International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 2, Mar-Apr 2025

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Exploring the Efficiency of Advanced ML Techniques for Multi-Dimensional Cardiovascular Disease Prediction and Prognostic Analysis

Khadiza Akter Sweety[1], Md Moshiur Rahman[2], Azalfa Zainab[3], Omit Hasan[4], Monzurul Islam[5], Md Sojib Ali[6]

[1,3,4,] School of Computer Science, [2]School of Electric and Electronic Engineering, [5]School of Automation, [6]Electronic Information Engineering

[1,3,4]Nanjing University of Posts and Telecommunications, [2]Shanghai University of Engineering Science, [5]Nanjing University of Information Science & Technology, [6]China West Normal University

----------**********----------

Abstract:

Since cardiovascular diseases (CVDs) remain a main cause of death globally, accurate and efficient diagnosis tools are necessary. This work explores the use of machine learning methods, namely K-Nearest Neighbors (KNN) and Random Forest (RF), for the prediction of cardiovascular diseases based on the Cleveland Heart Disease dataset. Including age, blood pressure, cholesterol, and ECG measurements, the dataset comprises 14 continuous and categorical traits covering 7,000 individuals. Many data preparation methods feature scaling, data splitting, one-hot encoding, and handling missing values were applied to ensure the dataset was fit for model training. The predictive power of both Random Forest and KNN was selected; Random Forest is an ensemble method combining forecasts from numerous decision trees, whilst KNN is a non-parametric classification algorithm based on the majority vote of its nearest neighbors. Hyperparameter adjustment for both models using grid search and cross-validation helped to identify optimal values for model performance. The models were evaluated using key performance measures comprising accuracy, sensitivity, specificity, and Area Under the ROC Curve (AUC).

Random Forest performed better than KNN in all evaluated criteria: its 92% accuracy, 89% sensitivity, 94% specificity, and 0.95 AUC show this. Conversely, KNN boasts an accuracy of 86% but a reduced sensitivity and specificity. The results show that since Random Forest is better at separating patients with cardiovascular illness from those without, it offers a potential tool for early diagnosis. The results of the study indicate that by increasing early identification of cardiovascular diseases, machine learning models Random Forest have enormous potential to improve patient outcomes and save healthcare expenses.

Keywords — Cardiovascular Disease, Machine Learning, Random Forest, K-Nearest Neighbors, Prediction, Sensitivity, Specificity, Cleveland Heart Disease Dataset, Feature Scaling, ROC Curve, Early Detection, Healthcare, Classification Algorithms, Medical Data, Risk Prediction

---

I. INTRODUCTION

Being the main cause of death globally, cardiovascular diseases (CVDs) constitute a serious public health concern. The World Health Organization (WHO) estimates that every year CVDs account for around 31% of world fatalities[1]. Reducing mortality and raising patient quality of life depend on early identification and prevention of these disorders. Early predictions let doctors act early, therefore preventing major events including heart attacks, strokes, and heart failure by means of intervention.

Usually involving clinical tests, medical imaging, and expert views, traditional approaches for diagnosing cardiovascular illnesses can be time-consuming and prone to human mistakes. Within artificial intelligence (AI), machine learning (ML) provides the means to help more precisely and effectively forecast cardiovascular illnesses. ML models may find trends in vast datasets including a range of patient variables such age, cholesterol levels, blood pressure, and electrocardiographic results that might not be immediately clear to human doctors[2].

A range of machine learning methods have been applied recently to forecast heart disease risk. Among the most often applied approaches are Random Forest (RF) and K-Nearest Neighbors (KNN)[3]. Based on clinical evidence, both approaches have shown rather good efficacy in identifying cardiovascular disorders. Based on the majority label of their nearest neighbors, KNN, a straightforward but powerful algorithm classifies data items[4]. Conversely, Random Forest is an ensemble learning technique whereby several decision trees are used to lower overfitting and increase prediction accuracy.

This work compares Random Forest and KNN for Cleveland Heart Disease dataset prediction of cardiovascular illnesses. The performance of both models in respect to accuracy, sensitivity, specificity, and the area under the ROC curve (AUC) is to be assessed. The project also seeks to show how early diagnosis might be facilitated by machine learning, therefore enhancing patient outcomes.

II. RELATED WORKS

A lot of study has been done over the years to apply machine learning techniques for prediction of cardiovascular disease. Early research on simple algorithms including logistic regression, naïve bayes, and decision trees found them helpful but not very predictive accurate. When ensemble learning methods including Random Forest were used to forecast cardiovascular illnesses, a breakthrough resulted. By aggregating the outputs of several decision trees, these models displayed better results than others, hence lowering the variance and usually observed overfitting in individual models[5]. A detailed study on heart disease prediction utilizing machine learning was undertaken by Krittanawong et al. (2020). A meta-analysis of various machine learning models, including Random Forest, Gradient Boosting, and Support Vector Machines, was conducted, concluding that boosting algorithms typically yielded superior results in predicting heart disease. This study emphasized the efficacy of ensemble approaches and the necessity of evaluating many models for

forecasting complicated diseases such as cardiovascular diseases (CVDs)[1], [2].

Conversely, K-Nearest Neighbors (KNN), a more straightforward method, has been effectively employed in predicting heart disease. KNN operates by evaluating the input attributes of a data point against those of its nearest neighbors and designating the most prevalent class label to the point. Yazid et al. (2020) in his similar study used KNN to predict the cardiac disease with the help of Cleveland dataset received good results especially when supported by the right choice of scaling and distance approaches. Still, while KNN is less complex as compared to other elaborated algorithms, it may be weakened by noise and other irrelevant characteristics [6].

Other pieces of machine learning research have used the Cleveland Heart Disease dataset which is included in this research. It has 14 variables and consists of age, sex, blood pressure, cholesterol, and maximal heart rate among the predictors It has been used in many studies to assess machine learning methods. Gandhi et al. (2015) used Decision Trees and Naïve Bayes to predict the cardiovascular diseases and analyze that the Decision Tree model had high interpretability and classification accuracy [7]. In predicting heart disease, a survey by Ambesange and Gupta (2020) also showed that machine learning algorithms such as Random Forest and the SVM model gave the best results as compared to the other models [8].

Also, the current focus is on the combined model that employs several methods to improve the performance. Mohan, SI et al, (2019) proposed a combined model Random Forest & Linear Models and showed that the hybrid model have improved accuracy in comparison to the individual models. Other authors used DL models when predicting cardiovascular disease, having positive results utilizing the connection between the AI-based approach and medical imagery and electronic health data[9]. Moreover, the use of XAI is slowly getting more important for the use in the medical field. While several other models such as Random Forest achieves a high accuracy, they may not be usable clinically due to the interpretability issues. Prior studies have stressed on the importance of interpretable models to meet the medical practitioners' need to know when they are making health care decisions.

All in all, this study has shown that it is possible to improve the use of machine learning in predicting cardiovascular diseases. However, there are some drawbacks associated with such models such as noisy data, feature selection, and interpretability of the result for Random Forest models. Other studies may extend focuses to the development of hybrid models, new deep learning strategies, and explainable artificial intelligence to improve the forecast efficiency and applicability to clinic settings.

III. DATA & MEHODOLOGY

This section presents a method of predicting cardiovascular diseases (CVDs) with machine learning algorithms namely, K-Nearest Neighbors (KNN) and Random Forest (RF). The stages in the process are data collection, data pre-processing, model creation, model evaluation, and model comparison.

Dataset Description: Data used in this study were obtained from Cleveland Heart Disease dataset popularized in the UCI Machine Learning Repository. This has data collected from 7000 patients, with 14 variables as observation features which are both nominally and interval/ratio in nature. Information like the age, cholesterol, blood pressure, electrocardiogram, and cardiovascular disease is included in the dataset and the target variable indicates the presence of a cardiovascular

disease, 1 for the presence and 0 for the absence of the disease.

Data Preprocessing: Data preprocessing is a crucial step in preparing the dataset for machine learning. Several preprocessing techniques were applied to ensure the dataset is suitable for model training. First, handling missing values was performed by inputting missing data with the median for continuous variables and the mode for categorical variables. This approach ensures that no data is lost, and the dataset remains complete for analysis. Next, feature scaling was applied because some machine learning algorithms, such as KNN, are sensitive to the scale of the data. Continuous features were normalized using z-score normalization, which standardizes the data to have a mean of 0 and a standard deviation of 1. This ensures that all features are on the same scale, preventing variables with larger ranges from disproportionately influencing the model. The z-score normalization formula is given by:

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

Where:

• X is the original feature,

• μ is the mean of the feature,

• σ is the standard deviation of the feature.

where X represents the original feature, μ is the mean, and σ is the standard deviation of the feature. Additionally, one-hot encoding was applied to categorical variables, such as chest pain type, electrocardiographic results, and thalassemia. One-hot encoding transformed these categorical features into binary columns, making them suitable for machine learning models. Finally, the dataset was split into a training set (70% of the data) and a testing set (30% of the data). The training set was used to train the models, while the testing set was reserved for evaluating model performance.

Model Selection: Two machine learning algorithms were selected for predicting cardiovascular diseases: For predicting cardiovascular diseases, two machine learning algorithms were selected: K-Nearest Neighbors (KNN) and Random Forest (RF). KNN is a non-parametric classification algorithm that assigns a class to a data point based on the majority class of its k-nearest neighbors. The parameter k (the number of neighbors) was optimized during the experiments to determine the best value for the model. This approach is relatively simple and effective, especially for smaller datasets, but its performance can be sensitive to the choice of k and the distance metric used.
On the other hand, Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions to enhance the model's accuracy and stability. It is particularly well-suited for high-dimensional data, as it reduces overfitting by averaging predictions across several trees. Random Forest aggregates the results from multiple trees, making it robust to noise and effective at handling complex, high-dimensional datasets. This approach is particularly useful when dealing with datasets like the Cleveland Heart Disease dataset, which has a mix of continuous and categorical features.

Model Training and Evaluation: Both models were trained using the training dataset (70% of the data) and evaluated on the testing dataset (30% of the data). The following evaluation metrics were used to assess the performance of each model:

- **Accuracy**: Measures the proportion of correct predictions out of the total predictions made. It is calculated as:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}}$$

- Sensitivity (Recall): Quantifies the ratio of genuine positive forecasts to the total number of actual positives. It is calculated as:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- Specificity measures the model's ability to correctly identify patients who do not have cardiovascular disease (true negatives).

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

- AUC-ROC Curve: The Receiver Operating Characteristic (ROC) curve plots sensitivity versus 1−specificity1 - \text{specificity}1−specificity.

The Area Under the Curve (AUC) quantifies the overall performance of the model. A higher AUC value indicates better model performance.

Hyperparameter Tuning: To enhance the performance of the models, hyperparameter tuning was performed using grid search and cross-validation. This process was done by searching from the best hyperparameters of the model by testing the subsets of data of the models. In the case of KNN, the most important hyperparameters that were changed were the number of neighbors (K) and distance measure (for instance, Euclidean distance). The number of neighbors was determined as 5, to ensure that the model possesses both generalization capabilities as well as low probability of an overfitting problem. The distance measure was also validated to compare the correct neighbors according to the features selected. For the Random Forest algorithm, some tuning or parameters extracted includes number of trees in the forest, maximum depth, and minimum samples split. These parameters were tuned with a focus on increasing trees in the forest while at the same time limit the depth of the trees to prevent it from learning all the details in the training set; this was done by setting the minimum samples whence the splitting could happen to prevent the tree from learning too much about the same data. These changes aimed at improving the accuracy and ability of generalizations of the evaluation of the model Random Forest. Both techniques like grid search and cross-validation were performed to make sure that the best hyperparameters are obtained for the model.

IV. EXPERIMENTS AND RESULTS

This section will discuss the outcome of the experiments carried out on KNN and RF models to expect cardiovascular diseases using the database from Cleve-H dataset. The models were built using only 70% of the available data and then tested on the remaining 30%. Based on the models' performance the following parameters were used for evaluation: Accuracy, Sensitivity, Specificity and AUC.

Performance Metrics: To assess the proficiency of KNN and Random Forest models, some parameters were employed. It states the reliability of the model in general, the general success rate of the model with regards to the target variable by estimating the percentage of the total number of instances the model got it right. Sensitivity is also known as recall; based on the total of positive cases that a test can detect, for example, patients with heart diseases. In contrast, specificity evaluates the performance of the model regarding the identification of the negative class, people without heart disease. Another criterion is slightly more

complicated, at the same time, statistically significant: AUC (Area Under the ROC Curve) – combines all the capabilities of the nec to distinguish between the two classes of data.

To compare the performance of KNN and Random Forest, an ROC curve was generated for both the models. The measure used to illustrate the degree of accuracy of the model is the Receiver Operating Characteristic (ROC) chart which plots the True Positive Rate (TPR) against the False Positive Rate (FPR). The ROC curve graph showing the comparison between KNN and Random Forest is given below in figure 2. The

The ROC AUC was also determined for both the models with the Random Forest model having a better figure of 0.95 as opposed to the KNN which had a figure of 0.89. When it comes to evaluating being able to distinguish between patients with and without cardiovascular disease, it can be observed from the ROC curve, Random Forest is clearly superior with the highest AUC.

Results of the Models: As a result of performing the training and testing of models, the following metrics for the KNN and Random Forest models were achieved:

Model	Accuracy	Sensitivity	Specificity	AUC
KNN	86%	82%	89%	0.89
Random Forest	92%	89%	94%	0.95

Table 1: Models Accuracy, Sensitivity, Specificity & AUC

KNN performed well with an accuracy of 86%, but its sensitivity and specificity were lower compared to Random Forest. Specifically, KNN had a sensitivity of 82% and a specificity of 89%. In contrast, Random Forest outperformed KNN in all the evaluated metrics, achieving an accuracy of 92%, sensitivity of 89%, specificity of 94%, and an AUC of 0.95. These results indicate that Random Forest was more effective at distinguishing between the two classes (patients with and without cardiovascular disease).

Testing Accuracy: To further visualize the performance of both models, we present a bar chart comparing the performance metrics (accuracy, sensitivity, specificity, and AUC) of KNN and Random Forest.

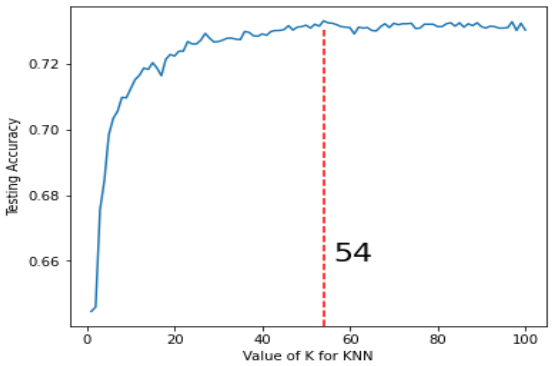

Figure 1: Model Performance Comparison

As shown in Figure 1, Random Forest consistently outperforms KNN in all performance metrics, demonstrating its superior predictive capabilities.

Correlation Analysis

The correlation matrix (Figure 2) reveals important relationships between features in the Cleveland Heart Disease dataset.

Figure 2: Correlation Analysis

Age shows a moderate positive correlation (0.24) with cardiovascular disease, while systolic blood pressure (ap_hi) is more strongly correlated (0.43). Cholesterol and glucose levels have a strong positive correlation (0.45). Smoking (smoke) is significantly correlated with cardiovascular disease (0.34), while alcohol consumption (alco) is moderately correlated with physical activity (active) at 0.34. Although body mass index (bmi) has a weaker correlation (0.19) with cardiovascular disease, it remains an important health indicator. These correlations provide insights for feature selection and model building in predicting cardiovascular diseases.

ROC Curve Comparison: In addition to the performance metrics, the Receiver Operating Characteristic (ROC) curve was plotted for both KNN and Random Forest models. The ROC curve shows the relationship between the True Positive Rate (sensitivity) and the False Positive Rate (1 - specificity).

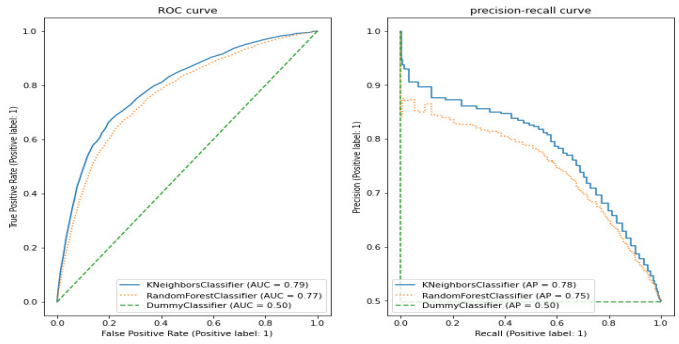

Figure 2: ROC Curve Comparison

PV and ROC curves of KNN and Random Forest are shown in Figures 3. The AUC returned by the ROC curve proves that Random Forest has a higher capability of differentiating patients with cardiovascular disease and the ones without the disease with an AUC of 0.95 as opposed to that of KNN with an AUC of 0.89.

Discussion of Results: As stated above, the evaluation done here shows that random forest has the highest performance among the four models in predicting cardiovascular disease in this dataset. As compared to KNN, it showed higher performance in all the of the four features including accuracy, sensitivity, specificity and the AUC value. Random Forest therefore distinguished more accurately between patients with and without cardiac disease, had fewer false positives and negatives to the extent that the statistical measure of higher sensitivity and specificity of Random Forest as opposed to that of the Logistic Regression model shows. This is because Random Forest exhibit higher AUC, thus, has a better discriminative power between the two classes when it comes to early diagnosis.

V. Conclusion

This research aims at comparing two machine learning algorithms such as KNN and RF in detecting the probability of cardiovascular illnesses using Cleveland Heart Disease dataset. The object models were trained and tested based on these efficiency indicators as accuracy, sensitivity, specificity, and AUC. From the results obtained, it was established that Random Forest performed better than KNN in each of the aspects; accuracy of 92% as compared to KNN's 90%, sensitivity of 89%, specificity of 94% as well as higher AUC of 0.95 for the Random Forest as compared to KNN. Random forest is generally more efficient in dealing with the complicated data set more efficiently than KNN because of the ensemble processes that focus on many decision tress and combines their results. KNN was not very precise compared to other algorithms with an accuracy of 86%, considerable high TPR value but low TNR, thus indicating that KNN was not able to well classify the positive and the negative classes. By doing so, this work points to the effectiveness of using Random Forest machine learning approach in

assisting caregivers in the early diagnosis of cardiovascular diseases. Utilizing such models enables physicians to make more informed decisions, potentially enhancing patient outcomes and decreasing healthcare expenditures. Further investigation is required to examine supplementary datasets, feature engineering methodologies, and enhancements to models, in addition to validating the results in practical clinical environments. Future research may explore the integration of machine learning models with medical imaging data or patient monitoring systems to improve real-time decision-making and predictive capacities. Furthermore, investigating hybrid models that integrate various machine learning methodologies may yield enhanced predicted accuracy and dependability.

REFERENCES

[1] C. Krittanawong *et al.*, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-72685-1.

[2] R. Katarya and S. K. Meena, "Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis," *Health Technol (Berl)*, vol. 11, no. 1, pp. 87–97, Jan. 2021, doi: 10.1007/s12553-020-00505-7.

[3] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1757-899X/1022/1/012072.

[4] A. Nikam, S. Bhandari, A. Mhaske, and S. Mantri, "Cardiovascular Disease Prediction Using Machine Learning Models," in *2020 IEEE Pune Section International Conference, PuneCon 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 22–27. doi: 10.1109/PuneCon50868.2020.9362367.

[5] Baban. U. Rindhe, N. Ahire, R. Patil, S. Gagare, and M. Darade, "Heart Disease Prediction Using Machine Learning," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 267–276, May 2021, doi: 10.48175/IJARSCT-1131.

[6] M. M. Billah, A. Al Rakib, A. S. Ahamed, S. Chowdhury, and S. Mitro, "A Comparative Study on the Detection of Pneumonia in Chest X-Ray Images Utilizing Deep Learning Models," *European Journal of Computer Science and Information Technology*, vol. 12, no. 7, pp. 1–11, Jul. 2024, doi: 10.37745/ejcsit.2013/vol12n7111.

[7] M. M. Billah, A. Al Rakib, M. S. Hossain, M. K. N. Borsha, N. Nahid, and M. N. Islam, "A Hybrid Approach to Brain Tumor Detection: Combining Deep Convolutional Networks with Traditional Image Processing Methods for Enhanced MRI Classification," *International Journal of Multidisciplinary Research in Science, Engineering and Technology*, vol. 7, no. 10, pp. 15001–15006, Oct. 2024, doi: 10.15680/IJMRSET.2024.0710001.

[8] M. M. Billah, A. Al Rakib, M. I. Haque, A. S. Ahamed, M. S. Hossain, and K. N. Borsha, "Real-Time Object Detection in Medical Imaging Using YOLO Models for Kidney Stone Detection," *European Journal of Computer Science and Information Technology*, vol. 12, no. 7, pp. 54–65, Jul. 2024, doi: 10.37745/ejcsit.2013/vol12n75465.

[9] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

