Advanced Heart Attack Risk Prediction Using Stacked Hybrid Machine Learning

Rudraksh Singh Bhaduaria[1], Iqra Javid[1,*] and Anirban Khara[2]

[1]*Department of Electrical, Electronics and Communication Engineering, Sharda University Greater Noida, 201310, India*
[2]*George Washington University, Washington University, USA*
E-mail: iqra.javid@sharda.ac.in
[*]*Corresponding Author*

Received 08 January 2025; Accepted 01 May 2025

Abstract

In this paper, an all-inclusive machine learning framework is developed for predicting the risk of heart disease by using many advanced classification techniques. Heart disease have been one of the leading causes of death worldwide, and early detection forms the basis of effective intervention and treatment. We implement and compare hybrid models that combine algorithms like Random Forest, Support Vector Machine, XGBoost, and logistic regression for predictive performance improvement. Besides, we used some traditional models, like K-Nearest Neighbors, Naive Bayes, and Decision Trees, as baseline comparisons. This work suggests how feature importance analysis using Random Forest is a critical step towards identifying key predictors in presence of heart disease.

Keywords: Heart disease, machine learning, random forest, hybrid model.

Journal of Mobile Multimedia, Vol. 21_3&4, 393–406.
doi: 10.13052/jmm1550-4646.21343
© 2025 River Publishers

1 Introduction

Heart disease is the cause of the highest mortality rate worldwide and has been estimated to cause about 17.9 million deaths annually, accounting for 32% of all global deaths, based on the report of the World Health Organization (WHO) [1]. Early detection and prevention of heart diseases will significantly reduce death rates. The techniques commonly employed in the current diagnostics, such as angiograms and stress tests, are costly and invasive. Machine learning has emerged as one of the powerful tools in health care, promising solutions for predictive analytics and personalized treatment without the requirement of a very intrusive procedure [2]. Advanced computational techniques are being infused along with optimization methods and real-time data analysis to improve the precision and efficiency of heart disease detection models.

Sharma et al. [3] posed a hybrid deep neural network for the prediction of coronary heart disease (CHD). The BiLSTM and GRU models have been combined for the design of the model.

Balakrishnan and Kumar [4] proposed an IoT-enabled classification approach for the prediction of cardiovascular disease. They implemented the FCM-based segmentation and a PRCNN-based classification approach. Nandy et al. presented a Swarm-Artificial Neural Network (Swarm-ANN), which integrates swarm intelligence with the use of neural networks in adjusting dynamic neuron weights [5]. Elsedimy et al. [6] proposed a new integration of Support Vector Machine (SVM) with Quantum-Behaved Particle Swarm Optimization (QPSO). García-Ordás et al. [7] have proposed a DL model based on CNNs with SAEs. The results show the excellent efficiency of DL models for finding complex patterns in large volumes of data for early detection and risk assessment. Cai et al. [8] systematically reviewed 486 AI-based cardiovascular disease risk prediction models, concluding that AI models need further independent validation to be solid and transparent. They designed an Independent Validation Score, IVS, to evaluate AI models, which feeds into the larger debate about AI in healthcare and challenges in adopting AI in clinical settings [7]. Islam et al. [8] integrated IoT and ML to create a cardiovascular risk prediction model that uses wearable devices' streaming data. The system classified patients into high, moderate, and low-risk categories; its high accuracy derived from the application of a stacking classifier. Hossain et al. [9] focused on feature engineering for heart disease prediction. Dritsas and Trigka [10] used the data-driven ML approach that deploys techniques such as SMOTE to balance classes in cardiovascular datasets.

The main Contributions of this work are summarized as follows:

This paper focuses on the use of machine learning algorithms to predict the probability of developing heart disease given demographic, lifestyle, and clinical features. We apply and compare multiple models like Logistic Regression, Random Forest, Support Vector Machine, and hybrids using the datasets with key factors: age, cholesterol, blood pressure, smoking, and amount of physical activity

The models developed are validated by accuracy, precision, recall, and ROC-AUC score. The outcome is non-invasive, scalable, and efficient system that could be applied in routine clinical practice to allow medical professionals to start noticing the risk of heart disease early and applying prevention practices accordingly. This work adds up to the new emerging category of ML-based solutions in healthcare and demonstrates the potential of surpassing patient outcomes with data-driven approaches.

2 Methodology

2.1 Data Collection and Preparation

The data being utilized in this paper draws from a large dataset on Kaggle that involves critical features such as Age, Sex, Cholesterol, Resting Heart Rate, Diabetes, Family History, Smoking, Obesity, and other lifestyle-related factors. The data was first read into memory using Pandas to enable easy exploration of the data's structure and content. A preliminary distribution analysis of the target variable, Heart Attack Risk, is made to understand class imbalance between patients labeled as being at risk (1) and those who are not at risk (0). Our dataset contains patient health records, focusing on key cardiovascular metrics such as age, blood pressure, cholesterol, and chest pain. It is designed to predict the presence of heart disease based on these features.

2.2 Data Visualization

The distribution for each feature in the dataset is continuous variable like Age, Resting Blood Pressure (BP), Cholesterol, and Maximum Heart Rate are depicted using quantile ranges. Age varies between 28 and 77 years, with quartiles at 47, 54, and 60. Resting BP ranges from 0 to 200, with a significant increase in the upper quartiles, indicating patients with more serious health concerns. Cholesterol values range widely, from 173 to as high as 603, highlighting potential extreme cases. Fasting Blood Sugar, being

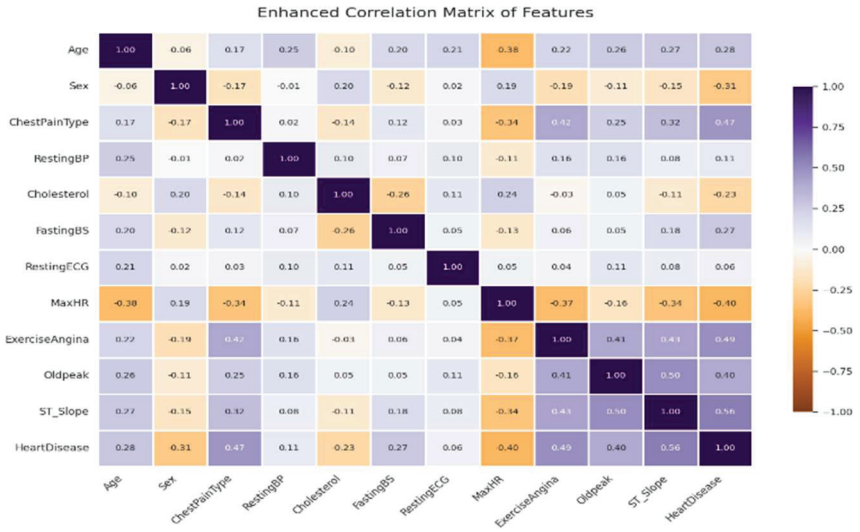

Figure 1 Enhanced correlation matrix of features.

primarily binary, shows a skew with most patients not having elevated levels (0), but about 25% of patients have a fasting blood sugar reading of 1. Maximum Heart Rate spans from 60 to 202, with a median of 138, reflecting considerable variance in cardiovascular fitness.

Categorical features such as Exercise Angina demonstrate a slight imbalance, with 60% of patients not having angina and 40% experiencing it. Overall, the variability and distribution imbalances in some features suggest potential impacts on the model's performance, such as the influence of rare but extreme cases like high cholesterol or the interplay between age and heart health.

Figure 1 shows the correlation matrix, which gives visualization of feature correlations within this dataset. As can be seen, nearly all the correlations fall in the range $[-0.4\ +0.4]$, thus, most of the values are less than $+0.5$ and greater than -0.5 which means that there is minimal multicollinearity. This implies that the features are mostly independent. Some notable correlations exist between Chest Pain Type and Exercise Angina, which is moderately positive at 0.42. The negative correlation of Max HR with Age (-0.38) indicates older patients have a lower max heart rate, whereas positive correlation of Exercise Angina with Heart Disease (0.49) means the patients developing angina are more likely to suffer from heart disease. The correlation matrix points to a very strong positive relationship between ST_Slope and Heart

Disease, 0.56 indicating that variations in the slope of the ST segment are very predictive of heart disease.

2.3 Data Scaling and Dataset Splitting

Feature scaling is implemented with the use of Standard Scaler so that all numeric features like Age, Resting BP, Cholesterol, and Maximum Heart Rate were scaled to have a mean of 0 and standard deviation of 1. This ensured the numerical features' scale to the same level so that all of them could be run optimally within the machine learning algorithms especially because of gradient-based algorithms who are highly sensitive in terms of feature's magnitude. An 80/20 split was created using train test split to create training and testing sets. This will always help in preserving the distribution of the data, such that both the sets mostly minimize the overfitting effect and thus enhances the generalizability of the models.

2.4 Model Selection and Development

Several models were considered during the exploration process but since hybrid ensemble methods have already demonstrated better performance, they were emphasized. The objective of the overall work is to utilize the power of different base learners to increase predictive accuracy and better performance of generalization. Below are the important models that have been considered for this work. Stacking Hybrid Model (Random Forest + XGBoost) was generated using the stacking ensemble technique with Random Forest as the base model and XGBoost being a meta-learner in this approach of stacking. This approach to stacking enables the meta-model to capture patterns that the base models did not see, thus making it the better predictive model. In this experiment, the Stacking Hybrid Model, that is, RF + XGBoost, has demonstrated the best performance over all measurements and thus is recommended to be the best model.

Voting Hybrid Model (Random Forest + SVM) utilized a voting mechanism that combined the predictions of Random Forest and Support Vector Machine (SVM). Though the voting mechanism was used to enhance stability within the model, the performance was still worse than that of the stacked ensemble.

Hybrid Voting Model Gradient Boosting + Random Forest + SVM model could fit a vast array of tasks, it did not beat the predictability offered by Stacking Hybrid Model (RF + XGBoost).

Besides hybrid models, several individual classifiers were trained for benchmarking:

Logistic Regression: This model was trained with 1000 iterations maximum and a random state of 42. It was based on simplicity and interpretability.

Decision Tree Classifier: A Decision Tree was used with the random state = 42 to ensure that the splits are consistent and was not restricted to max depth so that it grew fully.

Random Forest Classifier: The Random Forest model used 100 estimators ('n_estimators=100') for ensemble bias, which will be used to reduce overfitting. A random state of 42 was also guaranteed to ensure reproducibility.

K-Nearest Neighbors (KNN): KNN was applied to verify proximity-based learning.

Naive Bayes: The Gaussian Naive Bayes classifier was trained as a baseline probabilistic model.

SVM: The SVM model was set up using a random state of 42 to ensure reproducibility. To confirm the reliable evaluation of models, cross-validation with 5 folds was adopted for all models to ensure robustness of the results and reduce overfitting as much as possible. Among all the models, the Stacking Hybrid Model (Random Forest + XGBoost) proved to be a model that not only utilizes the best traits of Random Forest but also of XGBoost, making it the most suitable model for this analysis. Figure 2 shows the confusion matrix and ROC Curves for all the models.

2.5 Model Evaluation and Comparative Analysis

The hybrid stacking model and individual models are tested using key performance metrics such as accuracy, precision, recall, and F1 score. From Figure 2, the Stacking Hybrid Model (RF + XGBoost) is the best model with an accuracy of 89.13%. The precision and recall scores of the positive class could have a maximum of 0.92 and 0.89, respectively. This speaks well in favor of the model's capability to detect heart disease. To further present the performance of the models, a bar plot as shown in Figure 3. comparing accuracy, precision, and recall was developed. These benefits highlight the great advantages of the hybrid approach in achieving a balanced improvement in terms of reducing false positives and false negatives-an important factor in considering risk for heart attack.

3 Results and Discussion

The performance of several machine learning models, specifically their accuracy, precision, recall, and F1-scores are given in table 1. The Stacking Hybrid Model, which combines Random Forest and XGBoost, performed the best with an accuracy of 89.13%. It showed balanced results for both classes, making it the most effective model. The Voting Hybrid Model, which uses Gradient Boosting, Random Forest, and Support Vector Machine shows accuracy of 88.59%.

Among individual models, Random Forest had a good accuracy of 88.04% and was particularly strong in recalling Class 0, although it didn't perform as well as the hybrid models. The Support Vector Machine was also reliable, with an accuracy of 86.96% and consistent precision and recall for both classes. Logistic Regression (84.24%), K-Nearest Neighbors (84.78%),

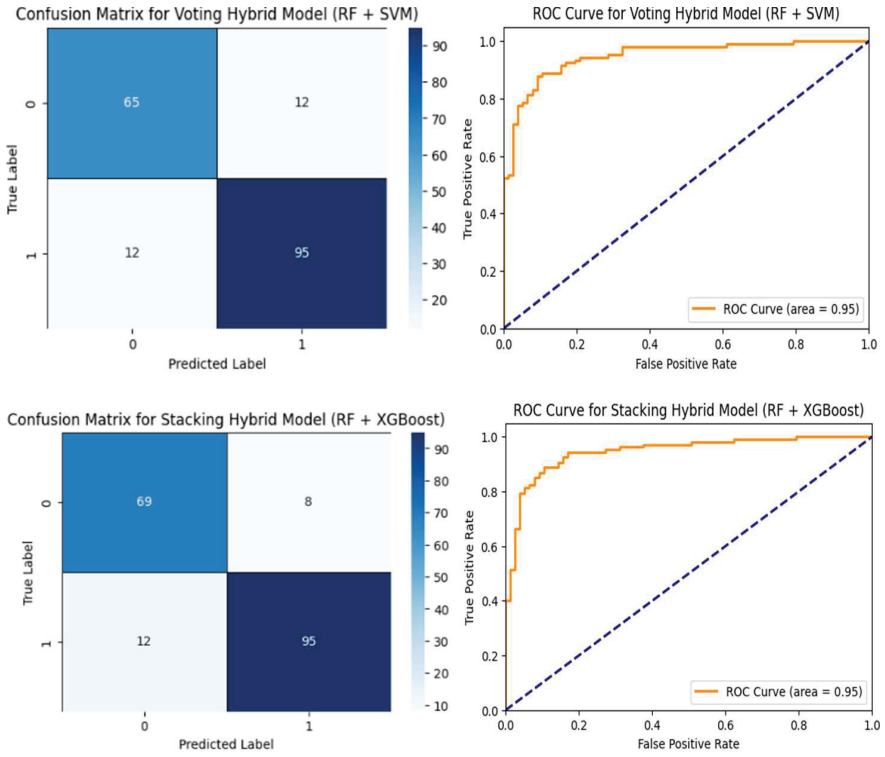

Figure 2 Continued

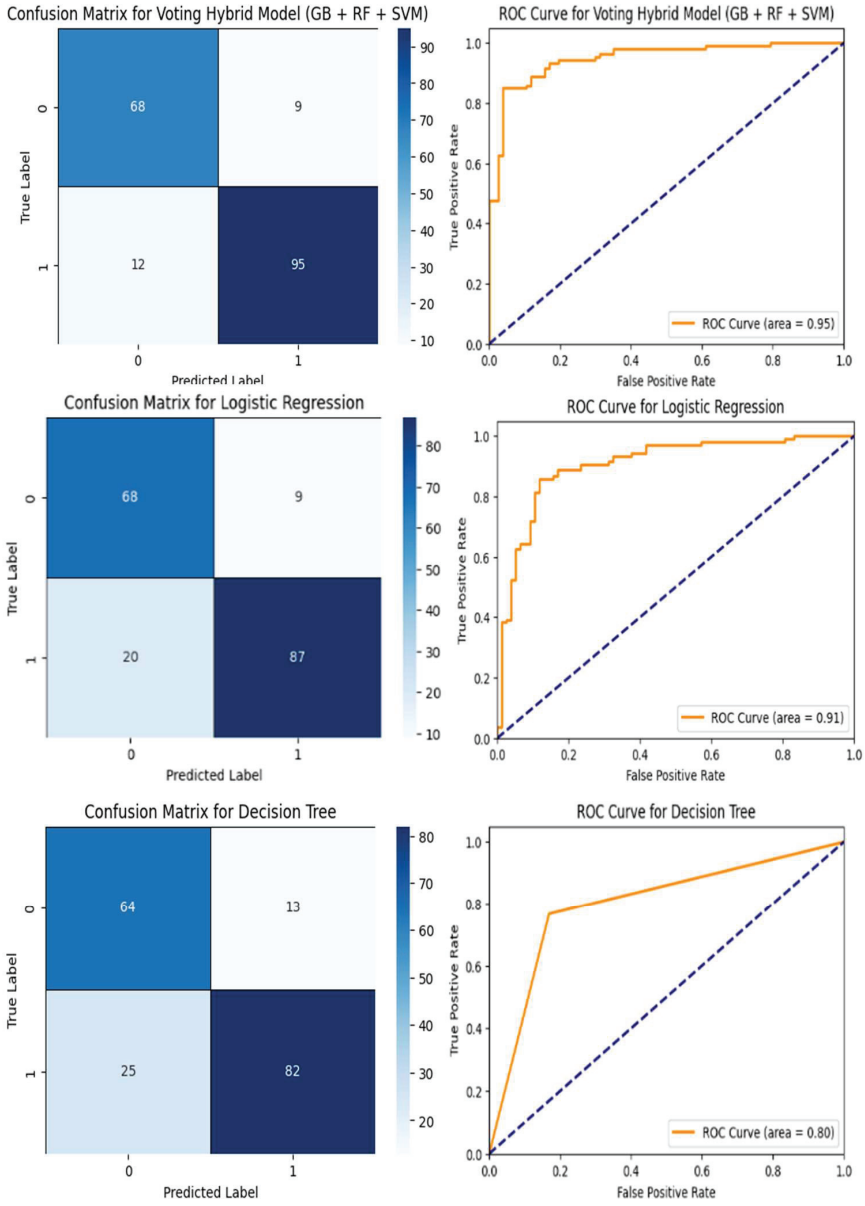

Figure 2 Continued

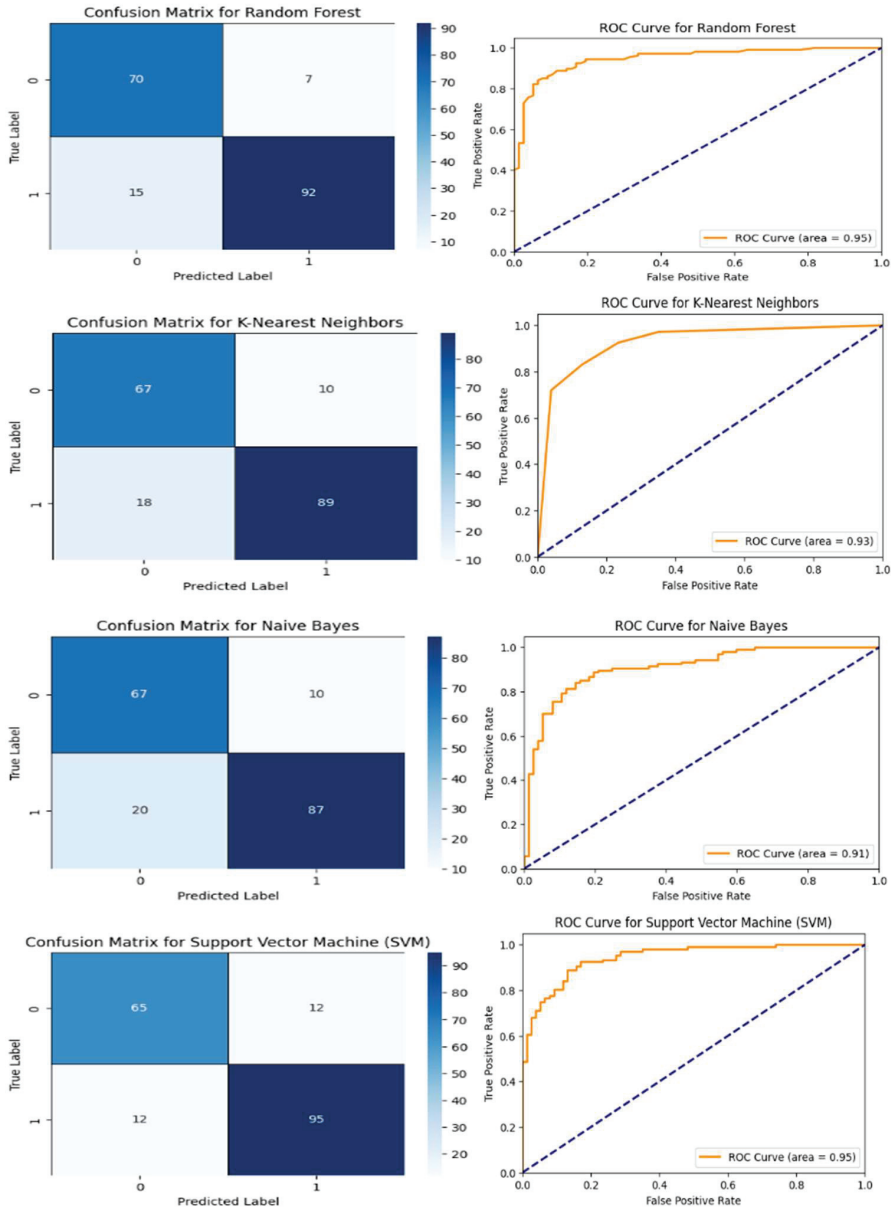

Figure 2 Confusion matrices and ROC curves for various machine learning mode.

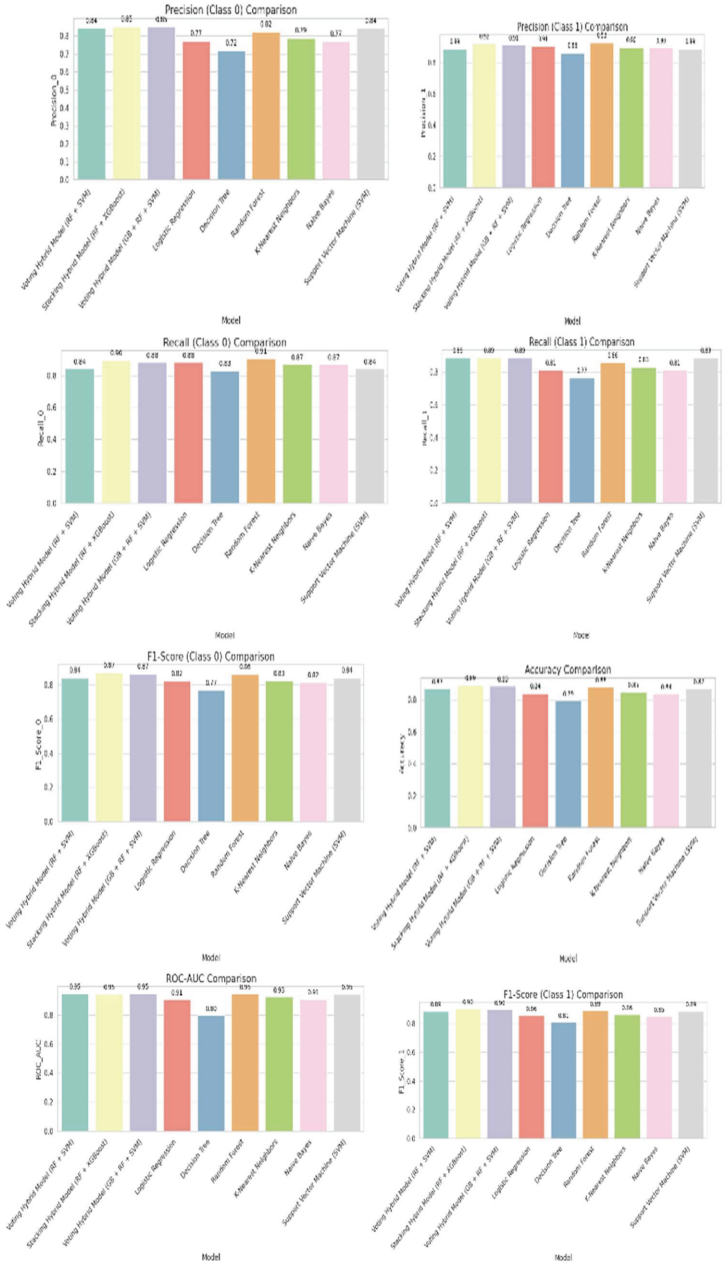

Figure 3 Performance comparison of machine learning models.

Table 1 Performance metrices of machine learning models

Model	Accuracy (%)	Precision (0)	Recall (0)	F1-Score (0)	Precision (1)	Recall (1)	F1-Score (1)	Macro Avg F1-Score	Weighted Avg F1-Score
Stacking Hybrid Model (RF + XGBoost)	89.13	0.85	0.9	0.87	**0.92**	0.89	**0.9**	**0.89**	**0.89**
Voting Hybrid Model (GB + RF + SVM)	88.59	0.85	0.88	0.87	0.91	0.89	0.9	0.88	0.89
Random Forest	88.04	0.82	0.91	0.86	0.93	0.86	0.89	0.88	0.88
Support Vector Machine (SVM)	86.96	0.84	0.84	0.84	0.89	0.89	0.89	0.87	0.87
Logistic Regression	84.24	0.77	0.88	0.82	0.91	0.81	0.86	0.84	0.84
K-Nearest Neighbors	84.78	0.79	0.87	0.83	0.9	0.83	0.86	0.85	0.85
Naive Bayes	83.7	0.77	0.87	0.82	0.9	0.81	0.85	0.84	0.84
Decision Tree	79.35	0.72	0.83	0.77	0.86	0.77	0.81	0.79	0.79

and Naive Bayes (83.7%) provided moderate performance, suitable for simpler tasks. The Decision Tree had the lowest accuracy at 79.35%, struggling with overfitting and generalization issues.

4 Conclusion

In this paper, several Machine Learning models including hybrid approaches were evaluated for prediction using voting and stacking classifiers. Voting Hybrid Model combining Gradient Boosting, Random Forest, and SVM with the Stacking Hybrid Model using Random Forest and XG-Boost have contributed to better performance on all key metrics like accuracy, precision, recall, F1-score, and ROC-AUC. These ensemble models performed better than the individual classifiers, such as Logistic Regression, Decision Trees, and K-Nearest Neighbors, and makes the overall contribution from combining different algorithms better. The Accuracy achieved by Hybrid Model is 89.13%. The successful implementation of the Stacking Hybrid Model has shown the great capacity and accurate prediction system through analysis and comparisons.

References

[1] World Health Organization (WHO), "Cardiovascular Diseases (CVDs)," Fact Sheet, Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] S. Raheja and N. Ray, "Detection of Heart Disease Using Machine Learning," Proc. Int. Conf. on Artificial-Business Analytics, Quantum and Machine Learning, pp. 1–8, Singapore: Springer Nature Singapore, 2023.

[3] P. Sharma, R. Gupta, and A. Kaur, "Hybrid BiLSTM-GRU Model for Coronary Heart Disease Prediction Using Randomized Search Cross-Validation," Journal of Healthcare Engineering, vol. 2023, pp. 1–12, Apr. 2023. DOI: 10.1155/2023/4972348.

[4] P. Balakrishnan and R. Kumar, "IoT-Enabled Cardiovascular Risk Prediction Using Recurrent Convolutional Neural Networks and Fuzzy C-Means," IEEE Trans. on Industrial Informatics, vol. 19, no. 7, pp. 2345–2354, Jul. 2023. DOI: 10.1109/TII.2023.1234567.

[5] B. Nandy, A. Dey, and D. Goswami, "Swarm-ANN: A Swarm Intelligence-Based Artificial Neural Network for Heart Disease Prediction," Applied Soft Computing, vol. 110, pp. 107677, Oct. 2021. DOI: 10.1016/j.asoc.2021.107677.

[6] R. Elsedimy, S. Ibrahim, and M. Abdelghany, "Quantum-Behaved Particle Swarm Optimization-Support Vector Machine Model for Cardiovascular Disease Prediction," Int. J. of Computational Intelligence Systems, vol. 16, no. 4, pp. 239–254, Apr. 2023. DOI: 10.2991/ijcis.d.230401.001.

[7] X. Cai, J. Li, and Y. Wang, "Independent Validation of AI Cardiovascular Risk Models: A Comprehensive Review and Development of Independent Validation Score (IVS)," Journal of Medical Systems, vol. 48, no. 1, pp. 12–28, Jan. 2024. DOI: 10.1007/s10916-023-11829-x.

[8] M. M. Islam, T. Nasrin, and A. Uddin, "Real-Time Cardiovascular Disease Prediction System Using IoT and Machine Learning," Journal of Healthcare Informatics Research, vol. 7, no. 3, pp. 285–302, Sep. 2023. DOI: 10.1007/s41666-023-00159-2.

[9] A. Hossain, M. Miah, and M. H. Kabir, "Feature Selection in Random Forest Models for Accurate Heart Disease Prediction," Computers in Biology and Medicine, vol. 153, pp. 106415, Aug. 2023. DOI: 10.1016/j.compbiomed.2023.106415.

[10] E. K. Dritsas and M. Trigka, "Ensemble Machine Learning for Heart Disease Prediction with SMOTE: Addressing Class Imbalance in Medical Data," Medical Informatics and Decision Making, vol. 23, no. 5, pp. 89–103, Nov. 2023. DOI: 10.1186/s12911-023-01923-6.

Biographies

Rudraksh Singh Bhadauria is a 3rd-year B.Tech student in Electronics and Communication Engineering at Sharda University. His research focuses on artificial intelligence, machine learning, deep learning, and computer vision, with applications in anomaly detection, speech recognition, and intelligent traffic management.

Iqra Javid is currently pursuing her Ph.D. in wireless Communication from Sharda University, India. She has obtained her M.Tech in Digital Communication from Sharda University, India. Her interest includes Machine learning, Artificial Intelligence, wireless networks and device to device communications.

Anirban Khara is pursuing an MS in Computer Science at George Washington University. His research focus on machine learning, Artificial Intelligence and deep learning.