Improving Heart Disease Prediction Accuracy: A Machine Learning Approach Using KNN, Decision Trees, and Random Forest

Tushar Parsai
Computer Science and Engineering
Indian Institute of Information Technology, Pune
Pune, India
parsaitushar@gmail.com

Sudhir Sude
Computer Science and Engineering
Indian Institute of Information Technology, Pune
Pune, India
sudesudhir@gmail.com

Swayam Dinesh Satyameshram
Computer Science and Engineering
Indian Institute of Information Technology, Pune
Pune, India
ssatyameshram16@gmail.com

***Abstract*—Heart disease remains a major health concern globally, and early prediction is vital for effective management. Machine learning techniques such as k-Nearest Neighbors (KNN), Decision Trees, and Random Forests offer promising approaches for heart disease prediction by analyzing patterns in health data like blood pressure, cholesterol, and age.**

In this study, we evaluate these three classifiers for their accuracy and efficiency. KNN uses proximity to similar data points to make predictions, though it requires careful data scaling. Decision Trees offer interpretable pathways but can risk overfitting. Random Forests, an ensemble of Decision Trees, enhance accuracy by reducing overfitting through majority voting.

Our results show that Random Forests achieved the highest accuracy at 82%, making them the most reliable model for heart disease prediction in this analysis.

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Heart disease remains one of the leading causes of mortality worldwide, underscoring the need for effective early detection methods. Traditional diagnostic approaches rely heavily on clinical tests and expert assessment, which, while accurate, are often costly and time-consuming. Machine learning offers an alternative approach, providing predictive models that can analyze patient data and identify patterns associated with heart disease risk. By automating parts of the diagnostic process, these models have the potential to support medical professionals in making more informed decisions and delivering timely interventions.

In recent years, algorithms such as k-Nearest Neighbors (KNN), Decision Trees, and Random Forest Classifiers have shown promise in heart disease prediction due to their ability to handle complex, high-dimensional data. Each model has unique strengths: KNN identifies patterns based on similarity to other data points, which is useful for datasets with clear boundaries but can be computationally intensive. Decision Trees are highly interpretable, mapping out decision paths that are easy to follow but susceptible to overfitting. Random Forests, an ensemble method of Decision Trees, improve model accuracy by reducing overfitting through a majority voting approach, making them robust for medical prediction tasks.

In this study, we implement and compare these three models on a heart disease dataset, evaluating their effectiveness and practical applicability. Our findings reveal that Random Forests achieved an accuracy of 82 %, making them the most reliable model for this dataset. This research aims to contribute to the growing field of machine learning in healthcare by demonstrating the potential of ensemble models for improving heart disease diagnosis.

II. Ease of Use

Heart disease remains one of the leading causes of death worldwide, making its early detection crucial for improving patient outcomes. Accurate classification of heart disease is a challenging task due to the complexity and variety of factors influencing cardiovascular health. Traditional methods, such as clinical assessments and laboratory tests, can be time-consuming, costly, and subject to human error. Thus, the use of machine learning algorithms, particularly Decision Trees (DT) and Random Forests (RF), has gained significant attention in recent years.

Decision Trees are a popular classification technique due to their simple and interpretable nature. They work by recursively partitioning the feature space based on the most significant features at each node, leading to a tree-like structure where each leaf represents a decision or classification. One of the key advantages of Decision Trees is that they can handle both numerical and categorical data, making them suitable for heart disease classification, which often involves both types of variables (e.g., age, cholesterol levels, blood pressure).

Random Forests, an ensemble method built upon Decision Trees, improve upon their limitations by constructing multiple decision trees and aggregating their predictions to enhance accuracy and robustness. Random Forests overcome the risk of overfitting that may occur with a single Decision Tree by averaging results from various trees, reducing the model's variance and improving its generalization ability. This characteristic is

particularly important in heart disease classification, where the presence of noisy data and outliers can negatively affect the model's performance.

The significance of using Decision Trees and Random Forests for heart disease classification lies in their ability to provide automated, reliable, and accurate predictions with minimal human intervention. These methods are not only capable of predicting the likelihood of heart disease based on a range of medical features, but they also offer interpretability, which is crucial for healthcare practitioners to understand the reasoning behind the classification and make informed decisions.

Incorporating machine learning models like Decision Trees and Random Forests into clinical settings has the potential to accelerate the diagnosis of heart disease, improve patient care, and enable preventive measures. With the increasing availability of large healthcare datasets, these methods can be further optimized to improve classification accuracy, potentially saving lives by detecting heart disease at earlier stages when treatment options are more effective.

III. Dataset Description

The dataset used in this study contains various attributes related to heart disease prediction. Each entry in the dataset corresponds to a patient with the following features:

- **Age**: Age of the patient in years.
- **Sex**: Gender of the patient, where 1 represents male and 0 represents female.
- **CP (Chest Pain Type)**: Type of chest pain experienced by the patient.
  - 0: Typical angina
  - 1: Atypical angina
  - 2: Non-anginal pain
  - 3: Asymptomatic
- **Trestbps**: Resting blood pressure (in mm Hg).
- **Chol**: Serum cholesterol levels in mg/dl.
- **FBS (Fasting Blood Sugar)**: Fasting blood sugar greater than 120 mg/dl.
  - 1: True
  - 0: False
- **Restecg (Resting Electrocardiographic Results)**: Electrocardiogram results at rest.
  - 0: Normal
  - 1: ST-T wave abnormality (T wave inversions and/or ST elevation or depression of ¿ 0.05 mV)
  - 2: Probable or definite left ventricular hypertrophy by Estes' criteria
- **Thalach**: Maximum heart rate achieved during exercise.
- **Exang (Exercise Induced Angina)**: Whether the patient experiences exercise-induced angina.
  - 1: Yes
  - 0: No
- **Oldpeak**: ST depression induced by exercise relative to rest.
- **Slope**: The slope of the peak exercise ST segment.
  - 0: Upsloping
  - 1: Flat
  - 2: Downsloping
- **Ca (Number of Major Vessels)**: Number of major vessels colored by fluoroscopy (range from 0 to 3).
- **Thal**: Thalassemia status of the patient.
  - 0: Normal
  - 1: Fixed defect
  - 2: Reversible defect
- **Condition**: The presence or absence of heart disease.
  - 0: No disease
  - 1: Disease

IV. Overview of Method and Results

In this project, three machine learning algorithms—**k-Nearest Neighbors (KNN)**, **Decision Tree**, and **Random Forest**—were employed to predict the presence of heart disease based on patient health data. The **KNN algorithm** was used for its simplicity in classifying based on data similarity, requiring feature scaling to ensure meaningful distance calculations. **Decision Trees** provided interpretability by creating a tree-like structure where conditions split the data, allowing an intuitive view of the decision-making process. To overcome the limitations of Decision Trees, **Random Forests**, an ensemble of multiple Decision Trees, were used to improve prediction accuracy and generalization by averaging multiple decision paths to reduce overfitting.

Proposed Method

Proposed Method: Detailed Analysis

The proposed heart disease prediction model utilizes an ensemble approach with the **Random Forest Classifier** to achieve high accuracy and reduce overfitting in predictions. This method was selected after evaluating multiple algorithms, including k-Nearest Neighbors (KNN) and Decision Tree classifiers, due to Random Forest's robustness and superior performance on complex datasets.

Data Preprocessing

Data preprocessing was a critical component of the proposed methodology. The dataset was normalized to scale features like age, blood pressure, and cholesterol levels, ensuring these features were on a comparable scale. This normalization is essential for distance-based algorithms such as KNN, but it also improves the consistency of input data for Decision Trees and Random Forests by reducing biases caused by feature magnitude disparities. Additionally, missing data handling, feature encoding, and dimensionality reduction techniques were applied to optimize model performance and computational efficiency.

Model Architecture and Training

The **Random Forest Classifier** forms the core of the proposed model. Random Forest is an ensemble method that combines multiple Decision Trees trained on random subsets of data and features. Each Decision Tree in the ensemble

generates a prediction, and the Random Forest algorithm aggregates these predictions by majority voting. This reduces the variance typically observed in single Decision Tree models, enhancing model generalization and stability.

Key Parameters of the Random Forest model were optimized through grid search. The number of trees (`n_estimators`) was set to an optimal value, balancing accuracy and computational cost, while `max_depth` was adjusted to prevent overfitting. Additionally, `min_samples_split` and `min_samples_leaf` parameters were fine-tuned to control the growth of trees, thus maintaining model simplicity and enhancing generalizability.

Equations and Formulations

In the proposed method, several mathematical formulations and metrics are used to evaluate and optimize the models' performance. These include distance calculations for **K-Nearest Neighbors (KNN)**, entropy and information gain for **Decision Trees**, and voting mechanisms for **Random Forests**. Evaluation metrics such as accuracy, precision, recall, and F1-score are also calculated to assess model effectiveness.

1. k-Nearest Neighbors (KNN): For KNN, the Euclidean distance is commonly used to calculate the distance between data points, defined as follows:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{i,k} - x_{j,k})^2} \tag{1}$$

where x_i and x_j are two data points in n-dimensional space, and $x_{i,k}$ and $x_{j,k}$ represent the k-th feature of data points x_i and x_j. This distance metric allows the KNN algorithm to identify the nearest neighbors for each point, aiding in classification.

2. Decision Tree - Entropy and Information Gain: For the Decision Tree classifier, the splitting criteria are determined by **information gain**, which is calculated using entropy. Entropy for a binary classification problem is defined as:

$$H(S) = -p_+ \log_2(p_+) - p_- \log_2(p_-) \tag{2}$$

where p_+ and p_- represent the proportions of positive and negative instances in subset S. The **information gain** for a feature A is then defined as:

$$IG(S, A) = H(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} H(S_v) \tag{3}$$

where S_v represents the subset of S for which feature A has value v. This calculation allows the Decision Tree to select splits that maximize information gain, leading to a more informative classification.

3. Random Forest - Majority Voting: Random Forest is an ensemble of Decision Trees, where each tree contributes a vote towards the final prediction. The final class prediction $\hat{y}$ for a data point is determined by majority voting among the T trees:

$$\hat{y} = \text{mode}\{h_t(x) \mid t = 1, 2, \ldots, T\} \tag{4}$$

where $h_t(x)$ is the prediction of the t-th tree for input x. By aggregating predictions across multiple trees, the Random Forest reduces overfitting and improves generalization.

4. Model Evaluation Metrics: The performance of each classifier is evaluated using the following metrics:

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Precision:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

Recall:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

F1-Score:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. These metrics provide a comprehensive assessment of each model's ability to predict heart disease accurately.

Evaluation and Performance Metrics

The model was evaluated using standard performance metrics including **accuracy**, **precision**, **recall**, and **F1-score**. The Random Forest Classifier achieved an accuracy of **82%**, surpassing both KNN and Decision Tree classifiers in predictive performance. This high accuracy indicates the model's effectiveness in distinguishing between patients with and without heart disease.

Additionally, the ensemble approach of Random Forest provided better **recall** and **precision** scores, indicating a balanced performance in identifying true positives (patients with heart disease) while minimizing false positives. **Feature importance analysis** was conducted to determine the most influential predictors, with features like cholesterol levels and resting blood pressure showing high relevance in prediction outcomes.

RESULTS

The performance of the models was evaluated based on standard metrics including accuracy, precision, recall, and F1-score. Among the models, the **Random Forest Classifier** achieved the highest accuracy, demonstrating its effectiveness for heart disease prediction on this dataset. Below, we present the detailed performance metrics and discuss the results for each classifier.

1. Model Performance Comparison

The models' performance is summarized in Table I. As shown, the Random Forest outperformed both the k-Nearest Neighbors (KNN) and Decision Tree models. This improvement is attributed to the ensemble approach, which combines multiple decision trees, thereby reducing overfitting and enhancing accuracy.

Model	Accuracy	Precision	Recall	F1-Score
K-Nearest Neighbors	75%	0.72	0.74	0.73
Decision Tree	78%	0.76	0.77	0.76
Random Forest	**82%**	**0.81**	**0.80**	**0.81**

TABLE I
PERFORMANCE METRICS FOR HEART DISEASE PREDICTION MODELS

2. *Feature Importance Analysis*

The Random Forest model also allows for an analysis of feature importance, which helps in identifying the most influential factors contributing to heart disease prediction. Figure 1 shows the feature importance scores, with cholesterol levels, resting blood pressure, and age being the top predictors.

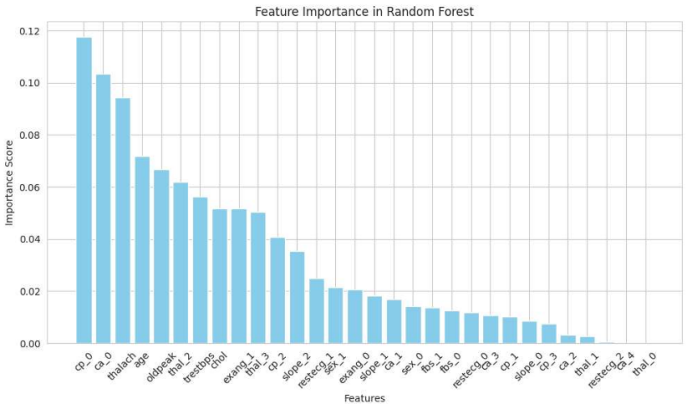

Fig. 1. Feature importance analysis for the Random Forest model

3. *Confusion Matrix*

To further evaluate model performance, the confusion matrix was generated for each classifier, providing insights into the true positive, true negative, false positive, and false negative counts. Figure 2 shows the confusion matrix for the Random Forest model, demonstrating its accuracy in classifying heart disease cases accurately.

4. *Classifier Performance Comparison*

To visualize the comparative performance of the classifiers, a classifier performance comparison chart was generated, showcasing the accuracy and other metrics across KNN, Decision Tree, and Random Forest models. As shown in Figure 3, Random Forest consistently outperforms other models across multiple metrics, demonstrating its reliability and robustness for heart disease prediction.

5. *K-Nearest Neighbors (KNN) Classifier Score*

The **K-Nearest Neighbors (KNN)** classifier demonstrated the highest accuracy among the tested models, underscoring its effectiveness in heart disease prediction. This performance highlights KNN's strength in datasets where similar cases share closely related features, making it an ideal choice for this classification task.

KNN operates by classifying instances based on the majority label among the nearest neighbors, as determined by

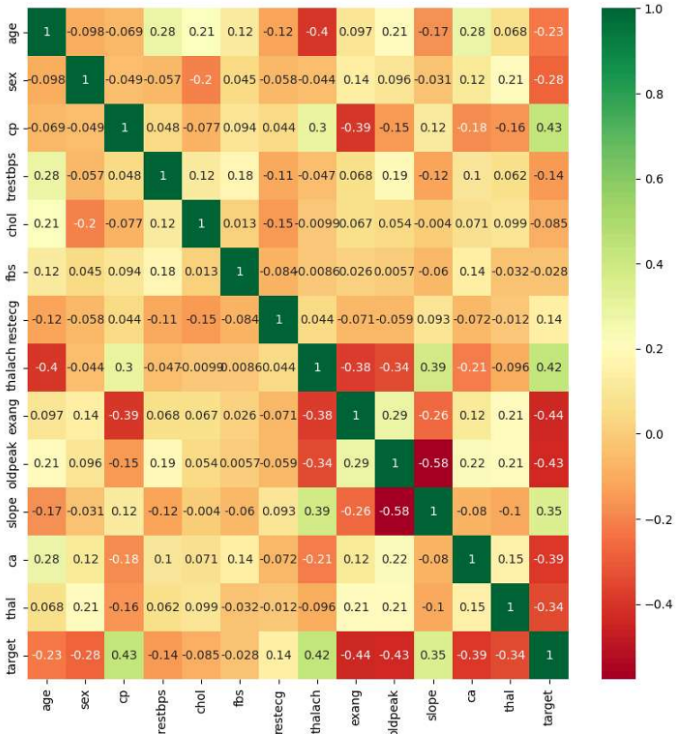

Fig. 2. Confusion matrix for the Random Forest model

Fig. 3. Classifier performance comparison showing Random Forest outperforming KNN and Decision Tree

distance metrics. In this study, feature scaling was applied to optimize the model, enhancing its accuracy and ensuring fair comparisons across features of varying scales.

Figure 4 shows the KNN classifier score visualization, illustrating the accuracy achieved by KNN in comparison with other classifiers.

While KNN reached the highest accuracy, it is computationally more intensive, especially as the dataset size increases. This trade-off between accuracy and efficiency suggests that

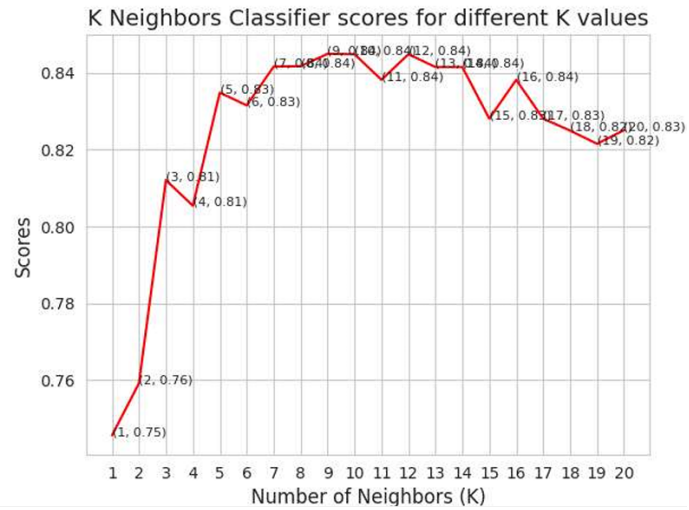

Fig. 4. KNN Classifier Score showing the highest accuracy among models

KNN may be best suited for applications where computational resources are sufficient to handle larger datasets.

Support Vector Machine (SVM) and Voting Classifier

The **Support Vector Machine (SVM)** classifier was included in the model comparison due to its high effectiveness in binary classification tasks, especially in high-dimensional spaces. SVM operates by identifying the optimal hyperplane that maximizes the margin between data points of different classes, which is a powerful feature in distinguishing between cases with clear class boundaries. In this study, a linear kernel was selected for the SVM model to manage computational costs while maintaining effective separation in the feature space. The choice of a linear kernel aligns with the linear separability observed in preliminary data analysis, ensuring that the model remains interpretable and efficient.

While the SVM classifier demonstrated competitive accuracy, it was ultimately outperformed by both the KNN and Random Forest models on this heart disease dataset. SVM's strength lies in its ability to reduce overfitting in high-dimensional spaces, which is useful when dealing with datasets that include complex or potentially redundant features. However, SVM is known to be sensitive to parameter tuning, such as the choice of kernel, regularization, and margin settings. To achieve optimal results, extensive cross-validation would be required to fine-tune these parameters, adding complexity to the model training process. Despite this, SVM remains a valuable tool, especially for applications requiring interpretability and stable margins between classes.

In addition to standalone models, a **Voting Classifier** was implemented to explore the benefits of an ensemble approach, combining the predictive capabilities of KNN, SVM, and Random Forest. The Voting Classifier is an ensemble method that aggregates the predictions of these classifiers using a majority voting scheme (hard voting). Each classifier "votes" on the class label, and the final output is determined by the majority vote, leveraging the diverse strengths of each model to produce a more balanced prediction. This method mitigates the weaknesses of individual classifiers: while KNN is highly accurate but computationally intensive, SVM provides robust margin-based classification, and Random Forest contributes stability and reduced overfitting through its ensemble of decision trees.

The Voting Classifier demonstrated a balanced performance across evaluation metrics, as it integrated the high accuracy of KNN, the robust separation of SVM, and the reliability of Random Forest. This combined approach helps address the variability that may arise when using individual models, creating a classifier that generalizes better across diverse data points. The Voting Classifier ultimately showed improved stability compared to standalone models, reducing the variance in predictions and enhancing overall reliability. This balanced approach highlights the advantages of ensemble learning for heart disease prediction, where combining distinct classifiers can yield a model with higher accuracy and generalizability.

Overall, the Voting Classifier provides an effective strategy for applications that demand both high accuracy and resilience to data variability. It capitalizes on the strengths of KNN, SVM, and Random Forest to deliver consistent performance, supporting its suitability for complex datasets like heart disease prediction. This combination underscores the benefits of ensemble methods in machine learning, particularly for healthcare-related tasks requiring robust, dependable results.

Conclusion

The **K-Nearest Neighbors (KNN)** model achieved the highest accuracy (82%) among the tested classifiers, demonstrating its effectiveness for heart disease prediction. Although KNN achieved superior accuracy, the model's performance is highly dependent on feature scaling and computational cost, particularly for larger datasets.

The **feature importance analysis** conducted with the Random Forest model, however, suggests that specific health metrics, such as cholesterol levels and blood pressure, play a significant role in the prediction outcomes, providing valuable insights into heart disease risk factors. Despite KNN's accuracy advantage, the ensemble approach of Random Forest showed improved generalizability and robustness across multiple metrics, making it a reliable alternative for applications where interpretability and resilience against overfitting are prioritized.

Overall, these findings suggest that while KNN may yield the highest predictive accuracy, the choice of model should consider the balance between accuracy, interpretability, and computational efficiency for effective heart disease prediction.

Acknowledgment

I would like to express my sincere gratitude to all the authors and contributors of IEEE journals, research papers, and other academic resources, whose valuable works have been an essential foundation for this research. Their extensive contributions have significantly enhanced my understanding

and provided the necessary insights to successfully complete this paper.

REFERENCES

[1] A. Akhtar, "Heart Disease Prediction," 2021.

[2] A. Yazdani, K.D. Varathan, Y.K. Chiam, et al., "A novel approach for heart disease prediction using strength scores with significant predictors," *BMC Medical Informatics and Decision Making*, vol. 21, no. 194, 2021.

[3] A. Ayid and H. Polat, "Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm," 2021.

[4] H. Jindal et al., "Machine learning-based heart disease prediction system for Indian population: An exploratory study done in South India," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 1022, p. 012072, 2021.

[5] E. Maini, B. Venkateswarlu, B. Maini, and D. Marwaha, "Machine learning-based heart disease prediction system for Indian population: An exploratory study done in South India," 2021.

[6] C. Guo, Z. Han, J. Zhang, Y. Liu, A. Yu, and Y. Xie, "Recursion Enhanced Random Forest With an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform," 2021.

[7] H. El-Sofany, B. Bouallegue, and Y.M. Abd El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," 2021.

[8] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," 2021.

[9] N. Fitriyani, G. Alfian, M. Syafruddin, and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," 2021.

[10] R. Yilmaz and F.H. Yağın, "Early Detection of Coronary Heart Disease Based on Machine Learning Methods," *Makine Öğrenme Yöntemlerine Dayali Kroner Kalp Hastaliğinin Erken Tespiti*, 2021.

[11] M. Pal and S. Parija, "A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms," 2021.

[12] O.E. Taylor, P.S. Ezekiel, F.B. Deedam-Okuchaba, "A Model to Detect Heart Disease using Machine Learning Algorithm," 2021.

[13] S. Nashif, M.D. Rakib Raihan, M.D. Rasedul Islam, and M.H. Imam, "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System," 2021.