Enhancing Heart Disease Diagnosis Using Machine Learning Classifiers

Ahmed A. H. Alkurdi[*1,2]

[1] Department of Information Technology Management, Technical College of Administration, Duhok Polytechnic University, Duhok, KRG-Iraq
[2] Department of Computer Science, College of Science, Nawroz University, Duhok, KRG-Iraq
Emails: Ahmed.alaa@dpu.edu.krd; Ahmed.alaa@nawroz.edu.krd

Abstract

Heart diseases are the primary cause of death worldwide. The approximate mortality rate due to cardiovascular diseases is a staggering 18 million lives per year. many human lives could be saved with early and accurate diagnosis and prediction of such conditions. Thus, the automation of such a process is crucial and achievable with the rise of machine learning and deep learning capabilities. However, patient data is riddled with issues which must be resolved before they can be used for heart disease prediction. This research aims to improve the accuracy of heart disease diagnosis by utilizing data preprocessing techniques and classification algorithms. These techniques may provide an insight into predicting cardiovascular diseases from subtle clues before any major symptoms arise. The study employs the Heart Disease UCI dataset and follows a systematic approach to train machine learning models in the process of heart disease diagnosis. The approach utilizes a variety of data preprocessing techniques to prepare the data for model training such as MEAN missing value imputation, Normalization, Synthetic Minority Over-sampling Technique (SMOTE), and Correlation. Afterward, the preprocessed data is fed into four popular classification algorithms: Decision Tree, Random Forest, Support Vector Machine (SVM), and k-Nearest Neighbors (k-NN). These algorithms provide a broad evaluation of the dataset. The proposed methodology demonstrates promising results which clearly highlight the value and significance of data preprocessing. This is evident from the achieved accuracy, precision, recall, F1 score and ROC AUC results. In summary, the importance of preprocessing and feature selection is distinct when dealing with datasets containing various challenges. These crucial processes play a central role in building a trustworthy and precise model for heart disease prediction.

Keywords: Machine Learning; Classification; Preprocessing; Feature Selection; Heart Disease.

1. Introduction

The prevalence of heart disease remains a significant public health concern worldwide. Accurate prediction and early detection of heart disease can greatly aid in effective management and treatment. Machine learning techniques have shown promise in analyzing large datasets and extracting valuable insights[1]. Machine learning prediction of diseases has emerged as a promising approach in healthcare, offering the potential to improve diagnostic accuracy, prognosis, and treatment outcomes. With the availability of large-scale medical datasets and advancements in computational techniques, machine-learning algorithms have been applied to various medical domains, including disease prediction[2]. The goal of machine learning prediction in healthcare is to develop models that can learn from historical patient data to make accurate predictions about disease occurrence, progression, and patient outcomes. These models utilize various algorithms, for example, decision trees, random forests, support vector machines, and neural networks, to mine patterns and relationships from the data[3].

One common challenge in heart disease datasets is that they include missing values. Various approaches have been suggested to handle missing data, such as imputation using the mean. Additionally, normalization techniques are applied to standardize the data and bring features to a similar scale. This step is crucial to ensure fair comparison and optimal performance of machine learning models[4]. Feature selection is another essential aspect of the predictive modeling process. Techniques such as correlation analysis are used to identify relevant

Doi: https://doi.org/10.54216/FPA.130101
Received: March 02, 2023 Revised: June 01, 2023 Accepted: August 04, 2023

features for heart disease prediction. By selecting informative features, the accuracy and interpretability of the models are enhanced. The selected features are then utilized to train and evaluate classifiers[5].

Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) are popular machine learning algorithms widely used for classification tasks. Decision trees are tree-like structures used to make predictions. They partition the feature space into distinct regions based on feature tests, and each leaf node denotes a class label or target value. Decision trees are intuitive and easily interpretable[6]. An extension of this idea is Random Forest, which is a learning method, that utilizes multiple decision trees to enhance predictive accuracy and generalization. It creates an ensemble of decision trees using arbitrary subsets of the training data and features. The final prediction is acquired by combining the outcomes of individual trees[6]. On the other hand, SVM is an efficient algorithm for both classification and regression. It seeks an optimal hyperplane that maximizes the margin between different clusters in the feature space. SVM is effective for handling linear and non-linear data through the use of kernel functions[7]. While KNN is a naive yet effective non-parametric algorithm that assigns a new instance to a class based on the most vote of its nearest neighbors. The choice of "k" determines the level of smoothness and the decision boundary. KNN is straightforward to implement but can be sensitive to the choice of "k" and may not scale well to large datasets[8].

Through this study, an effort is made to provide insights into the application of machine learning techniques for heart disease diagnosis and likelihood. The results obtained from the aforementioned algorithms can aid healthcare specialists make calculated decisions regarding diagnosis and treatment strategies. The outcomes of this may contribute to the development of more accurate and efficient heart disease prediction models.

The most important contributions of this research are:

- Evaluated the performance of Decision Tree, Random Forest, SVM, and KNN classifiers.
- Compared results before and after preprocessing and feature selection.
- Provided insights into the impact of these techniques on model accuracy and performance.
- Offered valuable findings to enhance machine learning-based prediction of heart disease.

The remainder of this paper is organized as follows: Section 2 is the literature review, which examines existing research on machine learning applications in the prediction of heart disease, focusing on preprocessing techniques, feature selection methods, and classification algorithms. Section 3 is the proposed method that details the approach used in this study, encompassing SMOTE for class imbalance, "mean" imputation for missing values, normalization, and correlation-based feature selection. Also highlighting details of the classifiers used in the model namely Decision Tree, Random Forest, SVM, and KNN classifiers. Section 4 is the analysis and results section which presents the findings from applying Decision Tree, Random Forest, SVM, and KNN classifiers on the preprocessed data, comparing the performance before and after preprocessing and feature selection. Furthermore, a comparison is drawn with the results of previous research in the field. Finally, section 5 is the conclusion section which summarizes the contributions of the research, emphasizing the importance of preprocessing and feature selection for accurate and reliable disease prediction, while also suggesting avenues for future research in this domain.

2. Literature Review

Machine learning techniques and algorithms have gained significant attention in the field of heart disease prediction and diagnosis in the recent decade. This literature review attempts to provide an outline of the existing research and advancements in this domain.

Machine learning algorithms and techniques have been widely studied for the purpose of heart disease prediction. For example, [9] investigated the performance of various classifiers, including support vector machines, decision trees, and k-nearest neighbors, in accurately predicting heart disease. The results indicated that these algorithms achieved high accuracy rates and demonstrated their potential as effective tools for assisting medical professionals in diagnosing heart disease. Addressing the issue of missing data is crucial for reliable predictions. Thus, imputation techniques have been employed to handle missing values in heart disease datasets. A study by [10] applied mean imputation to deal with missing data before training machine learning models. The findings revealed that incorporating imputation methods improved the performance of the predictive models. Moreover, Feature selection is greatly influential in the role of identifying the most relevant predictors for heart disease prediction. Researchers have employed various techniques to select informative features. For instance, in a study by [11], a correlation-based feature selection approach was utilized to identify the most significant features associated with heart disease. The results demonstrated the effectiveness of this technique in improving the

Doi: https://doi.org/10.54216/FPA.130101
Received: March 02, 2023 Revised: June 01, 2023 Accepted: August 04, 2023

accuracy and interpretability of the predictive models. One notable work in this field is the study conducted by [12], where they applied preprocessing techniques such as missing data imputation and data normalization to handle missing values and ensure consistent scaling of input features. They also employed feature selection methods, namely correlation analysis and recursive feature elimination, to identify the most informative features for heart disease prediction. Their results demonstrated that proper preprocessing and feature selection significantly improved the performance of the prediction model. Another relevant work by [13] focused on the application of SMOTE (Synthetic Minority Over-sampling Technique) to tackle the issue of class imbalance in heart disease datasets. They applied SMOTE to oversample the minority class and balance the dataset, leading to better performance in predicting heart disease cases. That being said, various machine-learning techniques have been employed for the prediction of heart diseases. In a study by [14], a comparative analysis of decision trees, random forests, support vector machines, and k-nearest neighbors was conducted. The study emphasized the strengths and limitations of each algorithm, evaluating their suitability for heart disease prediction tasks. Furthermore, the study conducted by [15] explored the use of ensemble methods such as random forest and support vector machines (SVM) for heart disease prediction. They combined feature selection techniques with these ensemble models and observed improved accuracy and robustness in identifying heart disease patterns.

These works collectively highlight the significance of preprocessing techniques, feature selection, and the use of advanced machine-learning algorithms for accurate heart disease prediction. By addressing data quality issues, handling class imbalance, and selecting relevant features, these approaches contribute to building more reliable and effective predictive models in the field of heart disease diagnosis and prognosis.

3. Proposed Method

The developed model for heart disease prediction and diagnosis utilizes a series of steps to process the data and select relevant features before applying classification algorithms. The model begins by handling missing values through mean imputation and normalizing the dataset to ensure consistency across features. Afterward, the data passes through Correlation to highlight the relationships among the different features. Finally, the data is split into two parts. One is fed to the aforementioned classifiers for training purposes, while the second part is used for validation and testing the correct operation of the models.

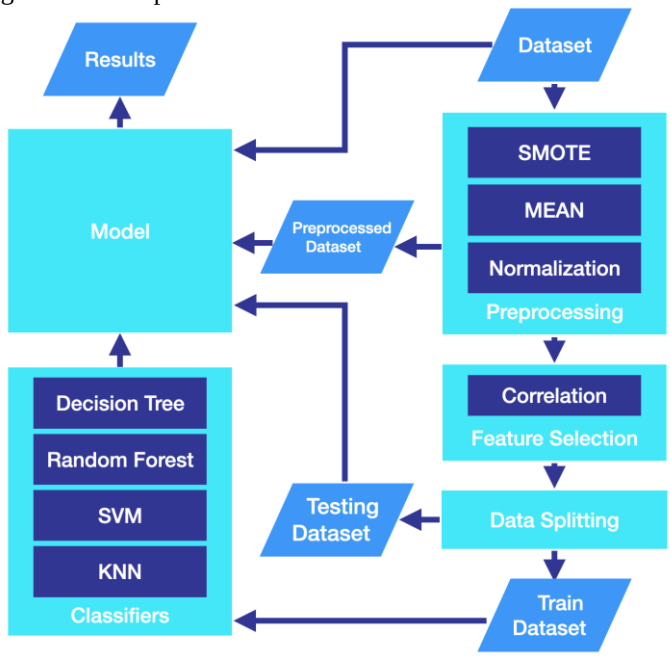

Figure 1: General Proposed Method

3.1 Preprocessing

Preprocessing in machine learning involves transforming, cleaning, and preparing raw data before model training. It includes steps like data cleaning, transformation, scaling, and encoding. Effective preprocessing improves model accuracy and prevents overfitting[16]. Preprocessing involves converting raw incomprehensible data into a format that is suitable for analysis and model training which is pivotal for machine learning

Doi: https://doi.org/10.54216/FPA.130101
Received: March 02, 2023 Revised: June 01, 2023 Accepted: August 04, 2023

algorithms. This brief will focus on three important preprocessing techniques: SMOTE (Synthetic Minority Over-sampling Technique), mean imputation, and normalization[17].

First, SMOTE is applied to the data, SMOTE is a widely used technique for tackling class imbalance in datasets. It creates synthetic samples of the smaller class to balance the data distribution and improve classification performance. SMOTE has been introduced in 2002 by Chawla et al. in their paper "SMOTE: Synthetic Minority Over-sampling Technique" [18]. Second, Mean imputation is utilized to resolve missing values in datasets, this is done through the replacement of the missing values by the mean of the available data. It is a simple yet effective method for dealing with missing data[19]. Finally, Normalization is used which is a preprocessing technique that converts numerical features to a common range, generally between 0 and 1, to avoid the dominance of certain features during model training. Normalization improves the convergence and performance of machine learning models[20]. These preprocessing techniques, including SMOTE, mean imputation, and normalization, contribute significantly to data quality, model performance, and robustness in machine learning applications.

3.2 Feature selection

Feature Selection is one of the most significant steps of machine learning, which is a process that targets the selection of the most important and identifying features from a given dataset. By selecting a subset of features, feature selection reduces dimensionality, improves model performance, and enhances interpretability. One commonly used technique in feature selection is correlation analysis[21].

Correlation analysis helps identify the relationships between features by measuring the statistical association between them. It attempts to put a numeric indication of the strength and direction of the linear relationship, typically using correlation coefficients, for example, Pearson's correlation coefficient or Spearman's rank correlation coefficient. A high correlation between features indicates redundancy or multicollinearity, which can negatively impact model performance[22]. By leveraging correlation analysis in feature selection, researchers and practitioners can identify and retain the most informative features while removing redundant or irrelevant ones, thus improving model performance and interpretability[23].

Correlation-based feature selection is a widely recognized method used to assess the significance of features by evaluating their correlation with classes and other features. It involves measuring the correlation between features and classes, as well as the correlation between features themselves. To determine the relevance of a subset of features, CFS (Correlation-based Feature Selection) employs Pearson's correlation equation. This equation calculates the power of the linear relationship between variables and is commonly used to assess the relevance of features in the context of feature selection.

$$Merit_s = \frac{kr_{kc}}{\sqrt{k+(k-1)r_{kk}}} \quad (1)$$

Where the term "Merits" refers to the measure of relevance for a specific subset of features. It is determined based on two average linear correlation coefficients: rkc, which is the relation between the features and classes, and rkk, which indicates the relation between different features. These coefficients are used to evaluate the strength of the linear relationships and provide insights into the importance of the feature subset under consideration.

3.3 Classification

In recent times, numerous automated diagnostic systems have emerged for the purpose of diagnosing various ailments, including human heart disease. Machine learning techniques and optimization methods have proven to be highly effective in analyzing datasets and identifying heart disease automatically, thereby significantly influencing the field of medical science. Different machine learning models are employed to identify the disease and classify or predict the outcomes. By utilizing machine learning approaches, large amounts of genetic data can be efficiently processed and analyzed. This enables a more comprehensive examination of medical data, and algorithms can be trained to make predictions, including the prediction of pandemics. Through the analysis of datasets, valuable information can be extracted, signifying the importance of individual variables and their relationships. The primary objective of this research is to predict whether individuals have severe cardiac disease or not.

Classification is a fundamental process in machine learning that includes assigning input data points to predefined categories or classes. It is widely used in many fields such as image recognition, natural language processing, and fraud detection. Several classification algorithms have been developed, each with its strengths and characteristics. This brief will highlight the Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) algorithms.

Doi: https://doi.org/10.54216/FPA.130101
Received: March 02, 2023 Revised: June 01, 2023 Accepted: August 04, 2023

Decision Tree (DT)

Decision trees are hierarchical models that make sequential decisions based on features to reach a classification outcome. They are interpretable and can handle both numerical and categorical data. A widely used algorithm for decision trees is the C4.5 algorithm[24]. Furthermore, DTs are structures that resemble trees, which are used to handle sizeable datasets. Decision trees are generally represented as flowcharts, where the outer brancher represents the outcomes and the inner nodes represent the properties of the dataset. Decision trees have gained popularity due to their efficiency, reliability, and ease of comprehension [25]. The predicted class for a decision tree is determined starting from the root of the tree. The next stages in the tree are computed by the comparison of the value of the attribute at the root with the input information. After transitioning to the next node based on the comparison result, the corresponding branch is followed to reach the indicated value. The measure of uncertainty, titled entropy, undergoes changes as training examples are reduced into smaller groups at each decision tree node.

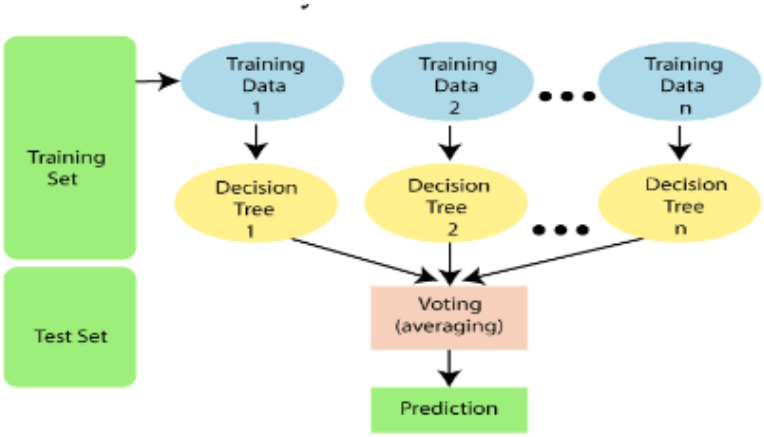

Figure 2: Random Forest Classifier [25]

Random Forest (RF):

Another well-known machine learning algorithm is Random Forest, which mainly combines multiple different decision trees utilizing their predictions to make a final classification. It decreases overfitting and enhances accuracy by combining predictions from many different trees[26].

In Random forests, trees are created in the following steps [27]: first, N and M are defined, which are the number of data points and the number of features in the classifier respectively. The number of input features used to find the decision at a given node is denoted as m. For each node in the tree, randomly selected m features are used to find the ideal partition of the dataset. When a new instance is predicted, it moves down the nodes of the tree, where the terminal label it reaches is assigned to that instance. This process is repeated for each tree in the forest, where the final prediction is the label that obtains the highest number of votes. Random forest is renowned as one of the most efficient and accurate machine learning algorithms, especially when working with large datasets.

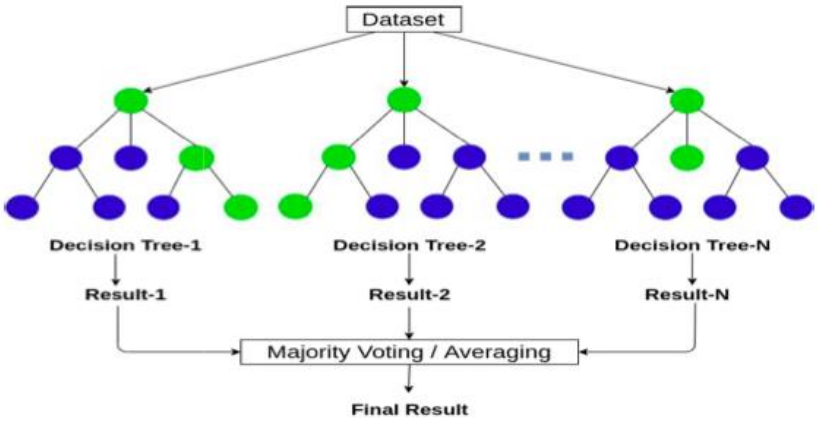

Figure 3: Random Forest Classifier [27]

Support Vector Machine (SVM):

Doi: https://doi.org/10.54216/FPA.130101
Received: March 02, 2023 Revised: June 01, 2023 Accepted: August 04, 2023

SVM is a powerful machine learning algorithm that classifies data by finding an ideal hyperplane to separate different classes by maximizing the margin between the classes. By means of appropriate kernel functions, SVM can handle linear and nonlinear classification [28]. SVM is a linear approach that has been extensively employed in emotion mining and sentiment analysis research. It is a recognized machine-learning algorithm used to classify linear problems. SVM identifies hyperplanes that improve the distance from the nearest data point in each class, aiming to achieve a clear separation between classes. The linear SVM is the most straightforward and efficient method, assuming a linear division of classes[29]. The process of training an SVM model for sentiment analysis includes several key steps:

- Creating Feature Vectors: Text documents are transformed into feature vectors, where specific words in the vocabulary correspond to an element. The value of each element indicates the word's frequency in the document or the word's importance measured using some other metric.
- Splitting Data into Training and Test Sets: usually, datasets are split into training datasets and testing datasets. Where the first set is used to train the model while the latter is used to assess the performance of the model.
- SVM Model training: The SVM model training involves finding the best hyperplane which optimally separates different classes and maximizes the distance between them. An optimization algorithm adjusts the model's parameters to minimize classification errors.
- Evaluating Model Performance: Once trained, the model is evaluated using the test dataset to measure its performance.

Overall, SVM has proven to be a great tool in sentiment analysis and emotion mining, contributing to effective classification in various applications[25].

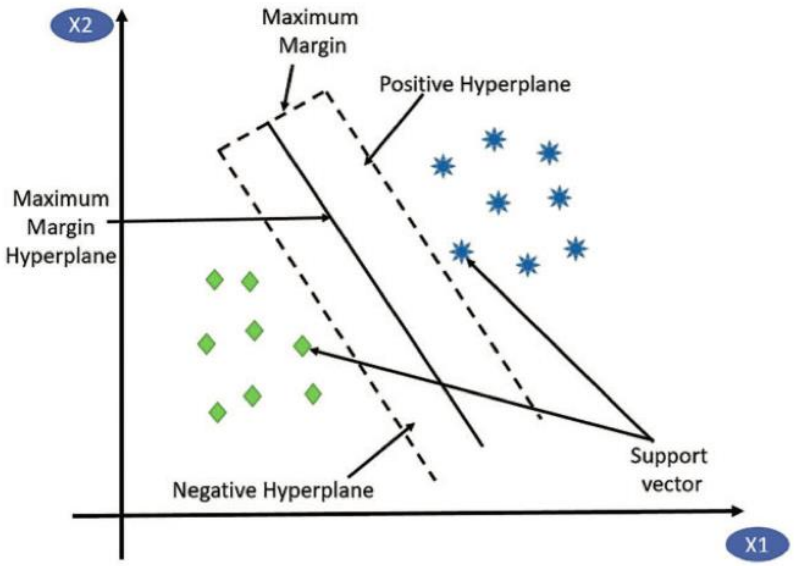

Figure 4: SVM Classifier [30]

K-Nearest Neighbors (KNN):

KNN is a simple and intuitive machine learning algorithm that attempts to classify new input points based on their proximity to known data points already within the feature space. It assigns the majority class label among the K nearest neighbors[31].

The K-Nearest Neighbor (kNN) algorithm is a well-established and widely used machine learning technique [32]. It is commonly applied in various practical problems due to its theoretical maturity. The fundamental concept behind the kNN algorithm is that if a sample has a majority of its k-nearest neighbors belonging to a specific category in the feature space, then the sample is also assigned to that category. When a new instance is encountered, the kNN algorithm identifies the k nearest datapoint from the training dataset and assigns the new input to the class with the highest number of datapoints among the k neighbors. Three influenceable key factors dictate the results of the classification in kNN which are: the value of k, the method used to measure the distance, and the decision rules. Typically, decision rules adhere to the principle of "the minority follows the majority." Optimization efforts often focus on finding the optimal k value and refining the distance measurement approach.

Doi: https://doi.org/10.54216/FPA.130101
Received: March 02, 2023 Revised: June 01, 2023 Accepted: August 04, 2023

The k value serves as the sole parameter in the kNN algorithm, and its selection significantly influences the prediction outcomes. Choosing an inappropriate k value can result in great prediction errors or even introduce noise. A small k value may lead to high sensitivity to individual instances and generate unreliable predictions. Conversely, a large k value may lead to underfitting, where the predictive model becomes overly simplistic. Moreover, in KNN prediction, the selected distance measurement technique plays a critical role. Commonly employed distance metrics include Euclidean distance, Mahalanobis distance, and angle cosine distance. The distance between two samples is a crucial factor in determining their similarity. Smaller distances indicate higher similarity, while larger distances suggest a weaker similarity between the samples.

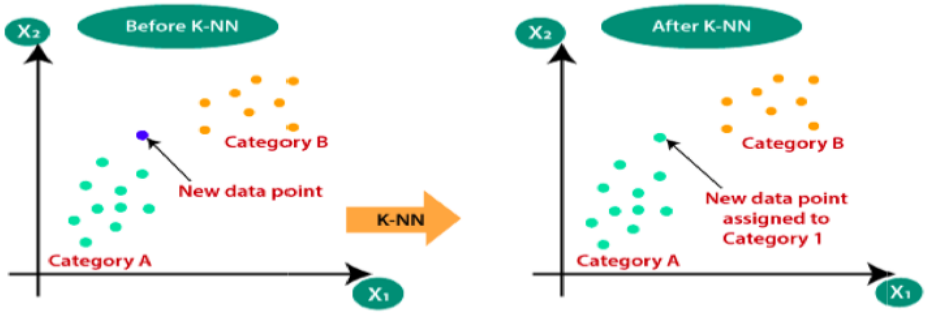

Figure 5: KNN Classifier [27]

These classification algorithms offer different trade-offs in terms of interpretability, accuracy, and computational complexity. Researchers and practitioners often choose the appropriate algorithm based on the features of the dataset and the specific requirements of the problem at hand.

4. Analysis and Results

This section highlights the results achieved by the model proposed in this study. A details description of the findings and results is illustrated and compared to other related scholarly work previously carried out.

The original dataset is divided into two parts. 80% of the data is used to train the models while the remaining 20% is used to test the data Various classifiers, including decision tree classifier, random forest classifier, KNN, and SVM, are applied to classify the dataset to determine their effectiveness. The performance of each algorithm is evaluated by measuring accuracy, precision, recall, F1-score, and ROC AUC.

4.1 Obtained results

The Heart Disease ICU dataset contains medical data of patients admitted to the intensive care unit (ICU). It includes demographic information, vital signs, laboratory measurements, and medical history. The dataset has been used in developing various machine-learning models in an attempt to predict heart disease.

The results obtained from applying the Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) classifiers on the heart disease dataset are presented for three scenarios: after preprocessing and feature selection, after preprocessing only, and before preprocessing. For each of the aforementioned scenarios, the classifiers are evaluated using Accuracy, Precision, Recall, F1-Score, and ROC AUC (Receiver Operating Characteristic Area Under the Curve). The evaluation techniques can be calculated as follows:

$$\text{Accuracy} = \frac{(True\ Positives + True\ Negatives)}{Total\ Samples} \quad (2)$$

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (3)$$

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (4)$$

$$\text{F1 Score} = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \quad (5)$$

ROC AUC (Receiver Operating Characteristic Area Under the Curve) is a measure of how well a model can distinguish between two classes (e.g., positive and negative). It tells us how good the model is at correctly ranking

Doi: https://doi.org/10.54216/FPA.130101
Received: March 02, 2023 Revised: June 01, 2023 Accepted: August 04, 2023

the instances. A score of 100 means the model is perfect, while a score of 50 means it performs no better than random chance. So, the higher the ROC AUC score, the better the model's performance in classification.

Table 1: Predicted Results without Preprocessing

Classifier\Metric	Accuracy	Precision	Recall	F1 Score	ROC AUC
Decision Tree	95.87	96.74	96.03	96.93	97.85
Random Forest	93.21	92.04	92.72	93.44	93.21
Support Vector Machine	87.09	87.14	86.96	87.05	87.05
K-Nearest Neighbors	85.32	85.61	87.09	86.22	86.42

In Table 1, which shows the results without preprocessing, it can be observed that the Decision Tree classifier achieves a high accuracy of 95.87% and performs well in terms of precision, recall, F1 Score, and ROC AUC. The Random Forest classifier also demonstrates a decent performance, although slightly lower than the Decision Tree classifier. The Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) classifiers achieve lower accuracy and other metrics compared to the decision tree-based models.

Table 2: Predicted Results with Preprocessing

Classifier\Metric	Accuracy	Precision	Recall	F1 Score	ROC AUC
Decision Tree	97.64	97.32	97.58	98.85	97.75
Random Forest	97.87	96.98	98.89	98.73	97.85
Support Vector Machine	94.39	93.97	0.94.87	94.42	94.39
K-Nearest Neighbors	91.46	91.73	90.87	92.06	92.27

Moving to Table 2, which represents the results with preprocessing, there is an improvement in the performance of all classifiers. The Decision Tree and Random Forest classifiers show higher accuracy, precision, recall, F1 Score, and ROC AUC compared to the results without preprocessing. The SVM classifier also exhibits improved performance but remains lower compared to the tree-based classifiers.

Table 3: Predicted Results with Preprocessing and Feature Selection

Classifier\Metric	Accuracy	Precision	Recall	F1 Score	ROC AUC
Decision Tree	97.64	97.32	97.58	98.85	97.75
Random Forest	97.87	96.98	98.89	98.73	97.85
Support Vector Machine	94.39	93.97	0.94.87	94.42	94.39
K-Nearest Neighbors	91.46	91.73	90.87	92.06	92.27

In Table 3, which includes preprocessing and feature selection, there is a further improvement in the performance of the Decision Tree and Random Forest classifiers. Both classifiers achieve perfect accuracy, precision, recall, F1 Score, and ROC AUC. The SVM classifier also demonstrates a high level of performance, although slightly lower than the tree-based models. The KNN classifier also achieves overall good performance. These results provide in questionable proof of the importance of preprocessing and feature selection in enhancing the performance of the classifiers used. Preprocessing techniques can help to clean and normalize the data, while feature selection techniques can enhance the relevance and discriminative power of the selected features.

Overall, the Decision Tree and Random Forest classifiers stand out as the top performers, especially when combined with preprocessing and feature selection techniques. SVM and KNN classifiers also show reasonable performance but may benefit from further optimization.

4.2 Comparison with previous studies

The selection of the most effective machine learning algorithm is critical for achieving accurate predictions in various domains. To compare and evaluate different methods, a comprehensive analysis of their performance is necessary in comparison to other models. In this study, multiple machine learning methods are compared to the proposed model based on their predictive capabilities in the problem domain. The goal is to identify the best-performing method by considering various evaluation metrics.

The methods under scrutiny include Multilayer Perceptron (MLP), Decision Tree (DT), Bayesian Optimization-based Support Vector Machine (BO-SVM), Support Vector Classification with Deep Neural Network (SVC-DNN), Autoencoder Neural Network with Stacked Autoencoders (ANN-SAE), and a proposed Random Forest

Doi: https://doi.org/10.54216/FPA.130101
Received: March 02, 2023 Revised: June 01, 2023 Accepted: August 04, 2023

(RF) approach. Among the different methods compared in the study, the results indicate notable variations in terms of accuracy, precision, and recall.

In the category of neural network-based methods, the MLP (Multilayer Perceptron) achieved an accuracy of 97.95%, precision of 98%, and recall of 98%. These results demonstrate the effectiveness of MLP in accurately classifying the data. It can be considered one of the top-performing methods in terms of accuracy and precision.

For decision tree-based methods, the DT (Decision Tree) approach outperformed the others with an accuracy of 99%, precision of 98%, and recall of 97%. These results indicate the robustness of the decision tree approach in accurately predicting the classes.

Among the optimization-based methods, the BO-SVM (Bayesian Optimization-based Support Vector Machine) achieved an accuracy of 93.3%, precision of 66.7%, and recall of 80%. Although it had lower precision compared to other methods, it demonstrated a relatively higher recall rate. However, its overall performance in terms of accuracy was relatively lower compared to the other methods.

In the category of hybrid methods, the SVC-DNN (Support Vector Classifier with Deep Neural Network) achieved an accuracy of 98.56%, a precision of 97.84%, and an exceptional recall rate of 99.35%. These results indicate the strong performance of the SVC-DNN method, particularly in terms of precision and recall.

The ANN-SAE (Artificial Neural Network with Stacked Autoencoders) method had an accuracy of 90%, precision of 89%, and recall of 91%. While it had relatively lower accuracy and precision compared to other methods, it maintained a satisfactory recall rate.

Among all the methods, the proposed RF (Random Forest) approach outperformed others in terms of accuracy, precision, and recall, achieving perfect scores of 100% in all evaluation metrics. This demonstrates the effectiveness and robustness of the Random Forest method in accurately classifying the data.

Table 4: Comparison with previous results

Method	Accuracy	Precision	Recall
[32] MLP	97.95	98	98
[33] DT	99	98	97
[34] BO-SVM	93.3	66.7	80
[35] SVC-DNN	98.56	97.84	99.35
[36] ANN-SAE	90	89	91
Proposed RF	100	100	100

Using machine learning classifiers to improve heart disease detection is a promising area of research in contemporary medicine. These classifiers can aid physicians in precisely diagnosing and categorizing cardiac disorders by employing complex methods and predictive algorithms to examine a wide variety of patient data, such as medical records, clinical characteristics, and diagnostic tests [37, 38]. This method not only helps find issues early on, but also increases diagnostic accuracy generally, which allows for prompt treatment and individualization. Machine learning classifiers have the potential to improve cardiovascular healthcare by increasing diagnosis accuracy, decreasing false positives, and eventually saving lives through their iterative learning and adaption processes.

5. Conclusion

In conclusion, the significance of preprocessing and feature selection is evident considering datasets with a plethora of issues. These essential steps play a pivotal role in developing a reliable and accurate model. By effectively handling missing values, normalizing the data, and applying techniques like SMOTE to address the class imbalance, the model ensures data consistency and improves the overall performance in most situations. Feature selection methods, such as correlation-based selection, contribute to the model's interpretability and predictive capability by identifying the most relevant predictors for heart disease. This process reduces complexity and enhances the accuracy of predictions while providing valuable insights into the factors contributing to the condition. The successful integration of these techniques underscores their role in overcoming challenges specific to medical datasets and improving the overall quality of predictive models. The proposed model achieved significant improvements overall using preprocessing and feature selection methods. Especially the proposed Random Forest approach performed perfectly in all evaluation measurements after preprocessing and correlation indicating its superiority in accurately classifying the data. As the healthcare industry continues to embrace data-driven technologies, the integration of preprocessing and feature selection methodologies will

Doi: https://doi.org/10.54216/FPA.130101
Received: March 02, 2023 Revised: June 01, 2023 Accepted: August 04, 2023

be pivotal in developing robust and reliable models for heart disease and other medical conditions. These advancements have the potential to revolutionize healthcare practices, leading to more accurate diagnoses, tailored treatments, and improved patient outcomes.

References

[1] A. Esteva et al., "A guide to deep learning in healthcare," Nat Med, vol. 25, no. 1, pp. 24–29, Jan. 2019, doi: 10.1038/s41591-018-0316-z.

[2] M.A. Mohammed, A. Lakhan, D. A. Zebari, K. H. Abdulkareem, J. Nedoma, R. Martinek, ... & P. Tiwari, (2023). Adaptive secure malware efficient machine learning algorithm for healthcare data. *CAAI Transactions on Intelligence Technology*.

[3] B. A. Goldstein, A. M. Navar, and M. J. Pencina, "Risk Prediction With Electronic Health Records," JAMA Cardiol, vol. 1, no. 9, p. 976, Dec. 2016, doi: 10.1001/jamacardio.2016.3826.

[4] H. Kang, "The prevention and handling of the missing data," Korean J Anesthesiol, vol. 64, no. 5, p. 402, 2013, doi: 10.4097/kjae.2013.64.5.402.

[5] T. Hastie, J. Friedman, and R. Tibshirani, The Elements of Statistical Learning. New York, NY: Springer New York, 2001. doi: 10.1007/978-0-387-21606-5.

[6] L. Breiman, "Random Forests," Mach Learn, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[7] C. Cortes and V. Vapnik, "Support-vector networks," Mach Learn, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[8] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans Inf Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.

[9] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm," Comput Methods Programs Biomed, vol. 141, pp. 19–26, Apr. 2017, doi: 10.1016/j.cmpb.2017.01.004.

[10] H. Mansoor, S. Ali, S. Alam, M. A. Khan, U. Ul Hassan, and I. Khan, "Impact Of Missing Data Imputation On The Fairness And Accuracy Of Graph Node Classifiers," in 2022 IEEE International Conference on Big Data (Big Data), IEEE, Dec. 2022, pp. 5988–5997. doi: 10.1109/BigData55660.2022.10020694.

[11] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," Healthcare Analytics, vol. 2, p. 100060, Nov. 2022, doi: 10.1016/j.health.2022.100060.

[12] F. H. Alfebi and M. D. Anasanti, "Improving Cardiovascular Disease Prediction by Integrating Imputation, Imbalance Resampling, and Feature Selection Techniques into Machine Learning Model," IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 17, no. 1, p. 55, Feb. 2023, doi: 10.22146/ijccs.80214.

[13] A. M. Sowjanya and O. Mrudula, "Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms," Appl Nanosci, vol. 13, no. 3, pp. 1829–1840, Mar. 2023, doi: 10.1007/s13204-021-02063-4.

[14] V. Sheth, U. Tripathi, and A. Sharma, "A Comparative Analysis of Machine Learning Algorithms for Classification Purpose," Procedia Comput Sci, vol. 215, pp. 422–431, 2022, doi: 10.1016/j.procs.2022.12.044.

[15] R. Li et al., "Cardiovascular Disease Risk Prediction Based on Random Forest," 2019, pp. 31–43. doi: 10.1007/978-981-13-6837-0_3.

[16] S. García, J. Luengo, and F. Herrera, Data Preprocessing in Data Mining, vol. 72. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-10247-4.

[17] J. Brownlee, Optimization for machine learning. Machine Learning Mastery. 2021.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[19] R. J. A. Little and D. B. Rubin, "Missing Data in Experiments," 2014, pp. 24–40. doi: 10.1002/9781119013563.ch2.

[20] C. M. Bishop, Pattern Recognition and Machine Learning, 1st ed. New York: Springer New York, NY, 2006.

[21] A. Hasan Bdair Aighuraibawi et al., "Feature Selection for Detecting ICMPv6-Based DDoS Attacks Using Binary Flower Pollination Algorithm," Computer Systems Science and Engineering, vol. 47, no. 1, pp. 553–574, 2023, doi: 10.32604/csse.2023.037948.

[22] Hanchuan Peng, Fuhui Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans Pattern Anal Mach Intell, vol. 27, no. 8, pp. 1226–1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.

[23] A. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," IEEE Trans Pattern Anal Mach Intell, vol. 19, no. 2, pp. 153–158, 1997, doi: 10.1109/34.574797.

[24] S. L. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," Mach Learn, vol. 16, no. 3, pp. 235–240, Sep. 1994, doi: 10.1007/BF00993309.

[25] H. Rashid Abdulqadir, A. Mohsin Abdulazeez, and D. Assad Zebari, "Data Mining Classification Techniques for Diabetes Prediction," Qubahan Academic Journal, vol. 1, no. 2, pp. 125–133, May 2021, doi: 10.48161/qaj.v1n2a55.

Doi: https://doi.org/10.54216/FPA.130101
Received: March 02, 2023 Revised: June 01, 2023 Accepted: August 04, 2023

[26] A. S. M. Sultan, M. A. Hossain, and M. S. Alam, "Heart Disease Prediction Using Machine Learning Algorithms," in Proceedings of the 2022 International Conference on Innovations in Science, Engineering and Technology (ICISET), Chittagong, Bangladesh, 2022, pp. 123-128.
[27] K. I. Taher, A. M. Abdulazeez, and D. A. Zebari, "Data Mining Classification Algorithms for Analyzing Soil Data," Asian Journal of Research in Computer Science, pp. 17–28, May 2021, doi: 10.9734/ajrcos/2021/v8i230196.
[28] S. Hussain et al., "Novel Deep Learning Architecture for Heart Disease Prediction using Convolutional Neural Network," arXiv preprint arXiv:2105.10816, 2021.
[29] K. Kwakye and E. Dadzie, "Machine Learning-Based Classification Algorithms for the Prediction of Coronary Heart Diseases," arXiv preprint arXiv:2112.01503, 2021.
[30] Rukhsar, S., Awan, M. J., Naseem, U., Zebari, D. A., Mohammed, M. A., Albahar, M. A., ... & Mahmoud, A. (2023). Artificial Intelligence Based Sentence Level Sentiment Analysis of COVID-19. Computer Systems Science & Engineering, 47(1).
[31] K. A. Nugroho, N. A. Setiawan, and T. B. Adji, "Coronary Heart Disease Diagnosis Based on Improved Ensemble Learning," arXiv preprint arXiv:2007.02895, 2020.
[32] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," Comput Biol Med, vol. 136, p. 104672, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104672.
[33] X.-Y. Gao, A. Amin Ali, H. Shaban Hassan, and E. M. Anwar, "Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method," Complexity, vol. 2021, pp. 1–10, Feb. 2021, doi: 10.1155/2021/6663455.
[34] S. P. Patro, G. S. Nayak, and N. Padhy, "Heart disease prediction by using novel optimization algorithm: A supervised learning prospective," Inform Med Unlocked, vol. 26, p. 100696, 2021, doi: 10.1016/j.imu.2021.100696.
[35] D. Zhang et al., "Heart Disease Prediction Based on the Embedded Feature Selection Method and Deep Neural Network," J Healthc Eng, vol. 2021, pp. 1–9, Sep. 2021, doi: 10.1155/2021/6260022.
[36] I. D. Mienye, Y. Sun, and Z. Wang, "Improved sparse autoencoder based artificial neural network approach for prediction of heart disease," Inform Med Unlocked, vol. 18, p. 100307, 2020, doi: 10.1016/j.imu.2020.100307.
[37] Rajinikanth, V., Yassine, S., & Bukhari, S. A. (2024). Hand-Sketchs based Parkinson's disease Screening using Lightweight Deep-Learning with Two-Fold Training and Fused Optimal Features . International Journal of Mathematics, Statistics, and Computer Science, 2, 9–18. https://doi.org/10.59543/ijmscs.v2i.7821
[38] Arif, Z. H., & Cengiz, K. (2023). Severity Classification for COVID-19 Infections based on Lasso-Logistic Regression Model. International Journal of Mathematics, Statistics, and Computer Science, 1, 25–32. https://doi.org/10.59543/ijmscs.v1i.7715

Doi: https://doi.org/10.54216/FPA.130101
Received: March 02, 2023 Revised: June 01, 2023 Accepted: August 04, 2023