Enhanced Classification Models for Heart Disease Prediction in Precision Medicine

Nataliya Boyko[1], Iryna Dosiak[1]

[1]*Lviv Polytechnic National University, Stepana Bandera Street, 12, Lviv, 79000, Ukraine*

Abstract

This investigation endeavors to scrutinize and enhance machine learning (ML) algorithms tailored for medical diagnostics, specifically in the prognostication of cardiac pathologies. The focal point of this study encompasses medical indicators such as electrocardiograms (ECGs), diagnostic outcomes, and symptomatic manifestations, examining their pivotal role in the efficacious identification of cardiac ailments. The research delves into mathematical data mining models and their classification synergies, constituting the core subject of inquiry.

The scholarly contribution of this work lies in the refinement of ML methodologies, thereby bolstering the precision in cardiac disease diagnostics. From a practical standpoint, the research aspires to forge a sophisticated classification schema, poised to mitigate the incidence of diagnostic inaccuracies.

The discourse underscores the pertinence of ML algorithm deployment in the medical domain, particularly for cardiac diagnostics. A meticulous review and critique of extant scientific literature and analogous studies have been conducted, alongside an evaluation of mathematical and informational frameworks. The study delineates fundamental classifiers and their collective ensembles, with a pronounced focus on discerning the associative patterns between cardiac conditions and various indicators.

The research illuminates the impact of demographic factors, such as age and smoking habits, as well as comorbidities, laboratory metrics, symptomatic expressions, test outcomes, and ECG readings on cardiac diagnostic processes. It also contrasts the symptomatic disparities between genders in both cardiac patients and healthy cohorts. An array of classification methodologies and their ensembles have been employed to facilitate the automated diagnosis of cardiac diseases.

Extensive trials utilizing diverse ML classification techniques have been executed. The investigation extends to feature selection algorithms and data transformation techniques. Each classification approach is critically examined, highlighting its merits and demerits. Consequently, a novel, modified model is introduced, which significantly enhances the quality of classifiers. This model is meticulously calibrated against selected metrics, including the crucial metric of completeness, which bears significant importance in the medical field.

Keywords 1

technology, processing, analysis, machine learning algorithms, classification, diagnosis, heart diseases.

1. Introduction

Cardiovascular diseases remain a leading cause of mortality globally, with their etiology rooted in a multitude of risk factors. Effective management of patient-specific indicators is paramount in mitigating the risk of heart disease. Medical institutions worldwide amass extensive datasets on health-related issues, providing a fertile ground for machine learning (ML) applications to enhance predictive analytics. However, the voluminous and often anomalous nature of this data presents significant

IDDM'2023: 6th International Conference on Informatics & Data-Driven Medicine, November 17 - 19, 2023, Bratislava, Slovakia

EMAIL: nataliya.i.boyko@lpnu.ua (N. Boyko); Iryna.Dosiak.KNM.2018@lpnu.ua (I. Dosiak)
ORCID: 0000-0002-6962-9363 (N. Boyko); https://orcid.org/0000-0002-4687-0541 (I. Dosiak)

analytical challenges. ML techniques, by virtue of their analytical prowess, have emerged as indispensable tools for the accurate prediction of heart disease presence or absence.

The complexity inherent in cardiac diseases necessitates meticulous analysis to prevent adverse outcomes, including premature mortality. Diagnosis typically hinges on a multifaceted interplay of clinical and pathological data, where the linkage between etiological factors and symptomatic expressions may be obscure. Consequently, medical data analysis in healthcare is a critical and formidable task, demanding precision and efficiency.

A prevalent conundrum in diagnostics is the similar symptomatic presentation of disparate diseases, complicating the diagnostic process. Physicians often navigate a non-linear diagnostic pathway, correlating observed symptoms with potential diseases before arriving at a definitive diagnosis. Decision support systems, augmented by expert medical opinion, are thus integral to this process. At its core, medical diagnosis is a classification challenge, where data must be assigned to one of N possible outcomes. Machine learning excels in such classification tasks, employing algorithms such as Convolutional Neural Networks (CNN), Naive Bayes (NB), Decision Trees (DT), Logistic Regression (LR), and others. Despite advancements, the quest for superior models with enhanced predictive capabilities is relentless, given the high stakes of patient outcomes.

In response to this need, a novel two-stage model has been introduced within a hospital management system to refine the prediction of heart disease. This model synergizes logistic regression stacking of Stochastic Gradient Descent (SGD) and XGBoost algorithms at the first level, with a second level comprising NB and DT classifiers aggregated through weighted bagging. Utilizing the Z-Alizadeh Sani dataset, the model leverages user-input risk factors to extract salient features pertinent to heart disease.

Despite ongoing research, the pursuit of more robust ML methodologies to curtail diagnostic inaccuracies remains critical. This work is dedicated to the development of an enhanced classification model for heart disease prediction. To achieve this objective, the following tasks have been delineated: a comprehensive review and analysis of fundamental ML algorithms and prevalent classifier ensembles in medical diagnostics; execution of preliminary data processing and identification of key features influencing heart disease; elucidation of metrics to evaluate classification quality, pinpointing those of utmost relevance to the domain; and the formulation of an advanced classification model predicated on the foundational models' results. Through these endeavors, this research aspires to contribute to the precision medicine paradigm, ultimately safeguarding patient health and preserving life.

2. Literature review

The exigency of enhancing cardiac disease diagnostics has catalyzed a plethora of scientific inquiries, leveraging data mining methodologies to discern the presence of cardiac pathologies, either in isolation or concomitant with other conditions. This corpus of research has employed a diverse array of machine learning (ML) techniques, yielding varying degrees of diagnostic accuracy. The following is a succinct literature review that encapsulates significant contributions in this domain.

Reference [1] delineates a system that amalgamates multiple classification methods through model stacking, utilizing sixty-four echocardiographic features alongside seven clinical indicators. This approach has notably augmented the accuracy of Coronary Heart Disease (CHD) classification from an approximate 70% to 87.7%. The analysis within this study provides critical insights for selecting optimal ML models for diagnostic purposes.

The investigation presented in [2] assesses the utility of clinical and cardiac rhythm data in forecasting patient relapses. A spectrum of classification algorithms was examined, including individual classifiers such as Support Vector Machines (SVM), Classification and Regression Trees (CART), k-nearest neighbors (KNN), and ensemble classifiers, achieving an accuracy of 82%. The findings from this study guide the strategic selection of datasets and pertinent indicators for disease prediction.

Paper [3] introduces a user-friendly tool aimed at primary care settings for the preliminary identification of risk factors associated with aortic aneurysms, employing naive Bayes and KNN algorithms on a dataset comprising 55 patients. The modest accuracy achieved, around 60%, suggests a need for expanding the patient dataset. This publication is a repository of information regarding diverse attributes influencing disease manifestation.

In work [4], the Structured Streaming module of the Apache Spark platform is utilized to construct an ML pipeline for real-time detection of cardiac arrhythmias. The study evaluates the efficacy of decision trees, multi-class random forests, and logistic regression classifiers using data from the MIT database, focusing on classification performance and detection latency.

Research [5] scrutinizes stroke diagnosis through ML methods, involving 52 stroke patients and 80 controls, encompassing a demographic spread. The application of Random Forest, SVM, logistic regression, and KNN models yielded a high classification accuracy of 96.6%. This study is instrumental in identifying diagnostic patterns across age groups and evaluating classification methodologies.

In summation, these scientific advancements in ML offer a gamut of tools for the enhancement of healthcare diagnostics. ML algorithms demonstrate considerable promise in disease diagnosis, particularly of cardiac conditions. Nonetheless, it is imperative to acknowledge that the studies reviewed provide but a snapshot of the capabilities of ML in heart disease prediction. My research is dedicated to unearthing significant features and patterns among the multitude of factors influencing diagnosis, with the ultimate aim of refining the quality assessments of foundational models through innovative modifications.

3. Materials and Methods

The pursuit of accurate heart disease prediction necessitates meticulous data selection and pre-processing to ensure the integrity and applicability of the dataset. This study utilizes the Z-Alizadeh Sani dataset [6], which encompasses 303 patient records, inclusive of 59 attributes across 216 diagnosed with coronary artery disease (CAD) and 88 non-affected individuals. The dataset is categorized into four distinct attribute clusters pertinent to patient diagnosis: demographic details, electrocardiogram (ECG) readings, symptomatic observations, and a range of laboratory and echocardiographic parameters.

To facilitate the analysis, categorical variables were numerically encoded using the Label Encoding method [7], thereby transforming qualitative data into a quantifiable format. An assessment of the target classes revealed an imbalance, with a higher prevalence of CAD instances compared to normal cases. This study also undertook a rigorous examination for missing values, imputing them with mean substitution to maintain dataset continuity.

Further data integrity checks included the identification and removal of duplicates, confirming the dataset's uniqueness. The Z-score method was employed to detect and exclude outliers, ensuring normal distribution of the data. Figure 1 illustrates box plots for age, body mass index (BMI), blood pressure, and lipoprotein levels, with outliers demarcated beyond the whiskers, providing a visual representation of the data's distribution.

This pre-processing pipeline is critical in refining the dataset, thereby laying a robust foundation for subsequent machine learning analysis aimed at enhancing the predictive accuracy of heart disease diagnostics.

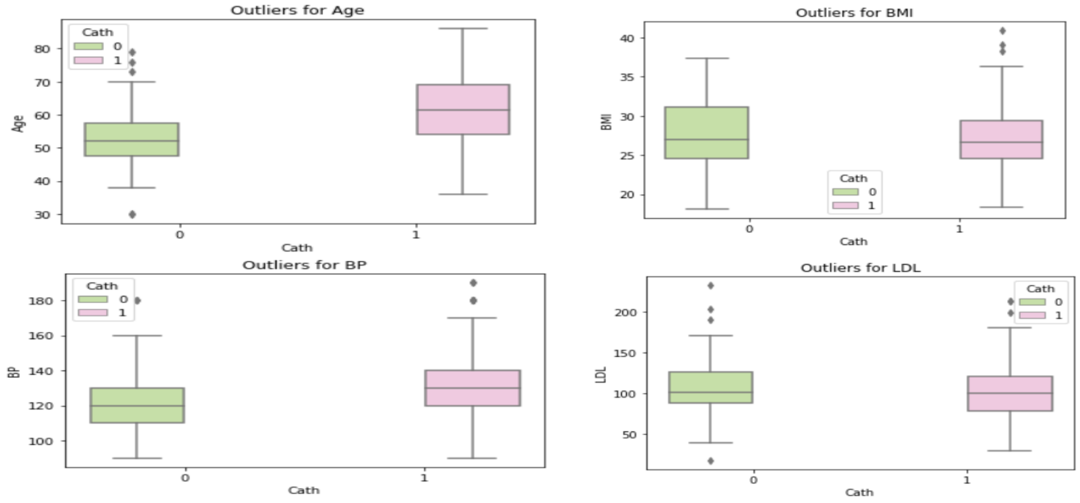

Figure 1: Visualization of outcomes

Although outliers may contain valuable information, it was investigated in [8] that most multivariate outliers are indeed medical outliers. In addition, removing outliers significantly improves the accuracy of classification models.

Therefore, it was decided to remove all outliers with a z-score greater than 5 (Formula 1).

$$z = \frac{(x - \mu)}{\sigma}, \tag{1}$$

where x is the data point you want to standardize; μ is the mean of the dataset; σ is the standard deviation of the dataset.

The refinement of datasets for the prediction of heart disease is a critical step in ensuring the reliability of subsequent machine learning analyses. This study details a rigorous data cleansing process applied to the Z-Alizadeh Sani dataset, which involved the computation of z-scores for each data point to identify and remove extreme outliers with z-scores exceeding 5. Prior to this outlier detection, feature transformation was performed using the minimax method, effectively scaling the dataset features from 0 to 1, a process that aids in standardizing the data for more effective machine learning application.

A correlation matrix was then constructed to elucidate the relationships between various parameters and the incidence of heart disease. This matrix revealed both positive and negative correlations, indicating how certain features may influence the likelihood of heart disease. Notably, age showed a positive correlation, aligning with literature that identifies increased age as a risk factor, particularly beyond the age of 65 [10]. Similarly, levels of low-density lipoprotein (LDL), or "bad" cholesterol, were significantly correlated with heart disease, supporting the well-established link between cholesterol and cardiovascular risk [11].

Chest pain emerged as a positively correlated symptom for both genders, with literature suggesting that women may experience more nuanced symptoms [12]. Hypertension was identified as another key factor, particularly given its role in arterial damage and its heightened correlation in women. Atypical symptoms, such as fatigue and shortness of breath, while not significant in isolation for diagnosis, were recognized as potential indicators of heart disease and should not be disregarded.

The study also highlighted obesity as a contributing factor to heart disease due to the increased demand on the circulatory system and the associated rise in blood pressure, a known precursor to cardiac events. Additionally, the adverse effects of smoking on blood composition and the increased risk of clot formation were acknowledged [9].

This data pre-processing and analysis endeavor underscores the importance of a meticulous approach to data handling and the insights that can be gleaned from a well-curated dataset. The findings from the correlation matrix not only reinforce established medical knowledge but also provide a data-driven foundation for the development of predictive models for heart disease.

In order to address the issue of high dimensionality in the dataset, the principal component analysis (PCA) method was employed to reduce the number of features. The PCA technique was selected due to its ability to generate a new set of uncorrelated variables that capture the maximum variance in the data. Given the limited size of our dataset but large number of features, we utilized 20 principal components to ascertain the degree of information retention. The resultant principal components are expressed as linear combinations of the original features, and are defined as follows: let X be the original feature matrix of size n x p, where n is the number of observations and p is the number of features. We can transform this matrix into a new matrix Z, which has n rows and k columns, where k is the number of principal components used. The k principal components can be calculated as follows (Formula 2):

$$Z = XW, \tag{2}$$

where W is a $p \times k$ matrix of weights, calculated by performing an eigenvalue decomposition on the covariance matrix of X.

The covariance matrix is defined as (Formula 3):

$$\sum = \frac{1}{n} X^T X \tag{3}$$

Using PCA, we were able to reduce the dimensionality of the dataset while retaining as much information as possible, allowing for more efficient and accurate analysis.

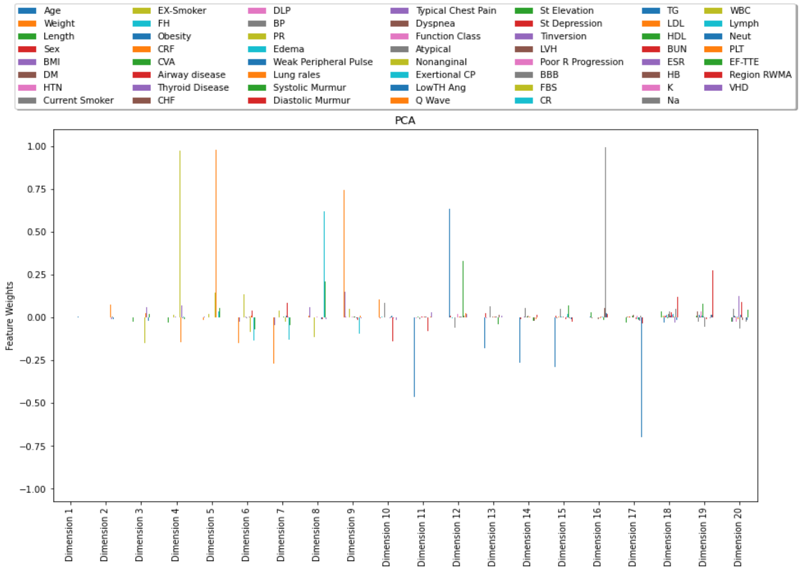

Figure 2: The results of the application of the method of principal components

The identification of critical features in heart disease prediction is paramount for the development of accurate diagnostic models. In our study, we have employed a multifaceted approach to feature selection, utilizing both statistical and machine learning techniques to discern the most influential factors. Figure 2 presents a component analysis where the x-axis represents the component number and the y-axis quantifies the feature's influence or weight. The analysis reveals that features such as WBC, TG, PLT, FBS, LDL, BP, ESR, Neut, Weight, HDL, PR, Age, Length, BMI, BUN, Na, Lymph, HB, Function Class, and Region RWMA have a substantial and positive weight, indicating a strong association with the likelihood of a heart disease diagnosis. To further refine feature selection, a decision tree algorithm was implemented to rank features by their importance. The output of this process informed the input for linear classification methods. A meta-estimator was utilized to assess feature significance, filtering out the most impactful ones based on a predefined threshold. Consequently, 13 features were identified as having the highest diagnostic value. Complementary to this, recursive feature elimination with cross-validation (Rfecv) and a linear support vector machine (SVM) with a meta-estimator were applied, yielding 30 and 25 significant features, respectively. Notably, Age and Region RWMA emerged as common and critical features across all methods, underscoring their diagnostic significance.

For the prediction of heart disease, we employed a suite of basic classifiers including the Naive Bayes classifier, Logistic Regression, Decision Trees, and Stochastic Gradient Descent, as well as ensemble methods such as Bagging, Random Forest, Boosting, and Stacking. These models demonstrated commendable diagnostic performance; however, given the high stakes of medical diagnostics, we proposed a modified model to further enhance predictive accuracy.

Figure 3 illustrates the architecture of the implemented model, showcasing the integration of various classifiers and feature selection techniques. This model aims to not only improve diagnostic outcomes but also to potentially extend the quality and longevity of life for patients. Through rigorous testing and validation, we strive to ensure that the model's predictions are both reliable and clinically relevant, thereby contributing to the advancement of precision medicine in cardiac care.

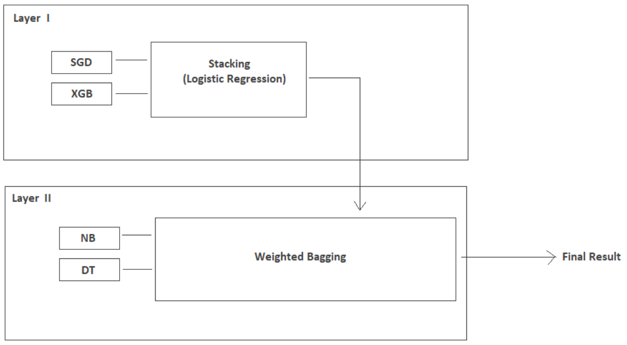

Figure 3: Architecture of the implemented model

In the realm of medical diagnostics, the development of sophisticated models for disease prediction is of paramount importance. This study introduces a novel two-stage classifier model designed to enhance the accuracy of heart disease prediction. The architecture of this model is bifurcated into two distinct levels, each employing a strategic amalgamation of classifiers to refine the predictive process.

At the inaugural level, the model synergizes the capabilities of Stochastic Gradient Descent (SGD) and XGBoost classifiers. The individual predictions from these classifiers are then harmonized through a logistic regression stacking technique, which serves as a meta-classifier to integrate the results into a cohesive prediction.

Progressing to the second level, the model incorporates a Naive Bayes classifier in tandem with a Decision Tree classifier. The predictions emanating from this duo are subsequently merged using a weight bagging approach. The weighting scheme is meticulously derived from the f1-scores of the classifiers' performance in the first stage, ensuring that the most accurate predictions exert greater influence on the final outcome.

This two-tiered classifier model is a testament to the innovative application of machine learning in healthcare, promising to deliver superior diagnostic precision. By leveraging the strengths of multiple classifiers and employing a hierarchical structure, the model aims to provide a robust tool for the early detection and treatment of heart disease, ultimately striving to improve patient prognosis and reduce the burden of cardiovascular diseases.

The model can be expressed mathematically as follows:

First stage:

- SGD classifier: $y_1 = SGD(X)$
- XGBoost classifier: $y_2 = XGBoost(X)$
- Logistic Regression Stacking: $y_3 = LogReg(y_1, y_2)$

Second stage:

- Naive Bayes classifier: $y_4 = NaiveBayes(X)$
- Decision Trees classifier: $y_5 = DecisionTrees(X)$

Weight Bagging: $y_6 = w_1 * y_4 + w_2 * y_5$, where w_1 and w_2 are the weights calculated based on the f1-score of the results from the first stage.

The quest for heightened accuracy in predictive models for heart disease diagnosis often leads to increased computational complexity. The model proposed in this study is a testament to this complexity, incorporating a multi-stage classification strategy that necessitates a series of intricate computations and optimizations. Each component within the model—Stochastic Gradient Descent (SGD), XGBoost, Naive Bayes, and Decision Trees—exhibits computational efficiency in isolation. However, the integration techniques of stacking and bagging, essential for combining classifier outputs, introduce additional computational demands.

The model's complexity is further compounded by the need to calculate weights based on the f1-scores from the first stage, adding another layer of computational intricacy. The overall complexity of the model is contingent upon a multitude of factors, including dataset size, feature count, and the nuances of each algorithm's implementation. While the individual time complexity of each component can be delineated, the aggregate time complexity of the model is a synthesis of these individual complexities.

To elucidate the model's computational demands, an estimation of the time complexity for each algorithmic component is undertaken. For instance, if the time complexities of the individual components are denoted as follows [8], the cumulative time complexity of the model can be inferred. This abstract sets the stage for a detailed discussion on the computational intricacies of the model, aiming to provide insights into the trade-offs between predictive performance and computational efficiency in the context of medical diagnostics.

- Stochastic gradient descent: $O(mn)$
- XGBoost: $O(kn \log n)$
- Naive Bayes: $O(kn)$
- Decision trees: $O(n \log n)$
- Stacking: $O(kn)$
- Bagging: $O(kn)$

where m - the number of training examples, n - the number of features, and k - the number of classifiers being combined. Then, the overall time complexity of the model would be roughly *O (kn log n + kn+ kn log n+ kn) = O(kn log n).*

The deployment of advanced machine learning models in medical diagnostics often necessitates a trade-off between accuracy and computational efficiency. The proposed two-stage classifier model for heart disease prediction exemplifies this balance, integrating ensemble methods and multiple classification algorithms to enhance predictive performance. While each constituent algorithm—Stochastic Gradient Descent (SGD), XGBoost, Naive Bayes, and Decision Trees—exhibits inherent computational efficiency, the ensemble techniques of stacking and bagging introduce additional layers of complexity due to their integrative computations.

This complexity is not merely theoretical but has practical implications, influenced by the specificities of the implementation. Ensemble methods, despite their computational demands, are chosen for their dual benefits: they not only bolster the model's accuracy but also serve as a bulwark against overfitting—a common pitfall in machine learning models.

Moreover, the model's complexity is not static; it can be modulated through the careful tuning of classifier parameters. This flexibility allows for the optimization of the model to suit varying computational constraints while maintaining a high standard of diagnostic precision. The abstract prefaces a discussion that acknowledges the intricate computational nature of the model and underscores the potential for its customization to achieve a desirable equilibrium between computational load and diagnostic accuracy.

4. Experiments

The challenge of handling unbalanced datasets in machine learning is a critical issue, particularly in medical diagnostics where the prevalence of disease classes can vary significantly. To address this, the current study employs a data balancing technique through the computation of weighting coefficients for each class within the target variable. This approach ensures that important information from the minority class is not overshadowed by the majority class.

Data partitioning adhered to the Pareto principle [12], allocating 80% for training and 20% for testing, to ensure a robust training set while retaining sufficient data for validation. Feature transformation was conducted using both minimax and standardization methods to prepare the dataset for various classification models. These models necessitate meticulous parameter selection to optimize accuracy, for which cross-validation was utilized. This method iteratively adjusts parameters to minimize the average cross-validation error.

The models were evaluated on both the original and feature-extracted datasets, with transformations applied. Five metrics were employed to assess classifier quality: accuracy, recall, precision, f1-score, and Cohen's kappa. Each metric provides insight into different aspects of classifier performance, from overall accuracy to the balance between precision and recall.

Table 1 illustrates the performance of the Naive Bayes classifier on the original dataset. Naive Bayes, a probabilistic classifier, leverages Bayes' theorem under the assumption of feature independence to predict class labels. It is computationally efficient due to its simplicity in calculating posterior probabilities. The classifier's efficacy is determined by its parameter estimation, which is derived from the training data.

Conversely, Table 2 showcases the performance post-feature selection, highlighting the impact of isolating the most salient features on the classifier's effectiveness. The feature selection not only streamlines the dataset but also potentially enhances the classifier's predictive power by focusing on the most informative attributes.

This abstract sets the stage for a comprehensive analysis of the Naive Bayes classifier's performance within the context of an unbalanced medical dataset, emphasizing the importance of data preparation and feature selection in the development of reliable diagnostic tools.

The formula 4 for Naive Bayes is:

$$P(y|x) = P(x\_1|y) * P(x\_2|y) * \ldots * P(x\_p|y) * P(y) / P(x), \quad (4)$$

where *P(y|x)* is the posterior probability of class label y given the features x, *P(x_i|y)* is the probability of the ith feature given the class label y, *P(y)* is the prior probability of class label y, and *P(x)* is the evidence probability of the features.

The performance metrics of the classifier are computed using standard measures such as accuracy, precision, recall, F1 score and Cohen's kappa [13], which are defined as follows:

- Accuracy: the proportion of correctly classified instances among all instances in the dataset.
- Precision: the proportion of true positives (correctly classified instances) among all instances classified as positive by the classifier.
- Recall: the proportion of true positives among all instances that actually belong to the positive class.
- F1 score: the harmonic mean of precision and recall, calculated as 2 * (precision * recall) / (precision + recall).

Cohen's kappa: a statistical measure of inter-rater reliability that takes into account the possibility of agreement occurring by chance. It is defined as follows (Formula 5):

$$k = \frac{p\_o - p\_e}{1 - p\_e}, \tag{5}$$

where p_o is the observed proportion of agreement between the classifier's predictions and the true labels, and p_e is the expected proportion of agreement that would occur by chance. Cohen's kappa ranges from -1 to 1, where a value of 1 indicates perfect agreement, 0 indicates agreement by chance, and -1 indicates complete disagreement.

These measures provide a comprehensive evaluation of the classifier's performance on the dataset, and are commonly used in machine learning applications. The results in Table 2 demonstrate the impact of feature selection on the classifier's performance, and highlight the importance of selecting relevant features for improving classification accuracy.

Table 1

Performance of the Classifier Based on the NB classifier on the original data set results of work of basic ensembles

Classifier	Accuracy	Recall	Precision	F1-Score	Cohen's kappa
NB	0.83	0.90	0.86	0.88	0.59

Table 2

The results of the classifier based on the NB classifier on data using feature extraction methods

Feature selection	MinMax Scaler	StandartScaler	Accuracy	Recall	Precision	F1-Score	Cohen's kappa
DT	+	-	0.85	0.88	0.9	0.89	0.61
DT	-	+	0.87	0.91	0.91	0.91	0.60
Rfecv	+	-	0.84	0.91	0.87	0.89	0.63
Rfecv	-	+	0.89	0.93	0.91	0.92	0.75
SVM	+	-	0.84	0.91	0.87	0.89	0.61
SVM	-	+	0.87	0.93	0.89	0.91	0.67
PCA	+	-	0.72	0.86	0.77	0.81	0.27
PCA	-	+	0.84	0.93	0.85	0.89	0.57

Table 3 shows the results of the Logistic Regression classifier on the original data set. Below, table 4 shows the results of the Logistic Regression classifier on the data set using feature selection methods. Logistic regression is a statistical method used for binary classification, commonly used in medical fields to predict diseases such as heart disease. It models the probability of an event occurring as a function of independent variables, by using a logistic function that maps a continuous range of inputs to a range between 0 and 1. The model estimates the coefficients of the independent variables using maximum likelihood estimation. The formula 6 for logistic regression is:

$$p = \frac{1}{1 + e^{-z}}, \tag{6}$$

where p is the probability of an event occurring, e is the mathematical constant, and z is the linear combination of independent variables (Formula 7):

$$z = b\_0 + b\_1x\_1 + b\_2x\_2 + \ldots + b\_p * x\_p, \qquad (7)$$

where b_0 is the intercept, b_i is the coefficient for independent variable x_i, and p is the number of independent variables.

In medical fields such as heart disease prediction, logistic regression can be used to estimate the probability of a patient developing heart disease based on their age, sex, blood pressure, cholesterol levels, and other relevant features.

Table 3

Performance of the classifier based on the lg classifier on the original data set

Classifier	Accuracy	Recall	Precision	F1-Score	Cohen's kappa
LG	0.87	0.93	0.89	0.91	0.67

Table 4

The results of the classifier based on the lg classifier on data using feature extraction methods

Feature selection	MinMax Scaler	Standar tScaler	Accuracy	Recall	Precision	F1-Score	Cohen's kappa
DT	+	-	0.89	0.91	0.93	0.92	0.75
DT	-	+	0.9	0.88	0.97	0.93	0.7
Rfecv	+	-	0.87	0.91	0.91	0.91	0.74
Rfecv	-	+	0.89	0.86	0.97	0.91	0.76
SVM	+	-	0.87	0.91	0.91	0.91	0.74
SVM	-	+	0.85	0.88	0.9	0.89	0.67
PCA	+	-	0.9	0.91	0.95	0.93	0.66
PCA	-	+	0.85	0.86	0.93	0.89	0.68

Table 5 shows the results of the Decision Tree classifier on the original data set. A Decision Tree classifier is a popular machine learning algorithm used for classification tasks that can handle both categorical and continuous data. It constructs a tree-like model of decisions and their possible consequences, where each internal node represents a test on an attribute, each branch represents an outcome of the test, and each leaf node represents a class label. The algorithm constructs the tree recursively by selecting the attribute that provides the most information gain or the best split, until a stopping criterion is met. The formula 8 for information gain is:

$$IG(D, A) = H(D) - H(D|A), \qquad (8)$$

where *G(D, A)* is the information gain of attribute *A* on dataset *D*, *H(D)* is the entropy of the dataset *D*, and *H(D|A)* is the conditional entropy of the dataset *D* given the attribute *A*. The attribute with the highest information gain is selected as the splitting criterion. The formula 9 for entropy is:

$$H(D) = -\sum (p\_i * log\_2(p\_i)), \qquad (9)$$

where *H(D)* is the entropy of dataset *D*, *p_i* is the proportion of samples in *D* that belong to class *i*, and *log_2* is the logarithm base 2.

Table 5

Performance of the classifier based on the dt classifier on the original data set

Classifier	Accuracy	Recall	Precision	F1-Score	Cohen's kappa
LG	0.82	0.86	0.88	0.87	0.58

Below, table 6 shows the results of the Decision Tree classifier on the data set using feature selection methods. To apply feature selection methods such as Principal Component Analysis (PCA) to a Decision Tree classifier, the algorithm selects the attributes that provide the most information gain or the best split, taking into account the reduced feature space obtained by PCA. The formula for information gain and entropy remains the same, but the dataset used for calculation is based on the principal components instead of the original features.

By using PCA to reduce the dimensionality of the dataset, the Decision Tree classifier can improve its performance by removing redundant or noisy features and focusing on the most important ones. This can lead to better accuracy and faster training times.

Table 6

The results of the classifier based on the dt classifier on data using feature extraction methods

Feature selection	MinMax Scaler	Standart Scaler	Accuracy	Recall	Precision	F1-Score	Cohen's kappa
DT	+	-	0.89	0.91	0.91	0.91	0.66
DT	-	+	0.85	0.88	0.9	0.89	0.71
Rfecv	+	-	0.82	0.88	0.86	0.87	0.65
Rfecv	-	+	0.8	0.88	0.84	0.86	0.74
SVM	+	-	0.84	0.93	0.85	0.89	0.59
SVM	-	+	0.84	0.93	0.85	0.89	0.62
PCA	+	-	0.7	0.77	0.8	0.79	0.72
PCA	-	+	0.69	0.74	0.8	0.77	0.6

Table 7 shows the results of the SGD classifier on the original data set. Stochastic Gradient Boosting (SGB) is a powerful ensemble method that combines multiple weak learners into a strong classifier by iteratively adding models that fit the residuals of the previous models. The formula 10 for the predicted probability of class *j* is:

$$p\_j = 1 / (1 + exp(-sum(f\_i(x)))), \tag{10}$$

where *f_i* is the *i*-th model that predicts the residual of the previous model, x is the input data, and exp is the exponential function. The algorithm optimizes the log loss function by using gradient descent to find the optimal values of the model parameters. Below, table 8 shows the results of the SGD classifier on the data set using feature selection methods. When combined with feature selection methods such as Principal Component Analysis (PCA), SGD selects the most important principal components as input features to improve performance and reduce training time. The formula 11 for updating the model parameters during each iteration is:

$$w\_j = w\_j - \alpha(\partial L/\partial w\_j + \lambda w\_j), \tag{11}$$

where *w_j* is the *j*-th weight parameter of the linear model, α is the learning rate, *L* is the loss function, and λ is the regularization parameter. By using PCA to reduce the feature space, SGD can achieve faster convergence and better generalization on high-dimensional data.

Table 7

Performance of the classifier based on the SGD classifier on the original d ta set

Classifier	Accuracy	Recall	Precision	F1-Score	Cohen's kappa
LG	0.72	0.86	0.77	0.81	0.53

Table 8

The results of the classifier based on the SGD classifier on data using feature extraction method

Feature selection	MinMax Scaler	StandartS caler	Accuracy	Recall	Precision	F1-Score	Cohen's kappa
DT	+	-	0.80	0.94	0.80	0.87	0.58
DT	-	+	0.87	0.91	0.91	0.91	0.57
Rfecv	+	-	0.89	0.88	0.88	0.92	0.64
Rfecv	-	+	0.84	0.88	0.88	0.88	0.56
SVM	+	-	0.77	0.94	0.77	0.85	0.52
SVM	-	+	0.87	0.88	0.93	0.9	0.53
PCA	+	-	0.89	0.91	0.93	0.92	0.57
PCA	-	+	0.85	0.91	0.89	0.9	0.54

5. Results

In the domain of medical diagnostics, the accuracy of predictive models is not the sole determinant of their clinical utility; the nature of the disease and the consequences of misdiagnosis play a crucial role in model evaluation. This study focuses on the binary classification problem of heart disease presence, where the absence of disease (a classifier output of 0) is of paramount importance. Ensuring a correct negative prediction is critical, as a missed diagnosis of heart disease can have dire implications for patient health and survival.

Consider a hypothetical cohort of 1000 patients screened for heart disease, with 100 true positive cases. A classifier with a 90% accuracy rate correctly identifies 900 patients. However, a deeper analysis reveals a recall of 80%, indicating that 20 patients with heart disease were overlooked (false negatives). These misclassified patients risk forgoing vital treatment, potentially leading to severe health consequences or fatality.

This scenario underscores the significance of recall in heart disease prediction models. High recall ensures that the majority of actual disease cases are identified, prioritizing patient safety and treatment efficacy over the inconvenience of additional tests for healthy individuals. Thus, the study advocates for a model evaluation framework that emphasizes recall, ensuring that the most critical clinical outcomes—accurate disease detection and patient care are achieved.

Table 9

The best results of basic classifiers

Classifier Method	Accuracy	Recall	Precision	F1-Score	Cohen's kappa
NB	0.89	0.93	0.91	0.92	0.75
LR	0.87	0.93	0.89	0.91	0.67
DT	0.84	0.93	0.85	0.89	0.62
SGD	0.80	0.94	0.80	0.87	0.58

Table 9 shows that the naive Bayes classifier gives the best accuracy. Let's look at the graphical representation of the performance of NB in Figure 4.

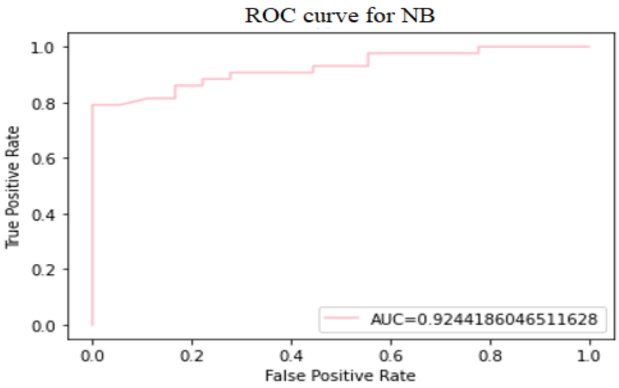

Figure 4: The ROC curve for NB

The efficacy of binary classifiers in medical diagnostics, particularly for heart disease prediction, is often assessed using the receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC). The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied, marking the trade-off between the true positive rate (TPR) and false positive rate (FPR). The AUC is a scalar representation of overall classifier performance, with a value of 1.0 denoting a perfect classifier.

In the context of heart disease prediction, the ROC curve is instrumental in evaluating the classifier's capacity to distinguish between patients with and without the disease. A classifier with a high AUC can reliably rank a patient with heart disease higher than a patient without, indicating strong predictive power. Figure 4 in this study demonstrates a classifier with an AUC of 0.92, suggesting high effectiveness in heart disease prediction.

Nonetheless, given the critical nature of recall in this domain - where missing a positive case can be life-threatening - additional metrics are considered. Stochastic Gradient Descent (SGD) is highlighted for its promising recall rates, despite a low Cohen's Kappa coefficient indicating potential performance due to chance.

To enhance classifier performance, ensemble methods are employed, leveraging the strengths of multiple models to improve accuracy, stability, and error variance. This study utilizes three ensemble techniques: Random Forest, XGBoost, and Bagging, each contributing to a robust predictive model. The performance of these ensembles is detailed in Table 10, showcasing the potential of combined approaches in advancing the predictive diagnostics of heart disease.

Table 10
Results of work of basic ensembles

Ensemble Method	Accuracy	Recall	Precision	F1-Score	Cohen's kappa
RF	0.9	0.94	0.89	0.93	0.71
XGBoost	0.84	0.93	0.85	0.89	0.68
Bagging	0.89	0.91	0.93	0.92	0.72

Table 10 shows that bagging and random forest perform better than XGBoost. Despite the higher recall score compared to bagging, XGBoost has a lower Cohen's ratio.

6. Discussion

The evaluation of the results for heart disease prediction showed that the logistic regression provides high accuracy, while stochastic gradient descent has the best recall metric. To further enhance the overall performance, a complex two-stage model was proposed, which includes stacking and weight bagging techniques. The first stage of the model combines two classifiers using stacking, which is a form of ensemble learning. The combined results are then used as input to logistic regression (Formula 12):

y_{LR} = \sigma(\beta_0 + \beta_1 x_{SGD} + \beta_2 x_{XGB}), (12)

where y_{LR} is the output of logistic regression, \sigma is the sigmoid function, x_{SGD} and x_{XGB} are the outputs of the stochastic gradient descent and XGBoost classifiers respectively, and \beta_0, \beta_1, and \beta_2 are the logistic regression coefficients. It improves the performance of basic classifiers.

The second stage significantly improves all performance metrics. In the second stage of the model, a naive Bayes classifier and decision trees are combined using weight bagging (Formula 13):

y_{WB} = \sum_{i = 1}^{N} w_i y_i, (13)

where y_{WB} is the output of the weight bagging ensemble, N is the number of base models (in this case, 2), w_i is the weight assigned to the i base model based on its f1-score from the first stage of the model, and y_i is the output of the i base model (either the naive Bayes classifier or decision trees).

The quest for precision in medical diagnostics has led to the development of complex machine learning models, such as the two-stage classifier for heart disease prediction discussed in this study. This model intricately weaves together multiple classifiers and ensemble techniques, each contributing its own layer of complexity and parameterization [15]. The resultant model stands out from simpler classifiers and ensembles, not just in complexity, but more importantly, in performance—particularly in metrics such as accuracy and recall.

The recall score is of utmost importance in medical diagnostics, where the cost of a false negative—overlooking a disease—can be life-threatening. The two-stage model demonstrates superior recall scores over its single classifier and ensemble counterparts, justifying its complexity. The computational complexity, while higher, is offset by the model's enhanced diagnostic capabilities.

This abstract introduces a comprehensive model that, despite its computational demands, provides significant improvements in predictive accuracy and recall for heart disease diagnosis. The subsequent sections will delve into the model's architecture, performance evaluation, and the implications of its complexity in clinical settings.

7. Conclusion

This research contributes to the critical understanding of the determinants of heart disease and presents a sophisticated model poised to enhance diagnostic precision in clinical settings. The model's architecture, which orchestrates a symphony of classifiers and ensemble methods across two distinct stages, is designed to maximize accuracy and robustness in predicting heart disease. By leveraging the strengths of various classifiers and integrating their outputs through advanced techniques like stacking and bagging, the model adeptly navigates the intricate feature-target relationships inherent in medical data.

Despite its promise, the model's complexity and computational intensity are non-trivial, necessitating a balance between performance gains and resource expenditure. The model's intricate design, while computationally demanding, is justified by the potential uplift in diagnostic accuracy and the consequent life-saving implications of timely and accurate heart disease detection.

The feasibility of model deployment is supported by the availability of necessary technologies and a detailed blueprint of the model's structure. The primary challenge lies in data collection, a time-intensive yet surmountable hurdle, as evidenced by the successful application of the Z-Alizadeh Sani dataset. This study lays the groundwork for future refinement and application of machine learning methodologies in the pursuit of medical advancements.

8. Limitations

The Despite the promising outcomes of the proposed model, it is imperative to acknowledge the limitations inherent in this study. Firstly, the complexity of the model, while beneficial for accuracy, poses significant computational demands. The integration of multiple classifiers and ensemble methods requires substantial processing power and memory, which may not be readily available in all medical institutions, particularly those in resource-constrained environments.

Secondly, the model's performance is heavily dependent on the quality and comprehensiveness of the input data. The reliance on a single dataset, such as the Z-Alizadeh Sani dataset, may introduce bias or limit the generalizability of the findings. The model's robustness must be tested across diverse populations and datasets to ensure its applicability to different demographic groups.

Furthermore, the model's interpretability is compromised by its complexity. The 'black box' nature of advanced machine learning models can be a barrier in clinical settings, where understanding the decision-making process is crucial for trust and adoption by healthcare professionals. Lastly, the time required for training and tuning the model, along with the necessity for continuous updates with new data to maintain its accuracy, presents an ongoing commitment. This requirement for sustained resources may be a limitation for institutions considering the long-term deployment of such a model.

In conclusion, while the model exhibits high potential for improving heart disease diagnostics, these limitations must be carefully considered and addressed in future research and application development.

9. Acknowledgements

The study was created within the framework of the project financed by the National Research Fund of Ukraine, registered No. 2021.01/0103, "Methods and means of researching markers of ageing and their influence on post-ageing effects for prolonging the working period", which is carried out at the Department of Artificial Intelligence Systems of the Institute of Computer Sciences and Information of technologies of the National University "Lviv Polytechnic".

References

[1] J. Zhang, H. Zhu, Y. Chen, Ensemble machine learning approach for screening of coronary heart disease based on echocardiography and risk factors, in: BMC Medical Informatics and Decision Making, vol. 21(1), 2021. https://doi.org/10.21203/rs.3.rs-120645/v1.

[2] J. Saiz-Vivo, V.D.A. Corino, R. Hatála, Heart rate variability and clinical features as predictors of atrial fibrillation recurrence after catheter ablation: a pilot study, in: Frontiers in Physiology, vol. 12, 2021. https://doi.org/10.3389/fphys.2021.672896.

[3] U. Hackstein, T. Krüger, A. Mair, Early diagnosis of aortic aneurysms based on the classification of transfer function parameters estimated from two photoplethysmographic signals, in: Informatics in Medicine Unlocked, Volume 25, 100652, 2021. https://doi.org/10.1016/j.imu.2021.100652.

[4] S. Ilbeigipour, A. Albadvi, & E. Akhondzadeh Noughabi, Real-time heart arrhythmia detection using Apache Spark structured streaming, in: Journal of Healthcare Engineering, Volume 2021, Article ID 6624829. https://doi.org/10.1155/2021/66248292021.

[5] K. Rathakrishnan, S. N. Min, & S. J. Park, Diagnostic approach to evaluating ECG features for the classification of post-stroke survivors, in: Applied Sciences, volume11(1), 2021, pp. 1-16. https://doi.org/10.3390/app11010192.

[6] UCI Machine Learning Repository: Z-Alizadeh Sani Data Set. (n.d.), 2021. URL: https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani.

[7] D. Yadav, Categorical Encoding using Label-Encoding and One-Hot-Encoder, 2021. URL: https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd.

[8] N. Boyko, N. Tkachuk, Processing of Medical Different Types of Data Using Hadoop and Java MapReduce, in: The 3rd International Conference on Informatics & Data-Driven Medicine (IDDM 2020), Växjö, Sweden, November 19-21, 2020, pp. 405-414.

[9] J. Laurikkala, M. Juhola, E. Kentala, (n.d.), Informal identification of outliers in medical data, in: Computer Science, volume 5, 2000.

[10] Heart Attack Age: Younger People Can Have Heart Attacks, Too, 2022. URL: https://www.healthline.com/health/heart-attack/heart-attack-age.

[11] Cholesterol Levels: By Age, LDL, HDL, and More, 2021. URL: https://www.healthline.com/health/high-cholesterol/levels-by-age.

[12] Chest Pain Types, Causes, Symptoms, Diagnosis & Treatment. (n.d.), 2017. URL: Retrieved from https://www.emedicinehealth.com/chest_pain_overview/article_em.htm.

[13] Pareto Principle Definition. (n.d.), 2020 URL: https://www.investopedia.com/terms/p/paretoprinciple.asp

[14] J. Cohen, A coefficient of agreement for nominal scales, in: Educational and Psychological Measurement, volume 20(1), 1960, pp. 37-46. https://doi.org/10.1177/001316446002000104

[15] T. Fawcett, An introduction to ROC analysis, in Pattern Recognition Letters, vol. 27, No. 8, 2006, pp. 861–874.

[16] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference, and prediction, 2nd ed. Publisher: Springer, 2009. https://doi.org/10.1007/978-0-387-84858-7

[17] N. Boyko, K. Boksho, Application of the Naive Bayesian Classifier in Work on Sentimental Analysis of Medical Data, in: The 3rd International Conference on Informatics & Data-Driven Medicine (IDDM 2020), Växjö, Sweden, November 19-21, 2020, pp. 230-239.