Investigating Different Classification Algorithms for Data-Driven Cardiovascular Disease Prediction Models using Machine Learning

Sonal Verma[1*] , Sanjeev Kumar[2]

[1]Shri Venkateshwara University, Department of Computer science 244235, Amroha-India

[2]Maharaja Agrasen Institute of technology, Department of computer science 110085, New Delhi-India

Abstract:

Chronic cardiovascular illnesses pose a substantial threat to the health of people all over the world, which indicates that there is an urgent need for the development of detection systems that are both more accurate and more efficient. It is still significant to enhance prediction models to solve the gaps in the latest detection methodologies, although a number of research studies have offered useful insights in this sector. The study developed a model using machine learning that can predict the risk of cardiovascular disease based on a dataset containing eleven different characteristics that may be utilized to make predictions. A dataset cantered on heart failure was utilized to assess a various of machine learning classification methodology. To determine the optimal model, the outcomes from support vector machines ,neural networks, K-nearest neighbours, random forests, and decision trees were evaluated. A specificity of 0.97, an area under the curve of 0.97, and a Matthews correlation coefficient of 0.92 were collected using the K-nearest neighbours approach. Due to the good percentage of accuracy, it was regarded as the most effective model. On the other hand, the accuracy rates of decision tree and support vector machine were much lower. The early identification of heart illness are potential benefits of the proposed prediction models. It may also support patients in controlling their sickness or life forms to improve recovery/survival. The study indicates that machine-learning algorithms may enhance illness detection and treatment outcomes with more precision.

Keywords: Cardiovascular Disease, machine learning model, classification, Prediction

1.Introduction

Cardiovascular disorders (CVD) are a prominent source of worldwide illness and mortality [1,2]. Early detection and diagnosis of these illnesses is a major healthcare problem [3]. Heart disorders, or cardiovascular illness, are a varied set of ailments that affect the heart and blood vessel arteries. Conditions affecting the heart include diseases impacting the coronary arteries, heart valves, and cardiac muscles, [4]. Worldwide, CVD is the leading cause of mortality for both male and female. While medical treatment has improved and risk factors for this disease, such as blood pressure, high cholesterol, smoking, and poor nutrition have reduced mortality rates [5]. In the early stages of periodontal disease, patients often do not experience any symptoms. This is because the illness is still in its early stages. It is possible that the illness might transmit to the other person under certain circumstances. The illness caused damage to the cardiac valves and coronary arteries, which might result in potentially severe consequences [6]. An accurate diagnosis requires a detailed description of the several signs that the illness might present. AI is a subfield of machine learning (ML), which is used in cardiovascular care. The ability of a computer to grasp data and classify tasks is the subject of this discussion. With the use of input data (such as photographs or text), the machine learning framework makes predictions about outputs by combining computational optimization and statistical evaluation. As noted [7], machine learning has a significant capacity to bring about improvements in the field of healthcare. Its remarkable improvements are due to its superior data processing capabilities. As a result, the healthcare sector has seen the emergence of several AI applications that use the rapidity and precision of machine learning, enabling groundbreaking solutions to various healthcare issues. The precise monitoring of patients at risk for heart failure is crucial for execution of preventative management programs [8,9]. These measures may related to lifestyle modifications, like as a nutritious diet and consistent physical activity, together with the management of menace factors, including hypertension as well raised cholesterol levels. Accurate categorization enables the allocation of resources and attention to those in greatest need. This study investigates the usage of the Kaggle platform's "Heart Failure" dataset, which contains roughly eleven critical characteristics for cardiovascular illness categorization. These traits are crucial for accurately classifying heart disorders using supervised learning and machine-learning approaches, since they supply important information about a patient's cardiovascular health. Robust measures like specificity, AUC, recall, and

precision highlight machine learning's promise for early diagnosis and clinical decision-making. The model not only accurately predicts cardiac failure but can also be incorporated into The Clinical Decision Assistance System (CDAS) identifies vulnerable patients in real time, prioritizes essential cases, and optimizes resources to improve hospital efficiency and clinical outcomes.

1.1 Contributions to the work

The purpose of this research is to present a robust machine learning (ML) framework for the prediction of CVD. This framework is developed by evaluating five distinct classifiers, including neural networks (NN), random forests (RF), support vector machines (SVM), K-nearest neighbours (KNN), and decision trees (DT), on a dataset that pertains to heart failure. An examination that is more comprehensive may be carried out with the help of the Matthews Correlation Coefficient (MCC), in addition to the area under the curve (AUC) and specificity. Because of its remarkable performance, the KNN model is an excellent choice for providing assistance in the process of clinical decision-making. The multi-metric evaluation and model comparison that is being presented here provide a realistic technique for the accurate and early identification of cardiovascular disease.

2.Related Work

This section provides technical foundation and overview of relevant literature on early prediction of heart disease using machine learning approaches. The authors [10] created two deep neural networks utilizing well-organized training datasets to accurately forecast cardiovascular heart disease probability. Process of prediction Inconsistent real-world data hinders learning. They advised isolating regular subsets from highly biased subsets for training instead of using whole datasets or randomly chosen ones. The proposed technique had an AUC of 0.882 and an accuracy of 0.892. A hybrid system for heart disease detection using medical voice recordings was reported in [11]. The hybrid structure was four-layered. In the first layer, segmentation was recommended, followed by feature extraction in the second layer. In the third layer, relevant configurations were chosen using optimization techniques, followed by validation in the fourth layer. The hybrid structure was believed to increase accuracy. Different methods are used for CVD detection. The first technique uses AI models to examine test findings and differentiate between CVD patients and healthy individuals. Electrocardiogram signals are used in the second method. Heart sound cues are crucial for ML models to categorize people as no CVD or having CVD [12]. While medical care systems for identifying hidden patterns and forecasting illnesses are currently absent, ML may help predict cardiac ailments early, enabling prompt intervention and therapy [13]. In the paper [14] analysed 299 heart disease patients from the Faisalabad Institute of Cardiology. The dataset had 13 attributes and a designated “Death Event” as the target variable for binary classification. The dataset was pre-processed for quality and steadiness. Post-preprocessing, the dataset was divided into train and test sets for model training and assessment. Different feature selection approaches were used to determine most important properties for heart failure prediction in the train set. This research used a dataset [15] including 14 different features. In this research, different preparation methods were applied to 14 attributes grouped into four main categories, with each group containing about 6 features, along with various machine learning models. Various research investigated the present developments and limitations of using machine learning for the detection of CVD. The study [16] presented many data mining and machine learning approaches. These approaches included heartbeat segmentation and selection methodologies, electrocardiogram imagery, carotid artery imaging, among other pertinent topics. Table 1 below summarizes the performance metrics associated with the current methodologies under assessment, with each item matching to particular assessment criteria.

Table 1. *Survey of the effectiveness of the current methodologies.*

Year	About Datasets	Best Model	Accuracy
2024 [17]	1 (Bangladesh patient dataset)	Random Forest	98.04%
2023 [18]	1 (Isfahan Cohort Study – 5,432 subjects)	Quadratic Discriminant Analysis (QDA)	75.50%
2024 [19]	5 (Heart Statlog, UCI, Stroke, Framingham, CHD)	Ensemble Model (Optimized RF)	98.48%
2019 [20]	1 (UCI Cleveland dataset)	Hybrid Random Forest	88.70%

2019 [21]	1 (UCI Heart Disease dataset)	Voting Classifier (Naive Bayes + Logistic Regression)	87.40%
2020 [22]	1 (UCI Cleveland dataset)	Random Forest	95.08%
2022 [23]	1 (Heart Disease dataset)	XGBoost (Bayesian Optimization)	91.80%

In the study author [24] used UCI Machine Learning Repository datasets to study CVD. Only 14 of their 76 traits—including the target trait—were significant for analysis. The researchers used the Cleveland Clinic Foundation dataset (three hundred three patients) and the Hungarian Institute of Cardiology dataset (294 patients). Many machine learning methods were used to predict cardiac disease.

3. Materials and Methods

In this study, the dataset on heart illness that was obtained from the open source platform in San Francisco, California, United States of America, and given the moniker "Heart Failure" aids as a fused resource for the study of cardiovascular disorders. This dataset is comprised of approximately 11 essential features for the classification of CVD. Utilizing MATLAB 2024a, the dataset analysis and ML model building were carried out in a collaborative setting. This was accomplished by utilizing MATLAB computing capabilities for effective data handling and testing.

3.1 Analysis of data and preprocessing

The dataset consist of 1018 items and 11 attributes, as seen in Table 2. The variables include age, sex, kind of chest discomfort, resting blood pressure, cholesterol levels, fasting blood sugar, and resting electrocardiogram (ECG).

Table 2. *Longitudinal characteristics used to classify heart diseases.*

SN	Input features	Type	Mean ± Std Dev	Min–Max	Units
1	Age	Numerical	53.51 ± 9.43	28.0–77.0	years
2	Sex	Categorical (f:0,m:1)	-	0-1	-
3	Chest Pain Type	Categorical	-	0-4	-
4	Resting BP	Numerical	132.40 ± 18.51	0.0–200.0	mm Hg
5	Cholesterol	Numerical	198.80 ± 109.38	0.0–603.0	mg/dL
6	Fasting BS	Binary	0.23 ± 0.42	0.0–1.0	-
7	Resting ECG	Categorical	-	0-2	-
8	Max HR	Numerical	136.81 ± 25.46	60.0–202.0	bpm
9	Exercise Angina	Binary	-	0-1	-
10	Old peak	Numerical	0.89 ± 1.07	−2.6–6.2	-
11	ST_ Slope	Categorical	-	0-3	-
12	Heart Disease	Binary	0.55 ± 0.50	0.0–1.0	-

The dataset includes ECG, heart rate, exercise-induced angina, ST depression, and ST slope as shown in figure 1. This dataset is the most comprehensive open-source accessible compilation on heart disease, offering a solid basis for research and examination. This extensive data promotes the examination of many risk variables linked to heart disease and allows researchers to create prediction models and find possible therapies. Utilizing this comprehensive knowledge, we can enhance our comprehension of cardiac diseases and optimize patient outcomes.

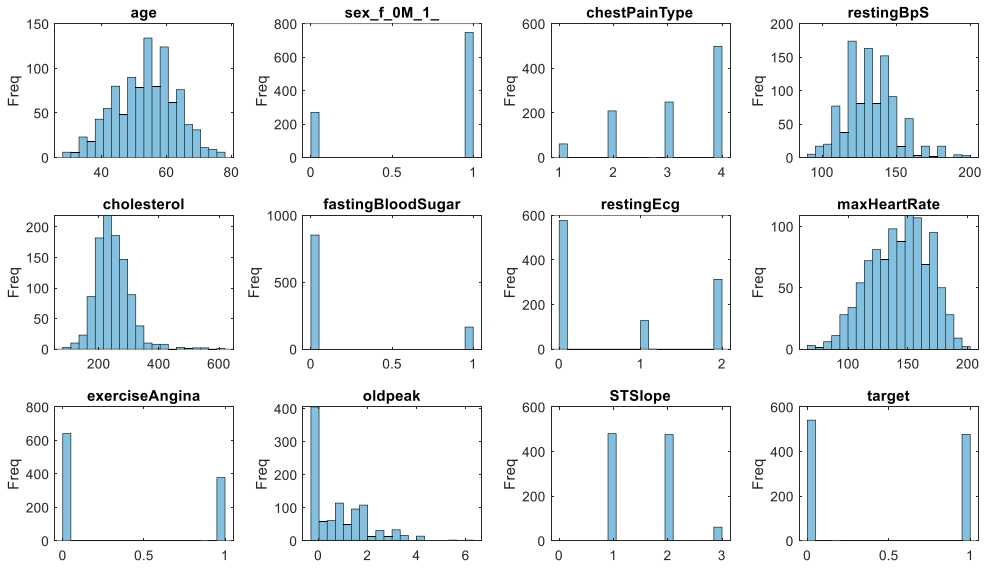

***Figure 1**: Frequency plots of the features*

Analysis revealed low-frequency categorical variables after the database cleaning process was complete. Because none of the categories in the columns that were examined met the 5% threshold for infrequent categories, each and every one of them was considered. The original classifications were preserved. For natural order categorical variables, we used ordinal coding, and for the other variables, we applied one-hot coding to change them into numbers for machine learning models. A correlation among feature shown in Figure 3, to investigate the links among the variables evaluated in the dataset. Each value inside the matrix denotes degree of correlation between two features, quantified on a scale ranging -1 to 1. A value of 1 shows positive correlation, -1 represents negative correlation, and 0 suggests no connection. The correlation matrix indicates that factors related to health concerns, including advanced age, hypertension, hyper lipidaemia, and diabetes, are also connected with an increased risk of heart disease. The target feature in this research denotes the existence or non-existence of CVD, with 1 signifying the presence of the disease and 0 indicating its absence. The distribution of the target variable shown in figure 2 ensures data representation of binary classes, thereby aiding the effective training of the model. This preprocessing approach improves the consistency of the model's evaluation and strengthens its capacity to generalize to new information. The bar chart in figure 2 describe about the distribution of target classes within the dataset for the purpose of predicting cardiovascular disease (CVD).

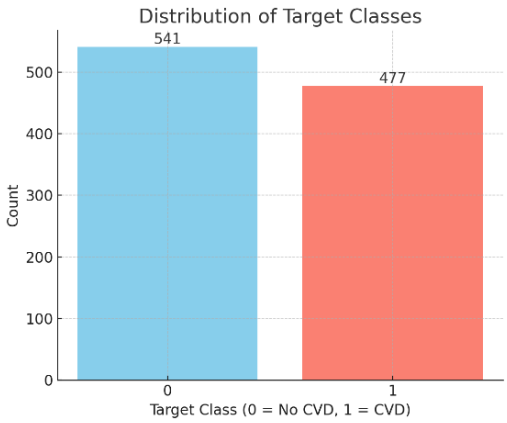

***Figure 2**: distribution of target data set*

Conversely, there were 477 cases indicating the presence of cardiovascular disease (class 1), whereas 541 occurrences evidenced the lack of cardiovascular illness (class 0). The dataset may be appropriate for training classification models due to its fairly balanced distribution. This would obviate the need for costly resampling methods to address class imbalance.

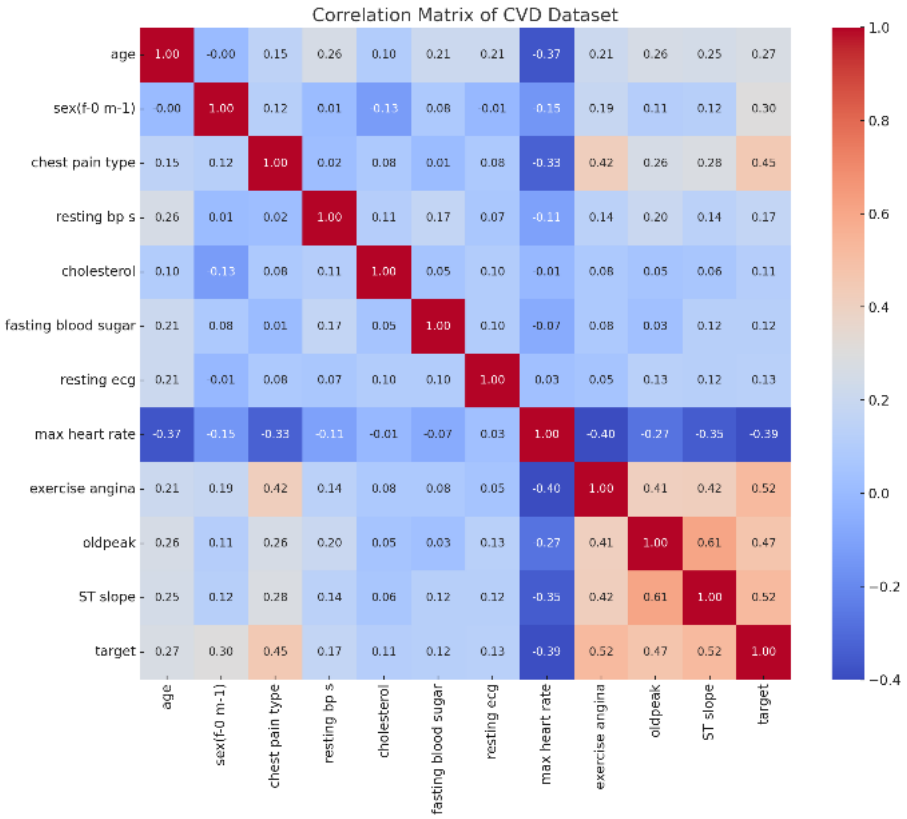

Figure 3 Corelation matrix of CVD

Data rescaling

Certain elements have large ranges, which might give the impression that they are more relevant than others over the course of our research. When it comes to classifying data, this might result in outcomes that are inconsistent. In order to resolve this issue, we must ensure that all of the elements are set to the same scale. In order to accomplish this, we rescale their values so that they fall inside the range of [0,1]. Making reliable comparisons and analyses of the data is made simpler as a result of this. This rescaling is applied to all of the characteristics in our dataset in order to guarantee uniformity and consistency.

$$A' = \frac{A - A_{min}}{A_{max} - A_{min}} \quad (1)$$

Where A_{max} and A_{min} are the maximum and minimum values of the respective features

Performance matrices of the model

The performance matrices of different models presented as shown in Eq(2) to (6) for the evaluation of various model.

$$\text{Precision(Pn)} = \frac{TP_n}{TP_n + FP_n} \quad (2)$$

$$\text{Recall } (R_n) = \frac{TP_n}{TP_n + FN_n} \quad (3)$$

$$\text{F1 score} = 2 \times \frac{\text{Pn} \times \text{Rn}}{\text{Pn} + \text{Rn}} \quad (4)$$

$$\text{Accuracy} = \frac{TP_n + TN_n}{TP_n + TN_n + FP_n + FN_n} \quad (5)$$

$$\text{Mcc} = \frac{Tcn \times Tnn - Fcn \times Fnn}{\sqrt{(Tcn+Fnn)(Tcn+Fnn)(Tcn+Fnn)(Tcn+Fcn)}} \quad (6)$$

Where,

TP: the test correctly identifies a patient with cardiovascular disease.

FP: the test incorrectly identifies a patient without cardiovascular disease as having the disease. TN: the test correctly identifies a patient without cardiovascular disease.

FN: the test incorrectly identifies a patient with cardiovascular disease as not having the disease

3.2 Machine learning models

This section provides a summary of ML methods for classification used in this study. These algorithms include NN,RF,DT,SVM and KNN

3.2.1 Artificial neural networks

The artificial neural networks (ANN) used in this study are non-linear functions that transfer an input values of data to a distinct value among a defined range of potential outputs. Each ANN is shown as a collection of linked nodes. The nodes, organized in layers, represent numeric values, while the connections denote multiplication and summation operations done sequentially, facilitating the interconnection of the neural node. Each operation has many weights that are calculated during the training phase. The outcome of the multiplication and summation operation at each node is processed by an activation function that determines whether the result is sent to the subsequent step or not. Each layer of the network may use a distinct activation function. By minimizing a network cost function, the weights for the network are calculated. Backpropagation via the network and gradient descent are necessary for the minimization procedure to fix the weights. In the event of multiclass classification, a similar procedure would be used, but with a categorical entropy loss function.

3.3.2 Random Forest

A more robust and precise prediction model may be created by the use of Random Forest (RF), which integrates numerous independent Decision trees (DTs) [25]. During the training process, it generates a large number of DTs and then outputs either the mode of the classes or the mean prediction of the individual trees. The randomness that is introduced during the process of tree-building, both in terms of the data samples that are used for training and the features that are considered at each split, improves the ability of the model to generalize and reduces the likelihood of overfitting, which ultimately results in improved overall performance and reliability when predicting outcomes for new data points [26,27].

3.3.3 Support vector machine

ML techniques that fall under the category of supervised learning are known as Support Vector Machines (SVM). These algorithms are often used for classification and regression problems. "Discriminant classifier is a term that is often used to describe it. The computational complexity is reduced as a result. Using a hyperplane, data is organized into distinct groups in order to facilitate classification. In accordance with the data that is supplied, a line or hyperplane is generated to differentiate between the classes. There are several applications for Support Vector Machines (SVM), including face recognition, handwriting recognition, the classification of news articles, the categorization of web pages, and the sorting of emails. On the basis of the input data, the SVM is used to establish the support vectors, which are the data points from each class that are closest to the line or hyperplane. To get the margin, one must first determine the distance that exists between each of these vectors and the line.

3.3.4 Decision Tree

Decision Trees (DTs) are sequential models that systematically integrate a series of elementary tests; each test evaluates a numerical features against a threshold value or a categorical attribute against a defined range of integer values[28]. Decision trees function as adaptable prediction and classification tools, systematically partitioning a dataset into subsets according to the values of relevant input variables or predictors [29]. This subdivision generates divisions and descendant nodes, referred to as leaves, which contain internally homogeneous goal values, while descending the tree reveals progressively disparate values across the leaves.

3.3.5 K-Nearest Neighbours

KNN is a classification and regression technique used in ML models. In classification, k-NN allocates a class to a point according to the classes of its closest neighbours inside a feature space [30]. In regression method, k-NN approximates the numerical value of a data point by calculating the mean of the data of its closest neighbours. The "k" in k-NN denotes the quantity of neighbours taken into account during prediction, and the selection of this value influences the model's accuracy [31]. K-NN is a simple method, particularly for data sets with complex patterns or shapes that are hard to describe with basic math.

4.Experimental Result

This section describes five machine learning methods for extracting features related to heart failure and classification. This includes ANN, RF, DT, k-NN, and SVM classifier approaches. The dataset was divided into two parts: 80% for training and 20% for testing. A high specificity model (≥0.90) is necessary to minimize false positives. Additionally, a high area under the curve (≥0.90) suggests almost perfect class discrimination.

4.1 Model implementation and Performance Evaluation

4.1.1 Artificial Neural Network

The confusion matrix of ANN model shown in figure 4, where this model perform well as if it identified 75 non CVD and 65 CVD cases, with 6 misclassifications each.

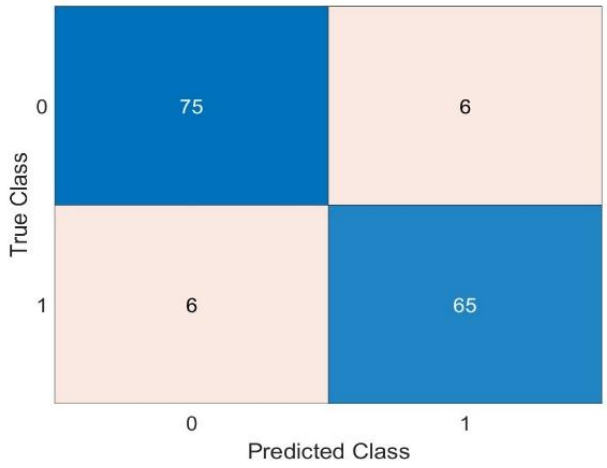

Figure 4: *Confusion matrix of ANN*

The ROC curve of ANN model represent the ability that it can differentiate well among CVD and non CVD cases , having area under the curve (AUC) of 0.9258 as shown in figure 5. On the basis of these performance matrix ANN model can be better option for the early detection of CVD.

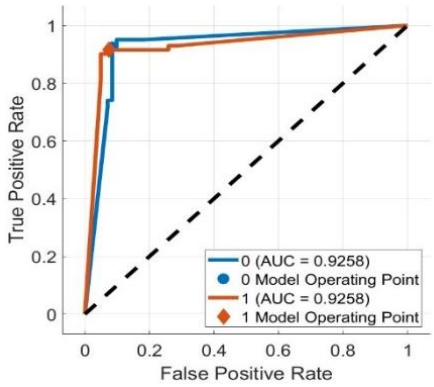

Figure 5: *ROC curve of ANN*

4.1.2 Random Forest

The confusion matrix of RF model shown in figure 6, where this model perform well as if it the total number of misclassification is only 9 .

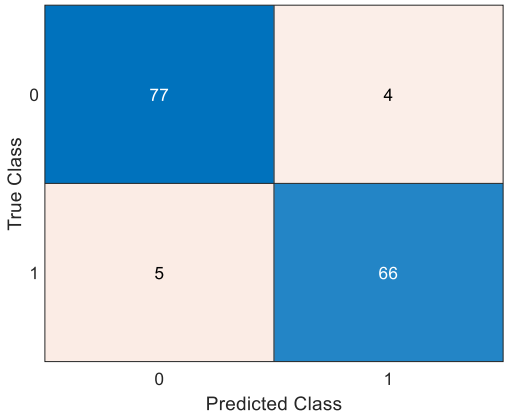

Figure 6: *Confusion matrix of Random Forest*

The ROC curve showed in figure 7 an Area Under the Curve (AUC) value of 0.9677 for both 0 and 1 classes. The RF model has good potential to identified disease quickly with lesser number of misclassification.

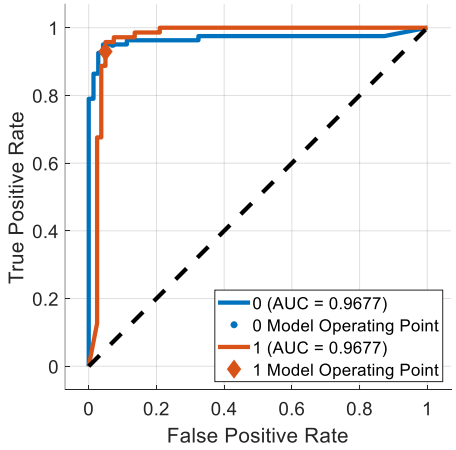

Figure 7: *ROC curve of RF*

4.1.3 Decision Tree

In the total samples the DT perform with moderate accuracy , it predicted 74 true negatives (non-CVD) and 65 true positives (CVD), the total misclassification values are 13 as shown in figure 8.

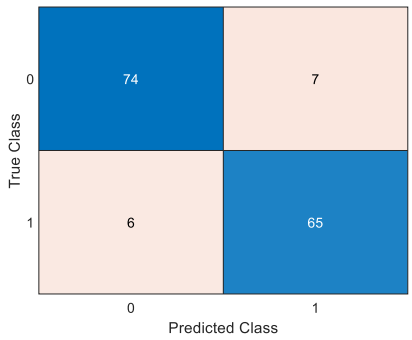

Figure 8: *Confusion matrix of DT*

The ROC curve shown in figure 9 the model's discrimination ability with an AUC value of 0.9371 for both classes. Decision Tree presents a simpler, more interpretable alternative to the Random Forest model with outstanding accuracy and reliability for clinical decision assistance in early CVD diagnosis

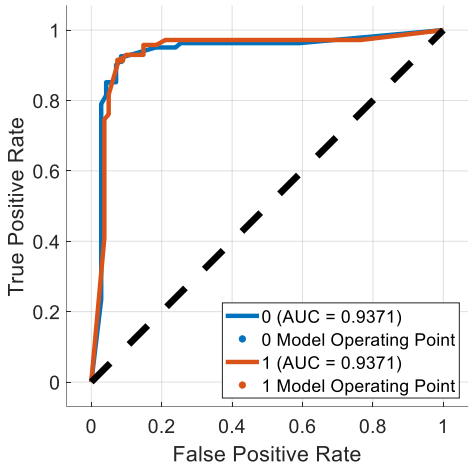

Figure 9: *ROC curve of DT*

4.1.4 K-NN

The KNN model perform better among all the mentioned model in this study having just 6 misclassified values which is shown in figure 10 This finding is further supported by the ROC analysis, which reveals shown in figure 10 that the model achieved better area under the curve (AUC) of 0.9654

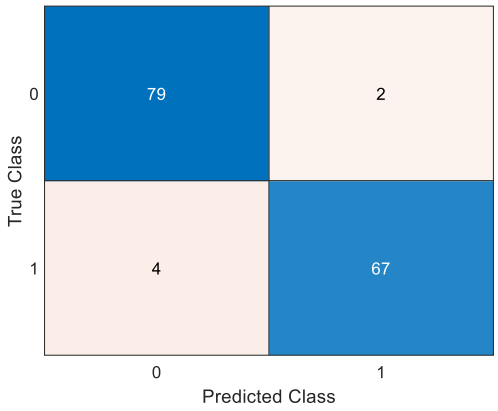

Figure 10: *Confusion matrix of KNN*

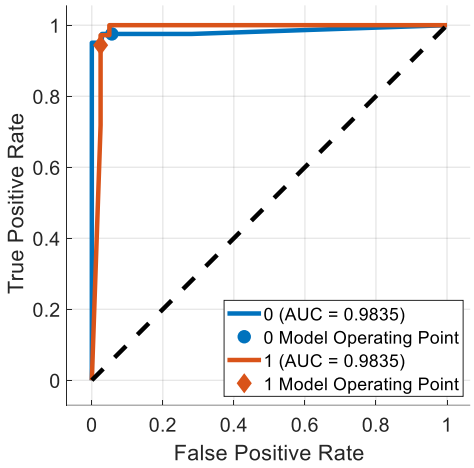

Figure 11: *ROC curve of KNN*

4.1.5 SVM

In the process of predicting cardiovascular illness, the Support Vector Machine (SVM) classifier displayed outstanding performance. It properly detected 76 instances of non-cardiovascular disease and 66 cases of

cardiovascular disease, with just five misclassifications in each category, demonstrating a sensitivity and specificity that is well-balanced shown in figure 12.

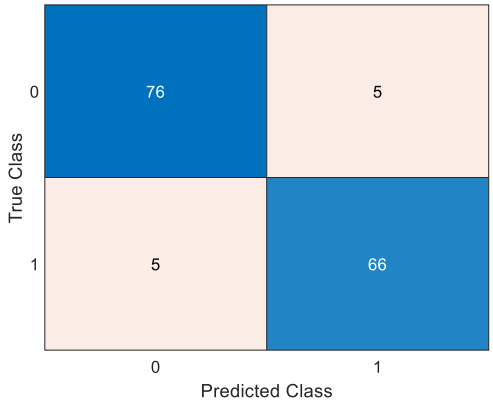

Figure 12: *Confusion matrix of SVM*

The robustness of the model is further reinforced by its ROC curve shown in figure 13, which produced a high Area Under the Curve (AUC) value of 0.9897. This result indicates that the model is able to discriminate between classes belonging to CVD and those that do not belong to CVD. These findings show the generalization power of the support vector machine (SVM) as well as its usefulness in managing high-dimensional clinical data, which makes it a trustworthy model for the early identification of cardiovascular disease (CVD).

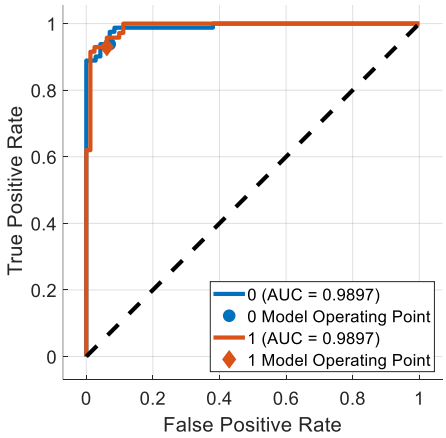

Figure 13: *ROC curve of SVM*

The results of the deployed approaches are summarized in Table 3, which includes the accuracy, recall, and F1 scores. By providing a summary of the performance of each approach with regard to accuracy and other essential characteristics, this table makes it possible to conduct a comparative assessment of the effectiveness of the various methods.

Table 3*. Comparative analysis of the models*

	Ac	**Pr**	**Re**	**F1**	**Sp**	**Se**	**MCC**
ANN	0.92	0.91	0.91	0.91	0.92	0.91	0.84
RF	0.94	0.94	0.93	0.93	0.95	0.93	0.88
DT	0.91	0.90	0.91	0.90	0.91	0.91	0.82
KNN	0.96	0.97	0.94	0.95	0.97	0.94	0.92
SVM	0.93	0.9	0.93	0.93	0.93	0.93	0.86

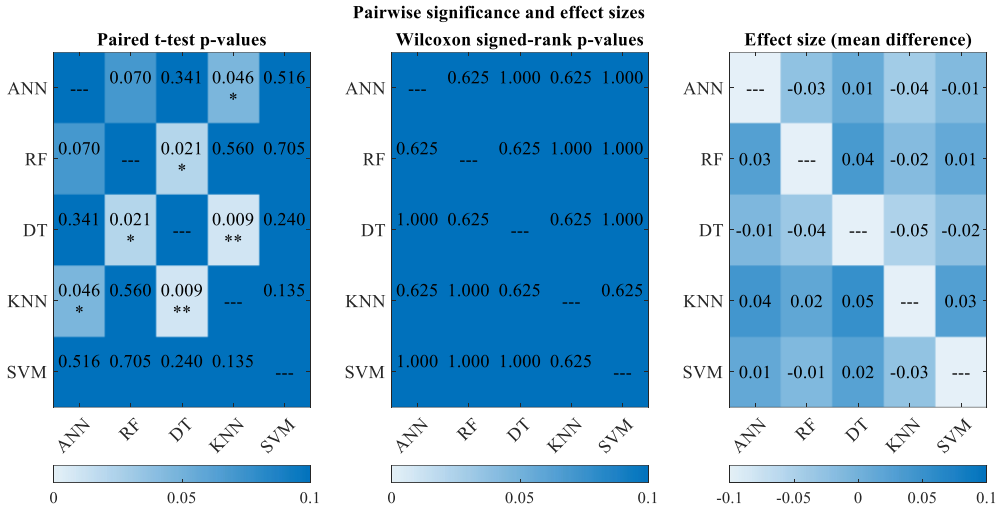

Figure 14:Pairwise significance and effect sizes

We conducted pairwise comparisons of cross-validation scores using both paired t-tests and Wilcoxon signed-rank tests with Bonferroni correction (Figure 14). This was done in order to determine whether or not the variations in performance between the models were statistically significant. The results of the paired t-test showed that KNN performed substantially better than DT ($p = 0.009$) and ANN ($p = 0.046$), while RF also performed significantly better than DT ($p = 0.021$). It was shown that there were no significant differences between KNN and RF or between KNN and SVM, which may be interpreted as indicating that these top-performing models provide outcomes that are statistically equivalent. After making the necessary adjustments, the Wilcoxon signed-rank test, which is considered to be more conservative, did not find any significant pairwise differences. These findings were validated by the effect size analysis, which revealed that KNN demonstrated a mean improvement of +0.05 in comparison to DT and +0.04 in comparison to ANN, whereas RF demonstrated a +0.04 advantage in comparison to DT. However, it is worth noting that the performance disparities across the leading models (KNN, RF, and SVM) were rather tiny (<0.02). This suggests that although KNN obtained the best overall accuracy, RF and SVM delivered approximately similar performance. When taken as a whole, these findings indicate that DT and ANN fared poorly in comparison to the other approaches, however KNN, RF, and SVM constitute a group of classifiers that are comparable in terms of their effectiveness for this dataset.

5.Conclusion

This research highlights the efficacy of machine learning algorithms in classifying heart failure, with the KNN model attaining the greatest accuracy of 96% and excelling in differentiating between positive and negative instances. The sophisticated preprocessing methods, including outlier elimination and class equilibrium, significantly enhanced the model's performance. The capacity of machine learning to process intricate data and provide precise diagnoses is essential for timely medical intervention and enhancing patients' quality of life.

This study is more accurate than others due to quick data processing and hyperparameter modification. Well-processed datasets outperform unprocessed ones, even with fewer attributes. In conclusion, machine learning improves medicine and might transform cardiovascular disease diagnosis and treatment. These types may be used with real-time monitoring equipment.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.

Data availability statement: The data that support the findings of this study are available on request from the corresponding author.

References

[1] Gaidai, O.; Cao, Y.; Loginov, S. Global Cardiovascular Diseases Death Rate Prediction; Elsevier: Amsterdam, The Netherlands, 2023; Volume 48, p. 101622.

[2] Ali´c, B.; Gurbeta, L.; Badnjevi´c, A. Machine learning techniques for classification of diabetes and cardiovascular diseases. In Proceedings of the 2017 6th Mediterranean Conference on Embedded Computing, MECO 2017—Including ECYPS 2017, Bar, Montenegro, 11–15 June 2017; Proceedings.

[3] Hawashin, B.; Mansour, A.; Fotouhi, F.; AlZu'bi, S.; Kanan, T. A Novel Recommender System Using Interest Extracting Agents and User Feedback. In Proceedings of the 2021 International Conference on Information Technology (ICIT), Amman, Jordan, 14–15 July 2021.

[4] Jinjri, W.M.; Keikhosrokiani, P.; Abdullah, N.L. Machine Learning Algorithms for the Classification of Cardiovascular Disease- A Comparative Study. In Proceedings of the 2021 International Conference on Information Technology, ICIT 2021—Proceedings, Amman,Jordan, 14–15 July 2021; pp. 132–138.

[5] Heron M. Deaths: leading causes for 2019. National vital statistics reports70. Hyattsville, MD: National Center for Health Statistics; 2021. 9.

[6] Dhadse P, Gattani D, Mishra R. The link between periodontal disease and cardiovascular disease: how far we have come in last two decades. J Indian Soc Periodontol 2014 Jul;14(3):148–54

[7] Ramesh, A.N.; Kambhampati, C.; Monson, J.R.; Drew, P.J. Artificial intelligence in medicine. Ann. R. Coll. Surg. Engl. 2004, 86, 334.

[8] Krittanawong, C.; Virk, H.U.H.; Bangalore, S.; Wang, Z.; Johnson, K.W.; Pinotti, R.; Zhang, H.; Kaplin, S.; Narasimhan, B.; Kitai, T.; et al. Machine learning prediction in cardiovascular diseases: A meta-analysis. Sci. Rep. 2020, 10, 16057. [CrossRef] [PubMed]

[9] Ahsan, M.M.; Siddique, Z. Machine learning-based heart disease diagnosis: A systematic literature review. Artif. Intell. Med. 2022, 128, 102289.

[10] O. Arslan and M. Karhan, "Efect of Hilbert-Huang transform on classifcation of PCG signals using machine learning," Journal of King Saud University-Computer and Information Sciences, vol. 34, 2022.

[11] Balaha, H.M., Shaban, A.O., El-Gendy, E.M., & Saafan, M.M. (2022). A multi-variate heart disease optimization and recognition framework. Neural Computing and Applications, 34, 15907 - 15944.

[12] Oresko, J.J.; Jin, Z.; Cheng, J.; Huang, S.; Sun, Y.; Duschl, H.; Cheng, A.C. A wearable smartphone-based platform for real-time cardiovascular disease detection via electrocardiogram processing. IEEE Trans. Inf. Technol. Biomed. 2010, 14, 734–740.

[13] Sharean, T.M.A.M.; Johncy, G. Deep learning models on Heart Disease Estimation—A review. J. Artif. Intell. 2022, 4, 122–130.

[14] Hasan, A.M.; Shin, J.; Das, U.; Srizon, A.Y. Identifying prognostic features for predicting heart failure by using machine learning algorithm. In Proceedings of the ICBET'21: 2021 11th International Conference on Biomedical Engineering and Technology, Tokyo, Japan, 17–20 March 2021; pp. 40–46

[15] Li, J.P.; Haq, A.U.; Din, S.U.; Khan, J.; Khan, A.; Saboor, A. Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. IEEE Access 2020, 8, 107562–107582.

[16] Sudha, V.K.; Kumar, D. Hybrid CNN and LSTM network For heart disease prediction. SN Comput. Sci. 2023, 4, 172.

[17]Hossain, S., Hasan, M. K., Faruk, M. O., Aktar, N., Hossain, R., & Hossain, K. (2024). Machine learning approach for predicting cardiovascular disease in Bangladesh: evidence from a cross-sectional study in 2023. *BMC Cardiovascular Disorders, 24*(1), 214.

[18]Mehrabani-Zeinabad, K., Feizi, A., Sadeghi, M., Roohafza, H., Talaei, M., & Sarrafzadegan, N. (2023). Cardiovascular disease incidence prediction by machine learning and statistical techniques: a 16-year cohort study from eastern Mediterranean region. *BMC Medical Informatics and Decision Making, 23*(1), 72.

[19]Mir, A., Rehman, A. U., Ali, T. M., Javaid, S., Ahmad, S., Khan, M. U., … & Jabbar, S. (2024). A novel approach for the effective prediction of cardiovascular disease using applied artificial intelligence techniques. *ESC Heart Failure, 11*(6), 3742–3756.

[20]Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access, 7*, 81542–81554.

[21]Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics, 36*, 82–93.

[22]Motarwar, P., Duraphe, A., Suganya, G., & Mariappan, P. (2020). Cognitive approach for heart disease prediction using machine learning. In *2020 Int. Conf. on Emerging Trends in Information Technology and Engineering (ic-ETITE)* (pp. 1–6). IEEE.

[23]Budholiya, K., Shrivastava, S. K., & Sharma, V. (2022). An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University – Computer and Information Sciences, 34*(7), 4514–4523.

[24] Deepika, K.; Seema, S. Predictive analytics to prevent and control chronic diseases. In Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, India, 21–23 July 2016; IEEE: New York, NY, USA, 2016; pp. 381–386.

[25] Saikumar, K.; Rajesh, V. A novel implementation heart diagnosis system based on random forest machine learning technique. Int. J. Pharm. Res. 2020, 12, 3904.

[26] Rigatti, S.J. Random Forest. J. Insur. Med. 2017, 47, 31–39.

[27] Biau, G.; Fr, G.B. Analysis of a Random Forests Model. J. Mach. Learn. Res. 2012, 13, 1063–1095.

[28] Suthaharan, S. Decision Tree Learning. In Machine Learning Models and Algorithms for Big Data Classification; Springer:Berlin/Heidelberg, Germany, 2016; pp. 237–269. [CrossRef]

[29] Kotsiantis, S.B. Decision trees: A recent overview. Artif. Intell. Rev. 2013, 39, 261–283. [CrossRef]

[30] Chethana, C. Prediction of heart disease using different KNN classifier. In Proceedings of the Proceedings-5th International Conference onIntelligent Computing and Control Systems, ICICCS2021, Madurai, India, 6–8 May2021; pp. 1186–1194.

[31] Ahmed,R.; Bibi, M.; Syed, S. Improving Heart Disease Prediction Accuracy Using a Hybrid Machine Learning Approach: A Comparative study of SVM and KNN Algorithms. Int. J. Comput. Inf. Manuf. (IJCIM) 2023, 3, 49–54.