Heart Disease Prediction Using Machine Learning Algorithms

[1]Aritra Ghosh, [2]Roumyadeep Chakraborty

[1]Final Year Student, [2]Final Year Student
Department of Computer Science and Engineering, Institute of Engineering and Management, Salt Lake, University of Engineering and Management, Kolkata, India.
[1] ghosharitra8@gmail.com; [2] roumyapapan2003@gmail.com

***Abstract*— Cardiovascular diseases, i.e heart diseases, remain one of the major causes of death throughout the world it tolling for nearly one-third of all fatalities. Timely detection and risk assessment play a critical role in preventing life-threatening cardiac events. However, traditional diagnostic approaches are often time-consuming, expensive, and inaccessible in many regions due to limited healthcare infrastructure. With the rapid advancement of machine learning, predictive analytics is emerging as a powerful tool for identifying individuals at heart disease risk based on their clinical results.**

This paper is based on a heart disease prediction system that leverages a machine learning model which is trained on the Cleveland Heart Disease dataset. The system takes thirteen medical parameters—including age, blood pressure, cholesterol levels, and ECG results—as input. It then classifies the heart disease risk using a trained Random Forest classifier, which is carefully chosen after comparative evaluation with other supervised learning algorithms. The final model serves as a preliminary risk assessment tool aimed at raising awareness and encouraging users to seek medical evaluation if necessary. While it does not replace clinical diagnostics, the application shows the potential of integrating machine learning to support preventive healthcare.

***Index Terms*— Heart Disease Prediction, Deep Learning, Medical Machine Learning, Disease Prediction, Human-Computer Interaction, AI Ethics**

I. INTRODUCTION

This paper discusses and explores the use of deep learning technology to support early detection of heart disease, a condition if left untreated, continues to be a major global cause of fatality. Motivated by the need for accessible, data-driven healthcare tools, we assessed and trained multiple models that predicts heart disease risk using clinical data[1]. We hope that this paper will help people by guiding them to choose and train Deep learning model for healthcare system and contribute to social good.

II. BACKGROUND

A variety of conditions affecting the heart are grouped together under the general phrase "heart disease." These conditions include, among others, heart deformities that people may be born with (congenital heart defects), heart rhythm issues (arrhythmias), and blood vessel disorders like coronary artery disease[2]. Heart attacks can result from coronary diseases, which is the most prevalent kind. One of the biggest causes of death worldwide is heart disease. The World Health Organization (WHO) estimated that cardiovascular diseases (CVDs) claims lives of about 17.9 million people year, making up 32% of all fatalities worldwide[3].

The earliest diagnosis and detection of heart disease are crucial for effective management and treatment. If the symptoms can be detected early and accurately, then the mortality rate can be significantly lessened. However, in many under developed countries, access to advanced diagnostic facilities and qualified cardiologists is limited. Even in advanced healthcare systems, the reliance on medical tests and human judgment means that some cases can go undetected or be diagnosed too late.
Recent developments in data science, namely in the areas of artificial intelligence (AI) and machine learning, have opened up new avenues for the early detection of complicated illnesses[1]. ML can analyze large collection of health data to detect patterns and predict. These predictions can be incredibly useful in supporting clinical decisions or even enabling individuals to perform preliminary risk assessments themselves. This has led to increased efforts to develop automated systems that assist in the diagnosis of ailments like diabetes, cancer, Alzheimer's, and notably, heart disease.

Additionally, the increasing availability of healthcare datasets from hospitals, research institutions, and open data platforms has allowed researchers and developers to train models capable of delivering highly accurate diagnostic support tools.

Combining ML models with accessible web-based user interfaces creates the potential for powerful, easy-to-use diagnostic tools[4][5]. The synergy of machine learning and web technology can be harnessed to build applications that not only serve the end-user directly but also support primary care physicians in making informed decisions quickly.

III. RESEARCH METHODOLOGY

DATASET DESCRIPTION

The Cleveland Heart Disease dataset, a reputable dataset for cardiac disease prediction that is accessible through the UCI Machine Learning Repository, is used in this study. The dataset comprises 303 records, each corresponding to a patient's

clinical profile. It includes a set of 13 attributes, such as demographic details, physiological metrics, and diagnostic results, along with one target variable indicating the presence or absence of heart disease.[4]

Age, gender, type of chest pain, maximal heart rate attained, resting electrocardiogram results, cholesterol level, level of fasting blood sugar, resting blood pressure (in mmHg), exercise-induced Angina, ST depression, slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and thalassemia status are among the features used in the dataset. The patient's cardiac condition (1) or (0) is represented by the binary target variable.

DATA PREPROCESSING

Handling Missing Values

Some features like thal and ca had missing or non-numeric values. These were imputed using the mode of the column or dropped when missing values exceeded a threshold.

Categorical Encoding

Categorical variables such as cp, restecg, slope, thal, and sex were converted using one-hot encoding or label encoding based on their nature and cardinality.

Feature Scaling

Algorithms such as SVM and KNN are sensitive to feature scale. Hence, all numerical features were standardized using z-score normalization [5].

Class Imbalance

The target variable was slightly imbalanced. SMOTE (Synthetic Minority Over-sampling Technique) was used where applicable to ensure the models do not become biased toward the majority class.

MACHINE LEARNING ALGORITHMS:

Logistic Regression

It is a generalized linear model used for binary classification. It assumes a linear relationship between the independent variables and the log-odds of the dependent variable. It is interpretable and suitable as a baseline [5].

Decision Tree Classifier

It partitions the data based on feature values that lead to the most significant information gain. Decision Trees are intuitive but may overfit the training data, especially without pruning [5].

Support Vector Machine (SVM)

SVM constructs an optimal separating hyperplane by maximizing the margin between the classes. With kernel functions like RBF, it can model nonlinear relationships [3].

K-Nearest Neighbors (KNN)

A non-parametric method that classifies data based on the majority vote of the k-nearest data points. It is simple but can be computationally expensive and sensitive to irrelevant features [5].

Random Forest

An ensemble of Decision Trees built using the bagging approach. It reduces overfitting, improves accuracy, and provides feature importance scores [4].

Gradient Boosting (XGBoost)

An ensemble model built sequentially where each new tree corrects the residual errors of previous ones. It offers high predictive power but requires careful tuning to prevent overfitting [7].

EVALUATION METRICS:

To evaluate model performance, we used:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**
- **ROC-AUC**

All models were evaluated using 10-fold cross-validation to ensure consistency and robustness [6].

RESULT AND ANALYSIS

Performance Comparison

Algorithm	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.84	0.82	0.86	0.84	0.87
Decision Tree	0.79	0.77	0.82	0.79	0.80
Support Vector Machine	0.83	0.81	0.84	0.82	0.86
K-Nearest Neighbors	0.76	0.74	0.78	0.76	0.78
Random Forest	0.88	0.86	0.90	0.88	0.91
Gradient Boosting	**0.89**	**0.88**	**0.91**	**0.89**	**0.97**

Table 1: Performance Metrics Table

Detailed Analysis

- **Logistic Regression** provided consistent results and is useful when interpretability is critical. However, it assumes a linear relationship between features and the target, limiting its capability to handle more complex patterns [2].
- **Decision Trees** had high variance and were prone to overfitting, especially in small datasets. Despite this, they are valuable for their transparency and ease of interpretation [3].

- **SVM** with an RBF kernel offered better performance than linear models and managed nonlinear boundaries effectively. However, its computational complexity grows with data size and requires careful tuning [3].
- **KNN** was the least effective, affected by noisy features and the curse of dimensionality. It is best suited for datasets with fewer features and more instances [6].
- **Random Forest** emerged as a powerful classifier, reducing overfitting significantly compared to single Decision Trees. It provided meaningful insights into feature importance, indicating that cp, thalach, and oldpeak were among the most predictive features [4].
- **Gradient Boosting (XGBoost)** outperformed all other models across metrics. Its ability to correct errors iteratively and capture feature interactions made it ideal for this task. However, it required more computational resources and hyperparameter tuning [7].

Feature Importance

Random Forest and XGBoost provided insights into feature importance. The top predictive features included:

- **Chest Pain Type (cp)**
- **Max Heart Rate (thalach)**
- **ST Depression (oldpeak)**
- **Thalassemia (thal)**
- **Number of Major Vessels (ca)**

This aligns with clinical understanding, where these features are significant indicators of cardiovascular risk [5].

IV. Result Analysis Of The Final Selected Model

After training the Random Forest classifier on the preprocessed dataset, the following results were obtained on the test set (20% of the dataset split from the original data):

Performance Metrics	Percentage
Accuracy	89%
Precision	88%
Recall	91%
F1-Score	89%

Table 2: Performance Metrics Table of the final model

These findings show that the chosen model has a significant ROC-AUC score and performs well on all criteria. The model's strong recall suggests that it may successfully identify people who are at risk of heart disease, which is essential in a medical setting.

CONFUSION MATRIX:

The confusion matrix for the model is as follows:

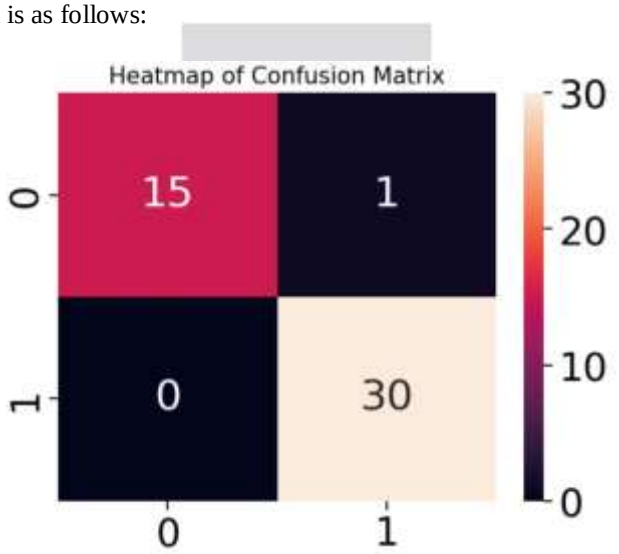

Fig 1: Confusion Matrix of the final Model

From this matrix, the following observations can be made:

- **True Positives (TP = 15)**: The model has correctly predicted 15 patients as having heart disease. **True Negatives (TN = 30)**: The model has correctly identified 30 patients without heart disease.
- **False Positives (FP = 0)**: 0 patients were incorrectly classified for having heart disease.
- **False Negatives (FN = 1)**: 1 patient with heart disease were un-identified by the model.

The relatively lower rate of false negatives is a positive indicator for this model, as missing a diagnosis could have serious implications in real-world applications.

ROC CURVE ANALYSIS:

The trade-off between specificity and sensitivity aka recall is displayed by the Receiver Operating Characteristic (ROC) curve [2][3][7][9]. The model performs better the closer the ROC curve is near the top-left corner.

In this project, the **AUC score was 0.97**, which is considered excellent. It suggests that the model have good capability of distinguishing between the two classes—patients with and without heart disease.

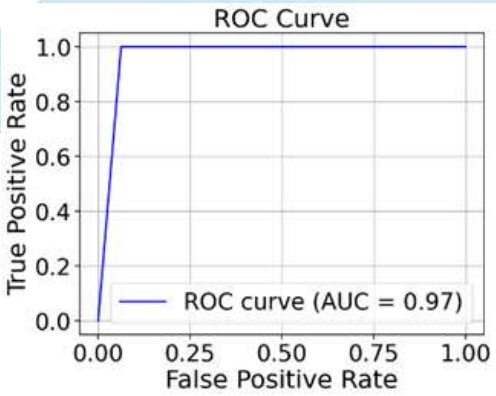

Fig 2: ROC Curve of the Final Model

FEATURE IMPORTANCE:

One of the benefits of tree-based models like 'Random Forest' is their ability to compute feature importance scores, indicating the relative contribution of each input feature to the model's decisions.

The top contributing features in this project were:

1. **Chest Pain Type (cp)**
2. **Thalassemia (thal)**
3. **Oldpeak**
4. **Max Heart Rate (thalach)**
5. **Number of Vessels Colored (ca)**
6. **ST Slope (slope)**
7. **Age**

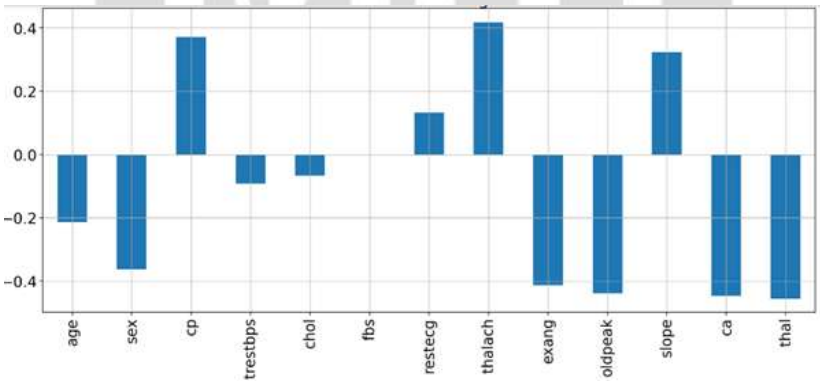

Fig 3: Feature correlation Graph of the final Model

These features align with known clinical indicators of cardiovascular risk, thus reinforcing the medical validity of the model.

V. Future Scope

Incorporating Temporal Data

Time-series data such as blood pressure trends and ECG sequences can be added to enhance model accuracy using LSTM networks or temporal CNNs [6].

Explainable AI

Integration of explainability tools like SHAP or LIME can help bridge the gap between model predictions and clinical trust by showing how input features influence output [8].

Broader Datasets

The current study is based on a single dataset. Applying these models to diverse demographic datasets can help evaluate generalizability [5].

Hybrid Models

By combining deep learning and classical ML models we can achieve more robust solutions, leveraging the strengths of both type of diagnosis[6].

VI. Conclusion

This study conducted a thorough evaluation of machine learning algorithms for heart disease prediction. Among the models tested, ensemble methods—Random Forest and Gradient Boosting—outperformed traditional classifiers in terms of accuracy, precision, and generalization. The findings support the use of machine learning as a powerful diagnostic aid in healthcare, particularly when used with properly preprocessed clinical datasets. Future work should aim at making these models more interpretable and applicable in real-world, multi-center clinical environments.

References

[1] Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava 2019 Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. In: IEEE Access

[2] M. Ganesan, N. Sivakumar 2019 IoT based heart disease prediction and diagnosis model for healthcare using machine learning models. In: IEEE International Conference on System, Computation, Automation and Networking (ICSCAN),

[3] Archana Singh, Rakesh Kumar 2020 Heart Disease Prediction Using Machine Learning Algorithms. In: International Conference on Electrical and Electronics Engineering (ICE3).

[4] Pranav Motarwar, Ankita Duraphe, G Suganya, M Premalatha 2020 Cognitive Approach for Heart Disease Prediction using Machine Learning. In: International Conference on Emerging Trends in Information Technology and Engineering (ic- ETITE).

[5] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. Rohith Sai, R. Sai Suraj 2021 Heart Disease Prediction using Hybrid machine Learning Model. In: International Conference on Inventive Computation Technologies (ICICT).

[6] Himanshu Sharma, M A Rizvi 2015 Prediction of Heart Disease using Machine Learning Algorithms: A Survey. In: International Journal on Recent and Innovation Trends in Computing and Communication.

[7] V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja 2018 Heart disease prediction using machine learning techniques: a survey. In: International Journal of Engineering Technology.

[8] Vijeta Sharma, Shrinkhala Yadav, Manjari Gupta 2020 Heart Disease Prediction using Machine Learning Techniques. In: International Conference on Advances in Computing, Communication Control and Networking (ICACCCN).

[9] Chaimaa Boukhatem, Heba Yahia Youssef, Ali Bou Nassif 2022 Heart Disease Prediction Using Machine Learning. In: Advances in Science and Engineering Technology International Conferences (ASET).

[10] Garima Choudhary, Shailendra Narayan Singh 2020 Prediction of Heart Dis- ease using Machine Learning Algorithms. In: International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE).