Archives available at journals.mriindia.com

International Journal on Advanced Electrical and Computer Engineering

ISSN: 2349-9338
Volume 14 Issue 01, 2025

Smart Detection for Healthy Heart

Prof. Rashmi Shende[1], Ms. Vaishali Ghodmare[2], Ms. Rina Bawankule[3], Ms. Pallavi Bhute[4], Ms. Tulsi Gaikwad[5]

[1]*Assistant Professor, Department of Computer Engineering, SCET, Nagpur, Maharashtra, India*
[2,3,4,5] *UG Student, Department of Computer Engineering, SCET, Nagpur, Maharashtra, India*
[1]*bhurleyrashmi1@gmail.com, 8446519053,* [2]*vaishalighodmare29@gmail.com, 7743851490,* [3]*rinabawankule85@gmail.com, 8010140055,* [4]*pallavibhute2003@gmail.com, 7249264597,* [5]*tulsigaikwad12@gmail.com, 7887694034*

Peer Review Information

Submission: 07 Feb 2025
Revision: 16 Mar 2025
Acceptance: 18 April 2025

Keywords

Machine Learning
Majority Voting Ensemble Method
Heart Disease
UCI Dataset

Abstract

This study presents a majority voting ensemble method aimed at predicting the likelihood heart disease. likelihood, providing a reliable tool for early detection. Predictions are based on common, affordable medical tests available at local clinics. The primary aim is to enhance diagnostic accuracy and confidence by offering machine-assisted insights. Trained on real-world data from both healthy individuals and heart disease patients, the model ensures diverse and realistic predictions. Patient classification is determined by the majority vote of multiple machine learning models, each trained on the available medical data. This ensemble approach improves accuracy by combining the strengths of different models, reducing errors associated with relying on a single algorithm. Results show that the method achieves an impressive 90% accuracy, proving its effectiveness in delivering reliable heart disease predictions. The findings highlight its potential as a valuable tool for doctors and patients, providing timely and trustworthy guidance for heart disease detection.

Introduction

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, responsible for over 70% of all fatalities. In developed countries, the prevalence of CVDs is high due to poor dietary habits, smoking, and obesity. Meanwhile, low- and middle-income nations are facing an increasing burden of CVDs as a result of changing lifestyles and limited access to healthcare. The economic impact of CVDs is also enormous, with an estimated USD 3.7 trillion spent between 2010 and 2015, which significantly affects healthcare systems and productivity.

Early detection of CVDs is crucial in reducing both the health and financial consequences of these diseases. Projections suggest that CVD-related deaths could reach 23.6 million by 2030, making early diagnosis more important than ever. Machine learning has shown great potential in predicting heart disease by analyzing critical risk factors such as high blood pressure, diabetes, and elevated cholesterol levels. However, achieving high

accuracy in these predictions remains a challenge. In this study, we evaluate several machine learning models, including random forest, decision trees, multilayer perceptrons, and XGBoost, to improve prediction accuracy. We also incorporate k-modes clustering as part of the data preprocessing process, using a large Kaggle dataset to ensure our models are both reliable and robust. All analyses were conducted using Python and Google Colab, ensuring an efficient and streamlined approach to the task.

LITERATURE SURVEY

Machine learning has made notable progress in cardiology, enabling the early detection of heart disease risk factors. For example, Narain et al. (2016) achieved 98.57% accuracy using a quantum neural network, far outperforming traditional methods like the Framingham risk score (19.22%). Shah et al. (2020) applied KNN to the Cleveland dataset and reached 90.8% accuracy. Drod et al. (2022) used logistic regression, PCA, and feature ranking to identify CVD risks linked to MAFLD, achieving an AUC of 0.87. Hasan & Bao (2020) found XGBoost with wrapper selection most effective, with 73.74% accuracy. Alotalibi (2019) reported 93.19% accuracy for decision trees, with SVM closely behind at 92.30%.

Despite these advancements, small datasets limit generalization. Our study addresses this by using a large Kaggle dataset of 70,000 patient records with 11 features, improving model reliability and reducing overfitting. Table 1 highlights key studies, emphasizing the importance of large datasets in predictive analytics.

Researcher(s) & Year	Methodology	Accuracy	Dataset
Shorewall (2021)	earning with KNN, RF, SVM, and logistic regression	75.1%	Kaggle (70,000 records, 12 features)
Maiga et al. (2019)	RF, Naïve Bayes, Logistic Regression, KNN	70%	Kaggle (70,000 records, 12 features)
Waigi et al. (2020)	Decision Tree	72.77%	Kaggle (70,000 records, 12 features)
Our and ElSeddawy (2021) [	Random Forest with Repeated Random Sampling	89.01%	UCI (303 records, 14 features)
Khan & Mondal (2020)	Neural Networks, Logistic Regression, SVM	71.82% - 72.72%	Kaggle (70,000 records, 12 features)

PROPOSED METHODOLOGY

This study uses machine learning to predict heart disease risk, offering valuable support for healthcare professionals and patients. Key preprocessing steps included removing outliers, integrating Mean Arterial Pressure (MAP) and BMI, and using k-modes clustering for better data segmentation. The model was trained on a Kaggle dataset of 70,000 records and 12 features, ensuring improved accuracy and reliability.

Data Source

The dataset [23] includes 70,000 patient records, with 12 attributes such as age, gender, and systolic and diastolic blood pressure. The target variable, "cardio," indicates whether cardiovascular disease (CVD) is present (1) or absent (0).

Data Preprocessing & Feature Engineering

Outliers in height, weight, and blood pressure were removed using the 2.5th–97.5th percentile range, refining the dataset to 57,155 records. To enhance classification, age, BMI, and blood pressure were grouped into bins.

MAP, an important indicator for CVD risk, was calculated using the following formula: MAP=(2×Diastolic BP)+Systolic BP3MAP = \frac{(2 \times \text{Diastolic BP}) + \text{Systolic BP}}{3}MAP=3(2×Diastolic BP)+Systolic BP MAP values were then categorized into ten ranges, such as 70–80, 80–90, to make the model more interpretable.

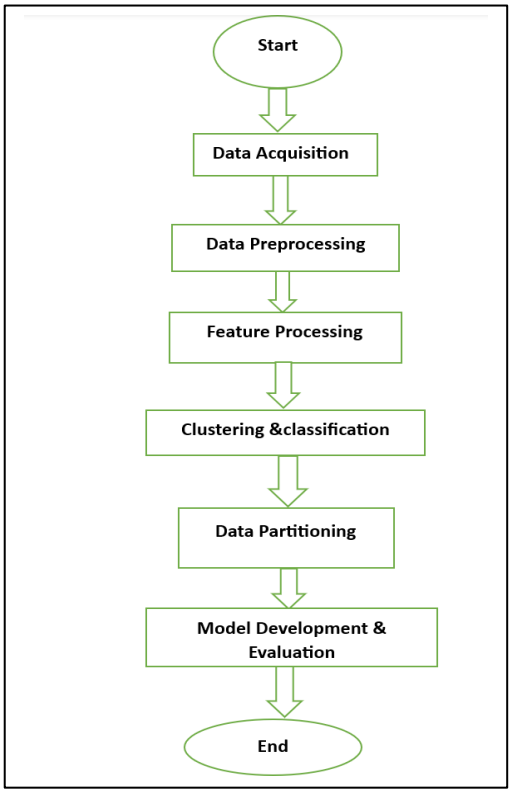

Fig. Flow Diagram of model

RESULT AND IMPACT ANALYSIS

The analysis was conducted using Google Colab on a Ryzen 7 4800-H with 16 GB of RAM. The original dataset, consisting of 70,000 records with 12 attributes, was refined to around 59,000 records with 11 categorical attributes after preprocessing and outlier removal to improve efficiency.
Model & Evaluation
Four machine learning models were tested for predicting cardiovascular disease: Random Forest (RF)
Decision Tree (DT) Multilayer Perceptron (MLP) XGBoost
The models were evaluated based on several metrics: accuracy, precision, recall, F1-score, and AUC. To ensure robustness, an 80-20 train-test split was used, and hyperparameter tuning was done using GridSearchCV with k-fold cross-validation.

Results & Performance
After tuning, the MLP model achieved the highest accuracy of **87.28%**, with: Precision: 88.70%
Recall: 84.85%
F1-Score: 86.71%
AUC: 0.95
There were also improvements in other models:
Random Forest: 86.48% → 86.90%
XGBoost: 86.40% → 87.02%

ROC Curve & AUC
The ROC curve helped visualize the True Positive Rate versus the False Positive Rate at different thresholds. The AUC score, which measures model performance, showed that higher values indicate more effective predictions.

Table 5: Classifier Performance Metrics (Before & After Cross-Validation)

Model	Accuracy	Precision	Recall	F1-Score	AUC
MLP	86.94 → 87.28	89.03 → 88.70	82.95 → 84.85	85.88 → 86.71	0.95

RF	86.92 → 87.05	88.52 → 89.42	83.46 → 83.43	85.91 → 86.32	0.95
DT	86.53 → 86.37	90.10 → 89.58	81.17 → 81.61	85.40 → 85.42	0.94
XGBoost	87.02 → 86.87	89.62 → 88.93	82.11 → 83.57	86.30 → 86.16	0.95

The results demonstrate significant improvements in model performance after cross-validation and tuning, with MLP and Random Forest showing the best outcomes.

CONCLUSION AND FUTURE WORK

This study utilized machine learning for heart disease classification, with a particular emphasis on enhancing data preprocessing through k-modes clustering. Key steps in preprocessing included age binning (in 5-year intervals), categorizing blood pressure into 10 intervals, and segmenting the data by gender. The elbow method was used to determine the optimal number of clusters, which improved data segmentation and model performance. Among the machine learning models tested, the Multilayer Perceptron (MLP) achieved the highest accuracy at 87.23%, highlighting the effectiveness of deep learning for this task. In contrast, Decision Trees achieved the lowest accuracy at 86.37%, which indicates some limitations of tree-based models in this scenario. Looking ahead, several areas for improvement and further exploration include comparing k-modes clustering with other methods like k- means and hierarchical clustering to identify the best approach for data segmentation. Additionally, analyzing the impact of missing data and outliers on model accuracy is crucial to ensure the robustness of predictions under various data conditions. Validating the models on independent datasets would help assess their generalizability and ensure consistent results across different populations. Lastly, improving cluster interpretability could provide clearer clinical insights, making the results more actionable for healthcare professionals. These enhancements have the potential to refine the model's performance and further its practical application in heart disease prediction.

References

"Cardiovascular Diseases (CVDs)." World Health Organization. Available at: https://www.who.int/news- room/fact-sheets/detail/cardiovascular-diseases-(cvds) (Accessed: January 28, 2024).

"Cardiovascular Disease: Types, Causes & Symptoms." Cleveland Clinic. Available at: https://my.clevelandclinic.org/health/diseases/21493-cardiovascular-disease (Accessed: January 28, 2024).

"Cardiovascular Disease - NHS." National Health Service. Available at: https://www.nhs.uk/conditions/cardiovascular-disease/ (Accessed: January 28, 2024).

Kakadiaris, I.A., Vrigkas, M., Yen, A.A., Kuznetsova, T., Budoff, M., & Naghavi, M. (2018). "Machine learning outperforms the ACC/AHA CVD risk calculator in MESA." Journal of the American Heart Association, 7(22), e009476.

Chowdhury, M.Z.I. et al. (2018). "Prevalence of cardiovascular disease in the Bangladeshi adult population: A systematic review and meta-analysis." Vascular Health and Risk Management, 14, 165.

Singh, A., & Kumar, R. (2020). "Heart disease prediction using machine learning algorithms." Proceedings of the International Conference on Electrical and Electronics Engineering (ICE3), 452–457.

Yadav, A.L., Soni, K., & Khare, S. (2023). "Heart disease prediction using machine learning." Proceedings of the 14th International Conference on Computing, Communications, and Networking Technologies (ICCCNT).

Salazar, L.H.A., Leithardt, V.R.Q., Parreira, W.D., Fernandes, A.M., Barbosa, J.L.V., & Correia, S.D. (2021). "Application of machine learning techniques to predict patient no-shows in healthcare settings." Future Internet, 14(1), 3.

Saha, S., Showrov, M.I.H., Rahman, M.M., & Majumder, M.Z.H. (2023). "VADER vs. BERT: A comparative performance analysis for sentiment analysis on the coronavirus outbreak." Lecture Notes in Computer Science (LNICST), 490, 371–385.

Rahman, M.M., Saha, S., Majumder, M.Z.H., Akter, F., Haque, M.A.S., & Anzan-Uz-Zaman, M. (2022).

"Design and development of an IoT-based smart system for monitoring and controlling laboratory environments." Proceedings of the 4th International Conference on Sustainable Technologies Industry 4.0 (STI 2022).

Kumari, J., Kumar, E., & Kumar, D. (2023). "A structured analysis of machine learning and deep learning applications in healthcare with big data analytics." Archive of Computational Methods in Engineering, 30(6), 3673–3701.

Lupague, R.M., Mabborang, R.C., Bansil, A.G., Lupague, M.M., & Marcus, R.J.M. (2023). "An integrated machine learning model for comprehensive heart disease prediction." European Journal of Computer Science and Information Technology, 11, 34458.

Krishna, C.S.R., Vasanthi, M., Reddy, K.H., & Jaswanth, G. (2023). "Heart disease prediction using machine learning." Proceedings of the International Conference on Advanced Computing, 589–595.

Jaya, T., Mohan, M., & Alam, M.S. (2023). "Effective heart disease prediction using machine learning—Modified KNN." Proceedings of the International Conference on Machine Learning and Applications, 479–489.

"UCI Machine Learning Repository." Available at: https://archive.ics.uci.edu/dataset/45/heart+disease (Accessed: January 28, 2024).

Jaya, T., Mohan, M., & Alam, M.S. (2023). "Effective heart disease prediction using machine learning—Modified KNN." Proceedings of the International Conference on Machine Learning and Applications, 479–489.

Krishnan, J.S., & Geetha, S. (2019). "Prediction of heart disease using machine learning algorithms." Proceedings of the 1st International Conference on Innovation in Information and Communication Technology (ICIICT).

Mohapatra, S., et al. (2023). "A stacking classifier model for detecting heart irregularities and predicting cardiovascular diseases." Healthcare Analytics, 3, 100133.

Reddy, K.V.V., Elamvazuthi, I., Aziz, A.A., Paramasivam, S., Chua, H.N., & Pranavanand, S. (2021). "Heart disease risk prediction using machine learning classifiers with attribute evaluators." Applied Sciences, 11(18), 8352.

Kumari, J., Kumar, E., & Kumar, D. (2023). "A structured analysis of machine learning and deep learning applications in healthcare with big data analytics." Archive of Computational Methods in Engineering, 30(6), 3673–3701.