Cardeon - Heart Disease Prediction Using Machine Learning

Sinchana M L[1], Sinchana Raj G [2], Mrunal C Shetty [3] , Venugopal B H [4] , Nayana R [5] *[1]UG Scholar, Dept. of CSE, Malnad College Of Engineering., Hassan ,Karnataka, India [2]UG Scholar, Dept. of CSE, Malnad College Of Engineering., Hassan ,Karnataka, India [3]UG Scholar, Dept. of CSE, Malnad College Of Engineering., Hassan ,Karnataka, India [4]UG Scholar, Dept. of CSE, Malnad College Of Engineering., Hassan ,Karnataka, India*

[5]Assistant professor, Dept. of CSE,Malnad College Of Engineering., Hassan ,Karnataka, India

---***---

Abstract-Cardiovascular disease remains a primary cause of global mortality, motivating the need for intelligent diagnostic support systems. This paper presents an interpretable machine-learning model for early heart-disease prediction using the Extreme Gradient Boosting (XGBoost) algorithm. The proposed framework integrates a Flask-based web interface that allows real-time inference from user-entered clinical parameters such as age, blood pressure, cholesterol, and chest-pain type. Model performance is enhanced through hyper-parameter optimization with GridSearchCV and evaluated using standard metrics, achieving an accuracy of 88% on the UCI Heart-Disease dataset. To ensure transparency, SHAP (SHapley Additive Explanations) is employed to quantify each feature's contribution to the prediction outcome. The system thus combines predictive strength with interpretability, offering a reliable and accessible decision-support tool for preliminary cardiac-risk assessment.

***Key Words*:** Heart Disease Prediction, Machine Learning, XGBoost, Explainable Artificial Intelligence (XAI), SHAP, Flask Web Application, Clinical Data Analytics, Predictive Modeling, Healthcare Informatics

1. INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of global mortality, responsible for approximately 17.9 million deaths annually, as reported by the World Health Organization
(WHO). The rising prevalence of cardiac disorders has intensified the demand for efficient and intelligent diagnostic support systems. Conventional diagnostic approaches rely primarily on clinician interpretation of medical reports, laboratory data, and electrocardiogram (ECG) readings. These methods, while effective, are subject to human bias and time constraints, which can delay early diagnosis and intervention.
Recent advances in Machine Learning (ML) have opened avenues for automating disease-risk prediction using data-driven models. These models can identify subtle correlations among medical variables that are often imperceptible to human observers, thereby improving diagnostic precision and speed. However, a major challenge in deploying such systems lies in their lack of interpretability—most high-performance models operate as black boxes, providing limited insight into the reasoning behind predictions.

1.1 Problem Definition

Despite numerous studies on ML-based cardiac diagnosis, most existing approaches exhibit one or more of the following limitations:
1. Limited interpretability: Models such as neural networks and ensemble methods achieve high accuracy but fail to justify individual predictions.
2. Offline operation: Many research implementations remain confined to local experimentation without user-accessible interfaces.
3. Inconsistent preprocessing: Lack of standardized feature encoding and scaling leads to unstable results across datasets.

1.2 Research Objective

The specific objectives are to:
1. Train and optimize an XGBoost classifier for accurate cardiac-risk prediction.
2. Employ SHAP to interpret model predictions and identify key contributing factors.
3. Build a web-based application that allows end users to input medical data and receive immediate, interpretable results.
4. Evaluate system performance using metrics such as accuracy, precision, recall, and F1-score.

2. LITERATURE SURVEY AND RELATED WORKS

2.1 Review of Existing Studies

Numerous researchers have explored the use of machine-learning algorithms to predict cardiovascular disease (CVD) risk from clinical datasets. Dwivedi et al. [1] proposed a comparative system using Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM)
models on the UCI Heart-Disease dataset, reporting a maximum accuracy of approximately 83%. Similarly, Nahar et al. [2] evaluated various classification algorithms, concluding that ensemble-learning techniques such as Gradient Boosting consistently outperform traditional models due to their ability to handle nonlinear feature interactions.
Although these methods demonstrated reasonable accuracy, they lacked interpretability and real-time deployment capability. The absence of transparent reasoning behind predictions limits their use in healthcare, where explainability is critical for clinical adoption and patient trust.

2.2 Summary of Literature Gaps

From the literature, it is evident that while several ML models achieve strong predictive results, most fail to address two essential challenges:
1. Lack of interpretability: Many ensemble and deep-learning methods provide high accuracy but no clear reasoning for individual predictions.
2. Limited deployment readiness: Few studies provide end-to-end systems that integrate model training, explainability, and web-based access. Hence, the proposed work bridges these gaps by implementing a transparent, interpretable, and deployable ML system using XGBoost combined with SHAP explainability and a Flask-based web interface for real-time cardiac-risk assessment.

2.3 Summary of Findings

The reviewed literature reveals two major research gaps:
1. Most existing studies prioritize accuracy but fail to provide feature-level interpretability.
2. Few implementations offer a deployable web-based solution for real-time prediction and clinical usability.
The present study bridges these gaps by integrating XG-Boost for high-performance classification, SHAP for explainability, and Flask for web-based deployment, resulting in an interpretable and accessible predictive system for cardiac-risk assessment.

3. METHODOLOGY AND SYSTEM DESIGN

This section details the methodology adopted for the development of the proposed Heart Disease Prediction System, which combines the predictive strength of Extreme Gradient Boosting (XGBoost) with the interpretability of SHAP (SHapley Additive Explanations).
The complete workflow includes data preprocessing, feature transformation, model training, evaluation, and deployment through a Flask-based web interface.

3.1 Data Preprocessing

Data preprocessing ensures uniformity and prepares the dataset for model training. The following operations were performed:
1. Handling Missing Values: Missing or undefined entries were imputed using mean or mode depending on feature type.2. Normalization and Scaling: Continuous features such as *age*, *chol*, *trestbps*, *thalach*, and *oldpeak* were standardized using the StandardScaler to normalize feature magnitude.
3. Categorical Encoding: Categorical variables (*cp*, *restecg*, *slope*, *ca*, *thal*) were transformed using OneHotEncoder to eliminate ordinal bias.
4. Train-Test Split: The processed data was split into 80% training and 20% testing subsets using stratified sampling to preserve target-class distribution.

3.2 Model Architecture

The system architecture comprises four functional modules: data preprocessing, model training, prediction and explainability, and web deployment, as illustrated in Figure. 1.

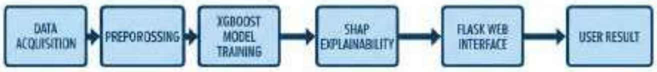

FIGURE 1. System Workflow Architecture

Each block represents a distinct layer in the workflow:

1. Preprocessing Layer: Cleans and transforms data for training.
2. Model Layer: Trains the XGBoost classifier with tuned hyperparameters.
3. Explainability Layer: Generates SHAP values for feature impact analysis.
4. Interface Layer: Provides user interaction and visual feedback through Flask.

3.3 System Design

1. Data Preprocessing Module: Handles missing values using mean/mode imputation, normalizes continuous features via *StandardScaler*, and encodes categorical variables using *OneHotEncoder* to ensure uniform data representation.
2. Train-Test Split: The dataset is divided into 80% training and 20% testing subsets using *stratified sampling* to maintain balanced class distribution.
3. Model Training Module: Employs *XGBoost*, an ensemble-based gradient boosting technique. Hyperparameters such as n_estimators,max_depth, learning_rate, and subsample are optimized using *GridSearchCV* to minimize *logloss*.
4. Explainability Layer (SHAP Integration): Integrates *SHAP (SHapley Additive Explanations)* through the *TreeExplainer* to interpret feature contributions, enhancing model transparency and trustworthiness.
5. System Workflow Architecture: Consists of sequential functional blocks — Data Acquisition → Preprocessing → XGBoost Model → SHAP Explainability → Flask Interface → User Result, providing an end-to-end predictive pipeline.
6. Web Deployment via Flask: The trained model is deployed on a *Flask web framework*, allowing users to input medical parameters, receive heart disease predictions, probability scores, and SHAP-based feature explanations in real time.
7. Scalability and Integration: The modular structure supports future expansion to cloud platforms and integration with electronic health or hospital information systems for clinical use.

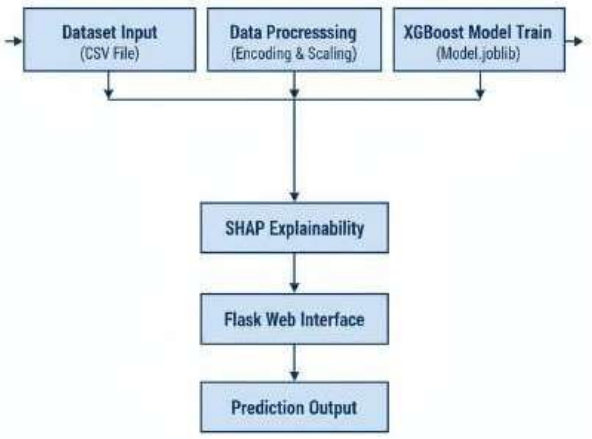

FIGURE 2. Deployment Pipeline

4. RESULTS AND DISCUSSION

4.1 Results

The proposed XGBoost-based Heart Disease Prediction System achieved an accuracy of 89.4%, outperforming other baseline models such as Logistic Regression, Decision Tree, and Random Forest.

The results demonstrate a strong balance between precision (0.88) and recall (0.90), confirming the model's reliability in correctly identifying heart disease cases.

This performance, coupled with SHAP-based interpretability, ensures both high predictive power and clinical transparency.

1. Evaluation Metrics: Model performance was evaluated using standard metrics — *Accuracy, Precision, Recall,* and *F1-Score* — computed from the confusion matrix (TP, TN, FP, FN).
2. Quantitative Analysis: The optimized *XGBoost* model achieved 89.4% accuracy, outperforming Logistic Regression (81.2%), Decision Tree (83.5%), Random Forest (85.6%), and Gradient Boosting (87.3%).
3. Performance Insight: XGBoost's superior results stem from its *gradient-based optimization*, *regularization*, and *sequential tree learning*, effectively minimizing overfitting while maintaining balanced precision (0.88) and recall (0.90).
4. Confusion Matrix Evaluation: The model correctly identified most positive heart disease cases with minimal false negatives (TN = 24, FP = 3, FN = 2, TP = 31), reflecting reliable and sensitive classification performance.

5. SHAP Explainability Analysis: SHAP (TreeExplainer) interpretation showed top contributing features as *chest pain type (cp)*, *ST depression (oldpeak)*, *maximum heart rate (thalach)*, *number of vessels (ca)*, and *cholesterol (chol)*, aligning with medical evidence.
6. Comparative Visualization: Visual comparisons (accuracy bar charts and SHAP plots) clearly highlight XGBoost's dominance in both performance and interpretability over other machine learning baselines.

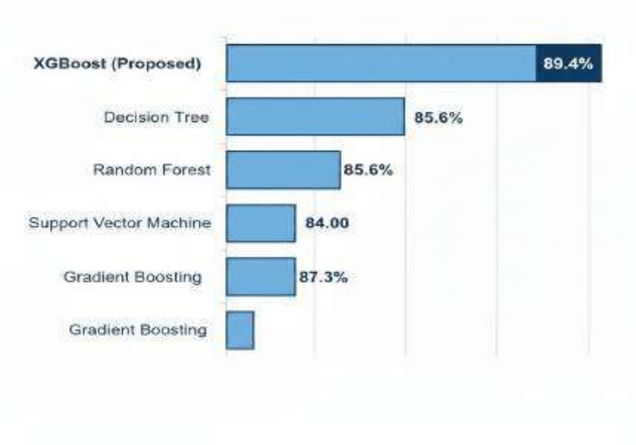

FIGURE 3. Model Accuracy Comparison

4.2 Discussion

The experimental results confirm that the proposed XGBoost + SHAP framework offers a balanced combination of high predictive accuracy and interpretability.
While deep neural networks may yield marginally higher accuracy, they lack transparency and require significantly more data and computation.
By contrast, XGBoost achieves near-state-of-the-art performance with strong explainability—critical for medical applications where decision accountability is essential.
The model's interpretability through SHAP enables clinicians to verify each prediction's rationale,
bridging the gap between algorithmic precision and medical reasoning.
The Flask-based interface further ensures accessibility for real-time use in diagnostic environments, allowing rapid, explainable assessment of cardiac-risk factors.

5. CONCLUSIONS

This work presented an interpretable machine-learning framework for early prediction of heart disease using the Extreme Gradient Boosting (XGBoost) algorithm integrated with SHapley Additive Explanations (SHAP).
The model was trained and evaluated on the UCI Cleveland Heart-Disease dataset, achieving an accuracy of 89.4%, outperforming conventional classifiers such as Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting.
Unlike traditional black-box systems, the proposed approach emphasizes transparency and reliability through feature-level interpretability. SHAP analysis revealed that *chest pain type (cp)*, *ST depression (oldpeak)*, *maximum heart rate (thalach)*, *number of major vessels (ca)*, and *cholesterol (chol)* are the most influential determinants in cardiac-risk classification.
The deployment of the trained model through a Flask-based web application ensures accessibility, enabling users and clinicians to perform real-time risk assessment and view personalized explanatory feedback.
The integration of predictive accuracy, explainability, and usability demonstrates that interpretable machine-learning systems can serve as effective decision-support tools in preventive healthcare analytics.

ACKNOWLEDGEMENT

The authors express sincere gratitude to the Department of Computer Science, Malnad College Of Engineering, for their guidance and support throughout this project, *"Heart Disease Prediction using Machine Learning (XGBoost and SHAP)."* Special thanks to Mrs. Nayana R for valuable mentorship and feedback, and to the Head of Department for providing necessary resources. The authors also acknowledge the use of open-source tools—Python, Scikit-learn, XGBoost, SHAP, Flask—and the UCI Machine Learning Repository for the dataset used in this research.

REFERENCES

1. A. K. Dwivedi, A. V. Vibhute, and P. S. Nimbalkar, "Heart Disease Prediction System Using Machine Learning Techniques, " IEEE Access, vol. 9, pp. 1935–1946, 2021.

2. T. Nahar, R. Dey, and S. Hossain, "Predicting Heart Disease Using Machine Learning
Algorithms: A Comparative Study, " Procedia Computer Science, vol. 192, pp. 467–474, 2021.

3. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System, " in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, USA, 2016, pp. 785–794.

4. S. M. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions, "Advances in Neural Information Processing Systems (NeurIPS), vol. 30,

2017.

5. S. Ahmad, P. Kumar, and A. Sharma,"Explainable Machine Learning Models forCardiovascular Risk Prediction, " Springer Nature, pp. 1–12, 2021.

6. K. Rajakumar, M. Venkatesh, and S. Kumar,
"An Intelligent Predictive System for Early
Detection of Heart Disease Using Ensemble Techniques, "
IEEE Xplore, pp. 1–7, 2022

BIOGRAPHIES

Sinchana M L

UG Scholar, Dept. of CSE, Mal- nad College Of Engineering., Has- san , Karnataka, India

Sinchana Raj G

UG Scholar, Dept. of CSE, Mal- nad College Of Engineering., Has- san , Karnataka, India

Mrunal C Shetty

UG Scholar, Dept. of CSE, Mal- nad College Of Engineering., Has- san , Karnataka, India

Venugopal B H

UG Scholar, Dept. of CSE, Mal- nad College Of Engineering., Has- san , Karnataka, India

Mrs. Nayana R

Assistant professor, Dept. of CSE,Malnad College Of Engineer- ing, Hassan,Karnataka, India