањето резултира со помали пресметковни потреби при анализа на податоците во споредба со податоци добиени при натсемплирање, но сепак процесот на потсемплирање е со ризик за загуба на информативните примероци од мнозинската класа, [124]. Понатаму, за време на процесот на селекција отстранетите примероци се дел од оригиналната популација и со тоа нивното отстранување може да резултира со селективна пристрасност, [125]. Една метода за потсемплирање е случајно потсемплирање (анг. random undersampling), [126], каде што примероците се отстрануваат случајно и оваа постапка може да го генерира истиот недостаток, [127]. Сепак, авторите на [128] го истражуваат ефектот што стратегијата на потсемплирање (анг. sampling strategy) може да го има врз перформансите на алгоритмите на МУ, а добиените резултати покажуваат робусност кај неколку алгоритми на МУ дури и во случаи на значително намалено податочно множество.

Спарување (анг. matching) е уште една техника на семплирање, каде примероците од мнозинската класа се спојуваат со примероците од малцинската класа, [129]. Оваа метода се заснова на обсервациони студии, кои имаат за цел да набљудуваат поединци или да мерат постоење на одредени исходи, без обид да се влијае на истите со цел да се постигне одреден исход, [130]. Постојат три главни типа на обсервациони студии: кохортни студии (анг. cohort studies), студии за контрола на случај (анг. case-control studies) и студии со вкрстен пресек (анг. cross-sectional studies), [131]. Кохортните студии следат група учесници, т.н. кохорта, кои сподедуваат заедничка карактеристика, како што е на пример одредена професија или друга демографска сличност (види **Слика 4-7а**). Студиите за контрола на случај поврзуваат две групи, група за студија на случајот, чии учесници сподедуваат одреден атрибут и контролна група, со учесници кои не делат одреден атрибут, по што ги споредуваат двете групи за да утврдат дали кај групата на случај се појавува специфичен исход од интерес повеќе пати во споредба со исходите кај контролната група (види **Слика 4-7б**). Студиите со вкрстен пресек го анализираат исходот и изложеноста на учесниците во студијата во иста, специфична временска точка, [132]. Без разлика дали спарувањето се изведува кај пациенти со или без изложеност од интерес (кохортна студија) или кај пациенти со или без исход од интерес (студија за контрола на случај), процесот на спарување има за цел да добие исти или слични дистрибуции на специфични карактеристики (на пр. возраст, пол, дијагноза) во групите формирани од главните податоци, [133]. Ова ги изедначува пациентите, што му дозволува на алгоритмот на МУ да ги базира одлуките на карактеристики кои не се од интерес за студијата. Понатаму, усогласените кохортни студии често се поефикасни од случајното земање примероци, [134]. Иако кохортните студии имаат помал ризик од пристрасност, не е нужно точно дека пристрасноста ќе биде целосно елиминирана, [125], но докажано е дека спарувањето како техника е успешно при тешко-измерливи збунувачки променливи, [135].



Слика 4-7 Типови на обсервациони студии

4.3.1 Селекција на карактеристики. Селекција на пациенти.

Иницијалниот чекор имаше за цел формирање на популација од пациенти и избор на значајни карактеристики за истражувањето од двете податочни множества, но и да се адресираат одредени слабости во податоците, за да се добие ретроспективна кохортна студија од критично болни пациенти кај:

- eICU-CRD (*Слика 4-8*): каде изборот беше направен од 200,859 приеми на критично болни пациенти помеѓу 2014 и 2015 во 335 ЕИИ од 208 болници во САД;
- MIMIC-III (*Слика 4-9*): каде изборот беше направен од 61,532 приеми на критично болни пациенти собирани од еден центар помеѓу 2001 и 2012.

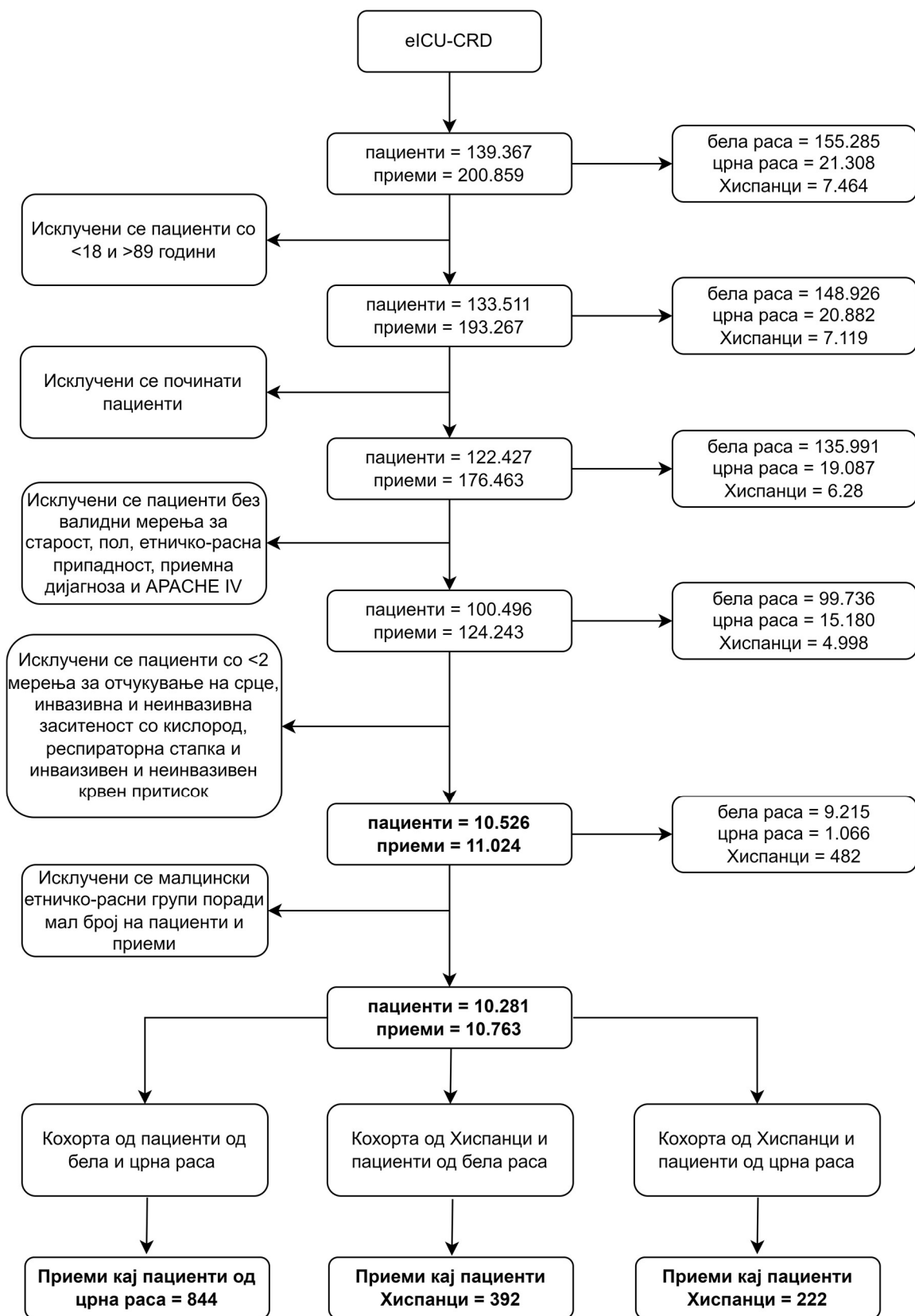
Кај двете податочни множества студијата ги разгледува сите возрасни пациенти (на возраст од 18 и повеќе години) кои биле живи во првите 24 часа по приемот во ЕИИ и имале најмалку две клинички валидни мерења за сите витални знаци разгледани во студијата: срцев ритам, респираторна стапка, неинвазивна кислородна сатурација (SpO₂), инвазивна кислородна сатурација (SaO₂), мерења на систолен, дијастолен и среден неинвазивен и инвазивен крвен притисок (NIBP и IBP), кои се детално прикажани во **Табела 4-10**. За секој од избраните витални знаци беа извлечени средна вредност, минимум, максимум и варијанса.

Дополнително, пациентите без информација за коефициентите APACHE IV и OASIS, приемна дијагноза, возраст или пол беа исклучени од студијата. Изборот на првите 24 часа за студијата е ограничувачки – едно, тој ја занемарува долгорочната прогресија на состојбата на пациентите, а второ, може потенцијално да влијае на бројот на пациенти; сепак, изборот беше заснован на фактот дека виталните знаци на пациентите во првите 24 часа ги информираат медицинските лица при следење на пациентите и во случај на итни интервенции. Ова значи дека пристрасните податоци за време на овој критичен период имаат потенцијал да носат најголема штета.

Табела 4-10 Карактеристики од интерес, единица мерка и опсег на валидни вредности

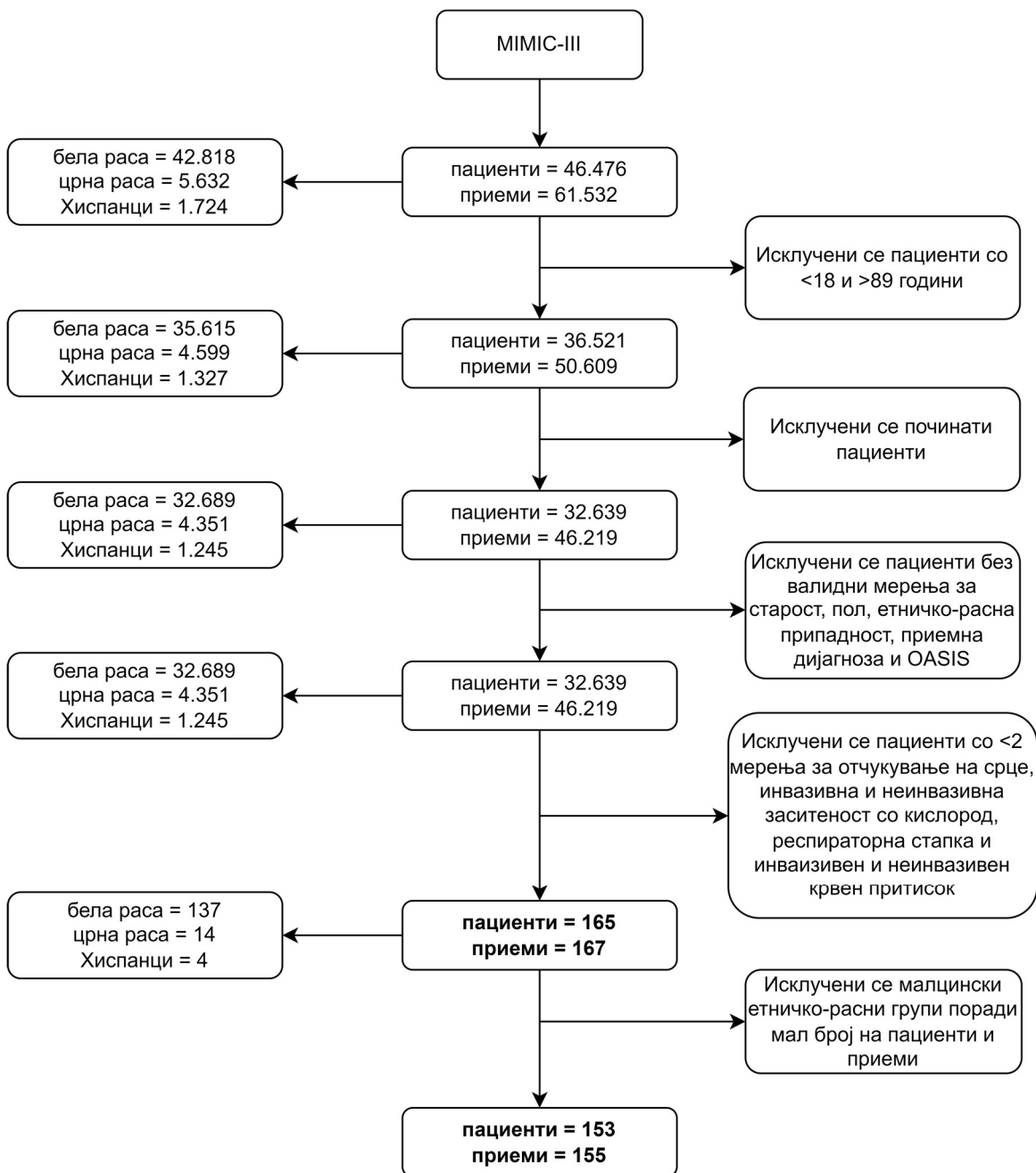
	Генералии		Медицинска состојба		Витални знаци			
Карактеристика	Возраст	Пол	Дијагноза	Статус на пациент	Срце	Респираторна стапка	Заситеност со кислород	Крвен притисок
Мерка	години	бинарна	текст	бинарна	bpm	insp/min	%	mmHg
Опсег	18–89	маж/жена	/	жив	0–300	0–200	0–100	0–300

Како што може да се види од *Слика 4-8*, процесот на селекција на пациенти за eICU-CRD резултираше со 10.763 приеми, од кои 9.215 (85%) беа пациенти од бела раса, потоа 1.066 (10%) беа пациенти од црна раса и 482 (5%) беа Хиспанци. Иако намалувањето на бројот на приеми е значајно, треба да се забележи дека преостанатите и испитуваните пациенти имаат слична етничко-расна распределба како онаа во оригиналното податочно множество (види *Слика 4-6*). Дополнително, бидејќи достапни се значително помалку податоци за пациентите од Азија и домородните Американци (вкупно 177 и 84 приеми на број, соодветно), овие етничко-расни групи не беа вклучени во студијата.



Слика 4-8 Дијаграм за селекција на пациенти: eICU-CRD

Но, иако процесот на селекција на пациенти не се разликуваше за MIMIC-III множеството, сепак тој резултираше со екстремно намалување на приеми, како што покажуваат бројките на **Слика 4-9**, т.е. само 155 приеми ги исполнија барањата на процесот на селекција, при што значајниот пад на бројот на пациенти се јавува при избор на валидни мерења на виталните знаци. Бројот на записи што недостасуваат за секој од виталните знаци кои се разгледуваат е даден во **Табела 4-11**, и тука може да се забележи дека над 20.000 приеми имале само едно или немале ниту едно мерење за инвазивна заситеност со кислород и инвазивен крвен притисок во првите 24 часа. Од оваа причина, процесот на селекција кај MIMIC-III беше променет за да зема предвид само четири од шесте првични витални знаци, по што добиената популација се состоеше од 7.287 (82%) пациенти од бела раса, потоа 1.228 (14%) пациенти од црна раса и 349 (4%) Хиспанци.



Слика 4-9 Дијаграм за селекција на пациенти: MIMIC-III

Табела 4-11 Преглед на празни полиња по витален знак во MIMIC-III

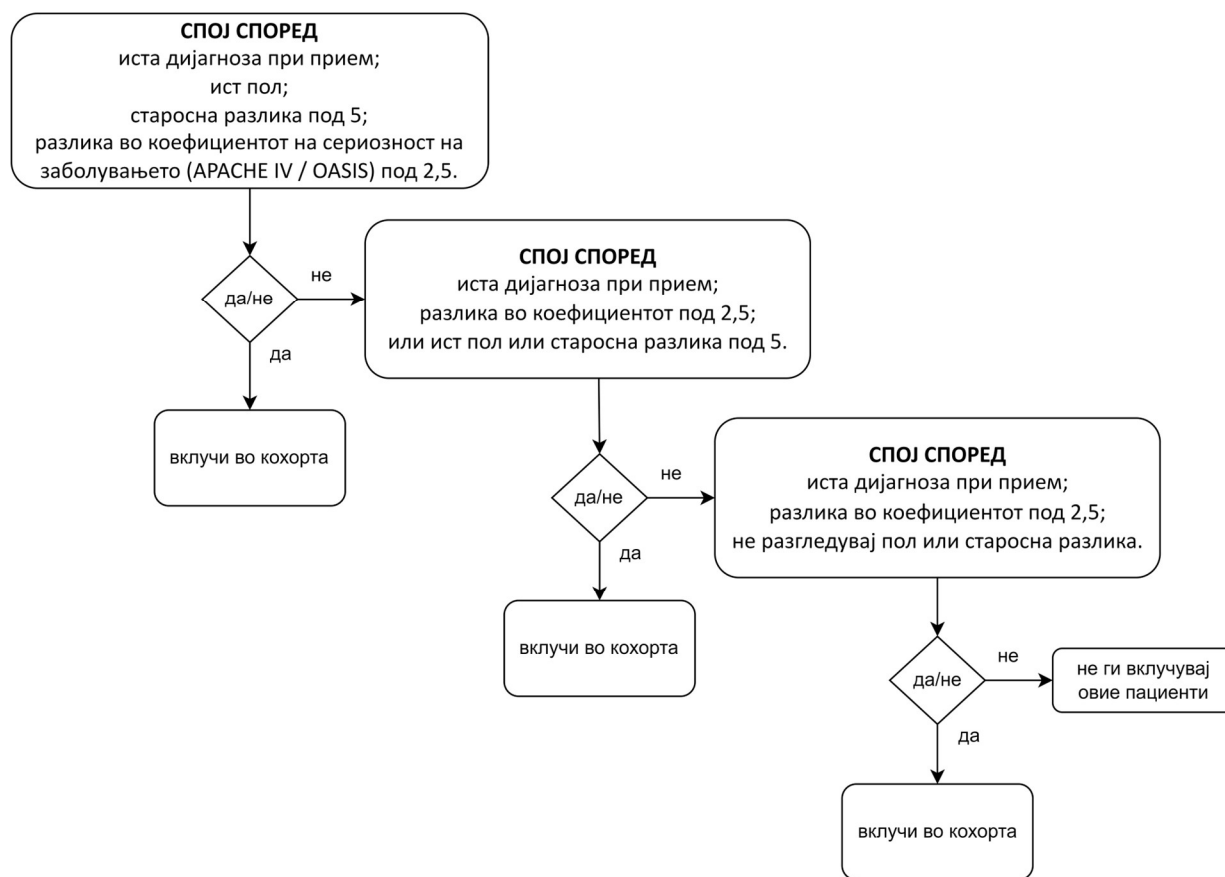
Витален знак	Број на приеми со под 2 мерења
Отчукување на срце	1.413
Респираторна стапка	1.544
Неинвазивна кислородна заситеност	1.597
Инвазивна кислородна заситеност	27.167
Неинвазивен крвен притисок	5.183
Инвазивен крвен притисок	22.754

4.3.2 Креирање на кохорта

Бројот на приеми покажува значителна нерамнотежа кај податоците претежно со пациенти од бела раса; затоа, со цел да се балансира бројот на пациенти по етничко-расна припадност, беше разгледано случајно потсемплирање на податоците, [136]. Сепак, истражувањето презентирано во [137] сугерира дека значителен дел од перформансите на алгоритмите при предвидување на демографски информации (на пр. раса) можат да бидат резултат на детекција на збунувачки фактори. „Случајниот“ аспект на случајното потсемплирање може ја создаде појавата и ненамерно да влијае на исходот како збунувачка променлива, што значи дека може да создаде група со особени разлики помеѓу пациентите кај потенцијални збунувачки променливи и последователно да создаде некоректна асоцијација со исходот. Затоа, со цел да се контролираат збунувачките променливи и да се ублажи пристрасноста од застапеноста на другите раси, беше направено формирање на парови од пациенти од недоволно застапените етничко-расни групи. Процесот на спарување беше точка на дискусија, каде што првично земените три карактеристики беа возраста, полот и дијагнозата при прием, [138], а пациентите беа спојувани во неколку различни кохорти на тој начин што за секоја кохорта соодветно една од трите карактеристики имаше предност пред другите. Сепак, пациентите со иста возраст, пол и дијагноза може да имаат различен коефициент за сериозност на болеста, што може да влијае на виталните знаци, и затоа конечните процеси на спарување ги земаа во предвид трите оригинални карактеристики, но и коефициентот за сериозност на болеста (APACHE IV за eICU-CRD и OASIS за MIMIC-III) како критериуми за совпаѓање, [139].

Процесот на спарување, илустриран на *Слика 4-10*, содржеше неколку чекори. Првично, пациентите од малцинската етничко-расна група беа споредувани со пациенти од позастапената етничко-расна група за еднаквост на приемна дијагноза и пол, додека разликата помеѓу возраста не требаше да надмине 5 години, а разликата помеѓу APACHE IV коефициентите да не надмине 2,5. Ако не беше пронајдено совпаѓање на пациентите, споредбата беше проширена за да се проверат пациентите со разлика под 2,5 во APACHE IV коефициентот и со иста дијагноза, но или со ист пол или со возрастна разлика под 5 години. Конечно, пациентот кандидат беше избран од оние пациенти кои имале истата дијагноза при прием и разлика во APACHE IV коефициентот која не е поголема од 2,5, без да се земе предвид полот или возраста. Доколку не беше пронајдено совпаѓање, тогаш пациентот од малцинската етничко-расна група не беше земен како дел од кохортата.

За eICU-CRD, спарувањето на пациенти резултираше со 844 пациенти од црна раса кои се совпаѓаат со ист број пациенти од бела раса, 392 пациенти Хиспанци кои се совпаѓаат со истиот број пациенти од бела раса и 222 пациенти Хиспанци кои се совпаѓаат со истиот број на пациенти од црна раса; за MIMIC-III, совпаѓањето на групата резултираше со 1.079 пациенти од црна раса кои се совпаѓаат со ист број пациенти од бела раса, 305 пациенти Хиспанци кои се совпаѓаат со истиот број пациенти од бела раса и 250 пациенти Хиспанци кои се совпаѓаат со ист број пациенти од црна раса.



Слика 4-10 Процес на спарување на пациенти

Основните карактеристики на пациентите за добиените кохорти од двете податочни множества се прикажани во **Табела 4-12**, каде што за секоја специфична карактеристика за секоја кохортна група податоците се анализирани со примена на интерквартален опсег кај континуираните променливи и проценти кај категориичните променливи, додека р-вредноста беше пресметана со Kruskal-Wallis тест (или еднонасочна ANOVA), [140].

Набљудуваните витални знаци се разликуваат помеѓу eICU-CRD и MIMIC-III, бидејќи кај MIMIC-III множеството SaO₂ и IBP карактеристиките не беа редовно мерени и затоа не беа вклучени во крајните вектори на карактеристики. Понатаму, критериумите за избор на валидни инстанци овозможува да не постои недостаток на податоци пред да се изврши анализата. Може да се забележи дека иако процесот на спарување го исклучи совпаѓањето на пациентите со ист пол во двата последни чекори, добиените кохорти се релативно балансирали во тој поглед, што сугерира дека постои мала веројатност полот на пациентите да биде збунувачка променлива. Резултатите претставени во **Табела 4-12** покажуваат дека скоро сите карактеристики се статистички значајни (р-вредности – англ. p-value – под 0,05 или 0,001).

Табела 4-12 Карактеристики на пациентите во добиените кохорти

Витални знаци	Кохорта од пациенти од бела и црна раса			Кохорта од Хиспаници и пациенти од бела раса			Кохорта од Хиспаници и пациенти од црна раса		
	црна раса	бела раса	p-value	Хиспаници	бела раса	p-value	Хиспаници	црна раса	p-value
eICU-CRD									
Пациенти	844	844	-	392	392	-	222	222	-
Пол (мажи)	457 (51,7%)	483 (57,23%)	-	236 (60,2%)	238 (60,7%)	-	141 (63,5%)	127 (57,21%)	-
Возраст	60 [51, 69]	65 [56, 73]	< 0,05	63 [53, 73]	66 [57, 73]	< 0,05	63 [54, 73]	62 [52, 70]	0,151
Отчукување на срце	87 [75, 99]	84 [73, 97]	< 0,05	84 [73, 97]	83 [73, 96]	< 0,05	84 [74, 96]	86 [76, 98]	< 0,05
Респираторна стапка	19 [15, 24]	19 [16, 23]	< 0,05	19 [16, 23]	19 [15, 23]	< 0,05	19 [16, 23]	19 [15, 24]	< 0,05
SpO2	99 [97, 100]	98 [95, 99]	< 0,05	98 [96, 100]	97 [95, 99]	< 0,05	98 [96, 100]	99 [96, 100]	< 0,05
SaO2	97.6 [95, 99]	97.4 [95, 99]	< 0,05	97 [95, 99]	97 [95, 99]	0,07	97 [95, 99]	97 [95, 99]	< 0,05
Систолен IBP	123 [107, 141]	120 [106, 138]	< 0,05	122 [108, 138]	123 [108, 140]	< 0,05	119 [106, 135]	122 [108, 140]	< 0,05
Дијастолен IBP	61 [53, 70]	58 [50, 66]	< 0,05	59 [51, 68]	59 [51, 68]	< 0,05	58 [51, 67]	60 [53, 68]	< 0,05
Среден IBP	80 [72, 91]	78 [69, 87]	< 0,05	79 [70, 90]	79 [70, 90]	< 0,001	78 [69, 88]	80 [72, 89]	< 0,05
Систолен NIBP	121 [105, 139]	118 [103, 135]	< 0,05	120 [106, 135]	119 [105, 136]	< 0,001	117 [104, 133]	123 [106, 141]	< 0,05
Дијастолен NIBP	66 [57, 76]	63 [55, 73]	< 0,05	64 [55, 74]	63 [55, 73]	< 0,05	62 [54, 72]	65 [56, 75]	< 0,05
Среден NIBP	82 [72, 94]	79 [69, 91]	< 0,05	78 [69, 89]	80 [70, 91]	< 0,05	76 [68, 87]	82 [72, 94]	< 0,05
MIMIC-III									
Пациенти	1.079	1.079	-	305	305	-	250	250	-
Пол (мажи)	462 (42,82%)	487 (45,13%)	-	203 (66,56%)	199 (65,25%)	-	164 (65,6%)	145 (58%)	-
Возраст	60 [47, 71]	62 [51, 73]	< 0,001	53 [41, 67]	57 [45, 68]	0,112	55 [43, 68]	57 [45, 67]	0,562
Отчукување на срце	87 [75, 101]	86 [73, 99]	< 0,001	89 [75, 103]	86 [72, 98]	< 0,001	89 [75, 104]	86 [73, 101]	< 0,001
Респираторна стапка	19 [16, 23]	19 [16, 22]	< 0,001	18 [15, 22]	18 [15, 21]	0,532	18 [15, 22]	18 [15, 22]	0,295
SpO2	99 [96, 100]	97 [95, 99]	< 0,001	98 [96, 100]	97 [95, 99]	< 0,001	98 [96, 100]	99 [97, 100]	< 0,001
SaO2	125 [108, 144]	116 [102, 133]	< 0,001	120 [105, 136]	119 [106, 136]	0,85	120 [106, 137]	126 [109, 144]	< 0,001
Систолен IBP	68 [57, 81]	63 [53, 74]	< 0,001	67 [57, 77]	66 [57, 78]	0,368	67 [57, 78]	69 [57, 83]	< 0,001
Дијастолен IBP	81 [70, 94]	75 [66, 87]	< 0,001	79 [70, 91]	79 [69, 90]	0,386	79 [69, 91]	83 [70, 96]	< 0,001

4.4 Алгоритми и оптимизација на параметри

4.4.1 Алгоритми

Откако податоците ќе бидат адекватно претпроцесирани може да се започне со тренирање на алгоритмите. Во зависност од проблемот и податоците кои се користат, учењето може да биде: надгледувано (анг. supervised, каде алгоритмите се тренираат со лабелирани подаци, т.е. точниот излез е достапен при тренирање на моделот), ненадгледувано (анг. unsupervised, каде алгоритмите се тренираат бел лабелирани податоци, т.е. кај податоците не е дефинирана корелација со точниот излез, па алгоритмот бара шаблони во податоците без никакво надгледување), полунадгледувано (анг. semi-supervised, кое е комбинација од претходните два типа, каде алгоритмот добива главно лабелирани податоци, но исто така има можност за анализа на податоците независно од лабелата), и учење со зајакнување (анг. reinforcement, кое е базирано на наградување на точни предвидувања на моделот и казнување на неточните), [141].

Алгоритмите за надгледувано машинско учење се применуваат на лабелирани (означени) податоци со цел да се откријат шаблони во карактеристиките и истите да се поврзат со одреден резултат (излезна информација). Доколку излезот од моделот е категорична вредност, процесот на учење се нарекува класификација. Категоричните вредности може да имаат едно или ограничен број на можни излезни вредности; во случај кога категоричниот излез добива само две вредности процесот на класификација е бинарен, а во случај кога категоричниот излез добива повеќе од две вредности процесот е мултикласна класификација. И покрај тоа што постојат голем број на популарни класификатори за МУ, како логистичка регресија (анг. Logistic Regression, скратено LR), Наивен Бејс (анг. Naïve Bayes Classifier), дрва на одлуки (анг. Decision Trees), случајна шума (анг. Random Forest), векторски машини за поддршка (анг. Support Vector Machines), и XGBoost, [142], [143], при што истражувањето во оваа дисертација ќе се фокусира на два од наведените алгоритми: логистичка регресија и XGBoost. Притоа, за секоја креирана кохорта од пациенти беше разгледувана бинарна класификација, имено дали виталните знаци може да се користат за предвидување на самопријавена раса во три засебни тестови за споредба на: пациенти од црна и бела раса, Хиспанци и пациенти од бела раса, и Хиспанци и пациенти од црна раса. За секоја од бинарните анализи, двата алгоритми се користеа за развој на моделите.

LR е веројатен класификатор, [144], и бара корпус за тренирање од n влезно/излезни парови $(x^{(j)}, y^{(j)})$, каде секоја влезна инстанца $x^{(j)}$ во момент j е вектор од карактеристики $[x_1, x_2, \dots, x_m]$, а секоја i -та карактеристика за соодветниот влезен податок ќе биде означена како $x_i^{(j)}$. Векторите се даваат на алгоритмот за време на фазата на тренирање и алгоритмот моделира врска помеѓу карактеристиките и резултантната променлива (лабела), по што користи класификациска функција за да одреди бинарен исход на проценетата класа $\hat{y}$. Врската се моделира преку член за пристрасност и тежински вектор, каде секоја тежина w_i се асоцира со единствена влезна карактеристика x_i , така што тежината ја репрезентира важноста на влезната карактеристика за одлуката при класификација. Добиениот излез може да биде или 0 (кога соодветниот примерок не припаѓа на соодветна класа) и 1 (инстанцата припаѓа на класата), а пресметката ја користи сигмоидната функција со цел да ја одреди веројатноста одредена инстанца да припаѓа на класата $P(y = 1|x)$, поточно:

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

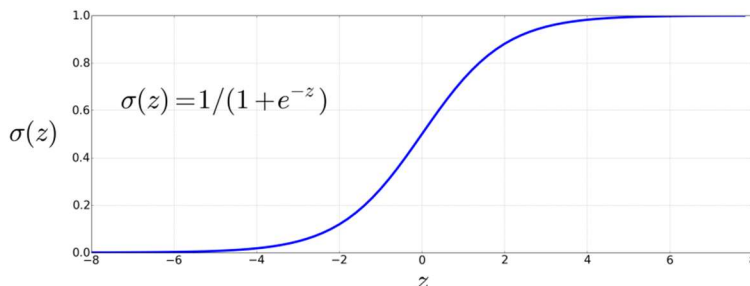
каде z е тежинска сума за класата (сума од скаларните производи за векторот на влезни променливи и тежинскиот вектор, по што се додава и член за пристрасност b), и изнесува:

$$z = \left(\sum_{i=1}^m w_i x_i \right) + b.$$

Во примена, бидејќи тежините се реални вредности, вредноста на z може да биде во опсег од $-\infty$ до ∞ , но сигмоидната функција (илустрирана на **Слика 4-11**) ги мапира вредностите во $(0, 1)$. За да вредноста стане веројатноста вредност, двата случаи $P(y = 1|x)$ и $P(y = 0|x)$ треба да имаат сума од 1, и од таа причина:

$$P(y = 1|x) = \sigma(z),$$

$$P(y = 0|x) = 1 - \sigma(z).$$

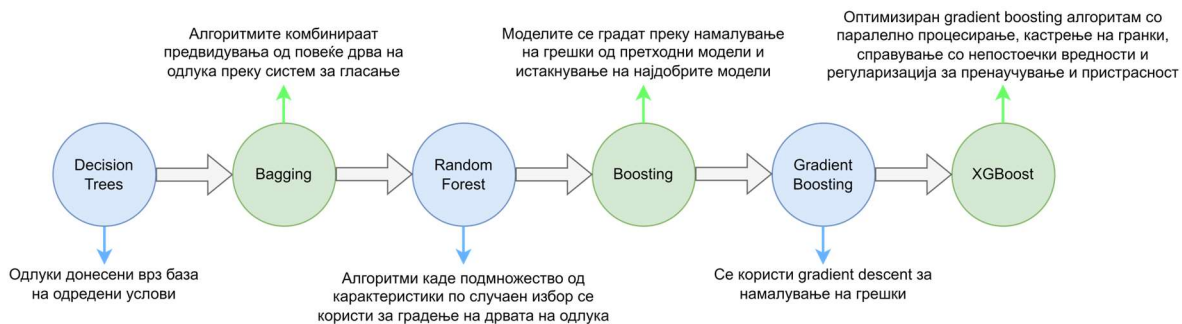


Слика 4-11 Сигмоидна функција

Кога се пресметува веројатноста за инстанца од тест множеството $x^{(j)}$, се применува следната граница на одлука:

$$decision(x^{(j)}) = \begin{cases} 1, & \text{if } P(y = 1|x) > 0.5 \\ 0, & \text{otherwise} \end{cases}.$$

XGBoost алгоритмот, [145], [146], е имплементација на gradient boosted trees, која се базира на функционална априксимација преку оптимизирање на специфична функција на загуба проследено со регулациски техники, што овозможува паралелно tree boosting за регресија, класификација и рангирање, што е изградено врз концепти на надгледувано МУ, дрва на одлука, ensemble learning, и gradient boosting, како што може да се види на **Слика 4-12**.



Слика 4-12 Од дрва на одлука до XGBoost

Со корпус од n влезно/излезни парови дефинирани како $D = \{(x_i, y_i)\}$ каде $|D| = n$, секој влезен примерок е вектор од m карактеристики, т.е. $x_i \in \mathbb{R}^m$, и имаа соодветен излез $y_i \in \mathbb{R}$, а tree ensemble model користи K адитивни функции (дрва) за предвидување на излезот:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^K f_k(x_i), \quad f_k \in \mathcal{F},$$

каде $\mathcal{F} = \{f(x) = w_{q(x)}\}$ е просторот на класификациски и регресиони дрва (CART = classification and regression trees space). Структурата на секое дрво може да се репрезентира со $q: \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^T$, што значи дека секоја независна структура на дрво q која содржи T листови, со

континуиран резултат од w_i за лист i , влезен примерок се мапира во соодветен излезен индекс за лист базирано на правилата за одлука кои ги дефинира q . Финалното предвидување се пресметува така што се сумираат сите вредности w во соодветните листови од сите дрва. Множеството од функции кои се користат во моделот се учат со минимизирање на објективната функција:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k),$$

$$\Omega(f) = \gamma T + \frac{\lambda \|w\|^2}{2}.$$

Првиот дел од објективната функција, т.е. l е диференцијабилна конвексна функција на загуба која го одредува растојанието помеѓу предвидувањето на моделот и оригиналната лабела, додека вториот дел ја казнува комплексноста на моделот, т.е. функциите на регресионото дрво. Објективната функција $\mathcal{L}(\phi)$ на моделот не може да се оптимизира во Евклидовиот простор користејќи традиционални оптимизациски методи, и последователно моделот се тренира на адитивен начин, односно се корегира наученото и се додава ново дрво во секој временски момент, имено:

$$\begin{aligned}\hat{y}_i^{(0)} &= 0 \\ \hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\ &\dots \\ \hat{y}_i^{(t)} &= \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)\end{aligned}$$

па со тоа, објективната функција која треба да биде минимизирана во одредена итерација t е:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t).$$

Со цел да оваа објективна функција се пренесе во функција во Евклидовиот домен, каде може да се користат традиционални оптимизациски техники, потребно е да се примени Тејлоровата формула за апроксимација, [147]. Според Тејлор, наједноставната линеарна апроксимација на функција $f(x)$ околу одредена точка a е: $f(x) \approx f(a) + f'(a)(x - a) = f(a) + f'(a)\Delta x$; во случајов, $f(x)$ е функцијата за загуба l , a е вредноста предвидена од чекор $(t - 1)$, и Δx е новиот модел кој треба да се додаде во чекор t . Користејќи ја Тејлоровата апроксимација од втор ред, формулата станува:

$$\mathcal{L}^{(t)} \cong \sum_{i=1}^n \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t),$$

каде, $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ и $h_i = \partial_{\hat{y}^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)})$ се градиентите на функцијата на загуба од прв и втор ред, а отстранувањето на константите од равенката резултира со поедноставена верзија на објективната функција:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t).$$

Со ова, објективната функција станува функција на дрво само во чекор t , во споредба со верзијата пред примена на Тејлоровата апроксимација во која истата беше сума од t CART дрва.

Ако се дефинира $I_j = \{i | q(x_i) = j\}$ како множество од индекси за податочните инстанци кои се доделени на листот, објективната вредност станува:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 = \sum_{j=1}^T \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T.$$

За фиксна структура $q(x)$, оптималната тежина на еден лист може да се пресмета со:

$$w_j^* = - \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

а соодветната оптимална вредност која го мери квалитетот на структурата на дрвото е:

$$\tilde{\mathcal{L}}^{(t)}(q) = - \frac{1}{2} \sum_{j=1}^T \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T.$$

Во пракса е невозможно да се енумерираат сите можни структури на дрвото q со цел да се одреди дрвото со максимално намалување кај загубата. Од таа причина, процесот кој се користи за градење на моделот започнува со единствен корен, кој ги содржи сите примероци од множеството за тренирање. Следно, сите карактеристики и вредности по карактеристика се изминуваат, па се евалуира секое намалување на загубата при поделба, односно се оценува добивката како разлика помеѓу загубата од инстанците кај јазолот родител и сумата од загубите од левата и десната гранка. Оваа добивка се користи како основа за поделбата и мора да биде позитивна вредност поголема од вредноста на параметарот за минимална добивка при делење на моделот со цел да дрвото може да се разграни. Ова е т.н. “exact greedy algorithm”, чија комплексност е $O(n * m)$, по што функцијата за загуба се дефинира како:

$$\mathcal{L}_{split} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma,$$

каде, I_L и I_R се множествата од примероци на левиот и десниот јазол по поделбата.

4.4.2 Оптимизација на параметри

Алгоритмите на МУ се склони на пренаучување, т.е. моделот може да предвидува точно за множеството за тренирање, но да има слаби перформанси за нови податоци; XGBoost не е исклучок. Техниките за контролирање на пренаучувањето кај XGBoost може да се групираат во четири категории: регулациски, методи на кастрење, семплирачки и рано запирање; регулациските параметри влијаат директно на тежините (т.е. lambda, alpha, eta, max_delta_step), методите за кастрење исфрлаат поделби кај дрвата за време и по процесот на градење на моделот (т.е. gamma, max_depth, min_child_weight), семплирачките методи ги прави дрвата помалку корелирани и поробустни на шум (т.е. subsample, colsample_*), а раното запирање ја мониторира промената на метриците од интерес на издвоени податоци и престанува да го гради моделот кога кај метриката нема да се забележи подобрување, [148]. Оптимизацијата на параметрите е еден од начините да се прилагодат дел или сите од овие параметри за да се добијат најдобри перформанси на моделот, како и да се спречи пренаучување кај моделот, [149]. Оптимизацијата во оваа дисертација е изведена со помош на RandomizedSearchCV, [150], кој по случаен избор избира вредности за параметрите и го пресметува резултатот, по што го враќа множеството параметри за кои се добива најдобриот резултат. Ова значи дека евалуиравме две верзии на алгоритмот XGBoost: една со стандардните

параметри и оптимизирана верзија (со параметри избрани со случајно пребарување за вкрстена валидација, при што разгледаните опсези за параметрите се дадени на *Слика 4-13*).

```
params = {
    'eta': np.arange(0, 1, 0.05),
    'min_child_weight': np.arange(0, 10, 0.5),
    'gamma': np.arange(0, 10, 0.5),
    'subsample': np.arange(0.1, 1, 0.1),
    'colsample_bytree': np.arange(0, 1.0, 0.1),
    'max_depth': np.arange(2, 7),
    'n_estimators': np.arange(100, 1500, 100),
    'learning_rate': np.arange(0.05, 1, 0.05),
}
```

Слика 4-13 XGBoost параметарска мрежа за RandomizedSearchCV

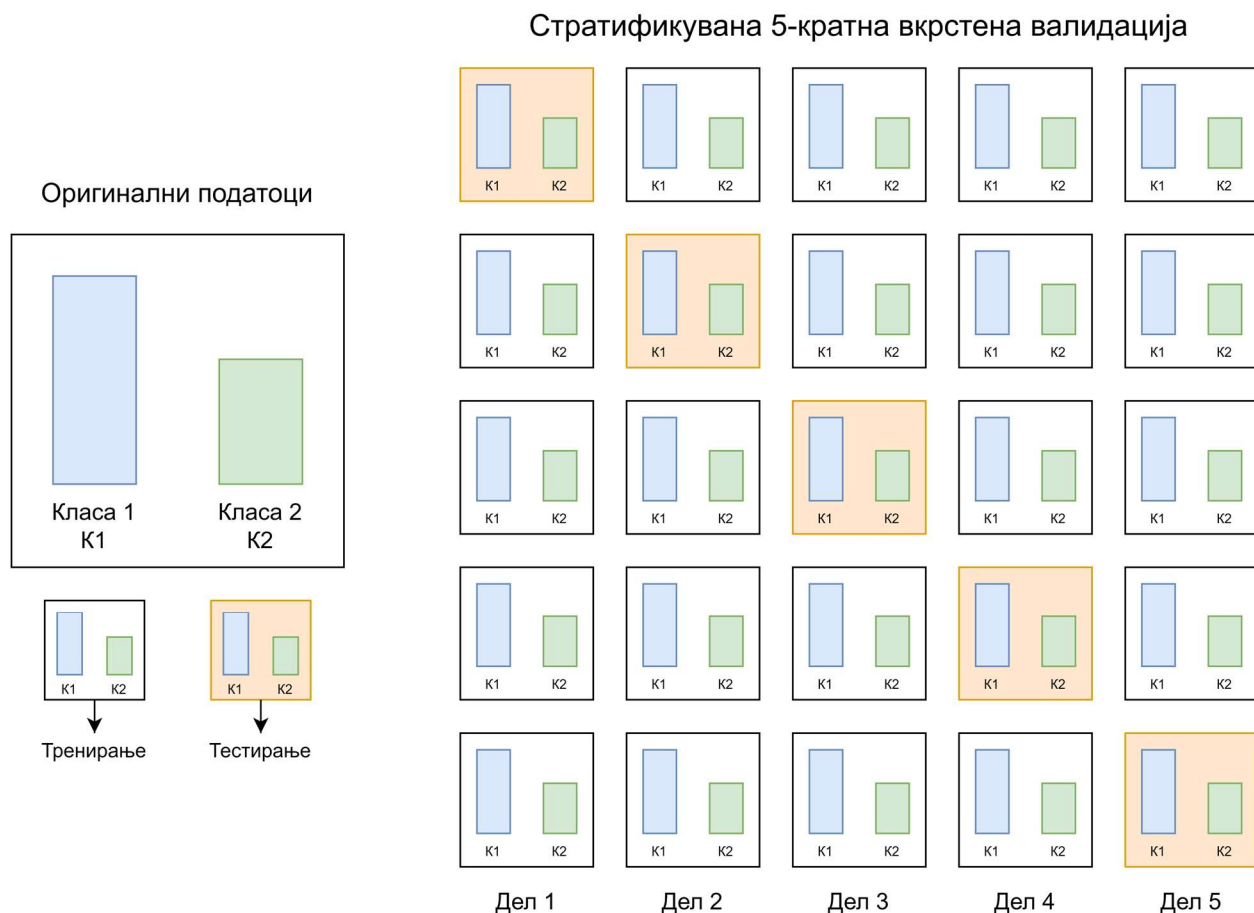
4.5 Валидација на моделот, евалуациски метрики, и интерпретабилност на моделот

4.5.1 Валидација на моделот

Откако моделот е обучен, неопходно е да се оцени квалитетот на системот за МУ и да се осигура дека добиениот модел функционира на предвидениот начин, а бидејќи сценаријата од реалниот свет користат примероци на податоци кои често не се вистински репрезентативци на популацијата, валидацијата на моделот се користи за да се добие чувство за грешката на моделот, [151]. Процесот на валидација се изведува по обуката на моделот и користи невидени податоци од множеството за тренирање за да добиениот модел направи предвидувања. Податоците треба да се усогласени со изворните податоци што се користат за обука на алгоритмите, а резултатите од излезот на моделот треба да бидат споредливи со вистинските резултати од податоците од тест множеството. Ова нуди информација која може да помогне во корекција на работата на моделот во случај на погрешна класификација поради изложеност на дел од податоците или грешка при мерење, [152].

Валидацијата може да биде внатрешна и надворешна. Внатрешната валидација користи дел од оригиналните податоци, додека надворешната валидација користи целосно посебна база на податоци; во контекст на медицината, внатрешната валидација ги набљудува перформансите на моделот низ дел од оригиналната популација на пациенти, додека надворешната валидација користи целосно одделна популација на пациенти за тестирање на перформансите на моделот, [153]. Внатрешната валидација може да се реализира со различни техники, дел од кои вклучуваат: ресупституција (анг. resubstitution), поделба (така што оригиналните податоци се поделени во две подмножества, анг. holdout), k-кратна вкрстена валидација (анг. k-fold cross-validation), вкрстена валидација со отфрлање на една класа (анг. leave-one-out cross-validation), случајно потсемплирање (анг. random subsampling) и валидација со случајно семплирање со замена (анг. bootstrapping), [154]. Ресупституцијата пресметува грешка при замена, што се оценува врз основа на исходот и вредноста од множеството за тренирање употребено за изградба на моделот. Со класична поделба на множествата се избегнува грешката при замена; за оваа цел оригиналните податоци се поделени на податоци за тренирање и тестирање на модели одржувајќи одреден сооднос, при што кај добиените подмножества може да постои нерамномерна распределба на различни класи, па затоа постапката може да бара стратификација на податоците за да се поправи нерамномерната распределба. И двете техники имаат ист недостаток; имено, евалуацијата се врши на специфично подмножество од оригиналните податоци, а избраните примероци во тоа подмножество можеби нема точно да ги отсликаат реалните перформанси на моделот, па од таа причина процеси на евалуација кои имаат само еден чекор при проверка на моделот често не нудат најдобар увид во квалитетот на моделот. Овој проблем претрпува корекција во преостанатите техники за валидација.

Со k -кратна вкрстена валидација, базата на податоци е поделена во k подмножества со еднаква големина или колку што е можно приближно еднаква големина. Процесот се повторува k пати додека секое поединечно подмножество не се употреби за валидација на моделот точно еднаш, додека соодветната итерација ги користи преостанатите $k-1$ подмножества за да го обучи моделот. Предноста овде е што сите податоци се користат за тренирање и тестирање на моделот, а бидејќи во процесот има k повторувања, конечната грешка на моделот се пресметува како аритметичка средина од грешките на сите повторувања. Кај k -патна поделба, бидејќи процесот на поделба на множеството се изведува случајно, добиените подмножества може да ја изгубат оригиналната дистрибуција на податоци, што може да биде проблем ако едно или повеќе од добиените подмножества не претставуваат одредена класа адекватно, што главно е проблем во неурамнотежени податочни множества, [155]. За да се зачува оригиналната дистрибуција на множеството, се користи стратификација, каде секое од k подмножествата се карактеризира со ист однос на класи како и оригиналното множество, што е илустрирано на **Слика 4-14**.



Слика 4-14 Стратификувана k -кратна вкрстена валидација

Вкрстена валидација со отфрлање на една класа има онолку итерации колку што е бројот на достапни примероци во податоците, при што секоја итерација ги користи сите податоци (освен еден примерок) за тренирање на моделот, додека преостанатиот примерок се користи за тестирање. Бидејќи процесот се повторува повеќекратно, грешката на моделот е аритметичка средина од грешките добиени во секоја од итерациите. Слично на тоа, случајно потсемплирање користи случајно избрани подмножества од податоците за да формира база на податоци за тестирање, додека преостанатите податоци се користат за формирање на множеството на тренирање, а аритметичката средина од сите грешки на секоја итерација ја формира грешката на моделот. Малку поинаква

техника е валидацијата со случајно семплирање со замена, каде што податоците за тренирање се избираат случајно со замена, така што во подмножеството за тренирање постојат дупликати на инстанци со цел да се заменат примероците кои биле избрани за тестирање на моделот. Големината на податоците од тест множеството се менува во секој следен чекор, но грешката на моделот сè уште е аритметичка средина од сите повторувања.

Бидејќи k-кратната вкрстена валидација нуди максимална искористеност на податоците без да бара тестирање на моделот на само една инстанца, може да ја задржи оригиналната дистрибуција на податоци за време на фазите на обука и тестирање и може да обезбеди еквивалентна поделба на податоците низ различни експерименти во споредба со случајното потсемплирање и случајното семплирање со замена, во ова истражување се одлучивме да користиме стратификувана k-кратна вкрстена валидација како техника за внатрешна валидација на моделот. Со тоа, секој бинарен тест претставуваше 5-кратна вкрстена валидација каде секој земен отсечок од податоците ја имаше оригиналната дистрибуција на ниво на етничко-расна припадност (т.е. класата од интерес). За тренирање на моделот се користеа четири од петте отсечоци, додека преостанатиот се користеше за валидација на перформансите на моделот. Процесот беше повторен пет пати, така што секој отсечок точно еднаш се користи за валидација на моделот. Финалните резултати се средна вредност од сите отсечоци.

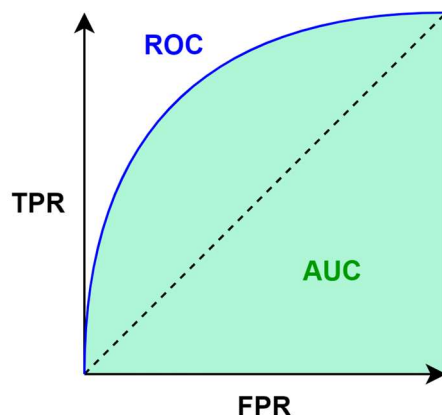
4.5.2 Евалуациски метрики

Метриката која се користи за евалуација на перформансите на моделот е критичен дел од процесот бидејќи нудат непристрасна информација за квалитетот на моделот, а изборот на метрика е значаен бидејќи во голем број случаи метриката може да залажува за реалниот квалитет на моделот, [156]. Една метрика која се користи кај проблеми од класификација е матрицата на грешка (анг. confusion matrix), [157], прикажана на **Слика 4-15**, и матрицата дава прави преглед на вистинските лабели на примероците и лабелите кои ги предвидува моделот. Основните термини кои се поврзуваат со матрицата на грешка се: true positives (TP) – точно предвидени како позитивни, true negatives (TN) – точно предвидени како негативни, false positives (FP) – неточно предвидени како позитивни (исто така познато како грешка од тип 1, анг. type 1 error), и false negatives (FN) – неточно предвидени како негативни (исто така познато како грешка од тип 2, анг. type 2 error). Матрицата на грешка нуди информации потребни за пресметка на дополнителни метрики, значајни за анализа на перформансите на моделот: точност (покажува колку често моделот предвидува точни лабели севкупно), recall (колку често моделот предвидува позитивни вредности во споредба со сите реални позитивни вредности), precision (колку често моделот точно превидува позитивни вредности), и F1 score (хармониска средина на precision и recall); сите прикажани на **Слика 4-15**.

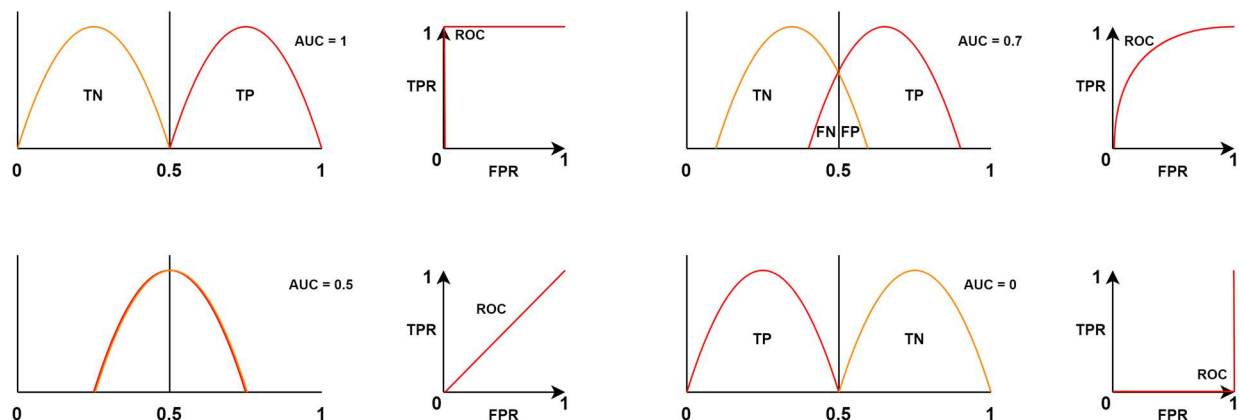
		True condition		
Total population		Condition positive	Condition negative	
Predicted condition	Predicted condition positive	True positive	False positive (Type I error)	$precision = \frac{\sum \text{true positive}}{\sum \text{predicted condition positive}}$
	Predicted condition negative	False negative (Type II error)	True negative	
		$\text{true positive rate, recall} = \frac{\sum \text{true positive}}{\sum \text{condition positive}}$	$\text{false positive rate} = \frac{\sum \text{false positive}}{\sum \text{condition negative}}$	$accuracy = \frac{\sum \text{true positive} + \sum \text{true negative}}{\sum \text{total population}}$ $F1 \text{ score} = \frac{precision * recall}{precision + recall}$

Слика 4-15 Матрица на грешка за бинарна класификација и дополнителни изведени метрики

Уште една метрика за класификациски проблеми која ги визуелизира перформансите на моделот за различни прагови на одлука е AUC (area under the curve)-ROC (receiver operating characteristic) кривата, [158], каде ROC е веројатносна крива и AUC го прикажува степенот на раздвојување помеѓу класите, и кривата се исцртува со false positive rate (FPR) на x-оската и true positive rate (TPR) на y-оската, како што може да се види на **Слика 4-16**. Вредност за AUC која изнесува 0,5 значи дека моделот не знае воопшто да разликува помеѓу класите, додека што поблиску вредноста на AUC е до 1 толку подобри се перформансите на моделот, како што е илустрирано на **Слика 4-17**.



Слика 4-16 AUC-ROC крива (ROC – веројатносна крива, AUC – степен на раздвојување меѓу класите)



Слика 4-17 Прагови на AUC-ROC крива. Секој пар фигури ја покажува матрицата на грешки и како тие вредности се пренесуваат на AUC-ROC кривата.

4.5.3 Интерпретабилност на модел

Кога моделот ќе биде истрениран, пред негова примена во реални услови, значајно е да се провери логиката која моделот ја користи за да ги направи потребните пресметки, за да се провери дека моделот донесува информирани одлуки на непристрасен начин со што е можно помала грешка, [159]. Алатки за објаснување на модели (анг. explainable AI tools), како SHapley Additive exPlanations [160] и Local Interpretable Model-agnostic Explanations [161], нудат низа процеси кои на развивачите на модели им даваат преглед на што точно се случува во моделот и како моделот го добива исходот кој го прикажува.

SHAP користи теорија на игри за да го пресмета придонесот на секоја карактеристика во исходот на моделот, т.е. важноста на секоја карактеристика за конечното предвидување. SHAP е

самостоен објаснувач на модели, кој претпоставува дека моделот е црна кутија, поради што објаснувањето кај SHAP се добива со користење само на влезните податоци и предвидувањето на моделот, т.е., објаснувачот учи да го имитира однесувањето на моделот и приближно да го направи истото предвидување за одреден пример како и оригиналниот модел. За алгоритми засновани на стебла на одлучување, како што е XGBoost, се користи објаснувач што е базиран на апроксимација на множество дрвја.

SHAP се базира на вредностите на Шепли, кои како математички концепт овозможуваат пресметување на очекуваниот маргинален придонес на поединецот во група. На пример, ако се разгледува кооперативна игра со M учесници, нумерирани од 1 до M , Шепли ќе го пресмета придонесот на секој играч во вкупната исплата. За таа цел, нека F го означува множество играчи кои се нарекуваат и голема коалиција, поточно $F = \{1, 2, \dots, M\}$, додека S ќе дефинира подмножество на F ($S \subseteq F$) наречено коалиција. Пrazно множество исто така се смета за коалиција без играчи. Бидејќи множеството на играчи содржи M учесници, постојат 2^M коалиции. Понатаму, за мапирање на коалицијата со вредност се дефинира карактеристична функција v , а за секоја коалиција S , вредноста на коалицијата – добивката која колективно ја остварува коалицијата – е претставена со $v(S)$. Празна коалиција има добивка 0.

За да се илустрира пристапот кон проблемот, ќе претпоставиме дека постои игра со 5 учесници, каде големата коалиција е $F = \{1, 2, 3, 4, 5\}$; меѓутоа, за да се формира големата коалиција играчите се додаваат на празно множество еден по еден и со додавање на секој нов играч се креира нова коалиција, а паралелно и се зголемува добивката на новата коалиција во споредба со претходната коалиција. Имено, ако првиот играч се додаде на празно множество тогаш ќе се формира коалицијата $\{1\}$ и нејзината вредност е $v(\{1\})$, па потоа се додава вториот играч на оваа коалиција по што новата коалиција станува $\{1, 2\}$ и вредноста која се добива е $v(\{1, 2\})$, и потоа ако трет играч се додаде на множеството се добива коалицијата $\{1, 2, 3\}$ и таа има добивка од $v(\{1, 2, 3\})$, итн. Гледајќи дека коалицијата и добивката на коалицијата се менуваат со секој додаден играч, придонесот на одреден играч може да се пресмета како разлика помеѓу придонесот на коалицијата откако играчот ќе биде додаден и придонесот на коалицијата пред играчот да биде додаде, па со тоа според наведените чекори, придонесот на играчот $\{3\}$ може да се пресмета како:

$$v(\{1, 2, 3\}) - v(\{1, 2\}).$$

Следно, коалицијата може да се прошири со додавање на играчите $\{4\}$ и $\{5\}$, што ќе ја смени вкупната придобивка кај новодобиената коалиција, но нема да има никакво влијание на придонесот што го дал играчот $\{3\}$. Меѓутоа, постои специфичен редослед по кој играчите се додаваат, а придонесот на играчот $\{3\}$ зависи од фактот дека учесникот $\{3\}$ е додаден по $\{2\}$, што може да влијае на придонесот на играчот $\{3\}$ ако и двата играчи имаат слични способности, па така придонесот на играчот $\{3\}$ може да биде значително повисок ако играчот $\{3\}$ бил додаден веднаш по додавање на играчот $\{1\}$; ова значи дека со цел да се добие коректна евалуација за реалниот придонес на секој играч потребно е да се разгледа и редоследот по кој играчите се додаваат во коалицијата гледајќи дека тоа може да го промени придонесот, т.е. за да се направи правична проценка на придонесот на $\{i\}$, треба да се разгледаат сите пермутации на F и крајната вредност на придонесот на $\{i\}$ да биде пресметана како средна вредност од сите индивидуални придонеси на играчот $\{i\}$ за секоја можна пермутација. За секоја пермутација P , важно е да се пресметаат две вредности: вредноста на коалицијата S пред додавање на играчот $\{i\}$ и вредноста на коалицијата формирана по додавање на играчот $\{i\}$ во S , што математички може да се означи како $S \cup \{i\}$, од каде придонесот на играчот $\{i\}$ може да се пресмета со следната формула:

$$\phi_i = \frac{1}{|F|!} \sum_P (v(S \cup \{i\}) - v(S)),$$

каде $|F|$ е бројот на елементи во големата коалиција и $|F|!$ е вкупниот број на постоечки пермутации.

Во примерот од претходно со истата голема коалиција, придонесот на играчот $\{3\}$ би имал иста вредност и во случај кога коалицијата е $\{1, 2, 3\}$ и кога е $\{2, 1, 3\}$, што покажува дека во случај на исти S и $S \cup \{i\}$ придонесот на играчот $\{i\}$ ќе биде еднаков и покрај тоа што соодветната пермутација е различна. Затоа, формулата може да се прилагоди така што ќе ги вклучува само дистинктивните придонеси и бројот на појавувања на секој од нив, што го дефинира средниот придонес на $\{i\}$ во сите пермутации F (односно вредноста на Шепли) како:

$$\phi_i = \sum_{S \subseteq F - \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} (v(S \cup \{i\}) - v(S)),$$

каде S е коалиција на F која не го вклучува играчот i , па последователно $|S|!$ е вкупниот број на сите пермутации пред играчот i да биде додаден во коалицијата, додека $(|F| - |S| - 1)!$ е бројот на пермутации по кои преостанатите членови на големата коалиција може да бидат додадени во S , по додавањето на играчот i .

Истиот процес се применува и при интерпретација на модел за МУ, така што секоја карактеристика се смета за играч. За податочно множество кое се состои од N редици и M карактеристики, може да се дефинира дека $x^{(j)}$ и $y^{(j)}$ се векторите од карактеристики и лабелите на j -тиот примерок, а вредноста на i -тата карактеристика во j -тиот примерок е $x_i^{(j)}$, додека X_i е i -тата карактеристика во целото податочно множество и учењето на моделот од податоците е дефинирано преку предиктивна функција на моделот $f(x) = f(x_1, x_2, \dots, x_M)$. Карактеристичната функција кога нема карактеристики за моделот е средна вредност од предвидувањата на моделот за k избрани примероци од податоците, или:

$$f(fna, fna, \dots, fna) = E[f(x)] = \frac{1}{k} \sum_{j=1}^k f(x^{(j)}),$$

каде fna заменува недостапни (непостоечки) карактеристики, додека карактеристичната функција за големата коалиција се пресметува со:

$$v(F) = v(\{X_1, X_2, \dots, X_M\}) = f(x) - E[f(x)] = f(x) - \frac{1}{k} \sum_{j=1}^k f(x^{(j)}).$$

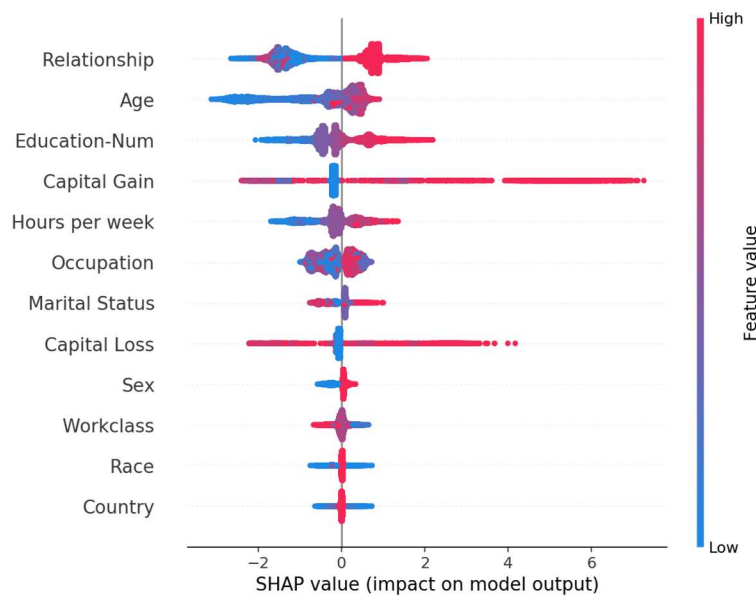
Тогаш, разгледувајќи ја коалицијата $S = \{X_{S1}, X_{S2}, \dots, X_{Sp}\}$, вредноста на f за овие карактеристики е $f_S(x_S) = f(x_{S1}, x_{S2}, \dots, x_{Sp})$, каде x_S е вектор кој ги содржи вредностите на карактеристиките присутни во S . Вредноста на f_S може да се добие со тренирање на истиот модел на новата коалиција или да се пресмета според оригиналната функција f , па вредноста на коалицијата може да се пресмета преку:

$$v(S) = v(\{x_{S1}, x_{S2}, \dots, x_{Sp}\}) = f_S(x_S) - E[f(x)],$$

од каде Шепли вредноста за карактеристиката X_i може да се пресмета како:

$$\begin{aligned} \phi_{xi} &= \sum_{S \subseteq F - \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} (f_{S \cup \{i\}}(x_{S \cup \{i\}}) - E[f(x)] - (f_S(x_S) - E[f(x)])) \\ &= \sum_{S \subseteq F - \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} (f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)). \end{aligned}$$

SHAP објаснувачот ги користи сите примероци од множеството за тренирање да го илустрира значењето на секоја карактеристика за моделот. Имено, на графикот од **Слика 4-18** прикажани се придонесите на најзначајните карактеристики за излезот на еден модел во опаѓачки редослед, така што секој примерок од податочното множество е претставено од една точка на графикот. На x-оската се прикажани SHAP вредностите на карактеристиката, така што примероците се дистрибуирани за да ја илустрираат густината на примероци за одредена SHAP вредност, додека бојата на точката ја прикажува реалната вредност на соодветната карактеристика кај секој примерок. Во конкретниов случај, повисоки вредности за години (точките се обоени во црвена боја) имаат позитивно влијание на целната променлива што се гледа од тоа што високи вредности за години може да бидат видени на позитивната страна од x-оската, додека помали вредности за години (точките се обоени во сина боја) имаат негативно влијание и можат да се најдат на негативната страна на x-оската.



Слика 4-18 SHAP график: пример [162]

5 Резултати

5.1 Пристрасност која произлегува од витални знаци, медицинска опрема и пракса

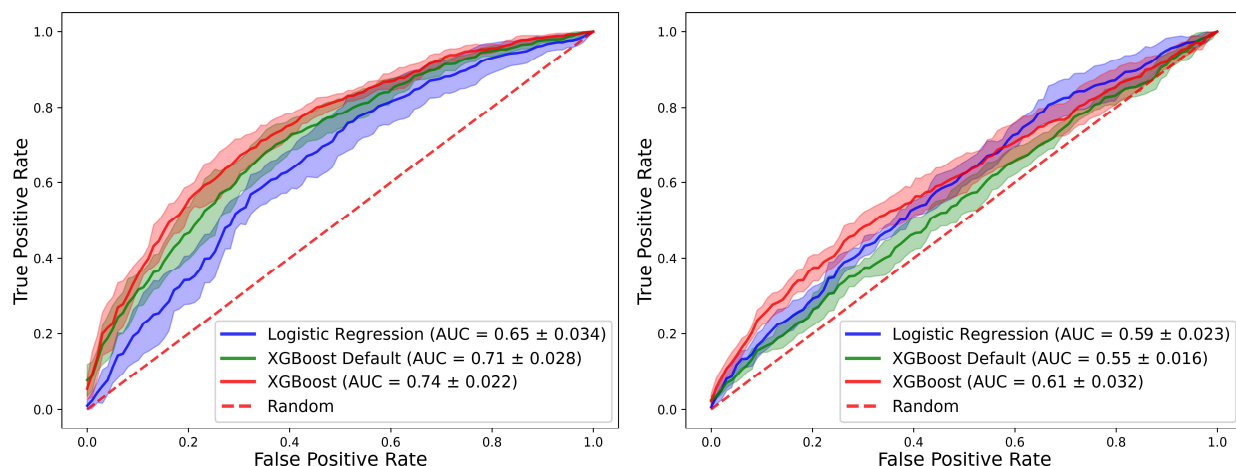
За да се постават појдовни точки за споредба за резултатите, иницијално беше истражено присуството на расна пристрасност во целокупната група на пациенти – прво, со сите шест витални знаци (означени како all), и второ, со два од шесте витални знаци (означени како basic), отчукување на срцето и респираторна стапка, како влезни карактеристики за моделот за eICU-CRD; и само со два од шесте витални знаци (исто отчукување на срцето и респираторна стапка) за MIMIC-III. Добиените резултати се детално опишани во **Табела 5-1**. Тестот за сите витални знаци не беше повторен за MIMIC-III бидејќи кај ова множество инвазивно мерени карактеристики беа исклучени за да се обезбеди големина на кохорта со која може да се работи

Табела 5-1 Резултати добиени од сите витални знаци и од основни витални знаци (отчукување на срце и респираторна стапка) за сите три кохорти

Алгоритам	Пациенти од црна и бела раса	Хиспанци и пациенти од бела раса	Хиспанци и пациенти од црна раса
eICU-CRD (all)			
LR	0,65 ± 0,034	0,63 ± 0,053	0,68 ± 0,079
XGBoost Default	0,71 ± 0,028	0,66 ± 0,071	0,71 ± 0,077
XGBoost Optimised	0,74 ± 0,022	0,68 ± 0,059	0,72 ± 0,045
eICU-CRD (basic)			
LR	0,59 ± 0,023	0,57 ± 0,041	0,64 ± 0,086
XGBoost Default	0,55 ± 0,016	0,53 ± 0,037	0,63 ± 0,031
XGBoost Optimised	0,61 ± 0,032	0,56 ± 0,048	0,69 ± 0,080
MIMIC-III (basic)			
LR	0,50 ± 0,027	0,46 ± 0,031	0,56 ± 0,052
XGBoost Default	0,49 ± 0,019	0,57 ± 0,027	0,61 ± 0,029
XGBoost Optimised	0,51 ± 0,030	0,59 ± 0,028	0,58 ± 0,026

Резултатите добиени со користење на сите витални знаци ја поставуваат горната граница за очекувањата од моделот, а резултатите добиени со користење само на отчукување на срцето и респираторна стапка како карактеристики ја поставуваат долната граница за очекувани резултати - имено, и двете анализи поставуваат информативен опсег за перформансите на моделите – и, како што може да се забележи од **Слика 5-1**, AUC кај резултатите се движи од 0,61 до 0,74. Анализата покажува дека алгоритмот XGBoost може да ја предвиди самопријавената раса на пациентите од шест рутински измерени витални знаци во група од 1.688 приеми на пациенти од црна и бела раса со перформанси од AUC 0,74 (± 0,022), во група од 784 пациенти приеми за Хиспанци и пациенти од бела раса со перформанси од AUC 0,68 (± 0,059), и во група од 444 приеми на Хиспанци и пациенти од црна раса со перформанси од AUC 0,72 (± 0,045). Алгоритмот може успешно да ја предвиди самопријавената раса на пациентите од eICU-CRD множеството дури и кога ги користи само основните витални знаци како влез, најизразено во споредбата на Хиспанците и пациентите од црна раса со AUC од 0,69 (± 0,08). Овие резултати се поткрепени и со резултатите од MIMIC-III, освен во случајот на кохорта од пациенти од црна и бела раса, каде што дури и покрај големата

група од 2.158 пациенти, добиената AUC од 0,51 сугерира дека нема забележителни разлики помеѓу пациентите од двете етничко-расни припадности.

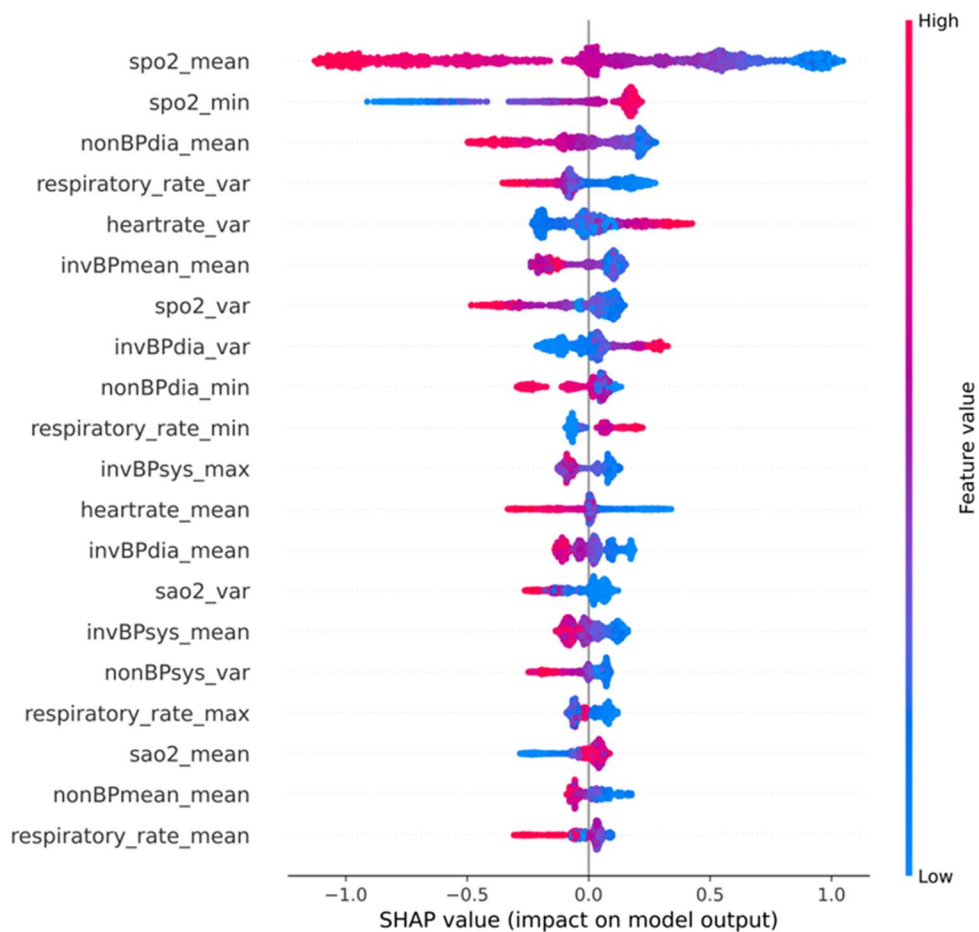


Слика 5-1 Резултати од пациенти од бела и црна раса за eICU-CRD со сите витални знаци (лево) и со основни витални знаци (десно)

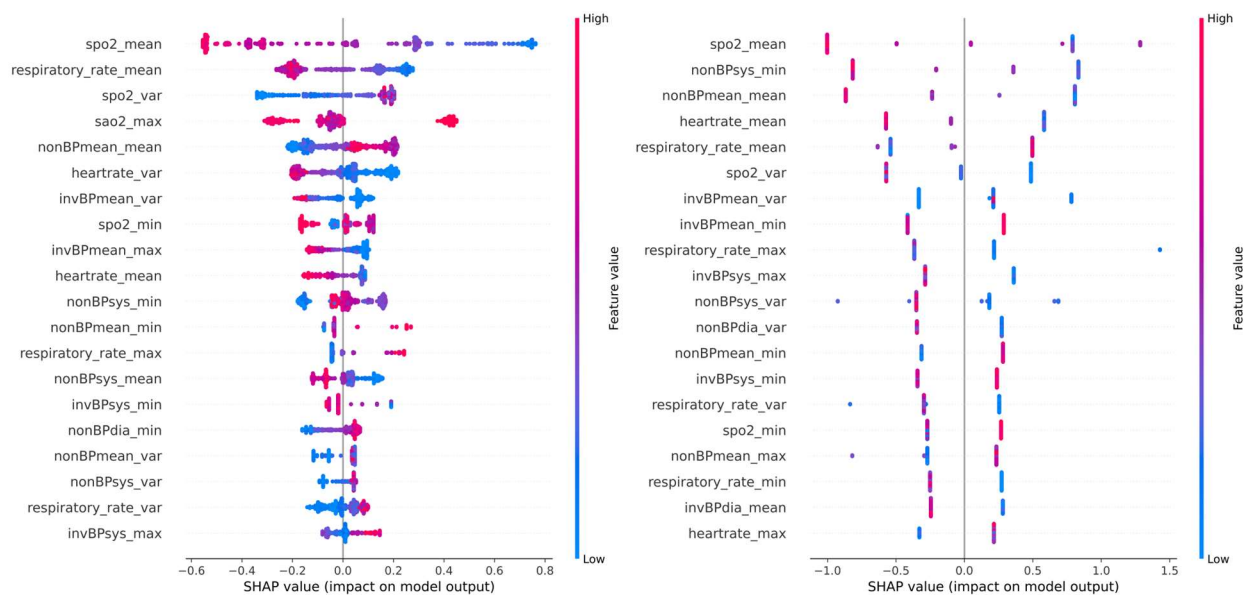
Иако има доказ за етнички разлики во кардиоваскуларниот морбидитет и морталитет, [163], [164], и при респираторните оштетувања, [165], контрадикторните резултати добиени од основните витални знаци за пациенти од црна и бела раса кај eICU-CRD и MIMIC-III множествата не водат до вистински заклучоци и бараат потемелна евалуација. Дополнително, AUC вредностите добиени за сите витални знаци бараат дополнително проучување за да се детектираат изворите на пристрасност.

За да се анализира можното потекло на расните информации, беше направена анализа на важност на карактеристики за сите тестови со помош на SHAP beeswarm график; меѓутоа, за да се добие концизност во графикот имињата на карактеристиките беа скратени, на тој начин што: неинвазивната заситеност со кислород е означена како `spo2`, инвазивната кислородна сатурација е `sao2`, инвазивниот крвен притисок го прифаќа префиксот `inv` и неинвазивниот крвен притисок го прифаќа префиксот `non`, по што следува кратенката BP која означува дека станува збор за крвен притисок и суфиксот е еден од три (`sys` за систолен, `dia` за дијастолен и `mean` за средна вредност), додека имињата за отчукување на срцето и респираторната стапка не се менуваат. Дополнителна наставка ја придружува секоја променлива за да покаже дали вредноста на карактеристиката е средна (`_mean`), минимална (`_min`), максимална (`_max`) или варијанса (`_var`).

Анализата на значајни карактеристики од eICU-CRD за кохортата од пациенти од црна и бела раса со сите витални знаци за eICU-CRD е прикажана на **Слика 5-2**, додека анализата на значајни карактеристики на Хиспанци и пациенти од бела раса и на Хиспанци и пациенти од црна раса е дадена на **Слика 5-3**. Може да се види дека средната вредност на неинвазивната заситеност со кислород е онаа со највисока средна апсолутна вредност на SHAP за сите кохорти за податочното множество, додека останатите карактеристики се разликуваат. SHAP графикот за моделот кој ги користи само основните витални знаци за пациенти од црна и бела раса, илустрира дека варијансите на отчукувања на срцето и респираторната стапка влијаат на перформансите на моделот, што исто така може индиректно да се забележи од **Слика 5-2**. Сите три графика покажуваат дека пониски вредности на неинвазивна заситеност со кислород имаат висока предиктивна вредност за моделот.



Слика 5-2 Анализа на карактеристики за кохорта од пациенти од бела и црна раса со сите витални знаци (eICU-CRD)



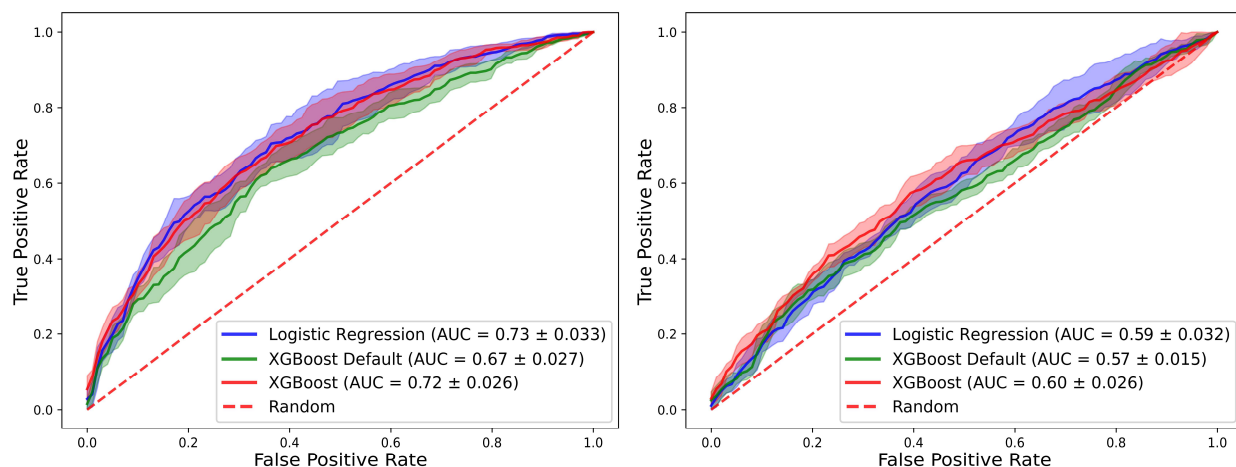
Слика 5-3 Анализа на карактеристики за кохорта од Хиспанци и пациенти од бела раса (лево) и кохорта од Хиспанци и пациенти од црна раса (десно) со сите витални знаци (eICU-CRD)

Бидејќи SHAP анализата покажа дека најзначајна карактеристика во предвидување на раса на пациенти е кислородна заситеност мерена преку пулсен оксиметар. Понатаму, анализата се фокусираше на зошто токму кислородната заситеност од пулсен оксиметар има толкав придонес за перформансите на моделот. Дополнително, имајќи ги во предвид многуте пријавени случаи на етничко-расна пристрасност во медицинската технологија и нејзините апликации во медицинска средина, [47], [48], [49], следниот логичен чекор е да се истражат разликите во медицинската технологија и опремата која се користи за мерење на витални знаци генерално како можни извори на расни информации.

Претходни истражувања покажале дека разликите во боја на кожата влијаат на мерењата направени од пулсен оксиметар, [166]. Понатаму, истражувањата откриваат значителни разлики во мерења на крвниот притисок добиени од апарат за мерење на крвен притисок и преку вградена артериска линија, [167], постојат индикации дека неинвазивните мерења покажуваат повисок крвен притисок кај хипотензивните пациенти и понизок кај хипертензивните пациенти, [168], како и извештаи за неточни отчитувања на крвен притисок со помош на апаратот за крвен притисок кај пациенти со висок индекс на телесна маса (BMI), [169]. Затоа, со цел да се процени дали овие познати разлики може да се откријат од страна на алгоритмите, извршените тестови разгледуваат потенцијални расни информации во мерењата на заситеноста со кислород со помош на пулсен оксиметар и потенцијални расни информации во мерењата на крвниот притисок со помош на апарат за мерење на крвен притисок.

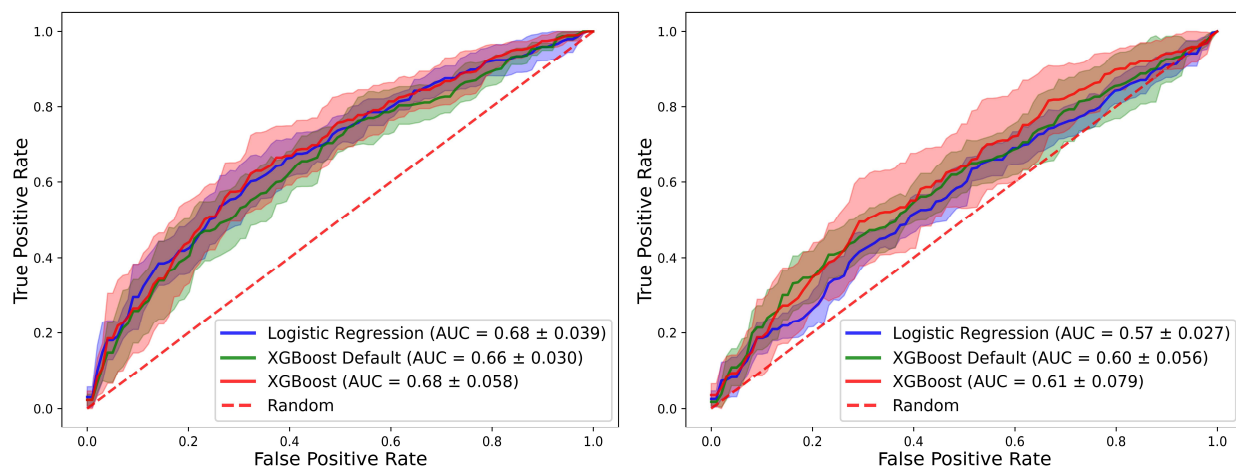
Бидејќи има доволно докази дека оксиметрите се расно пристрасни, главната цел на првиот тест е да се анализира дали оваа пристрасност постои во мерењата на заситеноста со кислород дури и ако тие се обработуваат и се агрегираат во текот на 24 часа. За да се постигне ова, анализата вклучуваше споредување на перформансите на предвидувањето на етничко-расната припадност со користење на вредностите на заситеност со кислород добиени неинвазивно од пулсна оксиметрија и оние добиени инвазивно со помош на крвна анализа. Тестовите исто така ги користеа отчукување на срце и респираторна стапка како влезни карактеристики, со додавање на неинвазивни или инвазивни мерења на заситеност со кислород за секој тест соодветно.

Резултатите за кохорта од пациенти од бела и црна раса се илустрирани на *Слика 5-4*, со резултати за неинвазивна сатурација лево и инвазивна сатурација на десната страна. Наодите покажуваат дека пулсната оксиметрија е значаен извор на расни информации со AUC од 0,72 ($\pm$ 0,026), во целосен контраст со резултатите од инвазивно измерената заситеност со кислород со предвидувања чија метрика е AUC од 0,60 ($\pm$ 0,026). Ова ги поддржува претходните наоди дека пулсната оксиметрија е зависна од боја на кожата, како во [170]; сепак, авторите на [170] ги групираат Хиспанците и пациентите од црна раса во една група кога ги споредуваат разликите во заситеноста со кислород, додека овие резултати нудат споредба само за пациенти од црна и бела раса.

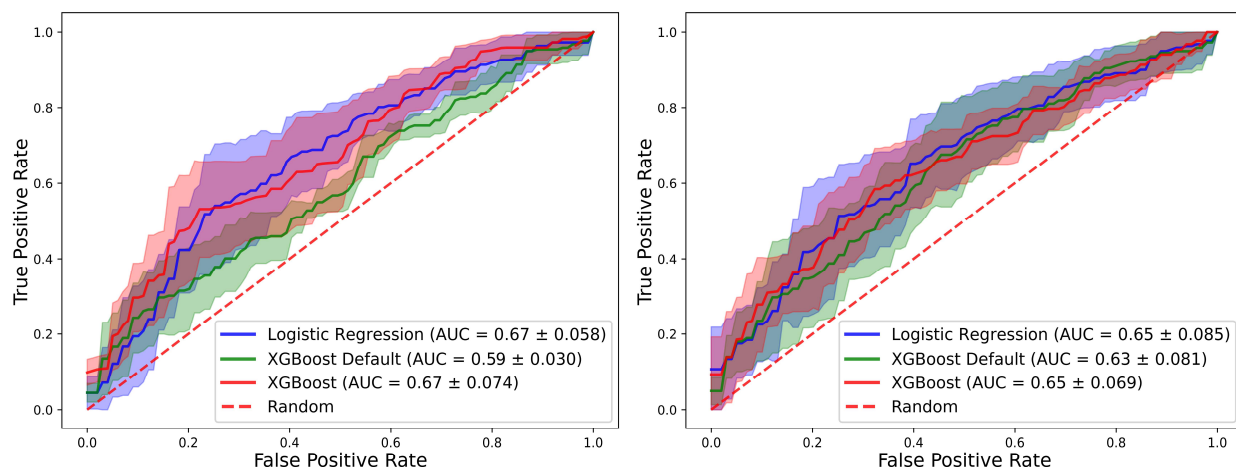


Слика 5-4 Резултати од пациенти од бела и црна раса за eICU-CRD со неинвазивен кислород (лево) и со инвазивен кислорот (десно)

За дополнително да се потврдат наодите на [170], но истовремено да се набљудуваат потенцијалните разлики помеѓу Хиспанците и пациентите од црна раса, истата анализа беше повторена за двете други кохорти, прво, Хиспанци и пациенти од бела раса (на **Слика 5-5**) со AUC од 0,68 ($\pm 0,058$) за неинвазивни и AUC од 0,61 ($\pm 0,079$) за инвазивни мерења на заситеност со кислород, и второ, Хиспанци и пациенти од црна раса (на **Слика 5-6**) со AUC од 0,67 ($\pm 0,074$) за неинвазивни и AUC од 0,65 ($\pm 0,069$) за инвазивни мерења. Имено, овие резултати покажуваат дека пристрасноста во мерења на неинвазивна заситеност со кислород може да се најде и кога се споредуваат Хиспанци и пациенти бела раса, бидејќи AUC се намалува од 0,68 на 0,61 кога неинвазивните мерења на заситеноста со кислород се исклучени од анализата. Сепак, падот на предвидувањата за AUC за Хиспанци и пациенти од црна раса не е толку изразен (0,67 до 0,65), што сугерира дека неинвазивните мерења на заситеноста со кислород не носат значајна разлика помеѓу двете етничко-расни групи, како што дополнително е илустрирано со анализата на карактеристики дадена во **Слика 5-7**, при што најзначајните три витални знаци од влезните параметри на моделот и во двата случаи покажуваат слично однесување и не вклучуваат никакви мерења на заситеност со кислород.

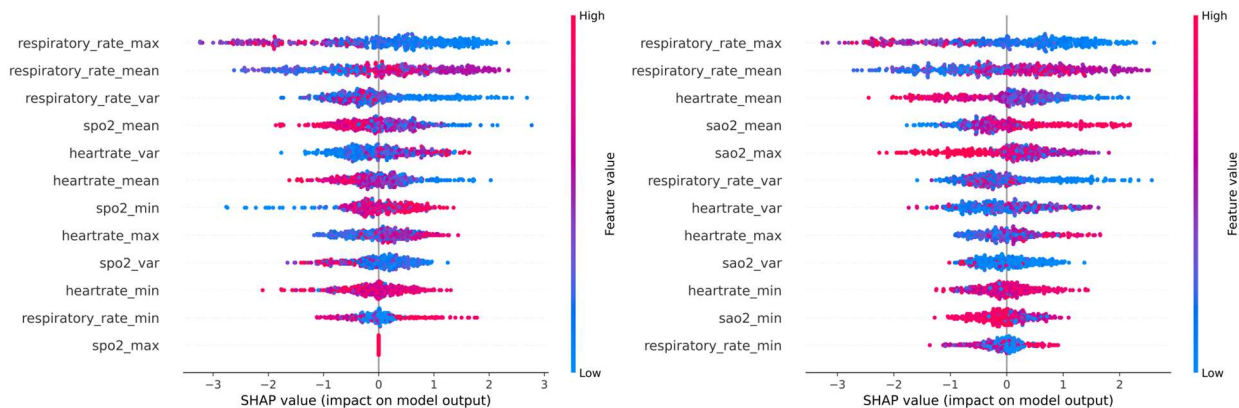


Слика 5-5 Резултати од Хиспанци и пациенти од бела раса за eICU-CRD со неинвазивен кислород (лево) и со инвазивен кислорот (десно)



Слика 5-6 Резултати од Хиспанци и пациенти од црна раса за eICU-CRD со неинвазивен кислород (лево) и со инвазивен кислорот (десно)

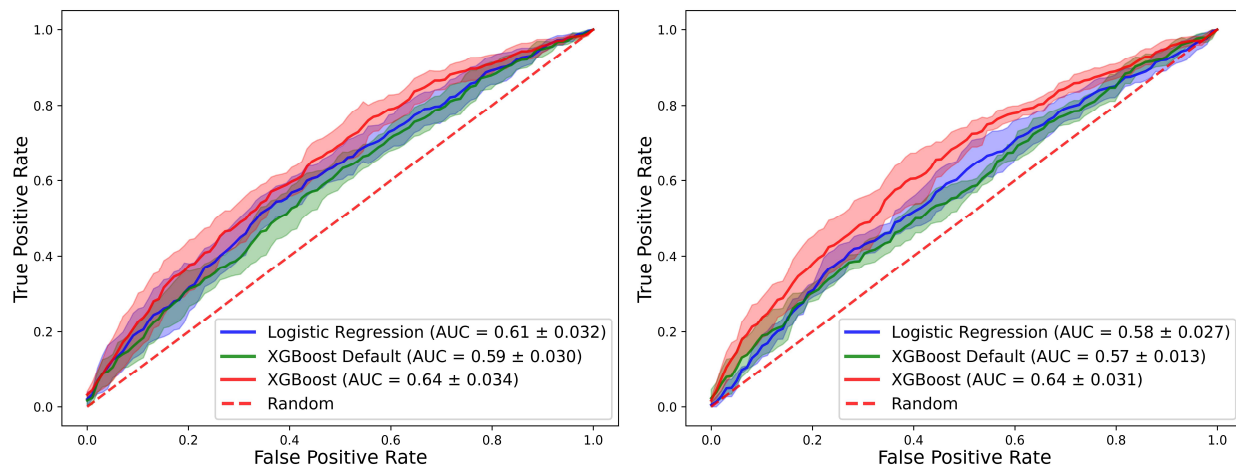
Како што е прикажано на **Слика 5-7** при тренирање за детекција на разлики помеѓу Хиспанци и пациенти од црна раса, двете одлучувачки карактеристики се средната и максималната вредност на респираторната стапка на пациентите. Иако се чини дека ова е делумно согласност со истражувањето кое наведува дека „други етничко-расни групи“ имаат помал капацитет на белите дробови од оние на пациенти од бела раса [171], ова се сепак значително неиздржани изјави. Сепак, презентираниите резултати може да произлезат од разликите во висината, [172], социоекономскиот статус, [173], или може да бидат делумно поврзани со откривањето на етничко-расна припадност од рендгенски снимки на граден кош, [70]. Резултатите за неинвазивни мерења исто така беа проверени со користење на MIMIC-III, а добиените вредности ги потврдија заклучоците добиени од eICU-CRD.



Слика 5-7 Анализа на карактеристики за кохорта од Хиспанци и пациенти од црна раса за eICU-CRD со неинвазивен кислород (лево) и со инвазивен кислорот (десно)

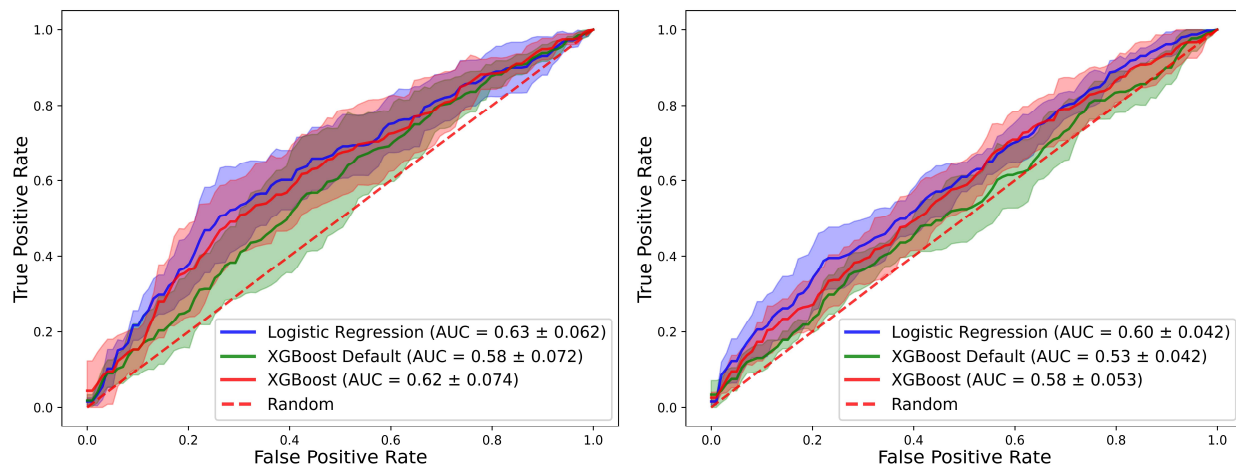
Литературата исто така покажува и неточности во технологијата за мерење на крвниот притисок. Со цел да се утврди дали ова важи и по агрегација на податоците од првите 24 часа, анализата се фокусираше на споредување на перформансите за идентификација на самопријавената етничко-расна припадност користејќи вредности на крвниот притисок измерени преку апарат за крвен притисок (неинвазивно мерење) со онаа при користење на артериска линија (инвазивно), како и резултатите добиени со користење на двете мерења заедно. Резултатите за пациентите од црна и бела раса се прикажани на **Слика 5-8**, со AUC од 0,64 ($\pm 0,034$) за неинвазивни мерења и AUC од 0,64 ($\pm 0,031$) за инвазивни мерења, без особени разлики кога тренирањето е спроведено и за неинвазивни и инвазивни мерења заедно со AUC еднаква на 0,64 ($\pm 0,029$). Постои мала до

непостоечка разлика помеѓу двете ROC криви за пациентите од црна и бела раса (главно, мала разлика во стандардната девијација), што покажува дека спротивно на претходни истражувања кои покажале технолошка разлика при мерење на крвен притисок во оваа споредба таа разлика не може јасно да се забележи. Затоа, за разлика од постоечката литература, резултатите од пациентите од црна и бела раса покажуваат дека иако има присуство на расни информации во мерењата на крвниот притисок, без оглед на користениот метод за мерење (AUC од $0,64 \pm 0,029$), нема големи разлики меѓу вредностите мерени со апарат за мерење на притисок и оние измерени преку артериска линија.

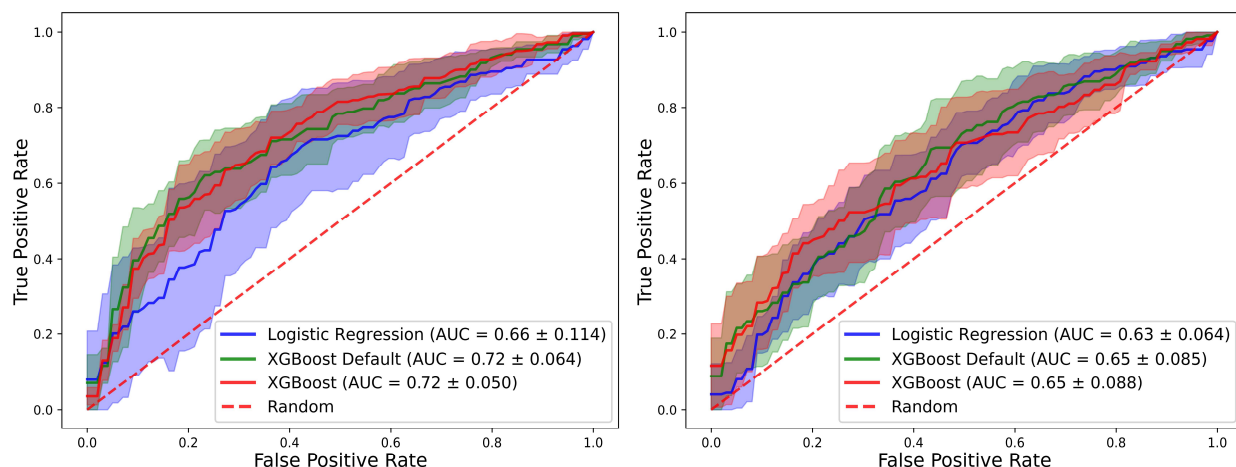


Слика 5-8 Резултати од пациенти од бела и црна раса за eICU-CRD со неинвазивен крвен притисок (лево) и со инвазивен крвен притисок (десно)

Резултатите добиени од споредбата на опремата за мерење на крвен притисок за Хиспанците и пациентите од бела раса се прикажани на **Слика 5-9**, со AUC од $0,62 (\pm 0,074)$ за неинвазивен крвен притисок и AUC од $0,58 (\pm 0,053)$ за инвазивен крвен притисок. Резултатите од Хиспанци и пациенти од црна раса се дадени на **Слика 5-10** и покажуваат дека неинвазивните мерења на крвниот притисок постигнуваат AUC од $0,72 (\pm 0,050)$, а со инвазивните мерења се добива AUC од $0,65 (\pm 0,088)$ – т.е., може да се види забележителна разлика помеѓу двете вредности. Иако способноста за предвидување на припадност во кохорта од Хиспанци и пациенти од бела раса е најниската од сите три споредби, видлива е мала разлика во предвидувањето помеѓу вредностите добиени од двата типа на мерења. Разликата е поизразена кај Хиспанците и пациентите од црна раса. Со ова, двата теста се усогласени со очекуваните мерни разлики помеѓу апаратот за крвен притисок и артериска линија.

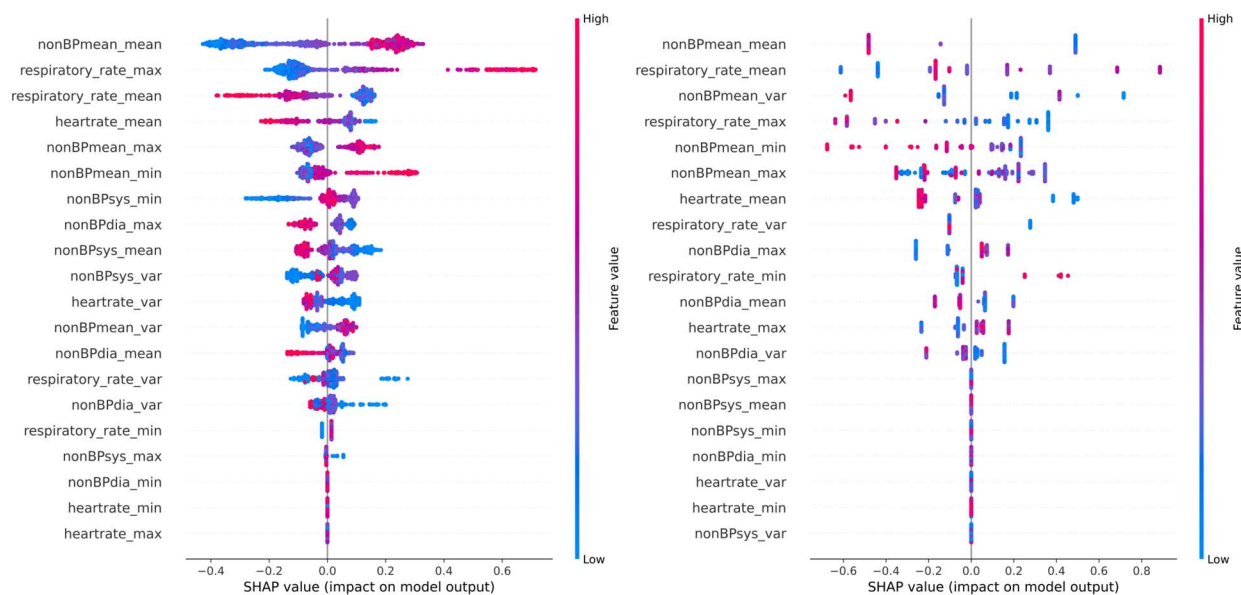


Слика 5-9 Резултати од Хиспанци и пациенти од бела раса за eICU-CRD со неинвазивен крвен притисок (лево) и со инвазивен крвен притисок (десно)



Слика 5-10 Резултати од Хиспанци и пациенти од црна раса за eICU-CRD со неинвазивен крвен притисок (лево) и со инвазивен крвен притисок (десно)

За дополнително да се потврди влијанието на неинвазивните мерења на крвниот притисок, беше извршена SHAP анализа на двете групи: Хиспанци и пациенти од бела раса, Хиспанци и пациенти од црна раса; графициите се дадени на **Слика 5-11**. Двата графика ја илустрираат средната вредност на неинвазивните мерења на крвниот притисок како најважна карактеристика. Понатаму, во кохортата од Хиспанци и пациенти од црна раса, каде што промената на AUC е поизразена, може да се види дека статистичките карактеристики за средниот крвен притисок се повисоки на листата на значајни карактеристики во споредба со кохортата од Хиспанци и пациенти од бела раса.

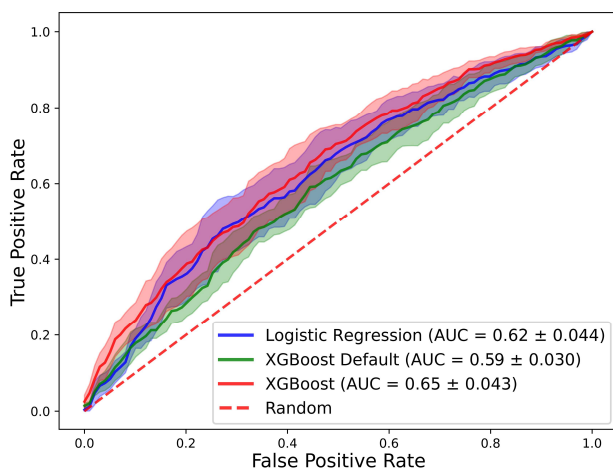


Слика 5-11 Анализа на карактеристики за неинвазивен крвен притисок за двете кохорти: Хиспанци и пациенти од бела раса (лево) и Хиспанци и пациенти од црна раса (десно)

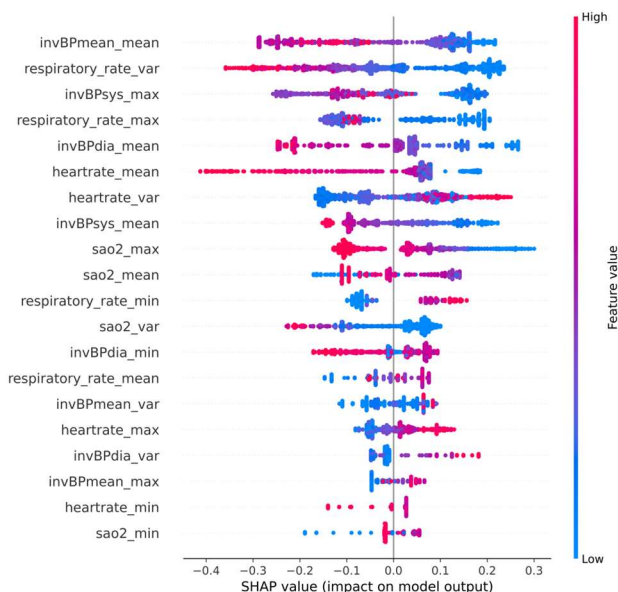
Тестовите за кохортите добиени од MIMIC-III множеството направени за неинвазивни мерења на крвниот притисок – што беше единствено можно поради процесот на селекција на кохорти за MIMIC-III – покажаа резултати споредливи со резултатите од eICU-CRD; имено, кохортата со пациенти од црна и бела раса кај моделот оствари предвидувања со AUC од 0,62 каде карактеристиките на крвниот притисок доминираа на врвот на SHAP анализата, Хиспанците и

пациентите од бела раса имаа најниска AUC, а пак кај Хиспанците и пациентите од црна раса имаше AUC од 0,61.

Бидејќи резултатите и анализата на важност на карактеристики ја покажаа технологијата за мерење на витални знаци како извор на расна пристрасност, потребна беше дополнителна анализа - каде што овие познати фактори се контролираат. Од таа причина, беше неопходно да се добие нова колекција на резултати, т.е., да се истражи дали самопријавената етничко-расна припадност може да се научи од виталните знаци откако технологијата за мерење веќе не влијаеше директно на процесот на учење. Од таа причина, беше спроведена анализа на само инвазивно измерени витални знаци на сите три кохорти. Перформансите на моделите за пациентите од бела и црна раса постигнаа AUC од 0,65 ($\pm 0,043$), како што е илустрирано на **Слика 5-12**. Анализата на важност на карактеристики за оваа кохорта е прикажана на **Слика 5-13**, и како што може да се види од сликата, средниот инвазивен крвен притисок е најважната карактеристика за оптимизираниот модел XGBoost, на врвот со статистичките информации добиени од респираторната стапка.



Слика 5-12 Резултати од eICU-CRD кохорта од пациенти од црна и бела раса со само неинвазивни мерења



Слика 5-13 Анализа на карактеристики од eICU-CRD кохорта од пациенти од црна и бела раса со само неинвазивни мерења

The figure displays two ROC curves comparing the performance of three models: Logistic Regression, XGBoost Default, and XGBoost. The y-axis represents the True Positive Rate, and the x-axis represents the False Positive Rate, both ranging from 0.0 to 1.0. A dashed red line indicates the performance of a random classifier.

Left Plot (Train Data):

- Logistic Regression (AUC = 0.59 ± 0.041)
- XGBoost Default (AUC = 0.56 ± 0.048)
- XGBoost (AUC = 0.58 ± 0.057)
- Random (AUC = 0.5)

Right Plot (Test Data):

- Logistic Regression (AUC = 0.65 ± 0.061)
- XGBoost Default (AUC = 0.65 ± 0.068)
- XGBoost (AUC = 0.69 ± 0.111)
- Random (AUC = 0.5)

Анализата на важност на карактеристики за Хиспанци и пациенти од бела раса и Хиспанци и пациенти од црна раса е дадена на *Слика 5-15*, каде може да се забележат различни важни карактеристики во споредба со SHAP графикот добиен за пациенти од бела и црна раса. Додека респираторна стапка е скоро секогаш блиску до врвот на важност на листата од карактеристики, точниот извор на пристрасност во резултатите и природата на пристрасноста откриена кај нив не може да се објаснат на јасен и директен начин.

Figure 1 displays two SHAP summary plots, one for each model, showing the impact of various features on the model output. The x-axis represents the SHAP value (impact on model output), ranging from -0.6 to 1.0 for the left plot and -1.00 to 0.75 for the right plot. The y-axis lists the features. The color scale indicates the feature value, ranging from Low (blue) to High (red).

Left Plot (Model 1):

- Features: sao2_max, respiratory_rate_mean, respiratory_rate_max, sao2_mean, heartrate_var, invBPmean_max, invBPsys_var, invBPsys_min, invBPsys_max, invBPdia_max, invBPmean_mean, respiratory_rate_var, heartrate_mean, sao2_var, heartrate_max, invBPdia_var, invBPsys_mean, invBPmean_var, respiratory_rate_min, sao2_min.
- SHAP value range: -0.6 to 1.0.

Right Plot (Model 2):

- Features: respiratory_rate_max, respiratory_rate_mean, heartrate_mean, sao2_var, invBPmean_max, invBPsys_min, respiratory_rate_var, invBPsys_mean, sao2_max, invBPsys_var, heartrate_var, invBPmean_min, sao2_min, invBPdia_min, invBPsys_max, respiratory_rate_min, invBPdia_var, invBPdia_mean, invBPmean_var, heartrate_min.
- SHAP value range: -1.00 to 0.75.

74

Откако се покажа дека технологијата на мерење е пристрасна, и, следствено, одредени мерења на виталните знаци се неверодостојни, дополнително потребно е да се истражи дали постои и пристрасност во редовноста и фреквенцијата со која се направени тие мерења. Имено, идејата беше да се анализира дали постои пристрасност во медицинската пракса врз основа на времето на мерење на виталните знаци. За таа цел беа земени предвид истите приеми од оригиналната кохорта од 1.688 пациенти на од црна и бела раса, а во тестирањето беа вклучени и веќе пресметаните статистички карактеристики од виталните знаци. Сепак, извлечени беа и дополнителни карактеристики, т.е., податоците беа подложени на дополнително инженерство на карактеристики. Процесот на инженеринг на карактеристики го зеде предвид поместувањето (временско) за тоа кога било направено секое мерење, што значи дека секое поединечно мерење во точно одреден временски момент е посебно разгледано и сите мерења се набљудувани за промени низ времето. Ова овозможи следење на виталните знаци и времето на мерење во текот на целиот престој во ЕИН.

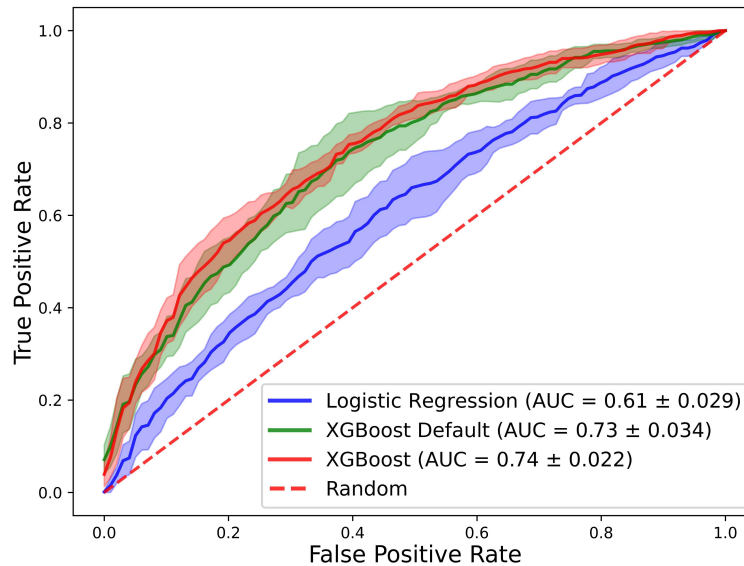
Иако постојат библиотеки кои обработуваат временски серии и екстраполираат важни информации (од кои некои дури може да изгледаат неостварливи и воопшто не се карактеристика што луѓето би сметале дека е потребна за пресметка, колку и да се квалитативни), како што е `tsfresh`, [174], за целите на оваа анализа сите карактеристики беа дискутирани, одредени и рачно пресметани без помош од алатките за инженерство на карактеристики.

При пресметка на карактеристиките фокусот беше на набљудување само на првите 24 часа од приемот, и бидејќи секое мерење беше придружено со временско поместување од времето на прием на пациентите, екстракцијата за секоја поединечна карактеристика (и во загради лабелата што се користи во графикот од SHAP анализата) вклучуваше:

- вкупен број на мерења во текот на првите 24 часа (`_total`),
- вкупно мерења во текот на првите шест часа (`_measurements_six_hours`),
- бројот на мерења на секој час за првите шест часа (`_measurements_hour_[1-6]`),
- максималниот број на мерења во прозорец од еден час во првите 24 часа (`_maximum_in_24_hours`) и
- поместувањето кога виталниот знак првпат бил измерен (`_first_feature_measurement`).

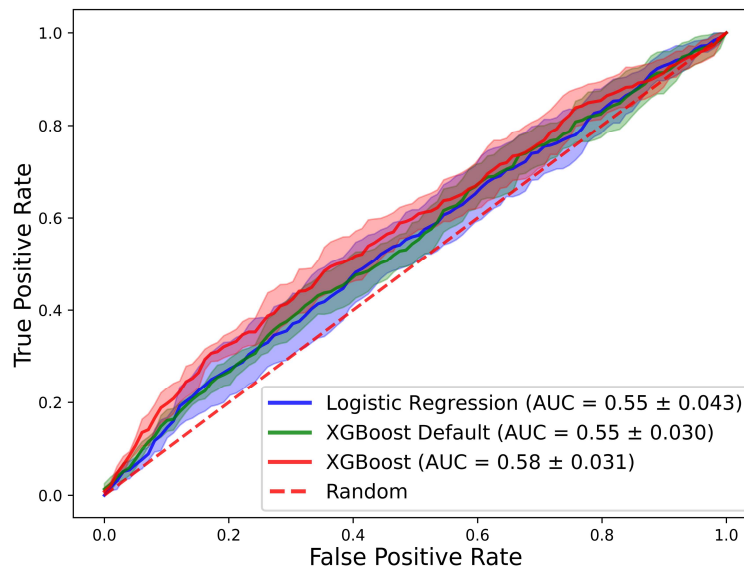
Образложението зад карактеристиките е да се процени дали има разлики во следењето на пациентите или дали потенцијалното време на првичниот одговор од страна на медицинските лица кон пациентите и фреквенцијата на мерењата варираат за различни етнички групи.

Беа извршени два различни експерименти со извлечените карактеристики: еден, каде што извлечените карактеристики за фреквенција на мерења беа користени во комбинација со статистичките карактеристики од виталните знаци, и другиот, каде што само извлечените карактеристики за фреквенција на мерења беа искористени за обука на моделот. И двата експерименти вклучуваа тренирање на истите модели, LR и XGBoost, на кохортата од 1.688 пациенти од црна и бела раса и беа изведени со користење на стратификувана петкратна вкрстена валидација и евалуирани со помош на AUC-ROC кривата. AUC-ROC кривата за комбинација на новите карактеристики и постоечките карактеристики од витални знаци е дадена на **Слика 5-16**, додека кривата за анализа на само новите карактеристики за фреквентност на мерења на виталните знаци е дадена на **Слика 5-17**.

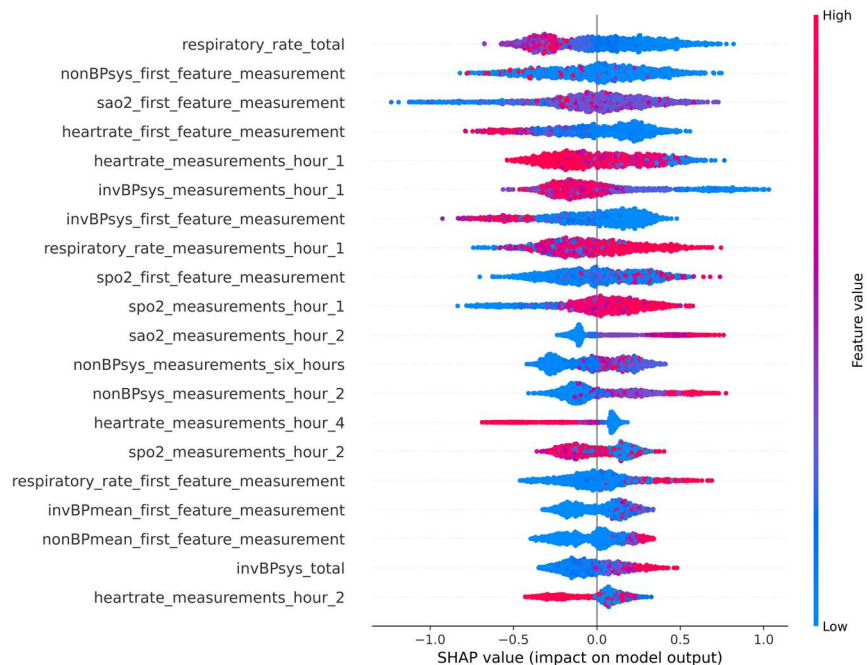


Слика 5-16 Резултати за кохортата од пациенти од црна и бела раса (eICU-CRD) од витални знаци и фреквенции на мерења комбинирано

Како што може да се забележи од добиените резултати, овој пристап покажува мала, скоро незабележлива пристрасност во практиките при испорака на нега, бидејќи пресметаните карактеристики не влијаат значително на предвидувањето на расата на пациентите, со оглед на AUC од $0,58 (\pm 0,031)$. Дополнително, беше извршена анализа на важност на карактеристики, дадена на **Слика 5-18**, и како што покажува сликата, повеќето од најважните карактеристики се првите временски поместувања од прием на пациенти за мерења на секоја карактеристика.



Слика 5-17 Резултати за кохортата од пациенти од црна и бела раса (eICU-CRD) од фреквенции на мерења само



Слика 5-18 Анализа на карактеристики за кохорта од пациенти од црна и бела раса (eICU-CRD) од фреквенции на мерења само

Овие резултати не се во согласност со иницијалните очекувања, особено имајќи ги предвид многубројните случаи на пристрасност во медицинската пракса, [83]. Сепак, по реevaluација и дополнително разгледување, резултатите требаше да бидат очекувани бидејќи виталните знаци се мерат многу порутински од други променливи. Понатаму, некои од разгледуваните витални знаци, поконкретно периодичните витални знаци, кои првично биле собрани во интервали од 1 минута, биле архивирани како медијана од 5 минути, [90]. Згора на тоа, процесот на селекција на пациенти за кохорта беше исклучително специфичен поради барањата за мерењата на виталните знаци што пациентот треба да ги има за да се квалификува за студијата. Така, процесот на селекција на кохорта создаде студиска популација на екстремни случаи на критично болни пациенти, бидејќи бараше диференцијација помеѓу артериската крвна линија и апарат за крвен притисок, и затоа вклучуваше пациенти со вметната артериска крвна линија која е применлива само за пациенти во тешки состојби, и како такво, времето за интервенција по пристигнувањето во ЕИИ е моментално или краткотрајно, веројатно намалувајќи ја пристрасноста во процесот на прием што можеше да биде детектирана овде доколку околностите и состојбата на пациентите дозволување отстапувања. Затоа, ова може да бидат дел од причините зошто ако навистина постои пристрасност во фреквенцијата на мерењата во базата на податоци, од избраните карактеристики и во избраната кохорта, би било тешко алгоритмите да ја утврдат таа пристрасност. Од таа причина, во иднина од интерес може да биде примена на пристапот на болнички пациенти, каде нема константна примена на процедури за итни случаи.

Експериментите исто така беше повторени со додавање на температура како витален знак. Понатаму, беа тренирани и евалуирани невронски мрежи за ДУ, и додека тие резултати покажаа присуство на пристрасност, немаше значително подобрување во споредба со класичното МУ, [139]. Додека додавањето на новата карактеристика го намали бројот на пациенти во кохортата, предвидувањата на моделите и за MIMIC-III и за eICU-CRD беа споредливи со резултатите добиени од поголемите кохортни групи.

5.2 Разрешавање на потенцијални збунувачки променливи

Претходни резултати покажаа дека пристрасност постои во технологија и во опремата која се користи за мерење на кислородна заситеност и крвен притисок. Сепак, со оглед на користениот процес на спарување на пациентите, постои потенцијал за псевдо влијание во резултатите кое потекнува од статистичките разлики во општите информации на пациентите во секоја кохорта. Како што се информациите за карактеристиките на пациентите од **Табела 4-12** покажуваат, постои статистичка разлика во возраста помеѓу етничко-расните припадности во секоја кохорта, како што е потврдено со p -вредност од $< 0,05$, и иако има мала разлика во бројот на пациенти по различен пол, сепак е можно општите карактеристики што се користат во спарување на пациенти во кохортата (возраст, пол и дијагноза при прием) да влијаат на резултатите и да предизвикаат значително влијание врз квалитетот на предвидувањето на расата, [137]. Затоа, за да се исклучи можноста овие потенцијално збунувачки променливи да се одлучувачки фактор за перформансите на моделите, неопходно беше да се разгледа присуството на возраст, пол и дијагноза на прием во целокупните кохорти и да се направи обид за ублажување на потенцијалните влијанија. За да се потврди дека збунувачките променливи немаат влијание, беа направени неколку експерименти.

За да се воспостави дека возраста и полот ненамерно не влијаат на перформансите на моделите, беше создаден дополнителен експеримент: каде, користејќи различен процес на спарување на пациенти (што бара спарување на еднаква возраст и пол) беше добиена нова кохорта за секоја етничко-расна споредба (каде што сите пациенти од едната етнички-расна група имале еднаква возраст и пол со нивните парови од другата), а процесот на МУ (оптимизација на параметри и тренирање) беше повторен за тие пациенти. Не постоеше значајна статистичка разлика ниту во пол, ниту во возраст, а добиените резултати се дадени во **Табела 5-2**. Дополнително, во табелата за сите споредби се дадени и резултатите од оригиналната кохорта (копија од **Табела 5-1**). Покрај тоа што стандардната девијација е зголемена кај одредени кохорти и алгоритми, сепак може да се забележи дека во резултатите кај новата кохорта постојат подобрувања на перформансите за повеќето модели. Со тоа, резултатите не покажуваат загуба во способноста за предвидување на моделите за сите три кохорти, што потврдува дека моделите не ја учеле ни возраста ни полот на пациентите.

Табела 5-2 Резултати од eICU-CRD за сите кохорти со сите карактеристики за оригиналната кохорта и нова кохорта реализирана со возраст и пол за проверка нивно влијание како збунувачки променливи

Алгоритам	Пациенти од црна и бела раса	Хиспанци и пациенти од бела раса	Хиспанци и пациенти од црна раса
eICU-CRD (оригинални кохорти)			
LR	$0,65 \pm 0,034$	$0,63 \pm 0,053$	$0,68 \pm 0,079$
XGBoost Default	$0,71 \pm 0,028$	$0,66 \pm 0,071$	$0,71 \pm 0,077$
XGBoost Optimised	$0,74 \pm 0,022$	$0,68 \pm 0,059$	$0,72 \pm 0,045$
eICU-CRD (рекреирани кохорти со користење на возраст и пол)			
LR	$0,63 \pm 0,104$	$0,84 \pm 0,074$	$0,70 \pm 0,068$
XGBoost Default	$0,74 \pm 0,034$	$0,83 \pm 0,085$	$0,74 \pm 0,075$
XGBoost Optimised	$0,75 \pm 0,031$	$0,84 \pm 0,075$	$0,74 \pm 0,064$

Бидејќи еднаква дијагноза при прием беше услов во текот на процесот на спарување на пациенти, според **Слика 4-10**, разликата во дијагнозата најверојатно не беше причина за прецизноста на моделот. Сепак, за дополнително да се потврди ова гледиште, и дополнително да се потврди недостатокот на пристрасност на полот и возраста, беа креирани три дополнителни

кохортни анализи, [138]. Имено по една за секоја од трите карактеристики на интерес, возраст, пол и дијагноза при прием, каде што секоја кохорта беше оформена врз база на една приоритетна карактеристика од трите наведени за совпаѓање. Секој процес на совпаѓање се обидувахе пациентите да ги спари така што пациент од малцинската етничко-расна група да биде споен со пациент од позастапената група, првично и доколку е можно, за сите три карактеристики, а потоа останатите пациенти беа усогласени врз основа на приоритетната карактеристика. Првиот процес на совпаѓање даде приоритет на дијагнозата, па според тоа совпаѓањето помеѓу двете етничко-расни групи следеше три фази: пациентите беа усогласени по сите три карактеристики, потоа ако не беше направено совпаѓање, пациентите беа усогласени на комбинацијата дијагноза-род или дијагноза-возраст, и на крај, ако не се најде совпаѓање во втората фаза, совпаѓањето беше направено само на дијагноза. Вториот процес на совпаѓање даде приоритет на полот, затоа совпаѓањето помеѓу двете етничко-расни групи следеше три фази: пациентите беа усогласени на три карактеристики, потоа ако не беше направено совпаѓање, пациентите беа усогласени со комбинација на пол-возраст или пол-дијагноза, и на крај, ако не се најде совпаѓање во втората фаза, совпаѓањето беше направено само на пол. Третиот процес на совпаѓање даде приоритет на возраста, затоа совпаѓањето помеѓу двете етничко-расни групи следеше три фази: пациентите беа усогласени на три карактеристики, а потоа ако не беше направено совпаѓање, пациентите беа усогласени според комбинацијата на возраст-дијагноза или возраст-пол, и на крај, доколку не се најде совпаѓање во втората фаза, совпаѓањето беше направено само на возраст. Во сите три стратегии за совпаѓање, ако дел од пациентите не беа спарени и останаа до крајот на трите соодветни чекори за спарување, тогаш тие пациенти беа исклучени од студијата. Процесот резултираше со три различни кохорти, при што секоја од нив беше основа за проверка на една од трите карактеристики како збунувачка променлива.

Резултатите за eICU-CRD и MIMIC-III, кои користеа отчукување на срце и респираторна стапка, SpO₂ и крвен притисок, се дадени во **Табела 5-3**. Може да се забележи дека и за двете податочни множества вредностите на AUC се стабилни во сите кохорти и за сите алгоритми. Алгоритмите обучени на MIMIC-III даваат полоши резултати на AUC во споредба со метриката што се добива кај eICU-CRD, но тие се должат на значајните разлики во големината на кохортите (за пациенти од црна и бела раса, за Хиспанците и пациентите од бела раса групите претрпуваат тројно намалување, за Хиспанците и пациентите од црна раса групите имаат четирикратно намалување). Меѓутоа, ако податоците се набљудуваат за совпаѓања за секоја од кохортите, според вредностите означени со задебелени букви во **Табела 5-3**, постоечките разлики се скоро незначителни. Двете анализи покажуваат дека набљудувањата направени во [137] не се валидни за ова истражување. Затоа, добиените резултати не се засноваат само на демографски информации, туку се засноваат на виталните знаци и технологиите за мерење (пулсна оксиметрија наспроти анализа на крв, апарат за крвен притисок наспроти артериска линија) користени за добивање на виталните знаци.

Дијагнозата при прием е од значителен интерес бидејќи тријажата на пациентите во ЕИН и во оддели за итни случаи се заснова на нивните витални знаци бидејќи му кажува на медицинското лице за степенот на нарушување што се случува, [175], и различна дијагноза може да влијае на виталните знаци на одреден начин. Затоа, неопходно е да се набљудува дали перформансите на моделот се менуваат за специфични дијагнози при прием. Анализата беше направена само на eICU-CRD и на кохортата од пациентите од црна и бела раса бидејќи другите две кохорти имаа мал број на пациенти за специфичните барања на пристапот. Притоа, од сите дијагнози присутни во податоците, изборот го опфати пациентите со четири највисоки дијагнози во кохортата од пациенти од црна и бела раса (есенцијална хипертензија, срцева слабост, акутна бубрежна инсуфициенција и хронична бубрежна болест) и дополнителна дијагноза (сепса) како што е предложено од страна на медицински експерт и оправдано преку специфичноста на состојбата сепса. Имено, секој експеримент се фокусираше на откривање на пристрасност во под-кохорта од оригиналната кохорта, при што секоја под-кохорта се состоеше од пациенти со соодветната од петте дијагнози.

Табела 5-3 Анализа на збунувачки променливи: Спарување на пациенти во кохорти со еднаква дијагноза при прием, пол и возраст

Спарување по	Алгоритам	Пациенти од црна и бела раса	Хиспанци и пациенти од бела раса	Хиспанци и пациенти од црна раса
eICU-CRD				
Дијагноза при прием	LR	0,62 ± 0,02	0,6 ± 0,05	0,66 ± 0,061
	XGBoost Default	0,7 ± 0,02	0,62 ± 0,03	0,67 ± 0,038
	XGBoost Optimised	0,75 ± 0,023	0,66 ± 0,038	0,69 ± 0,043
Пол	LR	0,64 ± 0,016	0,63 ± 0,056	0,70 ± 0,051
	XGBoost Default	0,71 ± 0,023	0,64 ± 0,04	0,69 ± 0,041
	XGBoost Optimised	0,75 ± 0,019	0,67 ± 0,047	0,72 ± 0,045
Возраст	LR	0,62 ± 0,025	0,61 ± 0,053	0,66 ± 0,048
	XGBoost Default	0,71 ± 0,018	0,61 ± 0,038	0,67 ± 0,025
	XGBoost Optimised	0,75 ± 0,022	0,65 ± 0,056	0,71 ± 0,044
MIMIC-III				
Дијагноза при прием	LR	0,58 ± 0,04	0,59 ± 0,042	0,63 ± 0,039
	XGBoost Default	0,62 ± 0,018	0,55 ± 0,044	0,56 ± 0,051
	XGBoost Optimised	0,64 ± 0,024	0,54 ± 0,038	0,61 ± 0,035
Пол	LR	0,62 ± 0,02	0,62 ± 0,033	0,62 ± 0,036
	XGBoost Default	0,62 ± 0,017	0,53 ± 0,035	0,61 ± 0,035
	XGBoost Optimised	0,65 ± 0,021	0,60 ± 0,026	0,64 ± 0,038
Возраст	LR	0,62 ± 0,024	0,62 ± 0,033	0,63 ± 0,058
	XGBoost Default	0,61 ± 0,029	0,53 ± 0,035	0,59 ± 0,045
	XGBoost Optimised	0,63 ± 0,032	0,60 ± 0,026	0,60 ± 0,036

Бидејќи од резултатите на сите претходни анализи можеше да се забележи дека кај оптимизираниот XGBoost имаше мало подобрување на перформансите во споредба со стандардниот XGBoost, процесот на оптимизација овде беше исклучен за да се намали времето потребно за развивање на моделот и потребната процесорска моќ. Резултатите се претставени во **Табела 5-4**, каде што може да се види дека слични перформанси во предвидувањето на расата на пациентите има и кога се разгледуваат пациенти со коморбидитети, освен кај пациенти со срцева слабост каде што предвидувачките способности на моделот се значително слаби. SHAP-от записи за сите дијагнози, исклучени од срцева слабост, покажуваат неинвазивна заситеност со кислород како најважна карактеристика. Графикот од SHAP анализата за експериментот со срцева слабост ги покажува карактеристиките на крвниот притисок како најважни за моделот, и ова посочи едно потенцијално објаснување зошто предвидувачките перформанси кај пациентите со срцева слабост се намалуваат, т.е. познато е дека срцевата слабост може да предизвика намалување на нивото на кислород, [176], што беше една од највлијателните карактеристики во самиот модел според претходните резултати.

Табела 5-4 Анализа на коморбидитети на пет избрани дијагнози (дадени во оригинална форма) за кохорта од пациенти од црна и бела раса

Алгоритам	Essential hypertension	Heart failure	Acute kidney failure	Chronic kidney disease	Sepsis
LR	0,62 ± 0,073	0,53 ± 0,113	0,60 ± 0,049	0,57 ± 0,099	0,57 ± 0,176
XGBoost Default	0,60 ± 0,064	0,51 ± 0,050	0,71 ± 0,068	0,72 ± 0,079	0,69 ± 0,093

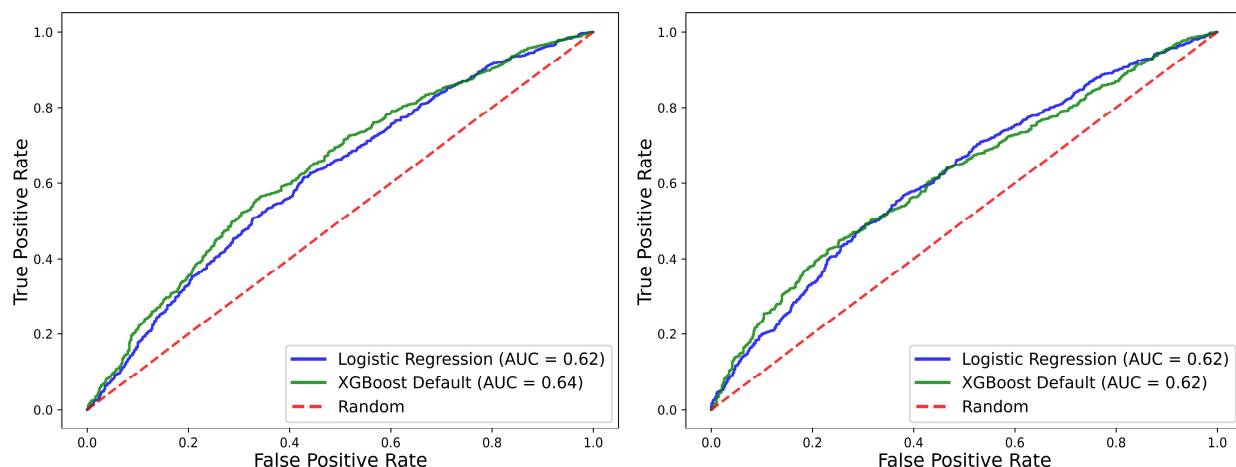
5.3 Екстерна валидација

За да се тестира стабилноста на перформансите за различни податочни множества, истражувањето се прошири и на надворешна валидација. Проценките за стабилноста на перформансите се попрецизни колку е поголем обемот на самите податоци, [177], и поради таа причина надворешната валидација беше извршена само на кохортата од пациенти од црна и бела раса. Валидацијата беше извршена во две насоки: едната, од eICU-CRD кон MIMIC-III, а другата, од MIMIC-III кон eICU-CRD. Имено, во првиот експеримент моделот беше обучен на податоците од eICU-CRD и евалуиран на MIMIC-III, а во вториот обратно, моделот беше обучен на податоците од MIMIC-III и евалуиран на eICU-CRD.

Бидејќи постои дел од карактеристиките кои се со мала застапеност кај MIMIC-III множеството, тестовите беа направени на достапните карактеристики, и затоа, спроведените експерименти вклучија проценка на: само неинвазивен крвен притисок, само неинвазивна заситеност со кислород и комбинација од неинвазивен крвен притисок и заситеност со кислород заедно. Сите AUC вредности добиени при надворешната валидација од сите експерименти во двете насоки на валидација се дадени во **Табела 5-5**, и како што може да се забележи перформансите на моделот - иако оптимизираниот модел не е обучен - за двата алгоритми важат за сите кохорти. Понатаму, најдобри перформанси се забележани со неинвазивна заситеност со кислород. Кривата AUC-ROC за тестовите каде што се користат двата вида на неинвазивни мерења (крвен притисок и заситеност со кислород) е дадена на **Слика 5-19**. И покрај тоа што може да се забележи мал пад во вредностите на AUC, разликата при екстерната валидација не е доволно драстична за да посочи нестабилност на моделите за различни податочни множества.

Табела 5-5 Резултати од екстерна валидација на кохорта од пациенти од црна и бела раса за три комбинации на карактеристики

Algorithm	Non-invasive blood pressure	Non-invasive oxygen saturation	Non-invasive measurements
eICU-CRD to MIMIC-III			
LR	0,57	0,6	0,62
XGBoost Default	0,59	0,6	0,64
MIMIC-III to eICU-CRD			
LR	0,59	0,69	0,62
XGBoost Default	0,56	0,6	0,62

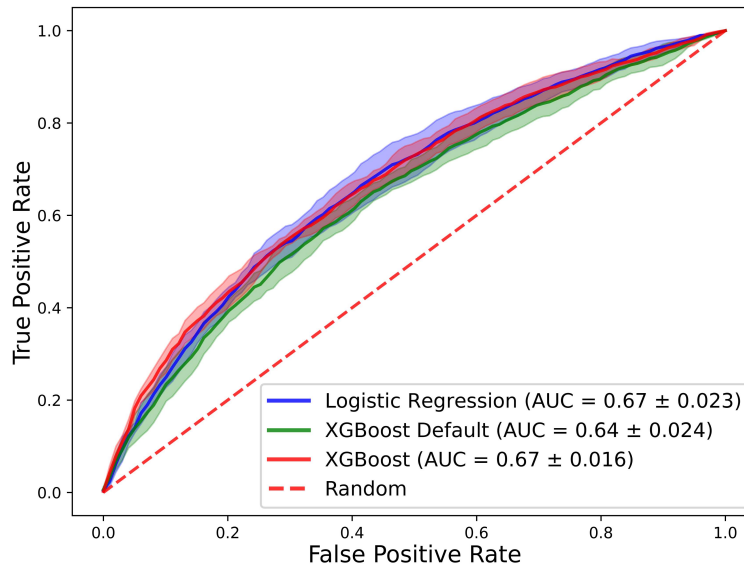


Слика 5-19 Резултати од екстерна валидација за неинвазивни мерења: од eICU-CRD до MIMIC-III (лево) и од MIMIC-III до eICU-CRD (десно)

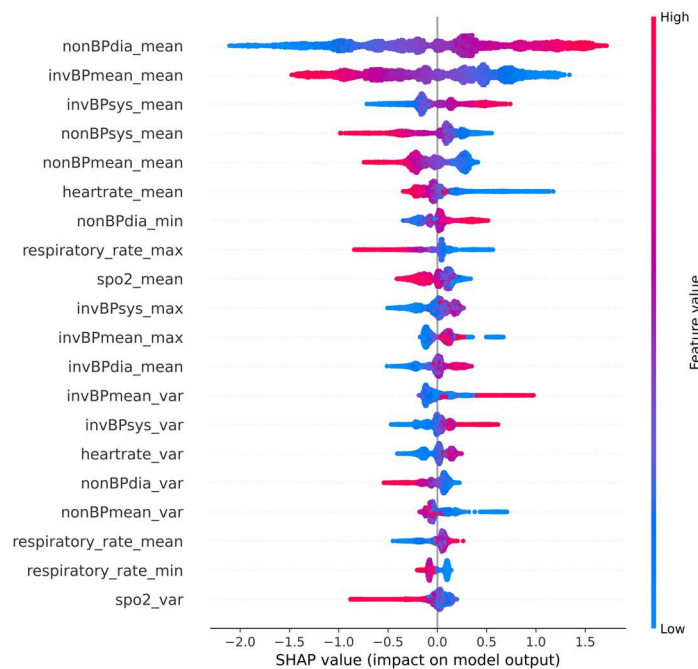
5.4 Трансферабилност: Полова пристрасност

Пристапот кој се користи за идентификација на етничка и расна пристрасност е реализиран за со мали модификации да биде трансферабилен. За да се илустрира ова, пристапот се примени за детекција на полова пристрасност во витални знаци. Беа направени мали промени за целиот процес да се пренасочи и да има нова лабела за учење. Поточно процесот на спарување не го вклучи полот како карактеристика за спарување (бидејќи полот е целната променлива која треба да се избалансира) и целната променлива која беше дадена на алгоритмот за учење беше променета во пол на место на етничко-расната припадност. Употребени беа истите витални знаци како и во оригиналното истражување за детекција на етничко-расна припадност, а и истите карактеристики, поточно средна вредност, минимум, максимум и варијанса на откучување на срце, респираторна стапка, инвазивна и неинвазивна кислородна заситеност, систолни, дијастолни и средни мерења од NIBP и IBP. Исходот кој моделот требаше да го научи и предвиди беше полова пристрасност преки класификација на пациентите како мажи или жени врз основа на виталните знаци. За таа цел беше креирана кова кохортна група во која полот не беше земен како променлива за спарување, бидејќи тоа би било контрадикторно со фактот дека полот е целна променлива. Вака добиената кохорта содржеше 5.940 пациенти, а резултатите добиени по разгледување на сите карактеристики се дадени на *Слика 5-20*.

Како што може да се види XGBoost постигнува AUC од 0,67 ($\pm 0,016$), што покажува дека моделот може да предвидува пол на пациентите врз база на статистички информации од мерења на витални знаци собирани во првите 24 часа од прием. Со цел да се разберат водечките фактори кои се кријат зад моделот, беше направена и салиентна анализа на карактеристиките, што резултираше со графикот прикажан на *Слика 5-21*. Како што може да се види, петте најзначајни карактеристики за моделот се мерења на крвен притисок, инвазивни и неинвазивни. Ова е во согласност со постоечката литература која покажува полов диморфизам во крвен притисок, кој отпочнува во пубертетот и опстојува и за време на процесот на стареење, [178], [179]. Гледајќи дека процесот на селекција на пациенти вклучува само пациенти постари од 18 години, но и земајќи предвид SHAP графикот, логично е да се заклучи дека истиот диморфизам е идентификуван од моделите.



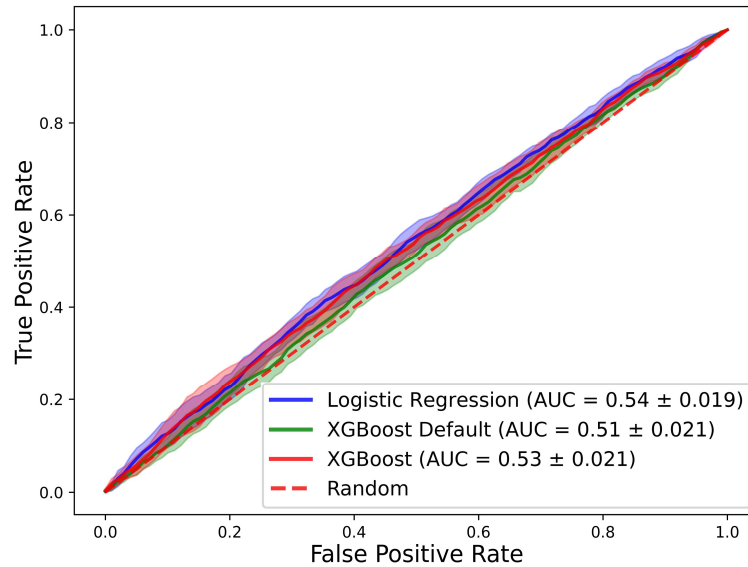
Слика 5-20 Полова пристрасност за eICU-CRD со сите витални знаци



Слика 5-21 Анализа на карактеристики за полова пристрасност за eICU-CRD со сите витални знаци

Бидејќи на SHAP графикот може да се воочат и други карактеристики, како отчукување на срце, респираторна стапка и кислородна заситеност, со цел да се провери дали влијание на моделот имал само крвниот притисок, како што беше очекувано бидејќи другите карактеристики не се имаат покажано како индикативни за пол, анализата беше проширена со дополнителен тест кој ги исклучува карактеристиките кои се однесуваат на мерења на крвниот притисок, а резултатите се дадени на **Слика 5-22**. Со добиената AUC од 0,54 ($\pm 0,019$) за LR и помалите вредности за двете верзии на XGBoost се покажува дека иако половата пристрасност може да се одреди при тренирање на моделот користејќи ги сите витални знаци и додека мерењата на крвен притисок носат маркери за дистинкција помеѓу машки и женски пациенти истото не може да се каже кога мерењата на

крвниот притисок исклучени, што значи дека останатите витални знаци не носат индикатори за пол. За резултатите беше направена и SHAP анализа, но бидејќи резултатите се екстремно слаби не е потребно да се врши анализа на карактеристиките значајни за моделот. Резултатите ја покажуваат преносливоста на пристапот и потребните модификации за да истата се пренамени за откривање на различни форми на пристрасност, како што е во случајов полова пристрасност.



Слика 5-22 Полова пристрасност за eICU-CRD по исклучување на мерења на крвен притисок

6 Дискусија

Со оглед на тоа што пристрасноста во медицината е вкоренета во ставовите на давателите на здравствена заштита, која се манифестира во варијации во третманите и терапијата низ етничко-расните групи и детално документирана во клиничките студии, важно беше да се разбере дали пристрасноста постои во неконтролиран аспект на медицинската пракса, особено онаму каде што не треба. Имено, виталните знаци и мерењата на виталните знаци.

Моделите за МУ тренирани на само шест витални знаци успеаја да ја предвидат самопријавената раса на пациентите со AUC од 0,74 ($\pm 0,022$) помеѓу пациенти од бела и црна раса, AUC од 0,68 ($\pm 0,059$) за Хиспанци и пациенти од бела раса и AUC од 0,72 ($\pm 0,045$) за Хиспанци и пациенти од црна раса. Технологиите што се користат за мерење на заситеноста со кислород се покажаа како значаен извор на пристрасност на самопријавени расни информации, а истиот тренд може да се види и во некои од тестовите за етничко-расна припадност при мерење на крвниот притисок. Бидејќи резултатите покажуваат како технологија за мерење на витални знаци крие расна пристрасност, за целосно да се провери дали виталните знаци се извор на пристрасност, беа разгледани дополнителни резултати каде за тренирање беа употребени само инвазивни мерења од виталните знаци. Постигнатите резултати од AUC 0.65 (± 0.043) за пациентите од црна и бела раса, од AUC 0.58 (± 0.057) за Хиспанци и пациенти од бела раса, и од AUC 0.69 (± 0.111) за Хиспанци и пациенти од црна раса покажуваат дека етничко-расна припадност може да биде научена од витални знаци и кога медицинската технологија не влијае на процесот на учење. Овие резултати и природата на пристрасноста која е забележана во нив бараат додатни анализи. Додека испитувањето на практики за пружање на нега и тренирање на моделите на карактеристики за фреквенција на виталните знаци обезбедија AUC од 0,58 ($\pm 0,031$), па затоа не беа означени како значаен извор на расни информации. Со оглед на тоа што моделите потенцијално можат да бидат под влијание на збунувачки променливи и учењето на етничко-расната припадност да виде врз основа на дистрибуцијата на тие збунувачки променливи во кохортите од интерес на испитуваната популација, беа создадени и испитани повеќе дополнителни групи, па дури и низ статистички неважни дијагноза при прием, возраст, и пол, резултатите сепак покажаа дека само-пријавената етничко-расна припадност може да се идентификува од статистичките информации вградени во виталните знаци добиени во првите 24 часа по приемот.

Гледајќи дека моделите на ВИ можат да ја научат етничко-расната припадност од инвазивно измерените витални знаци, без технологијата за мерење и медицинската пракса да влијаат на исходот, може да се заклучи дека само виталните знаци вградуваат расни информации што може да се научат со алгоритми на ВИ. Ова е неочекувано бидејќи досегашни истражувања сметаа дека расните информации не се присутни во вредностите на виталните знаци. Понатаму, способноста на алгоритмите за МУ да научат расни информации од виталните знаци собрани од 24-часовен прозорец, генерализирана за различни популации на пациенти и опстојување на таа способност дури и кога се контролираат познати извори на расна пристрасност, како што се отчитувањата на кислородна заситеност со пулсен оксиметар кои се погодени од боја на кожа, практиките при испорака на грижа на пациенти и во присуство на различни коморбидитети, докажуваат дека ВИ има потенцијал да научи со етничко-расна пристрасност таму каде што не треба да постои, што претставува значителен ризик за правичното клиничко одлучување. Друга голема проблематика што ја отвораат овие резултати е дека мерките и стратегиите за ублажување на пристрасноста може да претставуваат особен предизвик, имајќи ја предвид основната улога на виталните знаци во медицинската пракса.

За дополнително појаснување, учењето на самопријавени расни информации од виталните знаци создава значителна грижа за развој на алгоритми кои поддржуваат клиничко одлучување. Ако чувствителните атрибути, на пример, расата, се учат додека алгоритми се тренираат за предвидување или оптимизирање на исходи на модели, тогаш можно е алгоритмите да ја користат пристрасноста што ги детектира овие демографски карактеристики како информација во процесот

на донесување одлуки. Понатаму, уште позагрижувачка работа е што овие чувствителни атрибути се учат од суштински клинички информации, на кои медицинските лица во голема мера се потпираат за да го диктираат не само текот на лекувањето, туку и оперативните работни текови и во креирањето политики.

Веќе постојат огромни докази дека расата е значаен фактор при клиничкото одлучување, [180]. Во [181], авторите известуваат дека етничката усогласеност помеѓу пациент и лекар била инверзно корелирана со добивање на скенирање за проверка на постоење на рак меѓу Хиспанците, т.е., пациентите Хиспанци примени и прегледувани од лекари од друга етничка група добиле поголем број на скенирање за рак на дојка и колоректално скенирање од пациентите со шпанско потекло што ги гледале лекари Хиспанци. Лекарите припадници на црна раса системски ги надминуваат своите колеги кога се грижат за новороденчиња од црна раса, како што е соопштено од, [182], каде што авторите откриле 58% намалување на смртноста на новороденчиња од црна раса кога се под грижа на лекари од црна раса во споредба со кога се под грижа на лекари од бела раса. Друг аспект на медицината каде што расните разлики се многу присутни се третманите за болка. И покрај пријавувањето на поголема болка и болка поврзана со инвалидитети, малцинските пациенти појавуваат поголема веројатност за добивање на несоодветен третман за управување со болка во споредба со пациентите од бела раса, [183], [184], [185]. Покрај тоа, ова несовпаѓање опстојува со пациенти деца кои се примени во детскиот оддел за итни случаи поради акутна болка, [186], при што пациентите од црна раса без Хиспански примеси имаат помали шанси да добијат рецепт за опиоид дури и во случаи со најтешки фрактури. Друга студија, [187], покажува дека педијатриските лекари се подеднакво подложни како и другите лекари на имплицитна расна пристрасност, и дека пристрасноста кон деца од црна раса е на слични нивоа на пристрасноста кон возрасните пациенти од црна раса.

Овие случаи на пристрасност само докажуваат дека кога чувствителните атрибути, како раса, се применуваат во активна медицинска практика, постојат сериозни неоправдани последици за одредени популации на пациенти, што дополнително го зајакнува набљудувањето дека ако алгоритмите се способни да ги научат атрибутите како раса, дури и кога тие се намерно отстранети од податоците, не е невозможно да се очекува дека истите тие алгоритми, исто така, ќе ја забележат пристрасноста на медицинското лице и субјективноста во донесувањето одлуки. На крајот на краиштата, алгоритмите се обучени на податоци кои треба да ја документираат клиничката интуиција и расудување кодирани како одлуки, и ако тие одлуки се пристрасни, разбирливо е дека алгоритмите можат да ги соберат и да ги научат тие неправди, и следствено, да придонесат за нееднаквости во резултатите врз основа на раса, пол и други демографски фактори.

Иако резултатите тврдат дека не може да се добие дефинитивно предвидување на самопријавената раса, успехот на моделите во правилна класификација на два од три пациенти за сите три кохорти (пациенти од црна и бела раса, пациенти Хиспанци и од бела раса, пациенти Хиспанци и од црна раса) не е случаен. Покрај тоа, обучените модели за МУ користат само агрегирани информации од виталните знаци на пациентите во форма на статистички карактеристики, што сугерира дека овие карактеристики од рутински собрани информации во првите 24 часа од приемот содржат вградени информации долж расни димензии. Се покажа дека отчитувањата на пулсен оксиметар на заситеноста со кислород во крвта, кои се сметаа како важна непристрасна мерка за хипоксемија, зависат од бојата на кожата, што како информација излезе на виделина за време на кризата КОВИД, [188], [189]. По понатамошна истрага, беше откриено дека нивоата на заситеност со кислород имале поголема варијабилност кај пациентите кои се идентификувале како пациенти од црна раса, проследени со Хиспанци, Американци од азијско потекло и најмалку кај пациенти од бела раса. Резултатите што ги добивме го потврдија фактот од наведените студии, со тоа што, исто така, забележавме дека заситеноста со кислород, мерена преку пулсна оксиметрија, беше највисоката карактеристика во сите анализи за важност на карактеристики кај моделите, со тоа што најдоминантно присуство беше забележано помеѓу

пациенти од црна и бела раса. Овие резултати се целосно очекувани како што може да се докаже со варијабилноста од претходните истражувања; т.е., кај пациенти од црна раса има најголема варијабилност на мерењата на заситеноста со кислород, а кај пациенти од бела раса има најниска; така, се очекува да се забележи најзначајната промена во резултатите при обука со и без неинвазивна заситеност со кислород при анализата на овие две етничко-расни припадности.

Бидејќи анализата на важност на карактеристики го покажа и потврди влијанието на пулсна оксиметрија, беа извршени дополнителни тестови за контрола на овој фактор, а тие резултати, како и анализата на важност добиена за тие анализи, не понудија убедливи наоди за корелацијата помеѓу виталните знаци и самопријавената раса. Истото може да се каже и за мерењата на крвниот притисок. Имено, додека пулсните оксиметри и апаратите за мерење на притисок се докажаа како пристрасни дури и кај агрегирани витални знаци, и покрај тоа што нивното влијание е значајно, анализата на важност на карактеристики не ја објаснува целосно пристрасноста откриена во податоците при отстранување на нивното влијание.

Тешко е да се посочат потенцијалните извори на самопријавената етничко-расна припадност меѓу преостанатите карактеристики без понатамошно истражување. Постојат неколку хипотези, иако сите бараат дополнително испитување. Една хипотеза може да биде дека постојат разлики во калибрацијата во коефициентот APACHE IV што се користи во eICU-CRD множеството и коефициентот OASIS користен за MIMIC-III множеството, како што е прикажано во [190]. Имено, истражувањето наведува дека калибрациски мерења покажуваат постојани, а во некои случаи и статистички значајни, шаблони на разлика помеѓу пациентите Хиспанци и од црна раса наспроти Азијци и пациенти од бела раса, и додека калибрациите биле несовршени за сите групи, резултатите покажале конзистентно зголемено предвидување на смртност за пациентите од црна раса и Хиспанците. Варијација на хипотезата пак би ја доделила пристрасноста на разликите во сериозност на болеста кои не се прецизно рефлектирани во коефициентите APACHE IV, [191], и OASIS, [192]. Друга хипотеза се заснова на физичката средина и социоекономските фактори кои влијаат на индивидуалните фактори, што значи дека пристрасноста може да потекнува од факторите надвор од здравствениот систем, бидејќи критериумите што се користат за селекција на пациенти кандидати за вметнување на артериска линија, или дури и достапноста на пациентот може да придонесе за вградување на информации за самопријавена раса во виталните знаци, [193].

Покрај потенцијалните ризици, ова истражување ги истакнува и предизвиците во мерките за ублажување на пристап. Иако е критикуван, еден вообичаен пристап е селективно да се отстранат променливите што ги кодираат ваквите чувствителни атрибути, [194], со цел моделите на ВИ да не учат од нив и последователно да не станат дел од процесот на донесување одлуки. Придобивките од овој пристап кон медицинските истражувања би биле значајни, бидејќи методот нема да дозволи вложување на веќе вградените јазови помеѓу привилегираните и занемарените групи пациенти. Сепак, сеприсутната употреба на виталните знаци во клиничкото одлучување го прави овој пристап невозможен, и тоа не само затоа што потеклото на расните информации се чини дека е тешко да се изолира. Можеби е дојдено време да се примени обратен пристап, односно со користење на чувствителни атрибути како што е расата да се олесни ревизијата за потенцијална алгоритамска пристрасност и да се приспособат воспоставените политики за тоа како етички да се собираат, користат и споделуваат податоците за етничко-расната припадност, [195].

6.1 Ограничувања на истражувањето

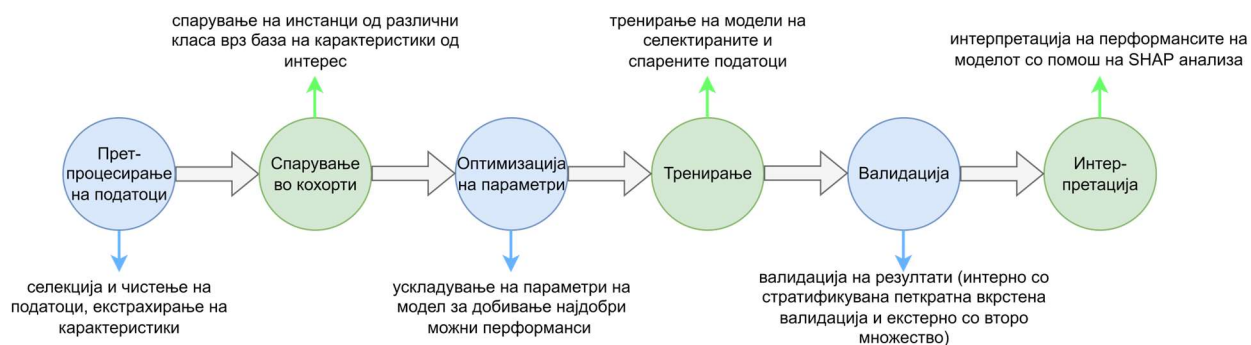
Ова истражување е силно концентрирано на решавање на потребата за транспарентност, еднаквост и правичност, па од особена важност е да се согледаат и да се дискутираат одредени ограничувања на студијата. Оваа студија вклучува две големи и разновидни податочни множества за податоци од ЕИ, едно мултицентрално множество со двегодишно собирање податоци од 335 ЕИН во 208 болници во САД и друго собирање во еден центар во тек на една деценија, избрани со цел да се ублажат разликите во начинот на собирање. Двете податочни множества се јавно достапни, добро

проучени и имаат неколку етничко-расни припадности, но и покрај тоа сепак постојат некои ограничувања.

Прво, употребата на самопријавената раса претставува предизвик, особено бидејќи студии имаат покажано дека генетската варијабилност е повисока во рамките на една раса отколку помеѓу расите, [196], што расата ја воспоставува како општествен наместо биолошки концепт. Понатаму, самопријавената раса во eICU-CRD множеството вклучува строги категории кои не ги земаат предвид пациентите со мешано потекло, додека самопријавената раса во MIMIC-III базата вклучува специфични и детални подкатегории на етничко-расни групи (на пр. од бела раса, од бела раса со руско потекло, од бела раса со источноевропско потекло, од бела раса со европско потекло, од бела раса со бразилско потекло) но со значително низок број на пациенти по поткатегорија. Покрај тоа, студијата опфати само пациенти од црна раса, Хиспанци и од бела раса, бидејќи другите расни идентитети (т.е. Азијци и домородните Американци за eICU-CRD и сите поткатегории за MIMIC-III) немаа доволно податоци за опфатна и реална анализа. Понатаму, последниве години поттикнаа дискусии за тоа дали при истражувањата се откриваат расни припадности или резултати од расизам. Додека студијата е фокусирана на категории на раса, објаснувањата дадени за наодите околу расата се фокусирани на влијанието на расизмот во општеството, а со тоа и во медицината (на пр. разликите во технолошките перформанси низ расата). Конечно, студијата се фокусираше на популација на пациенти со седиште во САД. Затоа, потребна е дополнителна истрага за да се утврди дали овие резултати може да се генерализираат за ЕИИ со седиште надвор од САД.

6.2 Генерализираност на пристапот

Генерализираноста го дефинира начинот на кој истражувањето може да се примени во услови различни од оригиналните тест околии. Пристапот за идентификација на етничко-расна пристрасност во витални знаци на пациенти, медицинска мерна технологија и медицинска пракса се состоеше од неколку чекори, кој и покрај тоа што беше приспособен и применет на две податочни множества, исто така беше направен со цел динамичност и трансферабилност, како и да биде лесна за модификација и примена во детекција на различни типови на пристрасност во различни податочни множества. Обопштената верзија на архитектурата која се користи во ова истражување е дадена на **Слика 6-1** и е организирана во шест фази. Ова поглавје ги детализира предложените насоки за секоја фаза во архитектурата, кои беа следени за време на развој на моделите, како сугестии и правила што треба да се применат со цел да се добијат правични и транспарентни резултати за откривање на пристрасност во податочни множества.



Слика 6-1 Предлог архитектура за детекција на пристрасност во податоци

Првата фаза го опфаќа претпроцесирањето на податоците: избор и чистење на податоците, како и извлекување на значајни карактеристики. Во оваа фаза, неопходно е да се биде сигурен дека избраните податоци се целосни и валидни, па од таа причина оваа фаза во архитектурата се фокусира на филтрирање на медицински неточни вредности. Откако беа избрани релевантните

податоци, процесот продолжи со екстракција на карактеристики, каде што мерењата на виталните знаци кои биле собрани во текот на првите 24 часа од прием на пациентите беа агрегирани и со тоа беа добиени карактеристики од интерес. Во нашиот случај, и двата аспекта беа опфатени со обезбедување и демографските информации и виталните знаци на пациентите да бидат валидни вредности во рамките на очекуваниот опсег, т.е. беа отстранети NaN податоците (скратено од not a number, полиња во кои нема никаков запис за вредност), негативни мерења од витални знаци, неконзистентни детали за демографските информации (на пр., старосни вредности над 200, раса/пол означени како непознати или неизјаснети, итн.). Ова ги чисти и структурира податоците за во понатамошните фази на архитектурата и е неопходен чекор кој осигурува дека моделот учи од точни и валидни информации, наместо да предвидува врз база на недоследности во податоците.

Втората фаза е фазата на спарување, каде што се создава конечната кохорта на пациенти, и во оваа фаза примероци (во случајов пациенти) од различни класи се спојуваат врз основа на карактеристики на интерес (или, карактеристики за спарување). Важното нешто што треба да се забележи во процесот е изборот на соодветни карактеристики на кои ќе се заснова самото спарување, бидејќи и вклучувањето и исклучувањето на одредени карактеристики од процесот на совпаѓање може да придонесе до или да предизвика негативно влијание врз перформансите на моделите и исходот, општо гледано. Вклучувањето на одредени карактеристики во совпаѓањето, особено ако тие карактеристики не се збунувачки, може да резултира со хомогена група каде што моделите ќе имаат малку или никакви информации кои може да ги научат. Од друга страна, исклучувањето на карактеристиката може да значи кај добиената група постои нерамнотежа во вредностите за таа одредена карактеристика и дека постојат скриени информации кои би можеле да го дефинираат исходот, со кои моделот ненамерно го учи тоа несовпаѓање и ги заснова предвидувањата на него. На пример, ако процесот на спарување ја исклучи дијагнозата при прием како карактеристика за спарување, можно е да постои доминантно присуство на специфична дијагноза за одредена целна група, т.е. пациентите од црна раса се повеќе склони кон срцеви заболувања, а моделот открива и ја учи таа информација и донесува одлуки врз основа на срцеви заболувања, наместо врз основа на вистинските витални знаци како што првичната идеја имала за цел. Во овој пример, дијагнозата на прием би била збунувачка карактеристика, бидејќи влијае на исходот на моделот, иако не треба, што ја докажува неопходноста да се обезбеди еднаква дистрибуција на карактеристиките за кои се верува дека се збунувачки.

Третата фаза се фокусира на подесување на параметрите на моделот за да се добијат параметри кои ги овозможуваат најдобрите перформанси на моделот. Иако ова не е задолжителен чекор од процесот, тој овозможува изнаоѓање на најдобрите резултати и перформанси, и како таков може да понуди вреден увид во податоците и нивната содржина. Овде се применува едноставно правило, со тоа што колку е пообемно пребарувањето, толку е подобар процесот на оптимизација. Четвртата и петтата фаза, т.е. тренирањето и валидацијата се тесно поврзани и во некои случаи може да се преклопуваат. Има неколку точки кои треба да се земат предвид: поделбата на податоците во податоци за тренирање и валидација на моделот, видот на методологијата за валидација и користените алгоритми. Предложената архитектура вклучува стратификувана петкратна вкрстена валидација која ја диктира и поделбата на податоците и самата валидација, т.е. моделите се интерно валидирани така што податоците се поделени во пет стратификувани подмножества, каде четири се користат за тренирање и едно се користи за валидација, а крајните резултати се средна вредност од резултатите од сите пет подмножества. Дополнително, беше извршена и надворешна валидација со цел да се воочат перформансите на моделот при работа со целосно неповрзани податоци од различно податочко множество. И двата пристапи за валидација имаат свои предности: внатрешната вкрстена валидација овозможува најдобро искористување на податоците и ја задржува оригиналната дистрибуција за време на тренирањето и тестирањето на моделот, додека надворешната валидација докажува дека моделот дава генерална слика за популацијата и покрај тоа што е трениран на само дел од неа. Шестата и последна фаза се фокусира на разбирање на перформансите на моделот и важноста што карактеристиките ја имаат за него, т.е., обезбедува интерпретабилност на моделот.

Ова е овозможено со користење на SHAP графици, но и со користење на опширни тестирања кои се реализираат со вклучување и исклучување на карактеристики за моделот по што се набљудуваат промените во перформансите на моделот откако ќе бидат отстранети влијанијата што одредени карактеристики ги носат.

Дополнителните резултати за полова пристрасност, дадени во резултатите, покажаа дека е потребна само мала промена за пристапот да се примени за детекција на различни типови на пристрасност. Имено, модификацијата на пристапот за негово пренасочување од идентификација на расна во идентификација на полова пристрасност бараше промена на лабелата за учење и нејзино исклучување од процесот на креирање на кохортата. Дополнително, што се однесува до применливоста на архитектурата на различни податочни множества, структурата на архитектурата треба да остане иста, но потребни се модификации во првата фаза (при избор на податоците и претпроцесирањето) и во втората фаза (т.е. избор на карактеристиките за спарување). Следните фази од архитектурата – оптимизација на параметрите, тренирање, валидација и интерпретабилност на моделот – остануваат непроменети.

Предложената архитектура, со нејзината динамичност и трансферабилност, нема краен корисник во класичната смисла на зборот, како што е на пример случај со модели за детекција на болести и дијагностички алатки, каде медицинските лица ги користат тренираните модели да ја анализираат состојбата на пациентите и да добијат компјутерски потпомогната дијагноза. Имено, доменот на примена на архитектурата е евалуација на пристрасност на други модели на ВИ и други податочни множества, што значи дека корисници на предложената архитектура се есенцијално други модели на ВИ или податочните множества. Моделите на ВИ генерално најчесто се оценуваат со мониторирање на средна грешка на моделот или преку следење на специфична метрика за истиот, што во некои случаи може да посочи дека моделот работи сосема задоволително и покрај тоа што има слаби перформанси во одредени ситуации. Ова се карактеристики на пристрасните модели, кога алгоритмите имаат севкупно одлични перформанси, а всушност моделот има одлични перформанси за една и слаби перформанси за друга популација од пациенти; расно пристрасните модели може да имаат одлични способности за предвидување кај пациенти од бела раса и слаби перформанси за пациенти од други етничко-расни групи.

Примената на предложената архитектура во реалниот свет се состои од евалуација и идентификација на пристрасност, и каде што се очекува пристрасност и каде што не е. Главниот аспект каде предложениот модел може да открие пристрасност е во податочните множества, т.е. оваа архитектура може да им овозможи на програмерите и истражувачите да го проценат квалитетот на податоците, пред да ги користат тие податоци во градењето модели за активна медицинска примена. Ова овозможува сигурност дека системите не се изградени на пристрасни информации и, последователно, не ја пропагираат таа пристрасност во самите модели. Секако, анализата нема да осигури дека моделите се целосно непристрасни, бидејќи многу дополнителни фактори можат да создадат пристрасност во процесот на градење модел (на пр., процес на селекција на податоци – што е пристрасност во начинот на кој се избираат податоците, а не самите податоци, пристрасност од поделба на податоците – податоците за обука се пристрасни ако се недоволно обемни или погрешно ја претставуваат реалната популација, пристрасност во креирање на карактеристики, пристрасност при обука на моделот – менување на изборот на алгоритам врз основа на дефиницијата на проблемот и посакуваниот исход, видот и количината на податоци, итн., и пристрасност при тестирање на моделот – преглед на перформансите на моделите врз невидени податоци). Сепак, оваа архитектура може да потврди дека податоците што се користат во креирање на моделот се непристрасни и сигурни, т.е. врши проверка дали податоците воведуваат кај моделот имплицитни предрасуди од медицинската пракса или од процесот на собирање податоци. Понатаму, архитектурата може да се користи за откривање на пристрасност во податоците по избор на податоци, нивното процесирање и екстракција на карактеристики, со што вака воведените предрасуди ќе може лесно да се препознаат. Дополнителна примена на архитектурата вклучува евалуација на пристрасност за време

на процесот на собирање на податоци и креирање на податочни множества; доколку правилно се користи и редовно се применува, архитектурата ќе им овозможи на собирачите на податоци да креираат нови, непристрасни множества. Овие податочни множества подоцна би можеле безбедно да се користат за создавање на модели за МУ од лица кои не се експерти во областа на истражување, а бидејќи податоците ќе бидат темелно оценети како непристрасни за време процесот на собирање ќе се намалат или целосно искоренат сите можности за пристрасност што произлегуваат од прикриени проблеми во самите податоци.

Затоа, со оваа реална примена на архитектурата, истражувачите, собирачите на податоци и развивачите на модели за ВИ може привремено да се наведат како крајни корисници на предложената архитектура, која ќе служи како детектор на пристрасност во податоците и кога се собираат податоците и за време на целиот процес при развој на моделот, за анализа по секоја фаза со обработка на податоци, за да се осигура дека податоците не носат имплицитни информации за пристрасност што може да се пропагираат и негативно да влијаат на корисниците на тие податочни множества и на моделите кои се тренирани на тие податоци.

Во различни фази за време на развој на моделот, сè од недоволни податоци, неконзистентно собирање податоци и лоши практики на податоци, до пристапот што го избираат развивачите може да доведе до пристрасност во одлуките на моделот. Со оглед на тоа што медицинската ВИ влијае на животите на пациентите, како и на понудениот пристап до здравствена заштита и нејзиниот квалитет, идентификувањето, оценувањето и решавањето на какви било потенцијални предрасуди што би можеле да влијаат на исходот е клучен услов при креирањето на нов модел за МУ. Како што ВИ наоѓа широка примена и бројот на истражувачи што ја применуваат во решавање проблеми се зголемува, одговорноста за проверка, следење, оценување, истражување и евалуација на системите базирани на ВИ пред тие да се применат во системот расте, особено како средство за избегнување на пристрасност што негативно ќе влијае на ефективноста на донесените одлуки. Се зголемува бројот на публикации кои предлагаат медицински апликации и анализи, а многу ретко резултатот од истражувањето (моделот) се анализира за недоследности и пристрасни практики. Само во изминатите неколку години свесноста за пристрасноста и неопходноста од ублажување на пристрасноста станаа актуелна тема, активно дискутирана дури и меѓу експертите. Како резултат на ова, постојат само мал број процедури за откривање на пристрасност, од кои повеќето се фокусираат на анализа на податоци, статистика и евалуација на моделот на различни подгрупи на пациенти (но сето тоа откако моделот ќе биде истрениран). Иако сите пристапи носат свои придобивки, ниту еден од нив не може да направи детекција на пристрасност во податоците и моделот, а во исто време точно да ги идентификува изворите на таа пристрасност. Пристапот предложен овде нуди единствена способност за детекција на пристрасност и идентификација на изворите кои ја предизвикуваат таа пристрасност (ако тие извори навистина се препознатливи) пред моделите да бидат истренирани и применети во медицинска околина; уште повеќе, пристапот не е применлив само за расна пристрасност во специфичните податочни множества што се користат за истражувањето, туку може да се примени и за различни форми на пристрасност во какви било табеларни податоци.

Имајќи предвид дека главната цел на истражувањето е предлагање на архитектура за откривање на пристрасност, би било прилично иронично доколку архитектурата предложена овде е ѝ самата пристрасна. Свесни за заблудите на ВИ, при развојот на архитектурата се применуваат неколку чекори и стратегии за ублажување на пристрасни ефекти со цел да се осигура дека предложената архитектура е без пристрасност.

Прво, целиот пристап е развиен и применет на две податочни множества, едно од една ЕИН и второ од повеќе ЕИН за да се земат предвид повеќе локации на собирање и варијабилноста на податоците меѓу ЕИН, но и дополнително едно собрано во период од нешто повеќе од 10 години од 2001 до 2012 година и второто во периодот од 2014–2015 за да се обезбеди применливост која не зависи од временски период на мерење на податоците. Второ, се врши и внатрешна и надворешна

валидација, при што надворешната валидација осигурува дека перформансите на моделите се задржуваат и при примена на невидени податоци собрани од различна ЕИИ и во различен временски период. Надворешната валидација беше извршена во две насоки, т.е. моделите обучени на eICU-CRD беа тестирани на MIMIC-III, и обратно, моделите обучени на MIMIC-III беа тестирани на eICU-CRD, и во двата случаи добиените перформанси на моделите покажаа задоволителни резултати. Двете податочни множества се значително неурамнотежени во нивната оригинална дистрибуција, што се проследи и по процесирање на податоците, па бројот на приеми на пациенти од бела раса далеку го надминуваа бројот на приеми на пациенти со потемна боја на кожата, и затоа пред да се отпочне со процесот на тренирање и евалуација на моделот беше неопходно да се балансираат податоците, бидејќи работа со небалансирани податоци се има покажано како значително влијателна на одлуките на моделот во корист на позастапената група. Беа земени предвид неколку постапки за семплирање на примероци со цел балансирање, при што како најсоодветна се покажа методата на спарување на пациенти од две етничко-расни групи. Процесот на спарување вклучуваше карактеристики од интерес кои може да влијаат на однесувањето на моделите, имено, возраст, пол, дијагноза при прием и коефициент за сериозноста на болеста. Добиеите групи беа совршено избалансирани меѓу етничко-расните припадности и имаа слични дистрибуции за потенцијалните збунувачки променливи.

Конечно, во процесот на селекција на видот на алгоритам кој ќе се користи беше направен избор кој ги адресира постојните ограничувања на податоците и, следствено, како техника за спречување на пристрасност во фазата на развој беше избран пристап заснован на дрво. Беа тестирани и невронски мрежи со цел да се демонстрира како структурата на архитектурата треба да се промени во случај на зголемен обем на податоци. Бидејќи добиените резултати укажаа на присуство на пристрасност дури и во случаи кои не може да се објаснат како влијание на технологијата или човечкиот фактор, неопходно беше да се провери дали резултатите се последица на мали разлики во возраста, полот и дијагнозата при прием помеѓу испитуваните етнички групи, и последователно како последен чекор за ублажување на пристрасноста беше извршена анализа на збунувачки променливи.

7 Заклучок

Како што примената на ВИ во медицината и за здравствена заштита расте, така се појавува зголемено внимание кон ефектот што алгоритмите го имаат врз недоволно застапената, маргинализираната или обесправената популација. Медицинската опрема се покажа како несоодветна за пациенти со потемна боја на кожа, јавно достапните податочни множества опфаќаат значително помал број на пациенти кај одредени етничко-расни групи во корист на темелна застапеност на пациенти од бела раса, па дури и алгоритмите за ВИ кои имаат примена во идентификација на сложени здравствени проблеми знаат да пропагираат етничко-расни разлики. Сето ова доведе до повик за поголема транспарентност кај алгоритмите од Сенатот на САД.

Иако е во иста насока, ова истражување оди подалеку од повикот за транспарентност кај алгоритмите и потемелно испитување на пристрасноста пред моделот да биде споделен за активна употреба, така што истражувањето цели да привлече внимание кон неочекуваните извори на пристрасност и потенцијалната штета со оглед на нивната сеприсутна употреба во клиничкото одлучување. Како средство за постигнување на оваа хипотеза, во докторската дисертација, главното истражување разгледува два алгоритми за МУ тренирани и валидирани за нивната способност во идентификување на етничко-расна припадност од мерењата на виталните знаци земени во првите 24 часа по приемот на возрасни пациенти во ЕИИ низ САД. Резултатите покажаа присуство на расна пристрасност во виталните знаци, особено во мерењата на заситеноста со кислород, а понатамошната анализа покажа дека дури и агрегирање на измерените вредности не ја ублажува пристрасноста создадена од мерењата од пулсен оксиметар. За да се ублажи пристрасноста создадена со пулсен оксиметар во резултатите, беше спроведена дополнителна анализа (каде што беа отстранети неинвазивни мерења на заситеноста со кислород), која покажа дека етничко-расната пристрасност постои дури и во инвазивни мерења на овие параметри, што претставува значаен проблем со оглед на широката употреба на мерења на виталните знаци во клиничкото одлучување воопшто, а особено, бидејќи истражуваните податоци се од особен интерес при генерирање на идни чекори во интервенциите опасни по живот и често се користат во градењето апликации базирани на ВИ. Покрај тоа, предложениот пристап, која е еден од производите на истражувањето, претставува рамка за анализа на пристрасност, со чекори што треба да се следат и потенцијални конфликти кои треба да се избегнат, и како таква може да се користи како алатка за ублажување на пристрасноста на тој начин што ќе врши анализа на податоците за маркери на пристрасност по секој чекор од процесот на развој на моделот. Затоа, пристапот овозможува детекција на пристрасност за различни податоци и нуди единствена можност да се идентификуваат изворите на пристрасност преку шема на исклучување на карактеристики и SHAP анализа.

Настрана од овој главен придонес, беа очекувани и постигнати дополнителни споредни придонеси, со тоа што истражувањето обезбедува класификација, анализа и импликации на различни форми и извори на пристрасност во медицинскиот амбиент, компаративна анализа на јавно достапните медицински податочни множества и етничко-расната застапеност и транспарентност во нив, но и детална анализа на можните етничко-расни разлики кои може да се забележат во избраните податоци. Со пристап што ја открива самопријавената етничко-расна припадност од виталните знаци на пациентите, дури и кога пристрасноста е неочекувана или не треба да постои и по контролирање на сите веќе познати фактори на етничко-расна пристрасност, постои потенцијал за применливост во широк опсег, а придобивките од моделите се сосема очигледни така што пристапот ќе им овозможи на истражувачите да го оценат квалитетот на нивните пристапи, и на тој начин да ги намалат ако не и да ги искоренат пристрасните медицински податочните множества и/или пристрасните апликации базирани на ВИ.

Со оглед на тоа што пристрасноста влијае на населението на глобално ниво, и со оглед на застрашувачкиот ефект што пристрасноста го има врз маргинализираните популации и квалитетот на грижа што им се пружа и нивниот живот, истражувањето за етничко-расна пристрасност е далеку од завршено. Иако ова истражување обезбедува добро заокружена тема така што дава транспарентен

преглед во динамичен и пренослив пристап за анализа на пристрасност, сè уште постојат аспекти за подобрување и проширување на истата. Ограничени од податоците, т.е. етничко-расните припадности вклучени во податочни множества и достапните приеми на пациенти, разгледани и споредени беа само три етничко-расни припадности. Покрај тоа, споредбите во кохортите каде една етничко-расна група беа Хиспанците опфатија помалку од 1.000 пациенти, и затоа, анализата ќе има корист од пообемна популација на пациенти. Дополнително, би било интересно и корисно да се анализираат дополнителни етничко-расни припадности и да се разберат имплицитните предрасуди и разлики меѓу нив. Двете податочни множества се собрани во САД, што е едно ограничување на студијата; имено, би било од голем интерес да се набљудува дали истите заклучоци може да се изведат од податоци кои не се базирани во САД.

Постојат неколку ограничувања на пристапот, кои се должат на потребата да се сосредоточи и контролира обемот на студијата, па затоа беа разгледани само ограничен број мерења на виталните знаци, поради што анализата може дополнително да се прошири на лабораториски тестови и потенцијалната етничко-расна пристрасност во нив. Анализата опфати четири основни статистички карактеристики од виталните знаци, кои може да се заменат или прошират со поспецифични карактеристики предложени од медицински експерти или добиени од алатки за екстракција на карактеристики; или, целиот процес на агрегација на виталните знаци во карактеристики може да се замени со анализа на цели временски серии, т.е. сите мерења по редоследот во кој биле мерени.

Иако етничко-расната припадност во медицината се поистоветува со генетското потекло, сепак двете доловуваат различни информации, бидејќи етничко-расната припадност е општествен концепт, додека генетското потекло е биолошки концепт. Од таа причина, како дел од идна работа може да се размисли за примена на развиениот пристап во откривање на генетско потекло врз основа на витални знаци и споредба на резултатите со резултати добиени од идентификација на етничко-расна припадност, што дополнително ќе ги илустрира разликите меѓу двата концепти.

Идната работа може да се фокусира на промени во процесот на развој на моделот. Во ова истражување, пациентите кај кои недостасуваше информација или само едно мерење на барем еден од разгледуваните витални знаци беа целосно отстранети од студијата. Меѓутоа, понекогаш недостатокот на податоци е исто така информативен за алгоритмите за МУ, поради што може да се размисли за алтернативен пристап кој би користел замена на вредностите што недостасуваат со -1, без отфрлање на пациенти од финалните кохорти. Ова ќе обезбеди можност да се разгледаат поголеми групи на пациенти од сите етничко-расни групи, што потенцијално би довело до зголемување на соодветните групи за пациентите Хиспанци или примена на пристапот и за Азијци и домородни Американци. Овој пристап, сепак, има скриени недостатоци кои можат да создадат пристрасност, па затоа бара внимателно разгледување на сите чекори и темелен, временски исцрпен пристап. Понатаму, споредбите извршени во ова истражување беа бинарни класификации, поради што може да се земе предвид повеќекласната споредба за сите етничко-расни припадности, иако резултатите веројатно нема да бидат подобри.

Референци

- [1] Park, C.W., Seo, S.W.: Artificial Intelligence in Health Care: Current Applications and Issues. *J Korean Med Sci* 35(42), 379 (2020).
- [2] Sutton, R.T., Pincock, D., Baumgart, D.C. et al.: An overview of clinical decision support systems: benefits, risks, and strategies for success. *Digit. Med.* 3(1), 17 (2020).
- [3] Oren, O., et al.: Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *The Lancet Digital Health* 2(9), 486–488 (2020).
- [4] Khetrapal, S., Bhatia, R.: Impact of COVID-19 pandemic on health system & Sustainable Development Goal 3. *Indian J Med Res.* 151(5), 395–399 (2020).
- [5] Phillips, M.C., Sarff, L., Banerjee, J., et al.: Effect of mortality from COVID-19 on inpatient outcomes. *J Med Virol.* 94(1), 318–326 (2022).
- [6] Aweimer, A., Petschulat, L., Jettkant, B. et al. Mortality rates of severe COVID-19-related respiratory failure with and without extracorporeal membrane oxygenation in the Middle Ruhr Region of Germany. *Sci Rep* 13(1), 5143 (2023).
- [7] Bottan, N., Hoffmann, B., Vera-Cossio, D.A.: The Unequal Impact of the Coronavirus Pandemic: Evidence from Seventeen Developing Countries. IDB Working Paper Series No. IDB-WP-1150, (2020).
- [8] Mathieu, E., Ritchie, H., Rod  s-Guirao, L., et al.: Coronavirus Pandemic (COVID-19). (2020). Retrieved from: <https://ourworldindata.org/coronavirus>.
- [9] Mohammed, M., Sha'aban, A., Jatau, A.I., et al.: Assessment of COVID-19 Information Overload Among the General Public. *J Racial Ethn Health Disparities* 9(1), 184–192 (2022).
- [10] Fawzy, A., Wu, T.D., Wang, K., et al.: Racial and Ethnic Discrepancy in Pulse Oximetry and Delayed Identification of Treatment Eligibility Among Patients With COVID-19. *JAMA Intern Med* 182(7), 730–738 (2022).
- [11] Kaur, D., Uslu, S., Rittichier, K. J., Durrezi, A.: Trustworthy Artificial Intelligence: A Review. *ACM Comput. Surv.* 55(2), 1–38 (2022).
- [12] European Commission. Ethics Guidelines for Trustworthy AI. (2018). Retrieved from: <https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>.
- [13] Henry, P.J.: Institutional Bias. In: *The SAGE Handbook of Prejudice, Stereotyping and Discrimination*, 426–440 (2013).
- [14] Green, A.R., Carney, D.R., Pallin, D.J., et al.: Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *J Gen Intern Med.* 22(9), 1231–1238 (2007).
- [15] van Ryn, M., Burke, J.: The effect of patient race and socio-economic status on physicians' perceptions of patients. *Soc Sci Med.* 50(6), 813–828 (2000).
- [16] Mayo, R.M., Sherrill, W.W., Sundareswaran, P., Crew, L.: Attitudes and perceptions of Hispanic patients and health care providers in the treatment of Hispanic patients: a review of the literature. *Hisp Health Care Int.* 5(2), 64–72 (2007).
- [17] Hoffman, K.M., Trawalter, S., Axt, J.R., Oliver, M.N.: Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proc Natl Acad Sci* 113(16), 4296–4301 (2016).
- [18] Rambachan, A., Noorulhuda, H., Fang, M.C., et al.: Pain Assessment Disparities by Race, Ethnicity, and Language in Adult Hospitalized Patients. *Pain Manag Nurs.* 24(4), 393–399 (2023).
- [19] Cooper-Patrick, L., Gallo, J.J., Gonzales, J.J., et al.: Race, gender, and partnership in the patient-physician relationship. *JAMA* 282(6), 583–589 (1999).
- [20] Hamel, L., Lopes, L., Mu  ana, C., et al.: The Undeclared Survey on Race and Health. KFF (2020). Retrieved from: <https://andscape.com/features/new-poll-shows-black-americans-put-far-less-trust-in-doctors-and-hospitals-than-white-people/>.
- [21] Amutah, C., Greenidge, K., Mante, A., et al.: Misrepresenting Race – The Role of Medical Schools in Propagating Physician Bias. *New England Journal of Medicine* 384(9), 872–878 (2021).
- [22] Aysola, J.: How Medical Schools Perpetuate Physician Bias. (2021). Retrieved from: <https://ldi.upenn.edu/our-work/research-updates/how-medical-schools-perpetuate-physician-bias/>
- [23] Ly, D.P.: Black-white differences in the clinical manifestations and timing of initial Lyme disease diagnoses. *J GEN INTERN MED* 37(10), 2597–2600 (2022).
- [24] Adekun, A., Onyekaba, G., Lipoff, J. B.: Skin color in dermatology textbooks: An updated evaluation and analysis. *Journal of the American Academy of Dermatology* 84(1), 194–196 (2021).

- [25] Vyas, D. A., Eisenstein, L. G., Jones, D. S.: Hidden in Plain Sight – Reconsidering the Use of Race Correction in Clinical Algorithms. *New England Journal of Medicine* 383(9), 874 (2020).
- [26] Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464), 447–453 (2019).
- [27] Kleinberg, G., Diaz, M.J., Batchu, S., Lucke-Wold, B.: Racial underrepresentation in dermatological datasets leads to biased machine learning models and inequitable healthcare. *J Biomed Res (Middlet)* 3(1), 42–47 (2022).
- [28] Lester, J.C., Taylor, S.C., Chren, M.M.: Under-representation of skin of colour in dermatology images: not just an educational issue. *Br J Dermatol* 180(6), 1521–1522 (2019).
- [29] Da Kumar, Y., Koul, A., Singla, R., Ijaz, M.F.: Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *J Ambient Intell Humaniz Comput* 14(7), 1–28 (2023).
- [30] Johnson, K.B., Wei, W.Q., Weeraratne, D., Frisse, M.E., Misulis, K., Rhee, K., Zhao, J., Snowdon, J.L.: Precision Medicine, AI, and the Future of Personalized Health Care. *Clin Transl Sci* 14(1), 86–93 (2021).
- [31] Rayan, Z., Alfonse, M., Salem, A.B.M.: Machine Learning Approaches in Smart Health. *Procedia Computer Science* 154, 361–368 (2019).
- [32] Weissler, E.H., Naumann, T., Andersson, T. et al.: The role of machine learning in clinical research: transforming the future of evidence generation. *Trials* 22, 537 (2021).
- [33] Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., Tekade, R.K.: Artificial intelligence in drug discovery and development. *Drug Discov Today* 26(1), 80–93 (2021).
- [34] Ghassemi, M.: Exploring Healthy Models in ML for Health. In: *AI for Healthcare Equity Conference, AI & Health*. MIT (2021).
- [35] Norori, N., Hu, Q., Aellen, F.M., Faraci, F.D., Tzovara, A.: Addressing bias in big data and AI for health care: A call for open science. *Patterns* 2(10), 100347 (2021).
- [36] Brooks, K.C.: A piece of my mind. A silent curriculum. *JAMA* 313(19), 1909–1910 (2015).
- [37] 2019 National Healthcare Quality and Disparities Report. Rockville, MD. Agency for Healthcare Research and Quality. AHRQ Pub. No. 20(21)-0045-EF (2020).
- [38] Fiscella, K., Franks, P., Gold, M.R., Clancy, C.M.: Inequality in quality: addressing socioeconomic, racial, and ethnic disparities in health care. *JAMA* 283(19), 2579–2584 (2000).
- [39] 2013 National Healthcare Disparities Report. Rockville, MD. Agency for Healthcare Research and Quality, US Dept of Health and Human Services. AGRQ Pub. No. 14-0006 (2014).
- [40] Dovidio, J.F., Eggly, S., Albrecht, T.L., Hagiwara, N., Penner, L.: Racial biases in medicine and healthcare disparities. *TPM-Testing, Psychometrics, Methodology in Applied Psychology* 23(4), 489–510 (2016).
- [41] Samulowitz, A., Gremyr, I., Eriksson, E., Hensing, G.: "Brave Men" and "Emotional Women": A Theory-Guided Literature Review on Gender Bias in Health Care and Gendered Norms towards Patients with Chronic Pain. *Pain Res Manag*, 6358624 (2018).
- [42] Casey, L.S., Reisner, S.L., Findling, M.G., Blendon, R.J., Benson, J.M., Sayde, J.M., Miller, C.: Discrimination in the United States: Experiences of lesbian, gay, bisexual, transgender, and queer Americans. *Health Serv Res* 54 (Suppl 2), 1454–1466 (2019).
- [43] Ben-Harush, A., Shiovitz-Ezra, S., Doron, I., Alon, S., Leibovitz, A., Golander, H., Haron, Y., Ayalon, L.: Ageism among physicians, nurses, and social workers: findings from a qualitative study. *Eur J Ageing* 14(1), 39–48 (2016).
- [44] VanPuymbrouck, L., Friedman, C., Feldner, H.: Explicit and Implicit Disability Attitudes of Healthcare Providers. *Rehabilitation Psychology* 65 (2), 101–112 (2020).
- [45] Phelan, S.M., Burgess, D.J., Yeazel, M.W., Hellerstedt, W.L., Griffin, J.M., van Ryn M.: Impact of weight bias and stigma on quality of care and outcomes for patients with obesity. *Obes Rev* 16(4), 319–326 (2015).
- [46] Arpey, N.C., Gaglioti, A.H., Rosenbaum, M.E.: How Socioeconomic Status Affects Patient Perceptions of Health Care: A Qualitative Study. *Journal of Primary Care & Community Health* 8(3), 169–175 (2017).
- [47] Bavli, I., Jones, D.S.: Race Correction and the X-Ray Machine — The Controversy over Increased Radiation Doses for Black Americans in 1968. *New England Journal of Medicine* 387(10), 947–952 (2022).
- [48] Bhavani S.V., Wiley Z., Verhoef P.A., Coopersmith C.M., Ofotokun I.: Racial Differences in Detection of Fever Using Temporal vs Oral Temperature Measurements in Hospitalized Patients. *JAMA* 328(9), 885–886 (2022).
- [49] Sjoding, M.W., Dickson, R.P., Iwashyna, T.J., Gay, S.E., Valley, T.S.: Racial Bias in Pulse Oximetry Measurement. *New England Journal of Medicine* 383(25), 2477–2478 (2020).
- [50] Tasci, E., Zhuge, Y., Camphausen, K., Krauze, A.V.: Bias and Class Imbalance in Oncologic Data-Towards Inclusive and Transferrable AI in Large Scale Oncology Data Sets. *Cancers (Basel)* 14(12), 2897 (2022).

- [51] Sun, M., Oliwa, T., Peek, M.E., Tung, E.L.: Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record. *Health Affairs* 41(2), 203–211 (2022).
- [52] Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences* 117(23), 12592–12594 (2020).
- [53] Ferryman, K., Pitcan, M.: Fairness in Precision Medicine. (2018).
- [54] ISDIS, <https://isdis.net/>, last accessed 2023/03/06
- [55] Adamson, A.S., Smith, A.: Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatol* 154(11), 1247–1248 (2018).
- [56] Kaushal, A., Altman, R., Langlotz, C.: Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *JAMA* 324(12), 1212–1213 (2020).
- [57] Sourlos, N., Wang, J., Nagaraj, Y., van Ooijen, P., Vliegenthart, R.: Possible Bias in Supervised Deep Learning Algorithms for CT Lung Nodule Detection and Classification. *Cancers (Basel)* 14(16), 3867 (2022).
- [58] Gordon, M., Di Bartolo, I.M.: Using Race with Caution in the ASCVD Calculator. *Am Fam Physician* 104(3), 292–294 (2021).
- [59] Taylor, L.M.: Race-Based Prediction in Pregnancy Algorithm Is Damaging to Maternal Health. (2021).
- [60] Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, IY., Ghassemi, M.: CheXclusion: Fairness gaps in deep chest X-ray classifiers. *Pac Symp Biocomput* 26, 232–243 (2021).
- [61] Wong, A., et al.: External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Intern Med* 181(8), 1065–1070 (2021).
- [62] Lutz, R.: Incident Number 110. in McGregor, S. (ed.) *Artificial Intelligence Incident Database*. Responsible AI Collaborative. Retrieved on February 16, 2023 from incidentdatabase.ai/cite/110.
- [63] Juhn, Y.J., Ryu, E., Wi, C.I., King, K.S., Malik, M., Romero-Brufau, S., Weng, C., Sohn, S., Sharp, R.R., Halamka, J.D.: Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the HOUSES index. *J Am Med Inform Assoc* 29(7), 1142–1151 (2022).
- [64] Fraser, K.C., Meltzer, J.A., Rudzicz, F.: Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *J Alzheimers Dis* 49(2), 407–422 (2016).
- [65] Yi, P.H., Wei, J., Kim, T.K., et al.: Radiology "forensics": determination of age and sex from chest radiographs using deep learning. *Emerg Radiol* 28(5), 949–954 (2021).
- [66] Eng, D.K., Khandwala, N.B., Long, J., et al.: Artificial Intelligence Algorithm Improves Radiologist Performance in Skeletal Age Assessment: A Prospective Multicenter Randomized Controlled Trial. *Radiology* 301(3), 692–699 (2021).
- [67] Rim, T.H., Lee, G., Kim, Y., et al.: Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms. *Lancet Digit Health* 2(10), e526–e536 (2020).
- [68] Munk, M.R., Kurmann, T., Márquez-Neila, P., Zinkernagel, M.S., Wolf, S., Sznitman, R.: Assessment of patient specific information in the wild on fundus photography and optical coherence tomography. *Sci Rep* 11(1), 8621 (2021).
- [69] Poplin, R., Varadarajan, A.V., Blumer, K., et al.: Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* 2(3), 158–164 (2018).
- [70] Gichoya, J.W., et al.: AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health* 4(6), e406–e414 (2022).
- [71] Hammaad, A., Ming, Y.Y., Kenrick, C., et al.: Write It Like You See It: Detectable Differences in Clinical Notes by Race Lead to Differential Model Recommendations. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*, 7–21 (2022).
- [72] Hochberg, M.C.: Racial differences in bone strength. *Trans Am Clin Climatol Assoc* 118, 305–315 (2007).
- [73] DiGangi, E.A., Bethard, J.D.: Uncloaking a Lost Cause: Decolonizing ancestry estimation in the United States. *Am J Phys Anthropol* 175(2), 422–436 (2021).
- [74] Ousley, S., Jantz, R., Freid, D.: Understanding race and human variation: why forensic anthropologists are good at identifying race. *Am J Phys Anthropol* 139(1), 68–76 (2009).
- [75] Zengin, A., Pye, S. R., Cook, M. J., et al.: Ethnic differences in bone geometry between White, Black and South Asian men in the UK. *Bone* 91, 180–185 (2016).
- [76] BelleMo, V., Lim, Z.W., Lim, G., et al.: Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *Lancet Digit Health* 1(1), 35–44 (2019).
- [77] Kim, H.E., Kim, H.H., Han, B.K., et al.: Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health* 2(3), 138–148 (2020).

- [78] Thorsen-Meyer, H.C., Nielsen, A.B., Nielsen, A.P., et al.: Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digit Health* 2(4), 179–191 (2020).
- [79] Gupta, A., Galinkin, E.: Green Lighting ML: Confidentiality, Integrity, and Availability of Machine Learning Systems in Deployment. Preprint abs/2007.04693, (2020).
- [80] Scannapieco, M., Catarci, T.: Data quality under a computer science perspective. *Computer Engineering* 2(2), 1–12 (2002).
- [81] Haug, A., Zachariassen, F., van Liempd, D.: The costs of poor data quality. *Journal of Industrial Engineering and Management* 4(2), 168–193 (2011).
- [82] Barocas, S., Hardt, M., Narayanan, A.: *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org (2019).
- [83] Velichkovska, B., Denkovski, D., Gjoreski, H., Kalendar, M., Osmani, V.: A Survey of Bias in Healthcare: Pitfalls of Using Biased Datasets and Applications. In: Silhavy, R., Silhavy, P. (eds) *Artificial Intelligence Application in Networks and Systems. CSOC 2023. Lecture Notes in Networks and Systems*, vol 724. Springer, Cham. (2023).
- [84] Hall, W.J., Chapman, M. V., Lee, K.M., et al.: Implicit Racial/Ethnic Bias Among Health Care Professionals and Its Influence on Health Care Outcomes: A Systematic Review. *Am J Public Health* 105(12), 60–76 (2015).
- [85] Vyas, D.A., Eisenstein, L.G., Jones, D.S.: Hidden in Plain Sight – Reconsidering the Use of Race Correction in Clinical Algorithms. *New England Journal of Medicine* 383(9), 874–882 (2020).
- [86] Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care; Smedley, B.D., Stith, A.Y., Nelson, A.R., editors.: *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. (2003).
- [87] Wang, R.Y., Storey, V.C., Firth, C.P.: A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering* 7(4), 623–640 (1995).
- [88] Komorowski, M., Marshall, D., Saliccioli, J., Crutain, Y.: Exploratory Data Analysis. In: *Secondary Analysis of Electronic Health Records*, 10.1007/978-3-319-43742-2_15 (2016).
- [89] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., et al.: UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med* 12(3), e1001779 (2015).
- [90] Pollard, T., Johnson, A., Raffa, J., et al.: The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data* 5(1), 180178 (2018).
- [91] Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data, <https://wwwn.cdc.gov/nchs/nhanes>, last accessed 2023/04/01
- [92] McCarty, C.A., Chisholm, R.L., Chute, C.G., et al.: The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 4(1), 13 (2011).
- [93] ICNARC, <https://www.icnarc.org/>, last accessed 2023/03/06
- [94] Johnson, A., Pollard, T., Shen, L., et al.: MIMIC-III, a freely accessible critical care database. *Sci Data* 3(1), 160035 (2016).
- [95] CERNER Health Facts, <https://uthsc.edu/cbmi/data/cerner.php>, last accessed 2023/04/01
- [96] Johnson, A.E.W., Bulgarelli, L., Shen, L., et al.: MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* 10(1), 1 (2023).
- [97] Goldberger, A.L., Amaral, L.A., Glass, L., et al.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101(23), e215–220 (2000).
- [98] Belmont Report, <https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html>, last accessed 2023/04/02
- [99] HIPAA, <https://www.cdc.gov/phlp/publications/topic/hipaa.html>, last accessed 2023/04/02
- [100] El Emam, K., Jonker, E., Arbuckle, L., Malin, B.: A Systematic Review of Re-Identification Attacks on Health Data. *PLoS ONE* 6(12), e28071 (2011).
- [101] eICU-CRD Online Documentation, <https://eicu-crd.mit.edu/>, last accessed 2023/04/02
- [102] eICU-CRD Repository, <https://github.com/mit-lcp/eicu-code>, last accessed 2023/04/02
- [103] MIMIC-III Online Documentation, <https://mimic.mit.edu/docs/gettingstarted/>, last accessed 2023/04/02
- [104] MIMIC-III Repository, <https://github.com/MIT-LCP/mimic-code>, last accessed 2023/04/02
- [105] PostgreSQL, <https://www.postgresql.org/>, last accessed 2023/04/02
- [106] Zimmerman, J.E., Kramer, A.A., McNair, D.S., Malila, F.M.: Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit Care Med* 34(5), 1297–1310 (2006).

- [107] Johnson, A.E., Kramer, A.A., Clifford, G.D.: A new severity of illness scale using a subset of Acute Physiology And Chronic Health Evaluation data elements shows comparable predictive accuracy. *Crit Care Med* 41(7), 1711–1718 (2013).
- [108] Shamambo, L.J., Henry, T.L.: Rethinking the Use of "Caucasian" in Clinical Language and Curricula: a Trainee's Call to Action. *J Gen Intern Med* 37(7), 1780–1782 (2022).
- [109] Smith, G., Nielsen, M.: ABC of intensive care. Criteria for admission. *BMJ* 318(7197), 1544–1547 (1999).
- [110] Virani, S.S., Alonso, A., Benjamin, E.J., et al.: Heart disease and stroke statistics—2020 update. *Circulation* 141(9), e139–e596 (2020).
- [111] Glynn, P., Lloyd-Jones, D.M., Feinstein, M.J., Carnethon, M., Khan, S.S.: Disparities in Cardiovascular Mortality Related to Heart Failure in the United States. *J Am Coll Cardiol* 73(18), 2354–2355 (2019).
- [112] Ziaieian, B., Kominski, G.F., Ong, M.K., et al.: National differences in trends for heart failure hospitalizations by sex and race/ethnicity. *Circ Cardiovasc Qual Outcomes* 10(7), e003552 (2017).
- [113] Bibbins-Domingo, K., Pletcher, M.J., Lin, F., et al.: Racial Differences in Incident Heart Failure among Young Adults. *New England Journal of Medicine* 360(12), 1179–1190 (2009).
- [114] von Elm, E., Altman, D.G., Egger, M., et al.: STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol* 61(4), 344–349 (2008).
- [115] Sarvi, F., Heuss, M., Aliannejadi, et al.: Understanding and Mitigating the Effect of Outliers in Fair Ranking. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 861–869 (2022).
- [116] Lemnaru, C., Potolea, R.: Imbalanced Classification Problems: Systematic Study, Issues and Best Practices. *Lecture Notes in Business Information Processing* 102, 35–50 (2012).
- [117] Weiss, G.: Mining with rarity: A unifying framework. *SIGKDD Explorations* 6(1), 7–19 (2004).
- [118] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16, 321–357 (2002).
- [119] Cunningham, P., Delany, S.: k-Nearest Neighbour Classifiers – 2nd Edition. Preprint arXiv:2004.04523, (2020).
- [120] He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks*, 1322–1328 (2008).
- [121] Hassanat, A., Tarawneh, A., Altarawneh, G.: Stop Oversampling for Class Imbalance Learning: A Critical Review. Preprint 10.21203/rs.3.rs-1336037/v1 (2022).
- [122] Elor, Y., Averbuch-Elor, H.: To SMOTE, or not to SMOTE?. Preprint arXiv:2201.08528, (2022).
- [123] Fixing Imbalanced Datasets, <https://medium.com/@ruinian/an-introduction-to-adasyn-with-code-1383a5ece7aa>, last accessed 2023/06/18
- [124] Brandt, J., Lanzén, E.: A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification. (2020).
- [125] Ramirez-Santana, M.: Limitations and Biases in Cohort Studies. In: *Cohort Studies in Health Sciences*. IntechOpen (2018).
- [126] Kotsiantis, S., Pintelas, P.: Mixture of expert agents for handling imbalanced data sets. *Annals of Mathematics, Computing and Teleinformatics* 1, 46–55 (2004).
- [127] Kotsiantis, S., Kanellopoulos, D., Pintelas, P.: Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* 30, 25–36 (2005).
- [128] Velichkovska, B., Cholakovska, A., Atanasovski, V.: Machine Learning Based Classification of IoT Traffic. *Radioengineering* 32(3), 256–263 (2023).
- [129] Wongvorachan, T., He, S., Bulut, O.: A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information* 14(1), 54 (2023).
- [130] Observational study, <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/observational-study>, last accessed 2023/06/18
- [131] Song, J.W., Chung, K.C.: Observational studies: cohort and case-control studies. *Plast Reconstr Surg* 126(6), 2234–2242 (2010).
- [132] Setia, M.S.: Methodology Series Module 3: Cross-sectional Studies. *Indian J Dermatol* 61(3), 261–264 (2016).
- [133] Iwagami, M., Shinozaki, T.: Introduction to Matching in Case-Control and Cohort Studies. *Annals of Clinical Epidemiology* 4(2), 33–40 (2022).
- [134] Kennedy, E.H., Sjölander, A., Small, D.S.: Semiparametric causal inference in matched cohort studies. *Biometrika* 102(3), 739–746 (2015).

- [135] Graaf, M.A., Jager, K.J., Zoccali, C., Dekker, F.W.: Matching, an Appealing Method to Avoid Confounding?. *Nephron Clinical Practice* 118(4), 315–318 (2011).
- [136] Velichkovska, B., Gjoreski, H., Denkovski, D., Kalendar, M., Celi, L.A., Osmani, V.: Investigating Presence of Ethnoracial Bias in Clinical Data Using Machine Learning. *Proceedings of XV International Conference ETAI* 2(1), 148–152 (2021).
- [137] Duffy, G., Clarke, S.L., Christensen, M., et al.: Confounders mediate AI prediction of demographics in medical imaging. *NPJ Digit Med.* 5(1), 188–193 (2022).
- [138] Velichkovska, B., Gjoreski, H., Denkovski, D., Kalendar, M., Mamandipoor, B., Celi, L.A., Osmani, V.: Vital signs as a source of racial bias. Preprint 10.1101/2022.02.03.22270291, (2022).
- [139] Velichkovska, B., Gjoreski, H., Denkovski, D., Kalendar, M., Mullan, I.D., Gichoya, J.W., Martinez, N., Celi, L.A., Osmani, V.: AI learns racial information from the values of vital signs. *Science Advances, Journal of the American Association for the Advancement of Science (AAAS)*, under second review (2023).
- [140] Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47, 583–621 (1952).
- [141] Sah, S.: Machine Learning: A Review of Learning Types. Preprint 10.20944/preprints202007.0230.v1 (2020).
- [142] Abro, A.A., Khan, A.A., Talpur, M.S.H., Kayijuka, I., Yaşar, E.: Machine Learning Classifiers: A Brief Primer. *University of Sindh Journal of Information and Communication Technology* 5(2), 63–68 (2021).
- [143] Tan, H.: Machine Learning Algorithm for Classification. *J. Phys.: Conf. Ser.* (1994).
- [144] Peng, J., Lee, K., Ingersoll, G.: An Introduction to Logistic Regression Analysis and Reporting. *Journal of Educational Research* 96, 3–14 (2002).
- [145] Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 785–794 (2016).
- [146] XGBoost Documentation, <https://xgboost.readthedocs.io/en/latest/index.html>, last accessed 2023/06/18
- [147] Taylor's theorem, https://mathinsight.org/taylors_theorem_multivariable_introduction, last accessed 2023/06/18
- [148] XGBoost overfitting, <https://www.capitalone.com/tech/machine-learning/how-to-control-your-xgboost-model/>, last accessed 2023/06/18
- [149] Yang, L., Shami, A.: On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice. *Neurocomputing* 415(1), 295–316 (2020).
- [150] Randomised search for cross-validation available in the Scikit Learn library: RandomizedSearchCV, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html, last accessed 2023/06/18
- [151] Wang, H., Zheng, H.: Model Validation, Machine Learning. In: *Encyclopedia of Systems Biology*. Springer (2013).
- [152] Lyles, R. H., Zhang, F., Drews-Botsch, C.: Combining Internal and External Validation Data to Correct for Exposure Misclassification: A Case Study. *Epidemiology* 18(3), 321–328 (2007).
- [153] Shung, D., Simonov, M., Gentry, M., Au, B., et al.: Machine Learning to Predict Outcomes in Patients with Acute Gastrointestinal Bleeding: A Systematic Review. *Dig Dis Sci* 64(8), 2078–2087 (2019).
- [154] Maleki, F., Muthukrishnan, N., Ovens, K., et al.: Machine Learning Algorithm Validation. *Neuroimaging Clinics of North America* 30(4), 433–445 (2020).
- [155] Hands-On Tutorial on Performance Measure of Stratified K-Fold Cross-Validation, <https://analyticsindiamag.com/hands-on-tutorial-on-performance-measure-of-stratified-k-fold-cross-validation/>, last accessed 2023/06/18
- [156] Hossin, M., Sulaiman, M.N.: A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process* 5(2), 1–11 (2015).
- [157] Introduction to Confusion Matrix, <https://towardsai.net/p/l/introduction-to-confusion-matrix>, last accessed 2023/06/18
- [158] Understanding AUC-ROC Curve, <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>, last accessed 2023/06/18
- [159] Arrieta, A.B., Díaz-Rodríguez, N., del Ser, J., Bennetot, A., et al.: Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* 58, 82–115 (2020).
- [160] Lundberg, S.M. Lee, S.: A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777 (2017).

- [161] Ribeiro, M., Singh, S., Guestrin, C.: “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 97–101 (2016).
- [162] SHAP library, <https://shap.readthedocs.io/en/latest/index.html>, last accessed 2023/06/18
- [163] Centers for Disease Control and Prevention (CDC): Disparities in premature deaths from heart disease—50 States and the District of Columbia. *MMWR Morb Mortal Wkly Rep.* 53(6), 121–125 (2001).
- [164] Hill, L.K., Hu, D.D., Koenig, J., et al.: Ethnic differences in resting heart rate variability: a systematic review and meta-analysis. *Psychosom Med.* 77(1), 16–25 (2015).
- [165] Vaz Fragoso, C.A., McAvay, G., Gill, T.M., et al.: Ethnic differences in respiratory impairment. *Thorax* 69(1), 55–62 (2014).
- [166] Valbuena, V.S.M., Merchant, R.M., Hough, C.L.: Racial and Ethnic Bias in Pulse Oximetry and Clinical Outcomes. *JAMA Intern Med.* 182(7), 699–700 (2022).
- [167] Storozh O.V., Ashraf, M.A., Smeltz, A.M.: Postoperative Hypertension: When Blood Pressure Cuff and Arterial Line Disagree. *Journal of Clinical Anesthesia and Pain Management* 3(1), 53–60 (2019).
- [168] Wax, D.B., Lin, H.M., Leibowitz, A.B.: Invasive and concomitant noninvasive intraoperative blood pressure monitoring: observed differences in measurements and associated therapeutic interventions. *Anesthesiology* 115(5), 973–978 (2011).
- [169] Palatini, P., Parati, G.: Blood pressure measurement in very obese patients: a challenging problem. *J Hypertens.* 29(3), 425–429 (2011).
- [170] Fawzy, A., Wu, T.D., Wang, K., et al.: Racial and Ethnic Discrepancy in Pulse Oximetry and Delayed Identification of Treatment Eligibility Among Patients With COVID-19. *JAMA Intern Med.* 182(7), 730–738 (2022).
- [171] Braun, L., Wolfgang, M., Dickersin, K.: Defining race/ethnicity and explaining difference in research studies on lung function. *Eur Respir J.* 41(6), 1362–1370 (2013).
- [172] Rossiter, C.E., Weill, H.: Ethnic Differences in Lung Function: evidence for proportional differences. *International Journal of Epidemiology* 3(1), 55–61 (1974).
- [173] Hegewald, M.J., Crapo, R.O.: Socioeconomic status and lung function. *Chest* 132(5), 1608–1614 (2007).
- [174] Christ, M., Braun, N., Neuffer, J., Kempa-Liehr, A.W.: Time Series Feature Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing* 307(1), 72–77 (2018).
- [175] Cooper, R.J., Schriger, D.L., Flaherty, H.L., Lin, E.J., Hubbell, K.A.: Effect of vital signs on triage decisions. *Ann Emerg Med.* 39(3), 223–232 (2002).
- [176] Ramachandran, G., Prasad, C.H.R.K., Garre, S., Sundar A.S.: Oxygen Management in Heart Failure Patients. *Indian Journal of Clinical Cardiology* 3(3), 150–156 (2022).
- [177] Why Data Scale Size Matters When It Comes to Improving Deep Learning Model Stability, <https://opendatascience.com/why-data-scale-size-matters-when-it-comes-to-improving-deep-learning-model-stability/>, last accessed 2023/07/03
- [178] Himmelman, A., Svensson, A., Hansson, L.: Influence of sex on blood pressure and left ventricular mass in adolescents: the Hypertension in Pregnancy Offspring Study. *J Hum Hypertens* 8(7), 485–490 (1994).
- [179] Stamler, J., Stamler, R., Riedlinger, W.F., et al.: Hypertension screening of 1 million Americans. Community Hypertension Evaluation Clinic (CHEC) program, 1973 through 1975. *JAMA* 235(21), 2299–2306 (1976).
- [180] Strumpf, E.C.: Racial/Ethnic Disparities in Primary Care. *Med Care* 49(5), 496–503 (2011).
- [181] Malhotra, J., Rotter, D., Tsui, J., et al.: Impact of Patient-Provider Race, Ethnicity, and Gender Concordance on Cancer Screening: Findings from Medical Expenditure Panel Survey. *Cancer Epidemiol Biomarkers Prev.* 26(12), 1804–1811 (2017).
- [182] Greenwood, B.N., Hardeman, R.R., Huang, L., Sojourner A.: Physician–patient racial concordance and disparities in birthing mortality for newborns. *Proceedings of the National Academy of Sciences* 117(35), 21194–21200 (2020).
- [183] Terrell, K.M., Hui, S.L., Castelluccio, P., et al.: Analgesic Prescribing for Patients Who Are Discharged from an Emergency Department. *Pain Medicine* 11(7), 1072–1077 (2010).
- [184] Todd, K.H., Deaton, C., D’Adamo, A.P., Goe, L.: Ethnicity and analgesic practice. *Ann Emerg Med* 35(1), 11–16 (2000).
- [185] Lee, H.H., Lewis, C.W., McKinney, C.M.: Disparities in Emergency Department Pain Treatment for Toothache. *JDR Clin Trans Res* 1(3), 226–233 (2016).
- [186] Langham, J., Holmes, S., Figueroa, J., et al.: Physician self-identified race and opioid prescription practices in upper extremity injuries in the pediatric emergency department. *Heliyon* 9(2), e13351 (2023).

- [187] Johnson, T.J., Winger, D.G., Hickey, R.W., et al.: Comparison of Physician Implicit Racial Bias Toward Adults Versus Children. *Acad Pediatr.* 17(2), 120–126 (2017).
- [188] Okunlola, O.E., Lipnick, M.S., Batchelder, P.B., et al.: Pulse Oximeter Performance, Racial Inequity, and the Work Ahead. *Respir Care* 67(2), 252–257 (2022).
- [189] Wong, A.I., Charpignon, M., Kim, H., et al.: Analysis of Discrepancies Between Pulse Oximetry and Arterial Oxygen Saturation Measurements by Race and Ethnicity and Association With Organ Dysfunction and Mortality. *JAMA Netw Open* 4(11), e2131674 (2021).
- [190] Sarkar, R., Martin, C., Mattie, H., et al.: Performance of intensive care unit severity scoring systems across different ethnicities in the USA: a retrospective observational study. *Lancet Digit Health* 3(4), e241–e249 (2021).
- [191] Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. *Science* (1979) 366(6464), 447–453 (2019).
- [192] EL-Manzalawy, Y., Abbas, M., Hoaglund, I. et al.: OASIS +: leveraging machine learning to improve the prognostic accuracy of OASIS severity score for predicting in-hospital mortality. *BMC Med Inform Decis Mak* 21(1), 156 (2021).
- [193] Kostick-Quenet, K.M., Cohen, I.G., Gerke, S., et al.: Mitigating Racial Bias in Machine Learning. *Journal of Law, Medicine & Ethics* 50(1), 92–100 (2022).
- [194] Ioannidis, J.P.A., Powe, N.R., Yancy, C.: Recalibrating the use of race in medical research. *JAMA - Journal of the American Medical Association* 325(7), 623–624 (2021).
- [195] Gervasi, S.S., Chen, I.Y., Smith-McLallen, A., et al.: The Potential For Bias In Machine Learning And Opportunities For Health Insurers To Address It. *Health Aff.* 41(2), 212–218 (2022).
- [196] Cooper, R., David, R.: The Biological Concept of Race and Its Application to Public Health and Epidemiology. *J Health Polit Policy Law* 11(1), 97–116 (1986).