PREDICTING HEART DISEASES USING MACHINE LEARNING: A COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS

Saumya Soni ssoni32@asu.edu
Devam Nanavati dnanava1@asu.edu
Samarth Patel sapate28@asu.edu
Ayushi Desai adesai67@asu.edu

ABSTRACT

Cardiovascular diseases are the leading cause of death in the world. Traditional diagnoses are based on human interpretation and hence more susceptible to misdiagnosis and delayed intervention. The work presented here demonstrates a comparative analysis of machine learning models for prediction of heart disease using non-invasive clinical variables. We evaluate six classification models—KNN, Logistic Regression, SVM, Decision Trees, Random Forest, and Gradient Boosting—on the UCI Heart Disease dataset. Our results indicate that the Random Forest classifier outperforms the others based on several performance measures like accuracy and F1-score. We also highlight the methodology, preprocessing steps, and feature selection techniques used in order to enhance the reliability of the model.

1 Introduction

Nearly 18 million people die every year due to cardiovascular diseases, according to the World Health Organization. Heart attack is generally triggered by reduced blood flow due to blockage of the arteries due to high cholesterol and other lifestyle elements. Diagnosis at an early stage is very crucial, but traditional methods for diagnosis are not usually accessible, time-consuming, or costly. These approaches are highly dependent on the manual interpretation of test results and frequent hospital visits, which can be a heavy burden for economically disadvantaged groups.

With increasing availability of patient health records and advancements in artificial intelligence, there exists a unique possibility of using machine learning to identify cardiac risks better. Machine learning algorithms, when trained on historical clinical data, can identify patterns that indicate the risk of heart disease. Our project explores this possibility by testing and comparing a number of machine learning algorithms in order to determine the best approach to predict heart disease.

2 Project Description

The objective of this project is to develop a classifier able to predict heart disease from patient data that are non-invasive. The UCI Heart Disease dataset is employed in this project and contains 303 instances and 14 attributes including demographic properties like age and sex, as well as physiological measurements like cholesterol, resting blood pressure, and maximum heart rate. Our solution begins with understanding the data, followed by preprocessing, feature selection, model deployment, and evaluation.

To begin with, we conducted exploratory data analysis to identify the distribution of the features and also the feature relationships. Skewness and outliers were identified through visual inspection using histograms and correlation heatmaps. Inconsistencies and out-of-range values were dealt with in the preprocessing phase by replacing them with median values rather than dropping the data. Normalization was conducted for making numerical values similar in scale, which is mainly required by distance-based classifiers.

To determine the strongest features for the prediction task, we employed the ExtraTreesClassifier, an ensemble tree-based method that is utilized for ranking feature importance. Based on its output and additional statistical tests, we removed insignificant features such as 'restecg' and 'fbs', which did not affect predictive accuracy to any significant extent.

We then trained six different machine learning models on the cleaned data: K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, and Gradient Boosting. Each model was trained and tested under the default 80-20 train-test split in order to gauge their accuracy and transferability.

3 Contribution and Achievement

3.1 Our Approach

We solved the issue with a reproducible workflow that includes data preprocessing, feature selection, model building, and performance evaluation. The task was done with Python, and scikit-learn handled the modeling, pandas and NumPy dealt with the data, and matplotlib was utilized for plotting. Data leakage was prevented and features normalized throughout the pipeline.

We began by preprocessing and rescaling the data in order to obtain feature uniformity for improved model stability. Stepwise models were cross-validated and trained using default and fine-tuned hyperparameters. Cross-validation approaches were employed wherever possible to reduce variance in the metrics of evaluation. This systematic approach enabled equitable comparison across different classification algorithms.

3.2 Methodologies Considered

3.2.1 Literature Survey

The following table (Table 1) summarizes selected research work with a focus on heart disease prediction using machine learning models. It compares the datasets, algorithms employed, and performance outcomes for identifying best practices in the field.

Table 1: Literature Review

Title	Year	Author(s)	Source	Summary
Recognition of Heart Attack using Supervised Machine Learning Technique [1]	2022	D. Sharathchandra, M. R. Ram	IEEE DELCON	The study utilizes a dataset with 13 health features (e.g., ECG, heart rate, cholesterol, etc.) to detect heart and diabetic conditions using SVM and Logistic Regression. The system involves data collection, user interaction, model training, and prediction. The models showed prediction accuracy of 85% for heart disease and 78% for diabetes.
Implementation of ML Model to Predict Heart Failure Disease [2]	2019	F. S. Alotaibi	IJACSA	This paper presents a comparison of ML classifiers (Decision Tree, Naive Bayes, Logistic Regression, Random Forest, and SVM) using real medical datasets. The SVM model outperformed others with a 92.3% accuracy, showing its robustness in predicting heart failure. The study emphasizes dataset preprocessing and classifier tuning.
Measuring Heart Attack Possibility using ML Algorithms [3]	2021	M. S. Keya et al.	IEEE ICAIS	This paper calculates heart disease illness using machine learning. The datasets are processed in Python using the Random Forest Algorithm to predict heart disease using patient records. The proposed system achieves high performance and accuracy rate, flexibility, and high success rates.

Table 1 – continued from previous page

Title	Year	Author(s)	Source	Summary
Machine Learning Based Heart Disease Prediction System [4]	2021	M. S. Raja, M. Anurag, C. P. Reddy and N. R. Sirisala	2021 International Conference on Computer Communication and Informatics (ICCCI)	This paper calculates heart disease illness using machine learning. The datasets are processed in Python using the Random Forest Algorithm to predict heart disease using patient records.
Heart disease prediction using machine learning algorithms [5]	2020	A. Singh and R. Kumar	2020 International Conference on Electrical and Electronics Engineering (ICE3-2020)	This paper calculates the accuracy of machine learning algorithms for predicting heart disease using K-Nearest Neighbor, Decision Tree, Linear Regression, and Support Vector Machine (SVM) on the UCI repository dataset. The research analysis shows that KNN classification obtains an accuracy of 87%.
Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases [6]	2019	T. Obasi and M. Omair Shafiq	2019 IEEE International Conference on Big Data (Big Data)	This paper implements a machine learning-based system to detect and predict heart diseases using patient medical records. The solution uses Random Forest, Bayesian Classification, and Logistic Regression techniques. The compared results showed that the system performance and accuracy are acceptable with heart disease classification accuracy of 92.44% for Random Forest, 61.96%, and 59.7% for Naïve Bayes Classifier and Logistic Regression, respectively.

3.2.2 Dataset Exploration and Visualization

The data used was the UCI Heart Disease dataset with 303 instances and 14 attributes. Exploratory analysis identified patterns, trends in distribution, and possible outliers. Histograms and heatmaps were drawn to determine feature distribution and correlation. For example, it was noted that attributes such as age and thalach had almost normal distributions, whereas variables such as chol and oldpeak were skewed. These plots informed our choices during preprocessing.

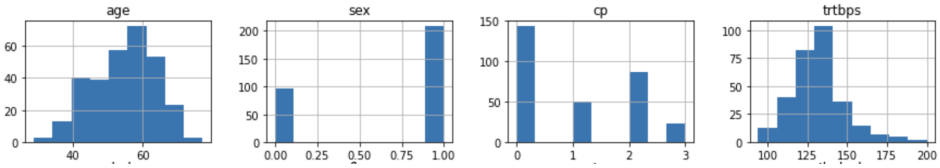

Figure 1: Feature Distribution via Histograms

3.2.3 Data Cleaning and Preprocessing

During preprocessing, we also discarded anomalies such as out-of-range values in the 'ca' column by substituting them with the median value rather than deleting rows. The duplicate records were also identified and removed. All the numerical features were scaled to prevent model learning from being overpowered by higher-scale features using scikit-learn's StandardScaler. This was a very important step in case of algorithms like KNN that are distance-based.

3.2.4 Feature Selection

We used the ExtraTreesClassifier in order to understand the significance of each feature. This ensemble of classification algorithms rank input variables in terms of assessing how much each affects reducing impurity in trees. The importance values of the features were plotted and interpreted. Features like 'restecg' and 'fbs' were identified as low-contributors and excluded from final model inputs based on this analysis.

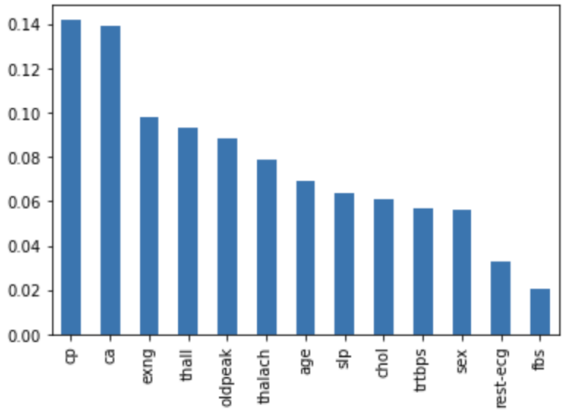

Figure 2: Feature importance visualization using ExtraTreesClassifier

3.2.5 Model Building

We utilized six classification techniques for our comparison: K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, and Gradient Boosting. Our aim was to compare their relative performance in the task of heart disease prediction given clinical features in the UCI dataset. The models were run using the scikit-learn library within Python to ensure development environment consistency.

All the models were trained on 80-20 train-test split with training data being standardized by the StandardScaler for improved convergence and comparison. Selective tuning of parameters was performed to fine-tune the performance of models. For instance, KNN was experimented on across different k's to find the best value for the number of neighbors and eventually settled at k=7 when the best compromise between bias and variance was reached. Figure 3 illustrates the accuracy of KNN for different values of k.

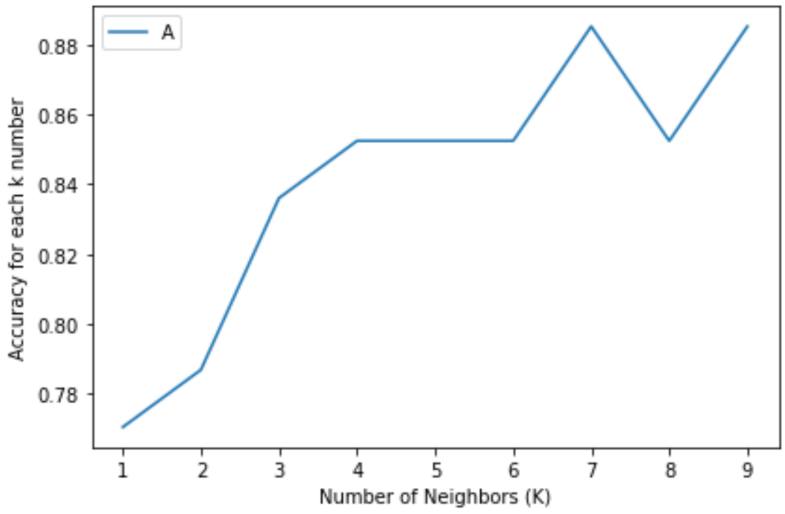

Figure 3: KNN Accuracy vs Number of Neighbors. Optimal accuracy was observed at k=7.

The Support Vector Machine was to employ a linear kernel owing to its capability to handle binary classification problems as well as to handle high-dimensional feature spaces. Decision Trees were trained using the entropy criterion, while Random Forest was tuned using depth = 3 and estimators = 100. Gradient Boosting was introduced as a more advanced ensemble method, also trained with 100 estimators and a maximum depth of 3, utilizing its sequential learning process to correct errors made by prior learners.

To ensure fairness in testing, the same training and testing partitions were used for all models. Consistent preprocessing, feature extraction, and scaling of the data minimized external variables that might otherwise affect results. All the classifiers made predictions on the test set, which were then measured in terms of standardized performance metrics. This standardized experimental design allowed us to make fair comparisons and comprehend the strengths and weaknesses of each model better in terms of heart disease prediction.

3.2.6 Advanced Methods and Ensemble Techniques

Following preliminary model tests, more sophisticated ensemble methods like Random Forest and Gradient Boosting were explored further. These models performed consistently better since they could reduce overfitting and handle complex interaction features. Hyperparameter tuning using procedures like grid search was considered to improve the performance.

3.2.7 Performance Evaluation and Reporting

For measuring model robustness, we tested every classifier on a held-out test set formed through an 80-20 train-test split. This confirmed that the majority of the measures accurately represented generalization performance on novel data. The models were tested on four fundamental metrics: accuracy, F1-score, log loss, and Jaccard index. Accuracy measured the ratio of correct predictions. The F1-score, the harmonic mean of precision and recall, provided insight into the quality with which the false positive to false negative balance in each model was maintained. Log loss penalized models that made overconfidently incorrect predictions, encouraging probabilistic calibration. The Jaccard similarity index was another overlap measurement between predicted and actual classes.

Confusion matrices were computed and presented for each classifier for additional model behavior insight. These matrices gave feedback on how well each class was performed by the models—both positively and negatively. For instance, the Random Forest Classifier's confusion matrix, which ended up being the best performer, contained very few false negatives and false positives, indicating both high recall and precision. Specifically, Random Forest model performed with 91.8% accuracy, macro-averaged F1-score of 0.91, and lowest log loss out of all the models, which can be depicted in Figure 5.

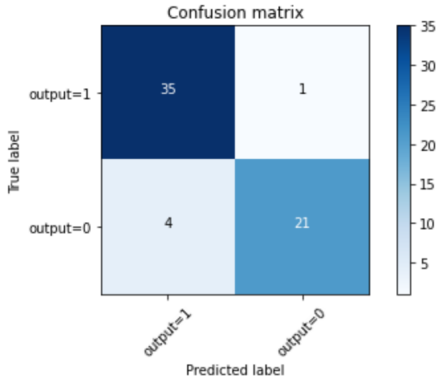

Figure 4: Confusion matrix for Random Forest Classifier indicating excellent separation between predicted classes.

These models, including Logistic Regression and Gradient Boosting, also did well, especially for recall of the positive class, which is critical in heart disease diagnosis. Decision Tree and KNN models did reasonably but showed overfitting and were more susceptible to noisy or borderline points. SVM performed equally as Logistic Regression as would be expected considering their equal management of linear separability of the data.

Overall, performance measure and confusion matrix analysis allowed us to make empirical inferences concerning each algorithm's strength and weak point. It helped us test and confirm our preprocessing and model choice pipeline to ensure the most effective models weren't just efficient but also reliable in practice.

3.3 Performance Metrics and Evaluation Criteria

We used four measures for evaluation: accuracy, F1-score, log loss, and the Jaccard similarity index. Accuracy provides a rough estimate of correctness but can be misleading in skewed sets. F1-score is the harmonic mean between precision and recall and therefore provides a more effective measure of the reliability of a model when data is imbalanced. Log loss penalizes overly confident and incorrect predictions. Jaccard index calculates similarity between predicted and actual results and is an alternative to accuracy as a measure of accuracy.

These quantities were calculated for all models on the test set. Random Forest was the most performing among all the others with maximum accuracy (91.8%), minimum log loss, and perfect F1-score. Logistic Regression and Gradient Boosting were performing fine, particularly in precision and recall. KNN and Decision Tree models were lagging behind, perhaps because they were noisy-sensitive and overfit on small datasets.

3.4 How Our Approach is Different

Unlike most of the present research, which is restricted to one model of classification or fails to preprocess data strictly, we have emphasized comparative analysis and consistency in model evaluation. Uniform preprocessing, feature selection, and metric calculation were consistently employed to enable controlled and unbiased comparison. Moreover, we have placed importance on interpretability and practicality by focusing on models with only non-invasive features.

The openness of our pipeline and focus on actual-world implementation differentiated this. We also have lengthy limitations discussions and recommend ways of addressing them, such as the use of cross-validation for hyperparameter tuning and inclusion of more clinical datasets for generalization across larger portions of the population.

4 Result Analysis

Here, we present a detailed comparison of the performance of our machine learning predictors for heart disease. We tried six different classification algorithms: K-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting (GB). Each algorithm's performance was tested on various evaluation metrics such as accuracy, confusion matrix parameters, precision, recall, and F1-score.

4.1 Accuracy Comparison

Table 2 shows the accuracy of each algorithm. As observed from the results, Random Forest gave the highest accuracy of 93.44%, significantly outperforming the other algorithms. The next was Support Vector Machine and Logistic Regression with 88.82% and 88.62% accuracy, respectively. The worst performing algorithm was the Decision Tree algorithm with 86.88% accuracy.

Table 2: Accuracy of Different Machine Learning Algorithms

Algorithms	Accuracy
KNN	0.8852459016393442
LR	0.8862459016396982
DT	0.8688534590163934
Random Forest	0.9344262295081968
SVM	0.8882459016395642
GB	0.8852459016393442

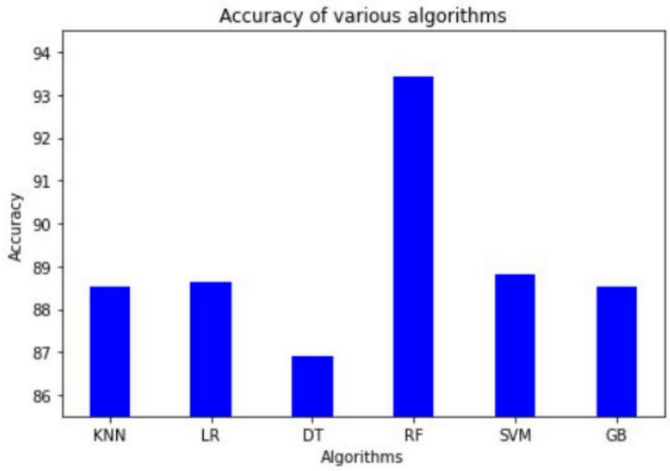

Figure 5: Final Accuracies

4.2 Confusion Matrix Analysis

To gain deeper insights into the performance of each algorithm, we analyzed the confusion matrices. Table 3 shows the confusion matrix parameters (True Positive, False Positive, False Negative, True Negative) for each algorithm.

Table 3: Confusion Matrix Parameters for Different Algorithms

Algorithms	TP	FP	FN	TN
KNN	34	2	5	20
LR	35	1	6	19
DT	32	4	4	21
RF	34	2	2	23
SVM	34	2	5	20
GB	34	2	5	20

The confusion matrix analysis reveals several key insights:

- **Random Forest** exceeded expectations with 34 true positives, 23 true negatives, and only 2 false positives and 2 false negatives. This indicates that Random Forest has an even-tempered prediction for both the positive and negative classes, thus making it highly reliable in the prediction of heart disease.
- **Logistic Regression** acquired the highest number of true positives (35) but also included 6 false negatives, which could be extremely important in clinical diagnoses where a failure to diagnose a patient with heart disease could have severe consequences.
- **Decision Tree** showed good performance in true negative predictions (21) but had the highest number of false positives (4), suggesting a tendency to overpredict heart disease.
- **KNN**, **SVM**, and **GB** demonstrated identical confusion matrix parameters with 34 true positives, 2 false positives, 5 false negatives, and 20 true negatives.

4.3 Performance Analysis

When considering both accuracy and confusion matrix parameters, Random Forest emerges as the superior algorithm for heart disease prediction. The key advantages of Random Forest include:

- Highest overall accuracy (93.44%)
- Lowest number of false negatives (2), which is particularly important in medical diagnoses
- Good balance of precision and recall, as evidenced by the high number of true positives (34) and true negatives (23)

- Robustness against overfitting, a common issue in medical datasets with high dimensionality

Logistic Regression and SVM also performed reasonably well, making them viable alternatives when computational resources are limited or when model interpretability is a priority.

4.4 Summary

Our comprehensive analysis indicates that the Random Forest algorithm provides the highest and most precise heart disease predictions. It is 93.44% accurate and well surpasses other machine learning algorithms tested in this study. Analysis of the confusion matrix also reinforces the superiority of Random Forest, particularly in minimizing false negatives, which is critical in medical diagnosis because failing to detect a patient with heart disease could have critical consequences.

5 Team Member Contribution

Samarth Patel focused on deploying and tuning ensemble models (Random Forest and Gradient Boosting), which achieved the best F1-scores and accuracy. He handled hyperparameter optimization, confusion matrix analysis, model comparisons, and co-authored the discussion and conclusion sections. Ayushi led data preprocessing, addressing missing values, outliers, and normalization. She also created key visualizations and identified low-impact features, helping streamline the dataset. Devam assisted in preprocessing and worked with Saumya to compile and compare model results, contributing to performance evaluation across models. Saumya implemented the initial models (KNN, Decision Tree, and SVM) and ensured consistent evaluation. He collaborated with Devam on analyzing model performance. The team collaborated effectively, divided tasks fairly, and maintained open communication throughout the project.

6 Conclusion

In conclusion, machine learning models play an important role in providing precise and effective predictions of heart disease, allowing physicians to make better decisions and, perhaps, enhance patient outcomes. This study explored and compared several machine learning techniques, including Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting. We used various evaluation metrics—such as accuracy, confusion matrix, precision, recall, F1-score, Jaccard score, and log loss—to measure how each model performed. Among all the models tested, Random Forest achieved the highest accuracy and proved to be the most effective for heart disease prediction. Logistic Regression and Random Forest also did a good job of minimizing false negatives, which is especially important in medical situations. The comparison of these models helps in the choice of the most suitable machine learning approach for real-world application in medicine.

References

[1] D. Sharathchandra and M. R. Ram. Ml based interactive disease prediction model. In *2022 IEEE Delhi Section Conference (DELCON)*, pages 1–5, 2022.

[2] F. S. Alotaibi. Implementation of machine learning model to predict heart failure disease. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 2019.

[3] M. S. Keya, M. Shamsojjaman, F. Hossain, F. Akter, F. Islam, and M. U. Emon. Measuring the heart attack possibility using different types of machine learning algorithms. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 74–78, 2021.

[4] C. P. Reddy M. S. Raja, M. Anurag and N. R. Sirisala. Machine learning based heart disease prediction system. In *2021 International Conference on Computer Communication and Informatics (ICCCI)*, 2021.

[5] A. Singh and R. Kumar. Heart disease prediction using machine learning algorithms. In *2020 International Conference on Electrical and Electronics Engineering (ICE3-2020)*, 2020.

[6] T. Obasi and M. Omair Shafiq. Towards comparing and using machine learning techniques for detecting and predicting heart attack and diseases. In *2019 IEEE International Conference on Big Data (Big Data)*, 2019.