Ensemble Learning-Based Predictive Analysis for Heart Disease Detection

Pasala Velangini Nishitha [1*], Vutla Nischitha [2*],
Bandhakavi Nitya Sriya [3*], Mr. K Subba Shankar [4],

[1] Student, Institute of Aeronautical Engineering, Hyderabad – 500080, India.
[2] Student, Institute of Aeronautical Engineering, Hyderabad – 500080, India.
[3] Student, Institute of Aeronautical Engineering, Hyderabad – 500080, India.
[4] Associate Professor, Institute of Aeronautical Engineering, Hyderabad – 500080, India.

*Corresponding author(s). E-mail(s): pvnishitha@gmail, vutlanischitha@gmail.com, b.nityasriya@gmail.com

Abstract. Heart disease remains one of the leading causes of illness and mortality worldwide, making quick and accurate diagnosis more essential. Many doctors rely on multiple clinical indicators such as patient's age, lifestyle habits, and laboratory test results to evaluate a patient's risk. Most existing computerized prediction systems use a single machine learning model and rely on only few parameters, which will likely limit the reliability of predictions and will not be able to give a complete picture of a patient's overall. The suggestion is to develop an ensemble machine learning method for predicting patient behavior. However, rather than relying on single predictive model, we use collective power of multiple different classifiers, which can be accomplished through a combination of bagging and boosting techniques. By evaluating population demographic information, biochemistry laboratory test results, and lifestyle habits, this method minimizes both bias and overfitting which means the model provides more stable predictions. The performance of the model is measured with standard metrics, thus providing the opportunity for earlier detection, better clinical decisions and timely diagnosis and individualized treatment options and preventative care, which allow for enhanced long-term heart health with improved outcomes.

Keywords: Ensemble Machine Learning, Decision Trees, Random Forests, Bagging & Boosting, Accuracy, Healthcare Diagnostics.

1 Introduction

Through the application of machine learning (ML), intelligent systems can be created that utilize a historical medical dataset to make accurate predictions concerning clinical outcomes. In the healthcare sector, ML techniques have the ability to analyze large amounts of data in which complex patient data such as age, cholesterol level, blood pressure and chest pain type provides a means of diagnosing heart disease. By examining these factors for patterns and correlations, machine learning will allow for improved accuracy and timeliness in making clinical decisions. Latest Research revealed that hybrid approaches combine oversampling techniques with adaptive boosting improve risk classification accuracy, and stacked ensemble methods

R. Vasanth Kumar Mehta et al. (eds.), *Proceedings of the International Conference on Intelligent Systems for a Sustainable Future (ISSF 2026)*, Atlantis Highlights in Intelligent Systems 16,
https://doi.org/10.2991/978-94-6239-693-7_6

that combines XGBoost and deep learning algorithms perform better than single classifiers on standard heart disease datasets. HeartEnsembleNet, for example, confirmed high accuracy rate of close to 93% using a combination of multiple classifiers.

Machine learning models are differ from traditional diagnostic models because they can keep learning as new data becomes available. It helps diagnosis by automating tasks, reducing human error, and allow early detection of heart disease in high-risk patients. Even with recent advancements, challenges remain regarding dataset imbalance, limited external validation, and inconsistent performance across diverse populations, indicating the need for further comparative studies.

Finally, using comparative studies of machine learning algorithms such as k-nearest neighbour,decision trees, support vector machines, logistic regression, etc. assist in determining which method is most appropriate for real-world clinical settings and enhancing the role of machine learning within predictive healthcare environments.

2 Literature Survey

Heart disease prediction has been extensively researched using machine learning algorithms to enhance early diagnosis. Alizadehsani et al. [1] introduced a data mining model for diagnosing coronary artery disease. It showed that systematic feature extraction and classification can greatly improve prediction accuracy. This study highlighted the importance of well-prepared clinical data in enhancing model accuracy.

Ismaeel et al. [2] used Extreme Learning Machines (ELM) to predict the disease, looking to train the model more quickly without losing accuracy. Polat and Güneş [5] combined fuzzy logic and neural networks to improve interpretability and decision support, explaining the importance of understandable models in medicine.

Liu et al. [4] proposed hybrid machine learning methods which uses combination of multiple classifiers to improve robustness and generalization. Sharma and Parmar [3] also proposed the use of deep learning neural networks for predicting heart disease, with high accuracy. These works justify the application of ensemble models in the current study.

3 Methodology

Data collection, data preprocessing, model building, evaluation, and result interpretation are part of the system's structured machine learning pipeline. The procedure begins with collection of required clinical data, which is then cleaned and prepared to ensure accuracy and consistency. The algorithms are employed to classify patients as either having heart disease or not. To compare the difference in the performance of the models, individual models as well as ensemble models are developed. The models are tested using various metrics such as accuracy, precision, recall, F1 score, cross-validation, and confusion matrix to provide reliable results.

3.1 Data Preparation

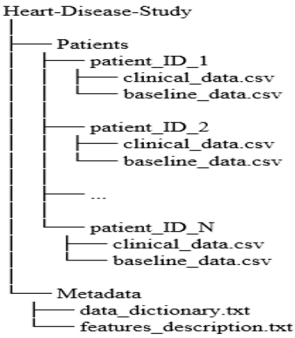
This study makes use of two structured datasets, which are frequently utilised in cardiovascular research. Important clinical characteristics like age, gender, blood pressure, cholesterol, type of chest pain, and other diagnostic features are included in the two datasets.

Table 1: Heart Disease Datasets Overview

Dataset Name	Source	Records	Features	Purpose
Heart-Disease-Study	Kaggle	70,000	14+	Model scaling and generalization
UCI Heart Disease	UCI Repository	303	14	Model training and evaluation

3.2 Data Preprocessing

The datasets are preprocessed to improve their quality prior to training. The feature names are standardised, the superfluous columns are removed, and median imputation is used to handle the missing values. To prevent multicollinearity, one category is removed after the categorical variables are converted into numerical variables using encoding techniques.



The data used for this study are structured hierarchically. All data reside under the Heart-Disease-Study directory and are organized at a patient level. Each patient has a unique identifier from patient_ID_1 to patient_ID_N. Every patient has two different data files: clinical_data.csv, with detailed clinical and laboratory measurements, and baseline_data.csv, with baseline or reference information collected before the collection of detailed clinical data.

Apart from the patient-specific data, there is also a separate Metadata directory to support the interpretation of the data. This metadata directory contains a data_dictionary.txt file that defines the variables used in the dataset and a features_description.txt file that describes the clinical relevance and meaning of each feature. The hierarchical structure defined here will uniformly organize the records of patients, and comparative analysis

of the baseline and clinical data for each individual will facilitate performing machine learning–based studies to predict heart diseases using this dataset.

3.3 Exploratory Data Analysis

To examine trends and correlations among the variables, exploratory data analysis is done. As seen in Figures 1. and 2. heatmaps of correlation are used to depict the significant and less significant correlations to identify the influential variables related to heart disease. After that, the data is split into 80:20 training and testing sets.

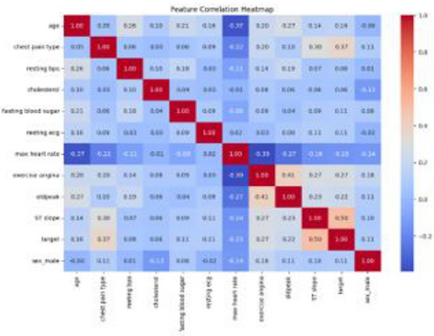


Fig. 1 Dataset-1 Correlation Heatmap

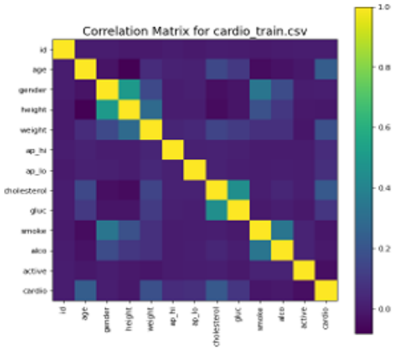


Fig. 2 Dataset-2 Correlation Heatmap

3.4 Feature Selection

Feature selection has been done using a technique involving correlation and ranking based on feature importance. After completing the preprocessor step, a correlation heat map has been developed to showcase correlations between features in a dataset and correlations above a certain threshold have been observed between some features.

Additionally, a Random Forest classifier was used to train the entire set of variables, and the feature importances were obtained from the trained classifier. The feature importances derived from the training of the model were utilized to derive the most important set of variables that are most important with respect to their relevance to the possibility of a person having a heart disease. The most important set of variables includes the outcome variable as well as the major study variable.

3.5 Model Development

To assess the predictive accuracy, various machine learning algorithms are used. These models range from simple to complex techniques to understand the trade-offs between interpretability, complexity, and accuracy. The models employed are:

Logistic Regression. Because of its ease of use and interpretability, logistic regression is used as a baseline model. It creates a benchmark for assessing the performance improvements attained by more intricate models.
Optimized Random Forest. Randomised search is used to improve the Random Forest model through hyperparameter tuning. The model enhances predictive stability and lowers variance by combining several decision trees.

Stacking classifier. The Stacking Classifier combines multiple base learners in a layered approach. It feeds the outputs of different models into a meta-classifier to improve generalization and take advantage of the strengths of individual classifiers.

Bagging Meta Estimator. Bagging uses bootstrap sampling to train multiple instances of a model on different subsets of data. The final prediction is obtained by aggregating individual outputs, reducing overfitting and enhancing robustness.

Gaussian Naïve Bayes. Gaussian Naive Bayes is added as a light probabilistic model that assumes independence between features. Although it is a simple model, it is very fast in making predictions.

4 Results and Discussion

The models that are employed for predicting heart disease on two different datasets are analyzed in this section. This includes confusion matrix analysis and performance metrics to assess the models' predictability. Several derived performance measures are computed using these measurements: precision, accuracy, F1-score, recall and confusion matrix.

4.1 Confusion Matrix Analysis

To visualize the distribution of true and predicted labels, a confusion matrix was generated for optimised random forest model.

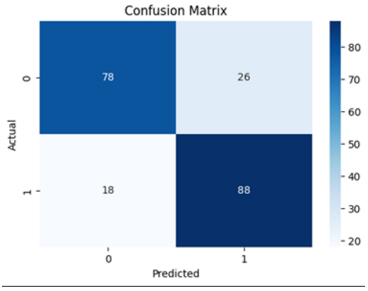


Fig. 3 Confusion Matrix Analysis - Dataset 1

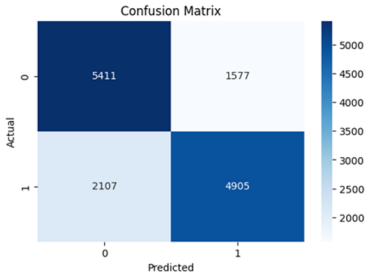


Fig. 4 Confusion Matrix Analysis - Dataset 2

Observations:

- Recall is a critical metric in healthcare, as low recall leads to false negatives and missed heart disease diagnoses.
- The Optimized Random Forest performed very well in terms of recall, making it a more accurate tool for early detection and screening.
- Prior cardiovascular prediction studies also emphasize recall as an essential measure in clinical machine learning systems.

4.2 Model Comparison

The results are summarized in Table 2(a) and Table 2(b):

Table 2(a). Model Comparison Table - Dataset 1

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Optimised Random Forest	80.04	77.19	83.02	80.00
Logistic Regression	75.71	75.72	75.71	75.71
Stacking Classifier	76.67	76.68	76.67	76.66
Bagging Meta Estimator	75.24	75.47	75.47	75.47
Gaussian NB	74.76	76.77	71.70	74.15

The accuracy of 80.04% achieved is comparable to the results obtained by Liu et al. (2020), who showed that hybrid ensemble and tree-based classifiers can achieve accuracy above 78% on cardiovascular disease datasets. This further reinforces the use of ensemble classifiers for heart disease prediction.

Table 2(b). Model Comparison Table - Dataset 2

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Optimised Random Forest	73.69	75.67	69.95	72.90
Logistic Regression	69.88	69.97	69.88	69.85
Stacking Classifier	73.66	73.78	73.66	73.63
Bagging Meta Estimator	71.61	72.95	68.85	70.84
Gaussian NB	57.39	75.60	22.05	34.14

The slight loss of accuracy on Dataset 2 could be attributed to the smaller size of the dataset (303 records) compared to Dataset 1 (70,000 records). This could be due to the smaller datasets generalizing less and possibly having class imbalance. On the other hand, Gaussian Naive Bayes was inadequate in larger or more unbalanced datasets because clinical features like blood pressure and cholesterol levels are not independent, making Naive Bayes a less effective classifier.

5 Conclusion and Future Scope

The system proposed in this study demonstrates a practical approach for the early prediction of heart disease using machine learning techniques. Experimental results show that ensemble-based models perform more reliably than individual classifiers, with the optimized Random Forest achieving the highest accuracy of 80.04% along with strong recall, which is essential for reducing missed diagnoses in clinical screening. Careful data preprocessing, including feature encoding, handling missing values, and class balancing, played an important role in improving model stability and consistency across datasets.

The findings also indicate that dataset size and class distribution significantly influence predictive performance, with better generalization observed on the larger dataset. These results are in line with existing studies that highlight the robustness of ensemble learning methods in healthcare prediction tasks. Future work can focus on incorporating advanced models such as XGBoost and deep learning techniques, validating the framework on real-world clinical data, and extending the system to support multi-disease prediction, thereby enhancing its applicability as a clinical decision support tool.

References

1. Alizadehsani, R., Habibi, J., Hosseini, M.J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Bahadorian, B., & Sani, Z.A., "A Data Mining Approach for Diagnosis of Coronary Artery Disease," Computer Methods and Programs in Biomedicine, vol. 111, no. 1, pp. 52–61, 2013.

2. Ismaeel, S., Miri, A., & Chourishi, D., "Heart Disease Diagnosis Using Extreme Learning Machines," IEEE Canada International Humanitarian Technology Conference, 2015.

3. Sharma, S., Parmar, M.: Heart diseases prediction using deep learning neural network model. International Journal of Innovative Technology and Exploring Engineering *(IJITEE)* **9**(3), 2020..

4. Liu, Y., Wang, Y., Zhang, J., & Li, Y., "Hybrid Machine Learning Techniques for Cardiovascular Disease Prediction," IEEE Access, vol. 8, pp. 18966–18979, 2020.

5. Polat, K., & Güneş, S., "Medical Decision Support Using Hybrid Feature Selection, Fuzzy Weighting, and Neural Networks," Digital Signal Processing, vol. 16, no. 4, pp. 489–496, 2006.