An Intelligent Machine Learning Framework for Early Prediction of Heart Disease

1st Ch.Keerthi mahalakshmi, 2nd K.V. Sai Harika,3rd P. latha Mounika, 4th B. Sai Durga, 5th S. Devi Poojitha

Dept. Computer Application, Aditya University, Surampalem, India

keerthimahalaxmi99@gmail.com
hariharika1431@gmail.com
lathamounika777@gmail.com
balisaidurga8@gmail.com
poojithaswamiredy@gmail.com

ABSTRACT

Cardiovascular diseases rank among the major causes of deaths in various parts of the world and therefore early detection is critical in terms of early diagnoses and treatment. This paper offers a smart machine learning model to predict heart disease early in clinical and demographic data. There are data preprocessing, exploratory data analysis, feature engineering, and running the machine learning models, which are Decision Tree, Random Forest, and Logistic Regression. An intelligent hybrid model is created to enhance prediction performance by use of a voting classifier which is a combination of Logistic Regression and an optimized random forest model. Hyperparameter optimization is implemented to improve the efficacy of models and feature importance analysis is implemented to single out crucial risk factors. The models will be measured by accuracy, precision, recall, and F1-score. It has been found that the hybrid model is more accurate and reliable than individual models. The suggested system can help healthcare providers in the early diagnosis and improved decision-making, which will contribute to a decrease in the risk of severe heart-related conditions.

KEYWORDS: Heart Disease Prediction, Machine Learning, Hybrid Model, Logistic Regression, Random Forest, Decision Tree, Voting Classifier, Hyperparameter Tuning, Feature Importance, Healthcare Analytics, Early Diagnosis

1. INTRODUCTION

The causes of cardiovascular diseases (CVDs) have become a significant burden on the global healthcare systems since they rank among the causes of mortality(Ali, F., El-Sappagh, S., Islam, S. R., Kwak, D., Ali, A., Imran, M., & Kwak, 2020). The rising cases of heart related diseases have led to the emergence of the necessity of early diagnosis and proper methods of prediction in order to decrease the mortality rates and enhance patient outcomes. Older methods of diagnosis tend to use either clinical judgment and manual examination, which may be time-consuming and subject to human error(Ramesh, B., & Lakshmanna, 2024). Thus, the need to use smart and automated system that will support the healthcare professional in making prompt and correct decisions is increasing.

The recent progress in machine learning and artificial intelligence has profoundly changed the healthcare industry by allowing predictions and diagnosis to be performed based on data. Machine learning models are able to process vast amount of clinical data and point to concealed patterns that can hardly be found with the help of traditional methods. A number of studies have shown that machine learning and deep learning methods can be used to predict heart disease more accurately and efficiently (Rahim, A., Rasheed, Y., Azam, F., Anwar, M. W., Rahim, M. A., & Muzaffar, 2021). In addition, hybrid and ensemble-based models have become the focus of attention due to their capacity to integrate various models and improve the performance of prediction (Elwahsh, H., El-Shafeiy, E., Alanazi, S., & Tawfeek, 2021).

1.1 Role of Intelligent Machine Learning in Healthcare

The introduction of smart machine learning systems in the field of healthcare has created new opportunities in the context of early disease diagnosis and preventive services. These frameworks employ state-of-the-art algorithms, optimization

processes, and feature selection to enhance prediction. Recent studies place an emphasis on deep learning, ensemble learning, and hybrid models to create smart healthcare systems, which can be used to monitor and predict diseases in real-time (Vincent Paul, S. M., Balasubramaniam, S., Panchatcharam, P., Malarvizhi Kumar, P., & Mubarakali, 2022). Also, machine learning optimized models have demonstrated encouraging performance with complex medical data and effective predictions (Jiao, 2024).

1.2 Motivation and Objective of the Study

Nevertheless, even with these developments, most of the current systems are still troubled by issues of overfitting, lack of interpretability and poor generalization. In order to overcome these weaknesses, this paper suggests a smart machine learning model to predict heart disease at an early stage. The suggested solution will combine several machine learning models, such as Decision Tree, Random Forest, and Logistic Regression as well as an optimized hybrid model with the help of a voting classifier(Pan, Y., Fu, M., Cheng, B., Tao, X., & Guo, 2020). This study aims at enhancing the quality of prediction, increasing the reliability of the model, and assisting healthcare professionals in the early diagnosis and decision-making process. Hyperparameter tuning and feature importance analysis are also included in the framework to guarantee the best performance and improved insight into the major risk factors related to heart disease (Pattanaik, S., & Nayak, 2024).

2. LITERATURE REVIEW

(Pachiyannan et al. 2024) suggested a machine learning approach for the early diagnosis of congenital cardiac disease through ECG signal processing. Their methodology emphasized the extraction of significant patterns from ECG signals and the implementation of classification algorithms to enhance diagnostic precision, underscoring the criticality of early-stage detection in clinical practice.

(Taylan et al. 2023) created a hybrid model that integrates machine learning, neuro-fuzzy, and statistical methodologies for the classification of cardiovascular diseases. Their methodology enhanced predictive performance by encompassing both linear and non-linear correlations, while also improving interpretability, a crucial factor in healthcare decision-making.

(Baghdadi et al. 2023) investigated sophisticated machine learning methodologies for the early detection of cardiovascular disorders. Their research highlighted that effective data preparation, feature selection, and model tuning markedly enhance prediction accuracy relative to conventional methods.

(Nandy et al. 2023) presented an advanced predictive method that combines swarm optimization with artificial neural networks. Optimization strategies improved model performance and efficiency in managing intricate healthcare data.

(Bertsimas, D., Mingardi, L., & Stellato, 2021) concentrated on real-time prediction of cardiac illness via machine learning models. Their research revealed that predictive systems can aid healthcare professionals in making swifter and more precise judgments in urgent contexts.

(Muhammad, Y., Tahir, M., Hayat, M., & Chong, 2020) introduced a computational framework for the early and precise identification of cardiac disease via several machine learning methods. Their findings indicated that hybrid and ensemble methodologies surpass individual models for accuracy and reliability.

(Tuli et al. 2020) introduced HealthFog, a healthcare system that use ensemble deep learning and is coupled with IoT and fog computing. Their approach facilitated real-time disease prediction and effective data processing, underscoring the significance of scalable and intelligent healthcare systems.

3. METHODOLOGY

This research presents an advanced machine learning framework for the early prediction of cardiovascular disease via structured clinical data. The methodology adheres to a methodical workflow encompassing data collection,

preprocessing, exploratory data analysis, model construction, optimization, hybrid modeling, and performance evaluation.

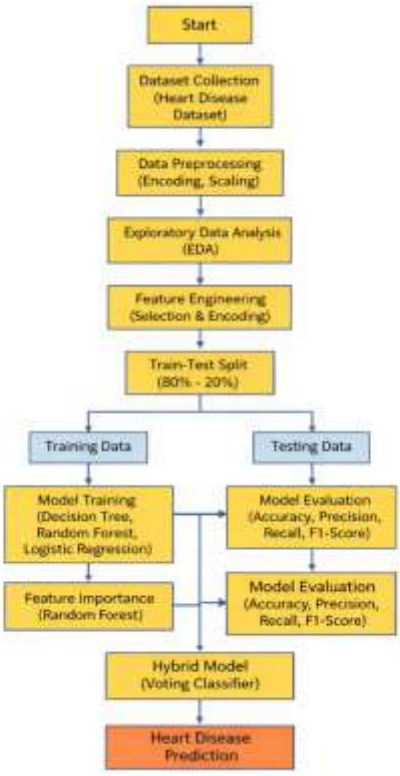

Figure 3.1 : Proposed Methodology for Heart Disease Prediction Using Hybrid Machine Learning Model

3.1 Dataset Collection

The dataset utilized is Heart_Disease_Prediction.csv, comprising anonymised patient records with clinical attributes including age, sex, chest pain kind, cholesterol levels, maximal heart rate, and ST depression.

The target variable is binary:

- **0 (Absence):** No heart disease
- **1 (Presence):** Heart disease present

3.2 Data Preprocessing

Preprocessing was performed to improve data quality and model performance:

- No missing values were found in the dataset.
- Categorical variables were converted into numerical format using Label Encoding.
- One-Hot Encoding was applied for better model compatibility.
- StandardScaler was used to normalize feature values.

3.3 Exploratory Data Analysis (EDA)

EDA was conducted to understand data patterns and relationships:

- Count plots and pie charts were used for categorical distribution.
- Histogram analyzed age distribution with heart disease.

- Correlation heatmap identified relationships between features.
- Pairplots showed feature interactions with the target variable.

3.4 Feature Engineering

Feature engineering included:

- One-Hot Encoding of categorical variables
- Selection of relevant features
- Separation of dataset into:
  - **X (features)**
  - **y (target variable)**

3.5 Train-Test Split

The dataset was divided into:

- **80% training data**
- **20% testing data**

A fixed random state ensured reproducibility of results.

3.6 Model Development

Three models were implemented:

- **Decision Tree:** Simple and interpretable but prone to overfitting
- **Random Forest:** Ensemble model with better accuracy and stability
- **Logistic Regression:** Efficient and widely used for binary classification

3.7 Feature Importance

The Random Forest model was employed to examine feature importance and discover critical elements affecting heart disease prediction.

3.8 Hyperparameter Tuning

GridSearchCV was utilized to improve Random Forest parameters, including the number of estimators and tree depth, thereby enhancing model performance.

3.9 Hybrid Model

A hybrid model was created with VotingClassifier by integrating Logistic Regression and optimized Random Forest. Soft voting was employed to enhance predictive accuracy.

3.10 Model Evaluation

Models were evaluated using:

- Accuracy
- Precision
- Recall

- F1-Score
- Confusion Matrix

A comparative analysis was performed to identify the best-performing model.

4. Results and Analysis

This section delineates the experimental outcomes of the suggested intelligent machine learning framework for the early prediction of heart disease. Multiple machine learning models, including Decision Tree, Random Forest, Logistic Regression, and a Hybrid Voting Classifier, were executed and assessed using conventional performance criteria.

4.1 Data Distribution Analysis

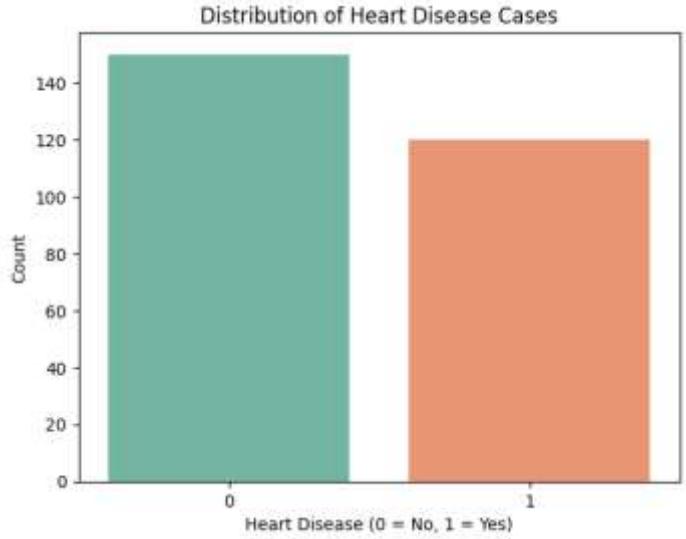

Figure 4.1: Distribution of Heart Disease Cases

The count plot depicts the distribution of patients with and without cardiovascular disease. The dataset exhibits a little imbalance between the two classes, which could affect model performance and necessitates further assessment.

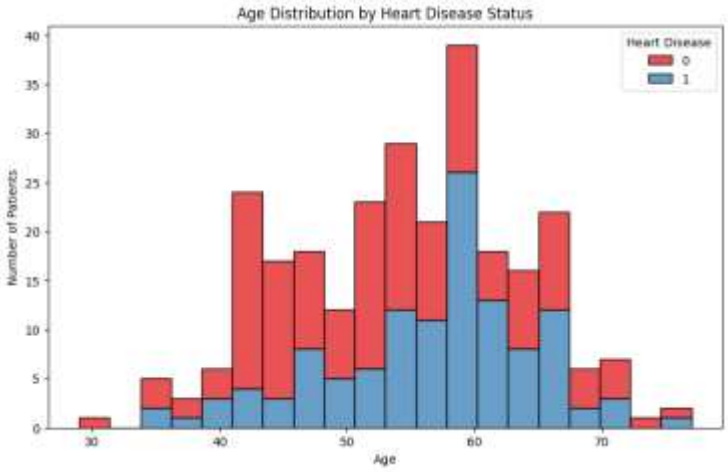

Figure 4.2: Age Distribution by Heart Disease

The histogram illustrates the age distribution of patients in relation to the existence of heart disease. Heart disease is more prevalent in middle-aged and older adults, signifying age as a critical risk factor.

4.2 Feature Analysis

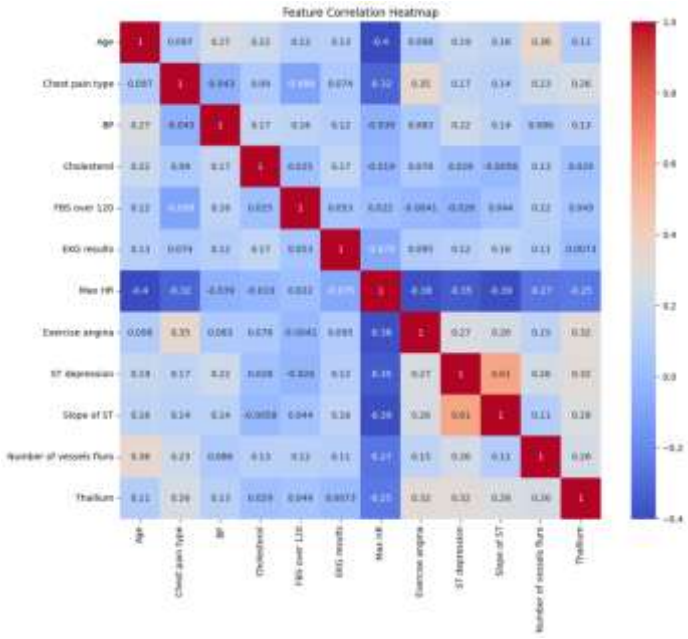

Figure 4.3: Correlation Heatmap

The heatmap illustrates the correlation among numerical features. Variables such as cholesterol level, maximum heart rate, and ST depression demonstrate significant connection with heart disease, underscoring its predictive value.

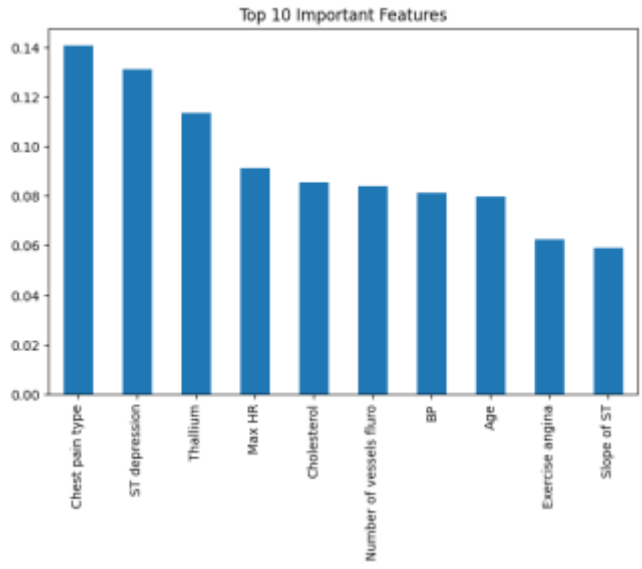

Figure 4.4: Top 10 Important Features

The feature importance graph derived from the Random Forest model delineates the most significant attributes influencing heart disease prediction. These properties are essential for enhancing model accuracy and decision-making.

4.3 Model Performance Evaluation

Numerous machine learning models were trained and assessed utilizing Accuracy, Precision, Recall, and F1-Score. A Hybrid Model integrating Logistic Regression and Random Forest was created to improve predictive performance.

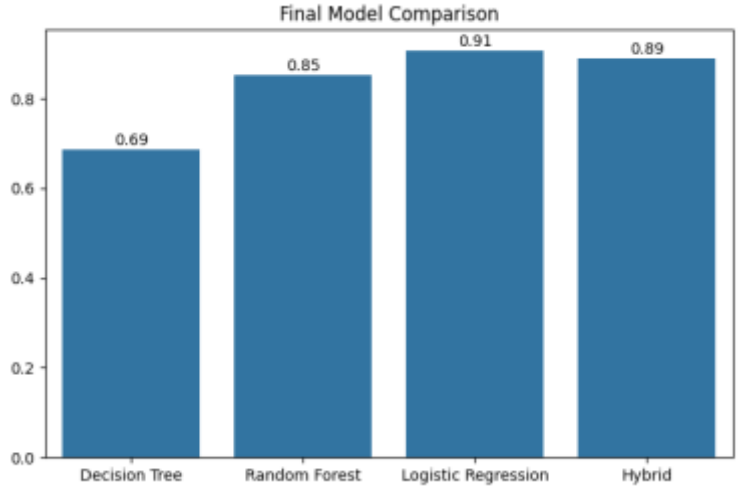

Figure 4.5: Model Accuracy Comparison

The bar chart illustrates the precision of all models. Logistic Regression demonstrates the highest accuracy among individual models, however the Hybrid Model performs competitively by integrating the strengths of many methods.

4.4 Performance Comparison Table

The table below presents a comparative analysis of all models based on key evaluation metrics:

Table 1 : Model Performance Comparison Table

Metric	Decision Tree	Random Forest	Logistic Regression	Hybrid Model
Precision	0.5714	0.8824	0.9000	0.8947
Recall	0.7619	0.7143	0.8571	0.8095
F1-Score	0.6531	0.7895	0.8780	0.8500
Accuracy	0.6852	0.8519	0.9074	0.8889

4.5 Analysis Summary

- Logistic Regression achieved the highest accuracy, indicating strong performance in linear classification tasks.
- Random Forest showed balanced performance with good precision and robustness.
- Decision Tree performed comparatively lower due to overfitting and limited generalization.
- The Hybrid Model provided improved stability and competitive performance by combining multiple models.

Overall, the results demonstrate that the proposed hybrid machine learning framework is effective for early prediction of heart disease and can support reliable decision-making in healthcare systems.

5. Conclusion

This study introduced an advanced machine learning framework for the early prediction of heart disease utilizing various classification models. Decision Tree, Random Forest, and Logistic Regression algorithms were executed and assessed, with a Hybrid Voting Classifier to enhance overall predictive efficacy. The findings indicate that Logistic Regression attained the highest accuracy among standalone models, however the Hybrid Model exhibited competitive and consistent performance by integrating the advantages of several techniques.

The examination of feature significance revealed that factors like age, cholesterol level, maximum heart rate, and ST depression are crucial in forecasting heart disease. Moreover, data visualization and correlation analysis facilitated the comprehension of the links among various clinical factors and the goal variable.

The proposed framework is effective, dependable, and appropriate for the early identification of cardiac disease. It can aid healthcare personnel in making prompt judgments and enhancing patient outcomes. The amalgamation of hybrid machine learning methodologies augments predictive accuracy, rendering the system more resilient and suitable for practical healthcare settings.

REFERENCES

1. Ali, F., El-Sappagh, S., Islam, S. R., Kwak, D., Ali, A., Imran, M., & Kwak, K. S. (2020). *A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Information Fusion, 63, 208-222.*
2. Baghdadi, N. A., Farghaly Abdelaliem, S. M., Malki, A., Gad, I., Ewis, A., & Atlam, E. (2023). *Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. Journal of Big Data, 10(1), 144.*
3. Bertsimas, D., Mingardi, L., & Stellato, B. (2021). *Machine learning for real-time heart disease prediction. IEEE Journal of Biomedical and Health Informatics, 25(9), 3627-3637.*
4. Elwahsh, H., El-Shafeiy, E., Alanazi, S., & Tawfeek, M. A. (2021). *A new smart healthcare framework for real-time heart disease detection based on deep and machine learning. PeerJ Computer Science, 7, e646.*
5. Jiao, N. (2024). *An efficient disease prediction framework based on optimized machine learning models for a smart healthcare application. Multimedia Tools and Applications, 83(17), 50825-50848.*
6. Muhammad, Y., Tahir, M., Hayat, M., & Chong, K. T. (2020). *Early and accurate detection and diagnosis of heart disease using intelligent computational model. Scientific reports, 10(1), 19747.*
7. Nandy, S., Adhikari, M., Balasubramanian, V., Menon, V. G., Li, X., & Zakarya, M. (2023). *An intelligent heart disease prediction system based on swarm-artificial neural network. Neural Computing and Applications, 35(20), 14723-14737.*
8. Pachiyannan, P., Alsulami, M., Alsadie, D., Saudagar, A. K. J., AlKhathami, M., & Poonia, R. C. (2024). *A novel machine learning-based prediction method for early detection and diagnosis of congenital heart disease using ECG signal processing. Technologies, 12(1), 4.*
9. Pan, Y., Fu, M., Cheng, B., Tao, X., & Guo, J. (2020). *Enhanced deep learning assisted convolutional neural network for heart disease prediction on the internet of medical things platform. Ieee Access, 8, 189503-189512.*
10. Pattanaik, S., & Nayak, K. (2024). *Heart diseases prediction using machine learning and deep learning models. In 2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT) (pp. 343-349). IEEE.*
11. Rahim, A., Rasheed, Y., Azam, F., Anwar, M. W., Rahim, M. A., & Muzaffar, A. W. (2021). *An integrated machine learning framework for effective prediction of cardiovascular diseases. IEEE access, 9, 106575-106588.*
12. Ramesh, B., & Lakshmanna, K. (2024). *A novel early detection and prevention of coronary heart disease framework using hybrid deep learning model and neural fuzzy inference system. IEEe Access, 12, 26683-26695.*
13. Taylan, O., Alkabaa, A. S., Alqabbaa, H. S., Pamukçu, E., & Leiva, V. (2023). *Early prediction in*

classification of cardiovascular diseases with machine learning, neuro-fuzzy and statistical methods. Biology, 12(1), 117.

14. Tuli, S., Basumatary, N., Gill, S. S., Kahani, M., Arya, R. C., Wander, G. S., & Buyya, R. (2020). *HealthFog: An ensemble deep learning based Smart Healthcare System for Automatic Diagnosis of Heart Diseases in integrated IoT and fog computing environments. Future Generation Computer Systems, 104, 187-200.*

15. Vincent Paul, S. M., Balasubramaniam, S., Panchatcharam, P., Malarvizhi Kumar, P., & Mubarakali, A. (2022). *Intelligent framework for prediction of heart disease using deep learning. Arabian Journal for Science and Engineering, 47(2), 2159-2169.*