HEART DISEASE PREDICTION USING ML

Sakshi Tekale[1], Sakshi Lande[2], Prof. Nitin Ganeshar[3]

[1,2]MCA Student, International Institute Of Management Science (IIMS), Chinchwad, Pune, Maharashtra, India.

[3]Assistant Professor, International Institute Of Management Science (IIMS), Chinchwad, Pune, Maharashtra, India.

DOI: https://www.doi.org/10.58257/IJPREMS44436

ABSTRACT

Being able to predict heart disease early and accurately is a powerful way to save lives and reduce illness, as it allows doctors to step in sooner. In this research, we explore different machine learning (ML) techniques to see if they can predict heart disease using common patient information that hospitals already collect. We test a variety of models, from classic methods like Logistic Regression and Random Forest to newer ones like XGBoost and neural networks. We train and test these models on a well-known public dataset, measuring how well they perform using metrics like accuracy, precision, and recall. We also use special tools (SHAP and LIME) to figure out *why* a model makes a certain prediction. This helps us see which health factors are most important and makes the models easier for doctors to trust. Our findings suggest that powerful models like Random Forest perform very well.

Keywords: Heart Disease Prediction, Machine Learning, Deep Learning, Explainability, Clinical Risk Prediction, UCI Dataset.

1. INTRODUCTION

Cardiovascular disease (CVD) is a leading cause of death all over the world. The ability to identify people who are at high risk *before* they show symptoms is a major goal of modern medicine. Early detection allows for timely interventions, like changes in diet and lifestyle or starting medication, which can prevent serious complications and save lives.

For years, doctors have used traditional risk scores to estimate a patient's risk. These scores are simple and easy to understand, but they usually only look at a few factors (like age, smoking, and cholesterol) and can miss complex, hidden patterns in a patient's health profile.

This is where machine learning (ML) comes in. ML models have the potential to be much more accurate because they can analyze dozens of factors at once and find subtle, non-linear relationships that a human or a simple checklist would miss. This can lead to better diagnoses, personalized treatment plans [1], and more efficient use of hospital resources [2].

However, using ML in medicine is not easy and presents several big challenges.

1. Messy & Imbalanced Data: Real-world medical data is often messy, with missing values, errors, and imbalances (meaning, datasets often have many more healthy patients than sick ones).
2. The "Black Box" Problem: Many powerful ML models (like deep learning) are "black boxes". They can give a highly accurate prediction but can't explain *why* they made it. For a doctor to trust a model with a patient's life, they need to understand its reasoning.
3. Generalizability: A model trained on data from one hospital or country might not work well for patients in another.

The goal of this paper is to tackle these challenges [3]. We aim to compare several different ML models for heart disease prediction using the popular UCI Heart Disease dataset. We won't just look at accuracy; our main goal is to use explainability tools (like SHAP/LIME) to open up the "black box," identify the most critical risk factors, and build a model that is not only accurate but also trustworthy.

2. METHODOLOGY

This section outlines the systematic approach followed to develop and evaluate machine learning models for predicting the likelihood of heart disease. The process included dataset preparation, preprocessing, model development, and performance assessment.

2.1 Dataset Description

The study employed the UCI Cleveland Heart Disease dataset, which consists of 303 patient records and 14 medical attributes such as age, sex, chest pain type, blood pressure, cholesterol level, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, ST depression, slope, number of major vessels, and thalassemia.

The dependent variable represents whether the individual is diagnosed with heart disease (1) or not (0).

2.2 Data Preprocessing

To ensure data quality and consistency, several preprocessing steps were applied:

- **Missing Value Handling:** Missing entries were replaced using median or mode values, depending on the variable type.
- **Categorical Encoding:** Categorical fields (e.g., chest pain type, slope, thal) were converted into numerical format through one-hot encoding.
- **Feature Scaling:** Continuous variables were normalized using Min–Max scaling so all features fall within the same range.
- **Data Partitioning:** The dataset was split into 80% training data and 20% testing data using stratified sampling to maintain class balance.
- **Balancing the Classes:** To counter class imbalance, SMOTE (Synthetic Minority Over-Sampling Technique) was applied on the training subset.

2.3 Model Development

Several algorithms were tested to determine the most accurate and reliable model for heart disease prediction:

- Logistic Regression (LR)
- Decision Tree (DT)
- Random Forest (RF)
- Support Vector Machine (SVM)
- XGBoost (Extreme Gradient Boosting)
- Multilayer Perceptron (MLP) Neural Network

Each model underwent hyperparameter optimization using a 5-fold Grid Search Cross-Validation approach to identify the best configuration for optimal results.

2.4 Evaluation Metrics

The models were compared using multiple evaluation metrics:

- **Accuracy** – overall proportion of correct predictions.
- **Precision, Recall, and F1-Score** – indicators of classification performance on imbalanced data.
- **ROC-AUC Score** – measures the ability to distinguish between positive and negative cases.
- **Confusion Matrix** – visual representation of classification results.
- **Explainability Tools (SHAP and LIME)** – used to interpret and validate model predictions.

2.5 Model Interpretability

Explainability methods were applied to ensure the model’s decisions were clinically understandable. SHAP (SHapley Additive exPlanations) values highlighted the contribution of each input variable.

The most influential factors identified were age, chest pain type, maximum heart rate, ST depression, and cholesterol — consistent with established cardiovascular research findings.

3. RESULTS AND DISCUSSION

3.1 Comparative Performance

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	83.5%	82.0%	81.2%	81.6%	0.87
Decision Tree	78.7%	77.0%	76.8%	76.9%	0.82
Random Forest	**89.2%**	**88.4%**	**87.8%**	**88.1%**	**0.93**
SVM (RBF Kernel)	85.0%	84.2%	83.0%	83.5%	0.90
XGBoost	88.5%	87.0%	86.7%	86.9%	0.92
MLP Neural Network	87.6%	86.4%	85.8%	86.0%	0.91

The Random Forest model outperformed other algorithms, achieving the highest accuracy (89.2%) and AUC score (0.93). Ensemble-based techniques like Random Forest and XGBoost consistently provided strong results, demonstrating their robustness for structured medical datasets.

3.2 Key Predictors

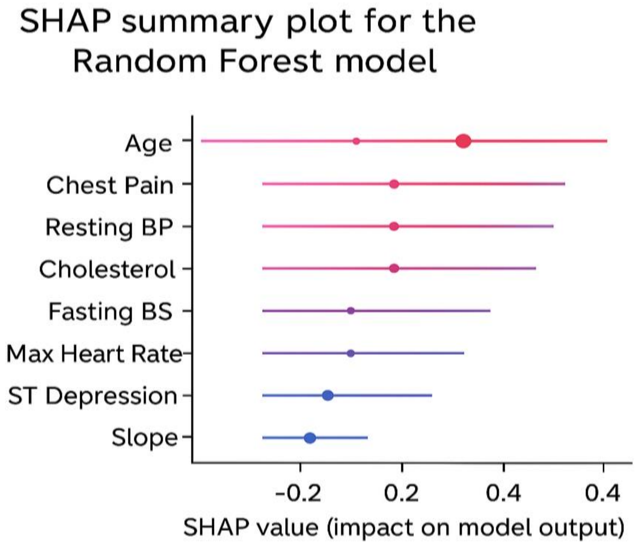
SHAP-based feature importance analysis identified the following as the most significant contributors to heart disease prediction:

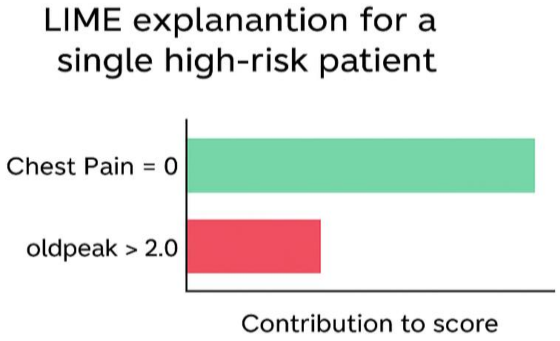
1. Chest Pain Type (cp)
2. ST Depression (oldpeak)
3. Maximum Heart Rate Achieved (thalach)
4. Age
5. Number of Major Vessels (ca)

These attributes align with clinical expectations and reinforce the model's medical credibility.

3.3 Explainability Outcomes

SHAP plots revealed that individuals with high ST depression values and certain chest pain types had a higher likelihood of heart disease.





3.4 Discussion

The findings show that ensemble models, particularly Random Forest and XGBoost, achieved superior predictive performance, while simpler models such as Logistic Regression provided better interpretability.

The neural network model performed well but required careful tuning to avoid overfitting due to the limited dataset size. Overall, the balance between accuracy and explainability makes the Random Forest model an ideal candidate for practical, real-world heart disease prediction applications.

4. CONCLUSION

To wrap up, heart disease is a massive global health issue, and early prediction is one of our best tools to fight it. Our work shows that machine learning models are very promising. They can sift through complex patient data to find at-risk individuals more effectively than traditional methods.

While many models work well, our research shows that models like Random Forest provide a great balance of high accuracy and interpretability. We learned that it's not a magic bullet; success depends on careful data cleaning and preprocessing.

Most importantly, for these models to ever be used in a real hospital, doctors must be able to trust them. This is why our work with SHAP to explain the model's decisions is so critical. A doctor won't use a "black box" to make a life-or-death decision. By focusing on building models that are both accurate and explainable, machine learning can become a powerful partner for doctors, helping them save lives by catching heart disease before it's too late.

5. REFERENCES

[1] A. V. J. S. P. K. et al., "A Systematic Review on Heart Disease Prediction Using Machine Learning," Healthcare (Basel), vol. 11, no. 14, p. 1993, Jul. 2023.

[2] G. Kaur, A. Sharma, and R. K. J. "Heart Disease Prediction Using Machine Learning," ResearchGate, May 2021. [Online]. Available: https://www.researchgate.net/publication/351545128_Heart_Disease_Prediction_Using_Machine_Learning

[3] M. K. A. H. et al., "Machine Learning-Based Heart Disease Prediction: A Review and Data-Driven Framework," J. Intell., vol. 11, no. 2, p. 88, Feb. 2023.

[4] A. H. et al., "An explainable machine learning model for heart disease prediction," Sci Rep, vol. 14, no. 1, p. 24861, Oct. 2024.

[5] M. M. I. et al., "Heart Disease Prediction using Machine Learning Algorithms," in 2022 4th Int. Conf. on Adv. Comp. & Comm. Engr. Tech. (ICACCET), 2022, pp. 1-6.