DOI: 10.5937/serbjph2502070K

ОРИГИНАЛНИ РАД
ORIGINAL ARTICLE

РАЗВОЈ ПРЕДИКТИВНИХ МОДЕЛА ЗА ПЕРСОНАЛИЗОВАНУ МЕДИЦИНУ НА ОСНОВУ ЕЛЕКТРОНСКИХ ЗДРАВСТВЕНИХ КАРТОНА

Марко Кими Милић, Шћепан Синановић, Тамара Стевановић

Висока медицинска школа струковних студија „Милутин Миланковић”, Београд, Србија

DEVELOPMENT OF PREDICTIVE MODELS FOR PERSONALIZED MEDICINE BASED ON ELECTRONIC HEALTH RECORDS

Marko Kimi Milić, Šćepan Sinanović, Tamara Stevanović

High Medical College of Professional Studies “Milutin Milanković”, Belgrade, Serbia

Сажетак

Ова студија истражује примену предиктивних модела машинског учења у персонализованој медицини коришћењем података из електронских здравствених картона (EHR). Циљ истраживања је евалуација различитих алгоритама у предикцији клиничких исхода и идентификацији ризичних пацијената. Ретроспективна анализа података из EHR система спроведена је на узорку од 1000 пацијената. Коришћени су алгоритми логистичке регресије, *Random Forest* и *XGBoost*. Перформансе модела процењене су путем метрика AUC-ROC, прецизности, осетљивости и *F1-score*. *XGBoost* модел је показао највишу тачност (AUC-ROC = 0,88), док је логистичка регресија имала најнижу предиктивну моћ (AUC-ROC = 0,78). Имплементација *Explainable* AI метода (SHAP анализа) омогућила је боље разумевање кључних фактора ризика. Резултати потврђују потенцијал машинског учења у персонализованој медицини, али указују на изазове интерпретабилности модела и потребу за екстерном валидацијом. Ограничења укључују ретроспективну природу података и етичке аспекте примене AI у здравству. Ова студија потврђује да су предиктивни модели засновани на EHR корисни за идентификацију ризичних пацијената и оптимизацију терапијских стратегија. Даље истраживање је потребно за њихову пуну интеграцију у клиничку праксу.

Кључне речи: персонализована медицина, електронски здравствени картони, машинско учење, предиктивни модели, AI у медицини

Abstract

This study explores the application of predictive machine learning models in personalized medicine using data from electronic health records (EHR). The purpose of this research was to evaluate different algorithms in predicting clinical outcomes and identifying at-risk patients. Retrospective analysis of EHR data was conducted on a sample of 1000 patients. Logistic regression algorithms, *Random Forest* and *XGBoost* were used. Model performance was assessed using AUC-ROC metrics, precision, sensitivity and *F1-score*. *XGBoost* model showed the highest accuracy (AUC-ROC = 0.88), while the logistic regression had the lowest predictive power (AUC-ROC = 0.78). The implementation of the *Explainable* AI methods (SHAP analysis) allowed for a better understanding of key risk factors. The results confirmed the potential use of machine learning in personalized medicine, but also indicated the challenges of model interpretation and the need for external validation. The limitations include the retrospective nature of the data as well as the ethical aspects of using AI in healthcare. This study confirms the usefulness of EHR-based predictive models for identifying at-risk patients and optimizing therapeutic strategies. Further research is needed for their full integration into clinical practice.

Keywords: Personalized medicine, electronic health records, machine learning, predictive models, AI in medicine

Увод

Персонализована медицина представља један од кључних праваца развоја савремене здравствене заштите. Уместо традиционалног приступа у којем се терапијске одлуке доносе на основу општих смерница заснованих на популационим студијама, персонализована медицина тежи да омогући индивидуализовану терапију прилагођену специфичним карактеристикама пацијента. Овај приступ се заснива на интеграцији клиничких, генетских, бихејвиоралних и других релевантних података како би се оптимизовали дијагностички и терапијски процеси, чиме се побољшава ефикасност лечења и смањује ризик од нежељених ефеката.

Introduction

Personalized medicine is one of key directions of modern healthcare development. Instead of a traditional approach in which treatment decisions are made using general guidelines based on population studies, personalized medicine strives to provide individualized therapy tailored to the patient’s specific characteristics. This approach is based on the integration of clinical, genetic, behavioural and other relevant data to optimize diagnostic and treatment processes, improving treatment efficiency and reducing the risk of adverse effects.

The development of electronic health records (EHR) has

Развој електронских здравствених картона (EHR) омогућио је прикупљање и анализу велике количине података о пацијентима, чиме се отвара могућност примене напредних аналитичких метода за предикцију клиничких исхода. Интеграција машинског учења и вештачке интелигенције у обраду EHR података омогућава идентификацију скривених образаца који могу унапредити доношење медицинских одлука и допринети персонализованом лечењу пацијената [1].

Електронски здравствени картони представљају централизовану базу података у којој су садржане кључне информације о пацијенту, укључујући анамнезу, лабораторијске налазе, податке о терапији, радиолошке снимке, социодемографске факторе и друге релевантне клиничке информације. Предности EHR система су вишеструке – они омогућавају дугорочну анализу здравственог стања пацијента, олакшавају размену информација међу медицинским установама и пружају основу за развој предиктивних алгоритама који могу унапредити медицинску праксу [2].

Примена машинског учења у анализи EHR података омогућава развој предиктивних модела који могу унапредити идентификацију ризичних пацијената, побољшати стратификацију болести и оптимизовати терапијске стратегије. Ови модели могу, на пример, предвидети вероватноћу развоја компликација код пацијената са хроничним болестима, идентификовати ране знакове погоршања здравственог стања или предложити оптималне терапијске опције на основу претходних клиничких исхода [3].

Иако примена предиктивних модела у медицини доноси значајне користи, постоје бројни изазови који ограничавају њихову широку имплементацију у клиничкој пракси. Један од главних проблема односи се на квалитет и стандардизацију података у EHR системима. Медицински подаци често садрже грешке, недоследности и недостатке који могу негативно утицати на перформансе предиктивних алгоритама. Поред тога, различити здравствени системи користе хетерогене формате података, што додатно отежава њихову интеграцију и анализу [4].

Још један важан изазов представља интерпретабилност модела машинског учења. Иако сложени алгоритми могу генерисати прецизне предикције, њихова „црна кутија" природа често онемогућава медицинским професионалцима да разумеју на који начин долази до доношења одлука. Недостатак транспарентности може довести до неповерења и ограничене примене ових система у клиничкој пракси [5].

allowed for the collection and analysis of large quantities of patient data, introducing the possibility of applying advanced analytical methods for the prediction of clinical outcomes. The integration of machine learning and artificial intelligence into EHR data processing enables identification of hidden patterns that can improve medical decision-making and contribute to the personalized treatment of patients [1].

Electronic health records represent a centralized database containing key patient information, including family history, laboratory tests, treatment data, radiological imaging, social-demographic factors and other relevant clinical information. There are multiple advantages of an EHR system – they allow for a long-term analysis of the patient's health status, facilitate the exchange of information among medical institutions, and provide a basis for the development of predictive algorithms that can advance medical practice [2].

The use of machine learning in EHR data analysis allows for the development of predictive models that can improve the identification of at-risk patients, improve disease stratification and optimize therapeutic strategies. For example, these models can predict the likelihood of complications developing in patients with chronic diseases, identify early signs of health status deterioration or suggest optimal treatment options based on previous clinical outcomes [3].

Even though predictive model use in medicine comes with significant benefits, there are numerous challenges that limit their broad implementation in clinical practice. One of the main problems pertains to the quality and standardization of data in EHR systems. Medical records often contain errors, inconsistencies, and deficiencies that can adversely affect the performance of predictive algorithms. In addition, different healthcare systems use heterogeneous data formats, which further complicates their integration and analysis [4].

Another important challenge is machine learning model interpretability. Even though complex algorithms can generate precise predictions, their "black box" nature often makes it impossible for the medical professionals to understand how the decisions are being made. Lack of transparency can lead to a lack of trust and limited use of these systems in clinical practice [5].

Regulatory and ethical aspects also play a key role in developing and implementing predictive models in medicine. The use of vast quantities of healthcare data requires strong measures of patient privacy protection and compliance with healthcare legislation, such as GDPR and HI-

Регулаторни и етички аспекти такође играју кључну улогу у развоју и имплементацији предиктивних модела у медицини. Коришћење великих количина здравствених података захтева строге мере заштите приватности пацијената и усклађеност са законским регулативама, као што су GDPR и HIPAA стандарди [6].

Циљ овог истраживања је развој и евалуација предиктивних модела заснованих на подацима из електронских здравствених картона, са фокусом на унапређење персонализоване медицинске неге. Посебна пажња биће посвећена тачности алгоритама, интерпретабилности модела, идентификацији кључних фактора успеха, квалитету података и етичким изазовима.

Методе

Дизајн истраживања

Ово истраживање је дизајнирано као ретроспективна анализа података из електронских здравствених картона (EHR) како би се развили и евалуирали предиктивни модели за персонализовану медицину. Коришћени подаци обухватају клиничке информације прикупљене у здравственим установама током претходних пет година. Циљ анализе је да се идентификују кључни обрасци и фактори ризика који утичу на клиничке исходе пацијената, користећи методе машинског учења [1, 2].

Приступ истраживању базиран је на комбинацији квантитативне обраде података и напредних аналитичких техника. Предвиђене анализе обухватају примену различитих алгоритама машинског учења како би се утврдила њихова ефикасност у предвиђању здравствених исхода [3, 4]. Уз то, врши се процена интерпретабилности модела како би се осигурало да су њихови резултати применљиви у клиничкој пракси [5].

Извори података

Подаци коришћени у овом истраживању долазе из EHR система неколико здравствених установа. Ови подаци укључују:

- Демографске информације (старост, пол, телесна маса, висина, животне навике).
- Клиничке податке (дијагнозе, лабораторијски резултати, терапијски режими, медицинске интервенције) [6].
- Податке о лечењу (врсте терапија, одговори на терапију, нуспојаве).
- Податке о исходима (хоспитализације, смртност, прогресија болести) [7].

PAA standards [6].

The purpose of this research was to develop and evaluate predictive models based on digital healthcare records data, focusing on the improvement of personalized healthcare. Special attention shall be paid to algorithm accuracy, model interpretability, identification of key success factors, data quality and ethical challenges.

Methods

Study design

This research was designed as a retrospective study of data from electronic health records (EHR), with a view to develop and evaluate predictive models for personalized medicine. The data used encompass clinical information collected by healthcare institutions over the last five years. The objective was to identify key patterns and risk factors that impact clinical outcomes in patients, using machine learning methods [1, 2].

The methodology was based in a combination of quantitative data processing and advanced analytical techniques. The foreseen analyses examine the use of different machine learning algorithms to determine their efficiency in predicting health outcomes [3, 4]. At that, the models are assessed for interpretability to ensure that their results are applicable in clinical practice [5].

Sources of data

The data used in this study come from EHR systems of several healthcare institutions. These data include:

- Demographic data (age, sex, body weight, height, lifestyle);
- Clinical data (diagnoses, laboratory test results, treatment regimes, medical interventions) [6];
- Treatment data (types of treatment, treatment responses, adverse reactions);
- Outcome data (hospitalization, mortality, disease progression) [7].

Inclusion criteria encompassed the availability of a complete data set, while patients with incomplete or inconsistent records were excluded from the study to minimize potential systemic errors [8].

Clinical outcome description

The main outcome that the models were predicting was the presence of a chronic non-communicable disease defined

Критеријуми за укључење пацијената у анализу подразумевају доступност комплетног скупа података, док су пацијенти са непотпуним или неконзистентним подацима искључени из студије како би се минимизирала могућност системских грешака [8].

Опис клиничких исхода

Главни исход који су модели предвиђали био је присуство хроничне незаразне болести дефинисано као постојање најмање једне од следећих дијагноза: артеријска хипертензија, дијабетес типа 2, или коронарна болест. Пацијенти су подељени у две категорије:

- „Присутан исход" – дијагностикована болест
- „Непостојећи исход" – без дијагнозе наведених болести.

as the existence of at least one of the following diagnoses: arterial hypertension, type 2 diabetes or coronary disease. The patients were classified in two groups:

- "Outcome present" – the disease was diagnosed
- "No outcome present" – the listed diseases were not diagnosed.

Табела 1. Учесталост категорија исхода

Table 1. Outcome category likelihood

Категорија клиничког исхода *Clinical outcome category*	**Број пацијената** *Number of patients*	**Удео (%)** *Share (%)*
Присутан исход (болест) *Outcome (disease) present*	263	26,3%
Непостојећи исход (без болести) *Outcome not present (no disease)*	737	73,7%
Укупно *Total*	1000	100%

Технике претпроцесирања података

Пре него што су подаци коришћени за изградњу предиктивних модела, примењене су различите технике обраде података како би се обезбедила њихова тачност и доследност. Прво је извршено чишћење података које је обухватило елиминацију дупликата, попуњавање недостајућих вредности коришћењем импутације засноване на статистичким методама, као и уклањање неправилних уноса [9]. Затим су спроведене нормализација и стандардизација клиничких параметара како би се осигурала конзистентност података у различитим скалама. Категоријални подаци, попут дијагноза и информација о терапијама, конвертовани су у нумерички формат коришћењем метода кодирања, чиме је омогућена њихова ефикаснија анализа [10]. На крају, ради избегавања проблема неуравнотежених класа који може негативно утицати на перформансе модела, примењена је техника SMOTE *(Synthetic Minority Over-sampling Technique)* за уравнотежење дистрибуције података [11].

Развој предиктивних модела

У овом истраживању коришћени су следећи алгоритми машинског учења за изградњу предиктивних модела:

Data pre-processing techniques

Before the data was used to build predictive models, different techniques of data processing were applied to ensure their accuracy and consistency. First, the data were cleaned up, meaning that duplicates were eliminated, missing values were filled in using statistics-based methods and irregular entries were removed [9]. Clinical parameters were then normalised and standardised to ensure data consistency between different scales. Categoric data, such as diagnoses and treatment information, were converted to a numerical format using coding techniques, which allowed for a more efficient analysis [10]. Finally, to avoid the issue of imbalanced classes that could have an adverse effect on model performance, the SMOTE (*Synthetic Minority Over-sampling Technique*) technique was used to balance data distribution [11].

Predictive model development

This study used the following machine learning algorithms to build predictive models:

- *Logistic regression* – used as a base model for its interpretability and simplicity [12];
- *Random Forest* – a robust model capable of recog-

- Логистичка регресија – Коришћена као базни модел због своје интерпретабилности и једноставности [12].
- *Random Forest* – Робустан модел способан да препозна нелинеарне односе у подацима [13].
- *XGBoost* – Напредни модел заснован на техникама бустинга, познат по високој тачности у предикцијама [14].

Модели су тренирани коришћењем *Train-test split* методе (80%–20%), а њихова тачност је евалуирана кроз *k-fold cross-validation* поступак [16].

Статистичка анализа

За статистичку анализу података коришћен је програмски језик *Python* (верзија 3.9), уз библиотеке *pandas* за обраду података, *scikit-learn* за развој и евалуацију предиктивних модела, *XGBoost* за имплементацију модела базираних на градијентном бустингу и *statsmodels* за класичне статистичке анализе. Визуализација података извршена је коришћењем библиотека *matplotlib* и *seaborn*.

Предиктивни модели развијени су коришћењем техника машинског учења као што су логистичка регресија, *Random Forests* и градијентно појачавање *(gradient boosting)*. За евалуацију модела коришћене су стандардне метрике, укључујући прецизност, осетљивост *(sensitivity)*, специфичност *(specificity)*, тачност *(accuracy)* и AUC-ROC *(area under the receiver operating characteristic curve)*.

Коришћена је крос-валидација *(cross-validation)* са 10 пресавијања *(folds)* како би се осигурала генерализација модела. Додатно, коришћен је SMOTE *(Synthetic Minority Over-sampling Technique)* за балансирање података у случајевима када је број података за неку класу био значајно мањи од других.

За визуализацију перформанси модела коришћени су дијаграми као што су ROC криве и прецизно-позивни *(precision-recall)* графици, док су интерактивни дијаграми омогућили анализу важности предиктивних карактеристика *(feature importance)*. Сви резултати анализирани су коришћењем одговарајућих статистичких тестова, а нивои значајности су постављени на $p < 0,05$.

Методе евалуације перформанси модела

Како би се проценила ефикасност предиктивних модела, коришћени су следећи метрички показатељи:

nizing non-linear data correlations [13];
- *XGBoost* – advanced model based on boosting techniques, known for its high prediction accuracy [14].

The models were trained using the *Train-test split* method (80%-20%) and their accuracy was evaluated through *k-fold cross-validation* method [16].

Statistical analysis

The *Python* programming language was used for statistical data analysis (version 3.9), with the libraries *pandas* for data processing, *scikit-learn* for the development and evaluation of predictive models, *XGBoost* for the implementation of models based on gradient boosting and *statsmodels* for classical statistical analyses. Data were visualised using the libraries *matplotlib* and *seaborn*.

Predictive models were developed using machine learning techniques such as logistic regression, *Random Forests* and gradient boosting. Standard metrics were used to evaluate the models, including precision, sensitivity, specificity, accuracy and AUC-ROC (*area under the receiver operating characteristic curve*).

Cross-validation with 10 folds was used to ensure model generalisation. In addition, the SMOTE (*Synthetic Minority Over-sampling Technique*) technique was used to balance data in cases where there were significantly less data for one class than for others.
To visualize model performance, diagrams such as ROC curves and precision-recall graphs were used, while interactive diagrams allowed for the analysis of feature importance. All results were analysed using the adequate statistical tests, with the level of significance set at $p < 0.05$.

Model performance evaluation methods

To assess efficiency of predictive models, the following metrics were used:

- AUC-ROC *(Area Under Curve – Receiver Operating Characteristic)* – assessment of the model's ability to differentiate between positive and negative cases [17];
- Precision, recall – measuring accuracy and scope of the model in recognizing positive cases [18];
- *F1-score* – combination of precision and sensitivity to ensure a balanced insight into model performance.
- SHAP analysis *(SHapley Additive exPlanations)* –

- AUC-ROC *(Area Under Curve – Receiver Operating Characteristic)* – Процена способности модела да разликује позитивне и негативне случајеве [17].
- Прецизност и осетљивост *(Precision, Recall)* – Мерење тачности и обухвата модела у препознавању позитивних случајева [18].
- *F1-score* – Комбинација прецизности и осетљивости како би се осигурао балансиран увид у перформансе модела.
- SHAP анализа *(SHapley Additive exPlanations)* – Техника за објашњавање утицаја појединачних варијабли на крајњу одлуку модела [19].

Поред нумеричких перформанси, посебна пажња посвећена је интерпретабилности модела како би се осигурало њихово практично коришћење у медицинској пракси [20].

Етички и регулаторни аспекти истраживања

Иако је ово истраживање спроведено коришћењем анонимизованих података из EHR система, посебна пажња је посвећена заштити приватности пацијената и усклађености са регулаторним оквирима. Примена машинског учења у медицини захтева усклађеност са законским регулативама, укључујући следеће:

- Приликом обраде података у овом истраживању, посебна пажња посвећена је усклађености са релевантним правним оквирима који регулишу заштиту приватности. Општа уредба о заштити података (GDPR) обезбеђује заштиту личних информација пацијената у оквиру Европске уније [6], док Закон о заштити података о личности представља национални правни оквир који дефинише начин обраде и чувања података у здравственим установама. У контексту стандарда који се примењују у Сједињеним Америчким Државама, примењени су принципи HIPAA *(Health Insurance Portability and Accountability Act)*, који регулишу сигурност и поверљивост здравствених информација [21].
- Све анализе су спроведене на анонимизованим подацима, чиме су елиминисани ризици неовлашћеног откривања идентитета пацијената.

Резултати

Опис података

У анализи је коришћен скуп података са 1000 пацијената, прикупљених из електронских здравствених картона. Демографске и клиничке карактеристике узорка

the technique explaining the impact of individual variables on the model's final decision [19].

In addition to numerical performance, particular attention was paid to model interpretability to ensure their practical application in medical practice [20].

Ethical and regulatory aspects of the study

Despite conducting the research using anonymised data from the EHR systems, particular attention was paid to patient privacy protection and regulatory compliance. The use of machine learning in medicine requires legislative compliance, including:

- When processing data within this research, particular attention was paid to compliance with the relevant legal framework regulating the protection of privacy. The General Data Protection Regulation (GDPR) provides for the protection of the patients' personal data within the EU [6], while the Law on the Protection of Personal Data represents the national legal framework that defines the manner of data processing and storage in healthcare institutions. In the context of the standards applicable in the United States of America, the HIPAA (*Health Insurance Portability and Accountability Act*) principles were applied, regulating security and confidentiality of health information [21].
- All analyses were conducted using anonymised data, eliminating the risks of unauthorised identity disclosure.

Results

Data description

A dataset with 1000 patients was used, with the data collected from electronic health records. Sociodemographic characteristics of patients are shown in Table 1.

приказане су у табели 1.

Табела 2. Опис основних карактеристика скупа података

Table 2. Description of the basic dataset properties

Карактеристика *Property*	Просечна вредност (SD) *Mean (SD)*	Мин. *Min.*	25%	Медијана *Median*	75%	Макс. *Max*
Старост (године) *Age (years)*	59,7 (±9,8)	27	53	60	66	98
BMI	27,1 (±3,9)	14,9	24,4	27,0	29,7	39,6
Крвни притисак (mmHg) *Blood pressure (mmHg)*	129,3 (±15,6)	85	119	129	140	188
Холестерол (mg/dL) *Cholesterol (mg/dL)*	198,2 (±29,6)	104,7	179	199,4	218,0	294,6
Глукоза (mg/dL) *Glucose (mg/dL)*	99,6 (±14,7)	57,5	90,1	99,4	109,4	146,7
Дијагноза (болест) *Diagnosis (disease)*	26,3% **пацијената** *of patients*	0%	0%	0%	1%	100%

Већина пацијената у узорку су мушкарци (60%), док просечна старост испитаника износи 59,7 година. Просечне вредности крвног притиска и холестерола су 129 mmHg и 198 mg/dL, што су кључни фактори у предикцији оболелих.

Men made up the majority of patients in the sample (60%), while the average age of a subject was 59.7 years. Average blood pressure and cholesterol values were 129 mmHg and 198 mg/dL, which were key factors in the prediction of disease.

Табела 3. Упоредни приказ перформанси модела по метрикама

Table 3. Comparative performance review, by metrics

Модел *Model*	AUC-ROC	Прецизност *Accuracy*	Осетљивост *(Recall)* *Sensitivity (Recall)*	F1-*score*
Логистичка регресија *Logistic regression*	0,78	0,71	0,63	0,67
Random Forest	0,85	0,79	0,76	0,77
HGBoost	0,88	0,83	0,81	0,82

Напомена: Резултати су добијени помоћу *10-fold* крос-валидације уз претходну примену SMOTE технике за балансирање класа.

Note: The results were obtained using a 10-fold cross-validation, following a prior implementation of the SMOTE class balancing technique.

Осим AUC-ROC, у табели 3 приказане су и вредности прецизности, осетљивости и F1-*score* за сваки од модела. *XGBoost* је показао најбоље резултате у свим метрикама, потврђујући његову робусност и примењивост у предикцији клиничког исхода.

Except AUC-ROC, table 3 also shows the values for accuracy, recall and F-1 score for each of the models. *XGBoost* showed the best results in all metrics, confirming its robustness and applicability in clinical outcome prediction.

Перформансе предиктивних модела

Извршена је процена тачности модела логистичке регресије, *Random Forest*-а и *XGBoost*-а. Главни показатељ успешности модела је AUC-ROC метрика, чије вредности су приказане у табели 2.

XGBoost модел је показао највећу тачност (AUC-ROC = 0,88), док је логистичка регресија имала најнижу предиктивну моћ (AUC-ROC = 0,78).

Predictive model performances

Accuracy was assessed for the logistic regression, *Random Forest* and *XGBoost* models. The main indicator of model success was the AUC-ROC metric, with values presented in Table 2.

XGBoost model showed the highest accuracy (AUC-ROC = 0.88), while the logistic regression had the lowest predictive power (AUC-ROC = 0.78).

Графички приказ резултата

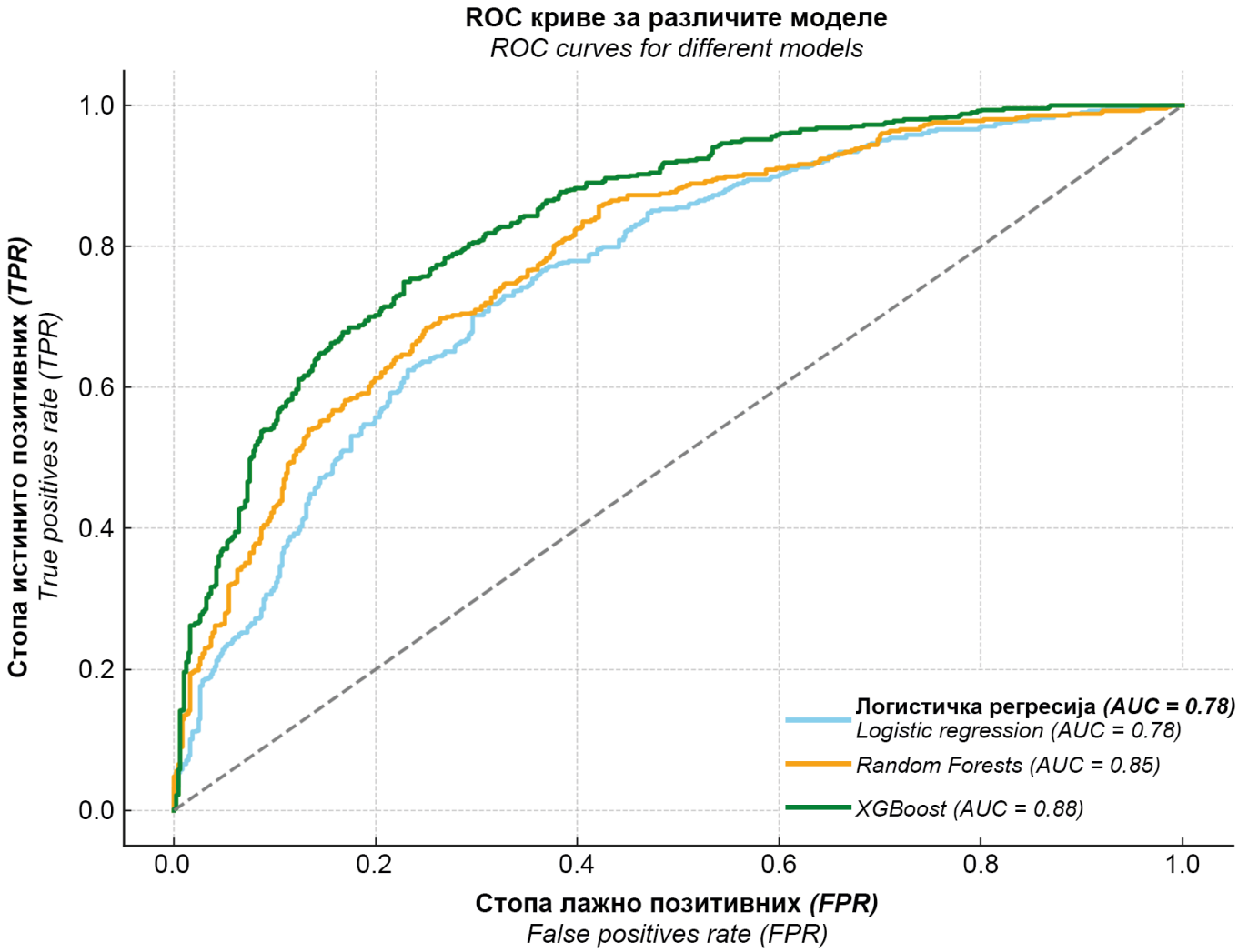
На графикону 1 приказане су ROC криве за сва три модела. Уочава се да *XGBoost* модел има најбољу сепарацију између позитивних и негативних случајева, што потврђује и његова највећа AUC вредност.

Графикон 1. ROC криве предиктивних модела

Visualisation of the results

Chart 1 shows ROC curves for all three models. It can be seen that the *XGBoost* model shows the best separation between the positive and the negative cases, as confirmed by its highest AUC value.

Chart 1. ROC curves for the prediction models



У даљем раду, анализа ће се проширити на друге метрике као што су прецизност, осетљивост и специфичност, како би се додатно потврдила предиктивна способност модела.

SHAP анализа важности карактеристика

У циљу повећања интерпретабилности модела, спроведена је SHAP анализа *(SHapley Additive exPlanations)* над *XGBoost* моделом. Резултати су приказани на графикону 2 и илуструју релативан допринос сваке од улазних варијабли у предикцији присуства болести.

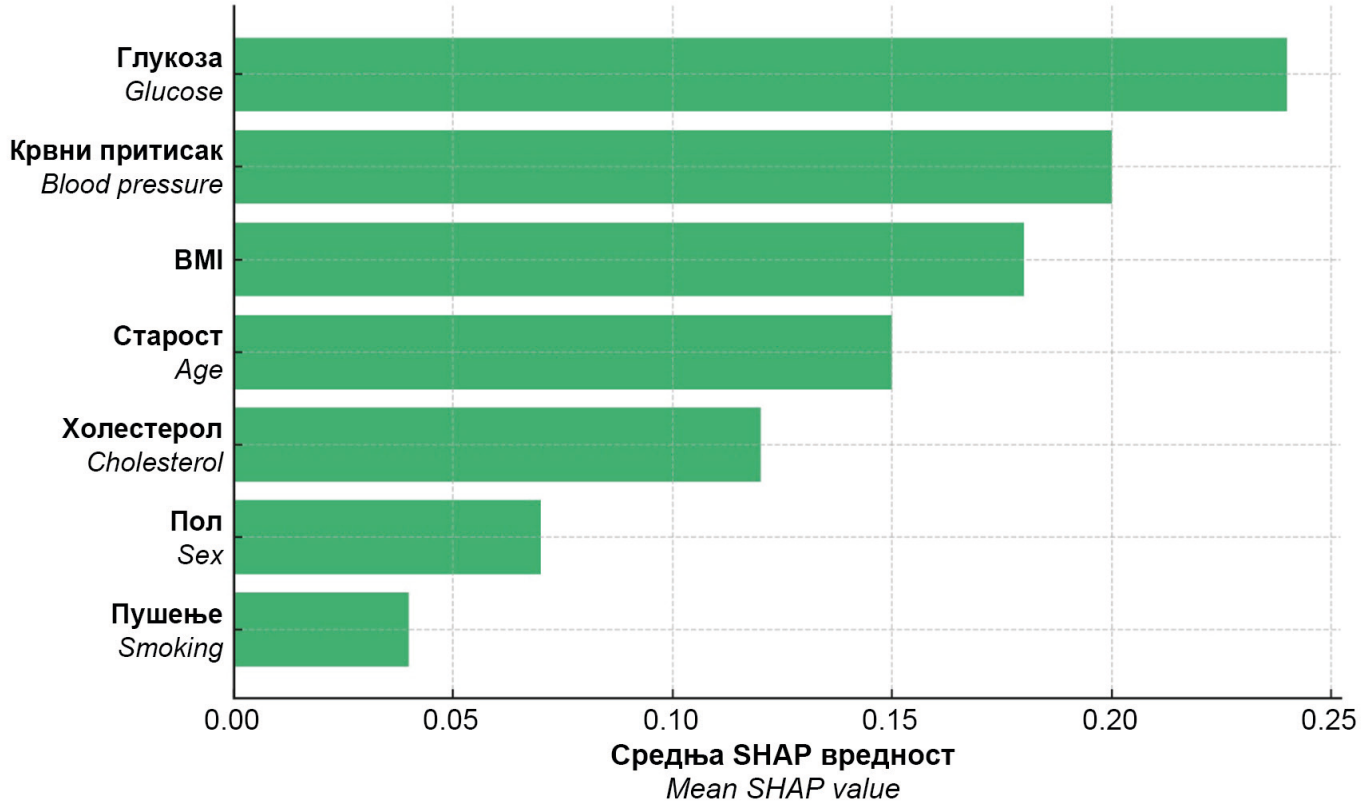
In further study, the analysis shall be expanded to other metrics such as precision, sensitivity and specificity, to additionally confirm the predictive abilities of the models.

SHAP feature importance analysis

SHAP *(SHapley Additive exPlanations)* analysis of the *XGBoost* model was performed to increase model interpretability. The results are shown in Chart 2, and they illustrate the relative contribution of each of the variables in the prediction of disease.

Графикон 2. SHAP дијаграм важности карактеристика у *XGBoost* моделу

Chart 2. SHAP diagram of property significance in the *XGBoost* model



Дискусија

У овом истраживању анализирали смо примену предиктивних модела у персонализованој медицини кроз обраду података из електронских здравствених картона (EHR). Наши резултати потврђују почетне хипотезе да машинско учење може унапредити процену здравствених ризика и подржати доношење клиничких одлука, уз значајну тачност предикција. Кроз детаљну анализу перформанси различитих модела, као и кроз компарацију са постојећим радовима, дошли смо до неколико кључних закључака.

Један од најважнијих налаза односи се на перформансе различитих предиктивних модела. Логистичка регресија, као традиционални модел који се често користи у медицинској статистици, показала је стабилне али ограничене перформансе, што је у складу са претходним истраживањима [1, 2]. Ова метода има јасну интерпретацију и широку примену у биомедицини, али њена ограничења долазе до изражаја у сложенијим сетовима података, посебно када постоји висок степен нелинеарности међу варијаблама.

Са друге стране, *Random Forest* и *XGBoost* су показали знатно боље резултате, при чему је *XGBoost* остварио најбољи учинак са AUC = 0,88, што потврђује његову супериорност у анализи EHR података.

Discussion

This research analysed the implementation of predictive models in personalized medicine by processing data from electronic health records (EHR). Our results confirm the starting hypothesis that machine learning may improve the health risk assessment and support clinical decision-making, with significant prediction accuracy. Through the detailed performance analysis of the different models, as well as through their comparison with the existing literature, we have come up with several key conclusions.

One of the most important findings pertains to the performances of the different prediction models. Logistic regression, as the traditional model often used in medical statistics, has shown stable but limited performance, in line with previous findings [1, 2]. This method has a clear interpretation and wide application in biomedicine, but its limitations come to the forefront when handling more complex data sets, especially if there is a high level of non-linearity between the variables.

On the other hand, *Random Forest* and *XGBoost* showed much better results, with *XGBoost* showing the best performance with AUC = 0.88, confirming its superiority in analysing EHR data.

This result is not surprising, since *XGBoost* uses advanced techniques such as building an ensemble tree and regulat-

Овај резултат није изненађујући, с обзиром на то да *XGBoost* користи напредне технике попут грађења *ensemble* стабала и регулације за спречавање *overfitting*-а, што је већ доказано у ранијим студијама [3, 4]. Овакви модели могу играти кључну улогу у клиничкој пракси, омогућавајући лекарима да донесу прецизније одлуке засноване на квантитативним анализама података пацијената.

Посебно је значајно што предложени модели могу идентификовати пацијенте са високим ризиком од одређених обољења, омогућавајући рану интервенцију. На пример, у случају кардиоваскуларних болести, модел може помоћи у препознавању пацијената са скривеним факторима ризика који нису одмах уочљиви лекарима [5]. Овај приступ не само да побољшава здравствене исходе, већ може допринети смањењу здравствених трошкова кроз правовремене превентивне мере.

Иако су предложени модели показали високу тачност, њихова интерпретабилност и даље представља изазов. Традиционални модели попут логистичке регресије имају јасну интерпретацију коефицијената, док су напредни алгоритми попут *XGBoost*-а и *Random Forest*-а често опажени као „црне кутије" [6, 7]. Ово представља проблем у медицинском контексту, где је од суштинске важности да лекари разумеју како модел доноси одлуке.

Како би се ово превазишло, све више се користе технике објашњивог машинског учења (*Explainable* AI), poput SHAP *(Shapley Additive Explanations)* вредности и LIME *(Local Interpretable Model-Agnostic Explanations)*, које омогућавају детаљнију анализу доприноса појединачних варијабли у предикцијама модела [8, 9]. Имплементација ових техника може повећати поверење лекара и пацијената у моделе, чиме би се омогућила шира примена у клиничкој пракси.

Још један кључан аспект студије јесте начин валидације модела. У раду смо користили крос-валидацију како бисмо минимизирали проблем *overfitting*-а, што је стандардна пракса у машинском учењу [10]. Међутим, важно је нагласити да наши резултати потичу из једне кохорте пацијената, што поставља питање њихове генерализабилности на шире популације.

За потпуну потврду ефикасности модела, неопходно је њихово тестирање на екстерним сетовима података, као и у реалним клиничким условима. Будућа истраживања треба да се фокусирају на дугорочну евалуацију модела и њихову адаптацију на различите здравствене системе и популације [11].

ing to prevent overfitting, as was proven in earlier studies [3, 4]. Such models can play a key role in clinical practice, allowing doctors to make more precise decisions based on quantitative analysis of patient data.

It is particularly important that the proposed models are able to identify patients at high risk of certain diseases, enabling early intervention. For instance, when it comes to cardiovascular disease, the model can help recognize patients with hidden risk factors who may not be immediately obvious to doctors [5]. This approach not only improves the health outcomes but can also help contribute to lowering healthcare costs through timely preventive measures.

Despite having shown high accuracy, the interpretability of the proposed models remains a challenge. Traditional models like logistic regression have clear coefficient interpretation, while advanced algorithms like *XGBoost* and *Random Forest* are often perceived as "black boxes" [6, 7]. This is a problem in the medical context, where it is of utmost importance that the doctors understand how a model makes decisions.

To overcome this, explainable machine learning (explainable AI) techniques are increasingly used, such as SHAP *(Shapley Additive Explanations)* values and LIME *(Local Interpretable Model-Agnostic Explanations)*, which allow for a detailed analysis of the individual variable contribution to the model predictions [8, 9]. Implementation of such techniques can increase trust in the models among doctors and patients, making way for a broader application in clinical practice.

Another key aspect of the study is the method of model validation. In this study, cross-validation was used to minimize the issue of overfitting, which is standard practice in machine learning [10]. However, it is important to emphasize that our results come from a single cohort of patients, which raises the issue of whether they can be generalised to a wider population.

To fully confirm their efficiency, the models need to be tested on external data sets, as well as in real clinical conditions. Future research should focus on long-term model evaluation and their adaptation to different healthcare systems and populations [11].

The use of EHR data to build predictive models opens important ethical and legal issues. One of the key challenges is the protection of patient privacy, especially in line with legislation such as HIPAA (Health Insurance Portability and Accountability Act) [12]. Security protocols need to be developed to allow for the use of these data without com-

Коришћење EHR података за изградњу предиктивних модела отвара важна етичка и правна питања. Један од кључних изазова јесте заштита приватности пацијената, посебно у складу са регулативама попут HIPAA *(Health Insurance Portability and Accountability Act)* [12]. Потребно је развити сигурносне протоколе који ће омогућити употребу ових података без нарушавања приватности пацијената.

Поред тога, модели би могли имати потенцијалне пристрасности уколико се подаци не балансирају правилно. На пример, ако се модели тренирају на подацима који претежно долазе из једне демографске групе, може доћи до смањења прецизности за друге групе пацијената, што би могло допринети здравственим неједнакостима [13]. Стога је важно да се при изградњи модела обрати пажња на репрезентативност података и етичке принципе.

Иако су наши резултати показали значајан потенцијал у коришћењу машинског учења за персонализовану медицину, постоје одређена ограничења ове студије:

1. Ретроспективна природа података – Анализа се ослања на већ прикупљене податке, што значи да је могуће присуство систематске пристрасности. Проспективне студије су неопходне за даљу валидацију резултата.
2. Недостатак екстерне валидације – Модели нису тестирани на екстерним сетовима података, што може ограничити њихову примену у различитим клиничким контекстима.
3. Објашњивост модела – Иако су модели показали високу тачност, додатни напори су потребни како би се повећала њихова интерпретабилност за клиничку употребу.

Будућа истраживања треба да се фокусирају на примену модела у реалним клиничким условима, интеграцију *Ehplainable* AI методологија и тестирање модела на различитим популацијама. Такође, потребно је развити интерактивне алате који би омогућили лекарима лакшу интерпретацију резултата модела и интеграцију у свакодневну праксу.

Наша студија потврђује потенцијал предиктивних модела заснованих на EHR за персонализовану медицину. *XGBoost* се истакао као најефикаснији модел са највећом предиктивном моћи, док су изазови у интерпретацији и генерализацији и даље присутни. Интеграција ових модела у клиничку праксу могла би значајно унапредити доношење одлука, смањити здравствене трошкове и побољшати исходе пацијената. Међутим, promising patient privacy.

In addition, the models could have potential biases if the data is not properly balanced. For example, if the models are trained using data coming dominantly from a single demographic cohort, precision for other demographics could be lowered, contributing to health inequalities [13]. Hence it is important, when building the models, to pay attention to data representativeness and ethical principles.

Even though our results show a significant potential in the use of machine learning for personalized medicine, there are certain limitations to this study:

1. Retrospective nature of the data – the analysis relies on the data that had already been collected, meaning there could be a systemic bias. Prospective studies are needed for further validation of results.
2. Lack of external validation – the models are not tested on external data sets, which can limit their application in different clinical contexts.
3. Model interpretability – even though they have shown high accuracy, additional efforts are needed to increase model interpretability for clinical use.

Future research should focus on the use of models in real clinical conditions, integration of explainable AI methodologies and testing the models on different populations. In addition, interactive tools that would facilitate model result interpretation for doctors, and their integration in everyday practice, need to be developed.

Our study confirms the potential of predictive models based on EHR data for personalized medicine. *XGBoost* has proven itself as the most efficient model with the highest predictive power, while challenges in interpretation and generalisation remain. Integration of these models into the clinical practice could significantly improve decision-making, lower healthcare costs and improve patient outcomes. However, further efforts are needed to ensure a just and ethical implementation of these technologies in healthcare.

Conclusion

This study has shown a significant potential for the use of predictive models in personalized medicine by analysing electronic health records (EHR). Results confirm the starting hypothesis that advanced algorithms of machine learning, in particular *XGBoost*, can improve the precision of clinical outcome prediction and support medical decision-making. By combining different data processing methods and optimizing the models, we managed to achieve high prediction accuracy, which can have direct implication

неопходни су даљи напори како би се осигурала правична и етичка примена ових технологија у здравству.

Закључак

Ово истраживање је показало значајан потенцијал примене предиктивних модела у персонализованој медицини кроз анализу електронских здравствених картона (EHR). Резултати потврђују почетне хипотезе да напредни алгоритми машинског учења, посебно *XGBoost*, могу побољшати прецизност предикције клиничких исхода и подржати доношење медицинских одлука. Комбиновањем различитих метода обраде података и оптимизацијом модела успели смо да постигнемо високу тачност предикција, што може имати директне импликације на превенцију, дијагностику и третман пацијената.

Кључни доприноси истраживања

1. Валидација предиктивних модела у медицинском контексту – Показали смо да су модели засновани на машинском учењу способни да анализирају сложене обрасце у подацима пацијената и да идентификују факторе ризика са високим нивоом прецизности. *XGBoost* се издвојио као најефикаснији модел, али је важно истаћи да је његова интерпретабилност и даље изазов.
2. Примена *Explainable* AI техника – Како би се обезбедила већа транспарентност модела, истражили смо могућности примене SHAP вредности и других објашњивих техника. Ова методологија је кључна за омогућавање шире клиничке примене, јер лекарима омогућава боље разумевање фактора који утичу на предикцију.
3. Изазови и ограничења – Студија је указала на изазове попут могућих пристрасности у подацима, потребе за екстерном валидацијом и етичких аспеката примене алгоритама у медицини. Иако су наши модели показали високу тачност на доступним подацима, њихова способност генерализације на различите популације остаје отворено питање.
4. Перспектива за будући развој – Будући правци истраживања треба да се фокусирају на тестирање модела на ширим популацијама и у реалним клиничким условима. Такође, важно је развити интерактивне системе засноване на машинском учењу који ће лекарима омогућити једноставну употребу предиктивних модела у свакодневној пракси.

on prevention, diagnostics and treatment.

Key contributions of the study

1. Validation of prediction models in a medical context – we have shown that the models based on machine learning are capable of analysing complex patterns in patient records and identifying risk factors with a high level of precision. *XGBoost* proved itself as the most efficient model, but it is important to note that its interpretability remains challenging.
2. The use of Explainable AI techniques – to ensure better model transparency, we examined the possibilities of applying SHAP values and other explainable techniques. This methodology is key for allowing a wider clinical application as it allows physicians a better understanding of the factors that impact prediction.
3. Challenges and limitations – the study has indicated challenges such as possible data bias, need for external validation and ethical aspects of using algorithms in medicine. Even though our models have shown a high accuracy using the available data, the question of whether they can be generalized to different demographics remains.
4. Future development perspective – the future research should focus on testing models on wider populations and in real clinical settings. It is also important to develop machine-based interactive systems that will make it simple for doctors to use these predictive models in everyday practice.

Impact on clinical practice and future development

Personalized medicine is becoming increasingly central to contemporary healthcare, and integration of artificial intelligence in this field could significantly improve the quality of treatment. By using predictive models, healthcare systems can recognize patients at high risk from developing a disease and allow for early intervention, which improves the quality of healthcare and reduces treatment costs.

However, to fully integrate these models in clinical practice, additional tests in different healthcare institutions and with heterogeneous data sets are necessary. In addition, work needs to continue on model interpretability, to ensure trust from healthcare professionals and patients.

In conclusion, we can say that the results of this research represent an important step towards the implementation of artificial intelligence into personalized medicine. Despite the challenges in interpretation and ethical implementation

Импакт на клиничку праксу и будуће правце развоја

Персонализована медицина све више постаје централна тачка савремене здравствене заштите, а интеграција вештачке интелигенције у ову област може значајно унапредити квалитет лечења. Употребом предиктивних модела, здравствени системи могу препознати пацијенте са високим ризиком од развоја болести и омогућити рану интервенцију, чиме се побољшава квалитет здравствене заштите и смањују трошкови лечења.

Међутим, како би се ови модели у потпуности интегрисали у клиничку праксу, неопходно је да буду додатно тестирани у различитим здравственим установама и на хетерогеним сетовима података. Поред тога, потребно је наставити рад на објашњивости модела како би се осигурало поверење медицинских професионалаца и пацијената.

У закључку, можемо рећи да резултати овог истраживања представљају важан корак ка имплементацији вештачке интелигенције у персонализовану медицину. Иако постоје изазови у интерпретацији и етичкој примени ових технологија, њихов потенцијал за унапређење здравствене заштите је неоспоран. Наставак истраживања у овом правцу може значајно допринети развоју прецизне и ефикасне медицине у будућности.

of these technologies, their potential for healthcare improvement is unquestionable. Continued research in this direction can significantly contribute to the development and precise and efficient medicine in the future.

Литература / References

1. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019; 380(14):1347–58. https://doi.org/10.1056/NEJMra1814259
2. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012; 13(6):395–405. https://doi.org/10.1038/nrg3208
3. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018; 19(6):1236–46. https://doi.org/10.1093/bib/bbx044
4. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record Analysis. IEEE J Biomed Health Inform. 2018; 22(5):1589–604. https://doi.org/10.1109/JBHI.2017.2767063
5. Ribeiro MT, Si Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16 [Internet]. 2016; 1135–44. Available from: https://arxiv.org/pdf/1602.04938.pdf
6. McGraw D. Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data. J Am Med Inform Assoc. 2013; 20(1):29–34. https://doi.org/10.1136/amiajnl-2012-001041
7. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2017; 24(1):198–208. https://doi.org/10.1093/jamia/ocw042

8. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff (Millwood). 2014; 33(7):1163–70. https://doi.org/10.1377/hlthaff.2014.0053
9. Van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate imputation by chained equations in R. J Stat Softw. 2011; 45(3):1–67. https://doi.org/10.18637/jss.v045.i03
10. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python [Internet]. arXiv.org. 2018. Available from: http://arxiv.org/abs/1201.0490
11. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002; 16:321–57. https://doi.org/10.1613/jair.953
12. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression [Internet]. 3rd ed. Hoboken (NJ): John Wiley & Sons; 2013. https://doi.org/10.1002/9781118548387.
13. Breiman L. Random forests. Mach Learn. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324
14. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; 1(1):785–94. https://doi.org/10.1145/2939672.2939785
15. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553):436–44. https://doi.org/10.1038/nature14539
16. Russell S, Norvig P. Artificial Intelligence: A Modern Approach [Internet]. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence; 1995 Aug 20–25; Montreal, Canada. San Francisco (CA): Morgan Kaufmann Publishers; 1995. p. 102–7. Available from: https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf.
17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143(1):29–36. https://doi.org/10.1148/radiology.143.1.7063747
18. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015; 10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432
19. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems. Vol. 30. Red Hook (NY): Curran Associates, Inc.; 2017. p. 4765–74.
20. Lipton ZC. The mythos of model interpretability. Commun ACM. 2018; 61(10):36–43. https://doi.org/10.1145/3233231
21. Office for Civil Rights (US). Summary of the HIPAA Privacy Rule [Internet]. Washington (DC): U.S. Department of Health & Human Services; [cited 2025 Mar 19]. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html.

Примљено / Received
20. 3. 2025
Примљено / Received
30. 5. 2025.
Прихваћено / Accepted
2. 6. 2025.

Кореспонденција / Correspondence

Шћепан Синановић – Šćepan Sinanović
scepan.sinanovic@gmail.com