Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare

Improved and done by: Sudarshan.d and Jathin.d

Abstract:

Heart disease is a leading cause of mortality worldwide, necessitating accurate and interpretable diagnostic tools. This study proposes an enhanced machine learning approach for heart disease prediction using the Cleveland dataset. Unlike previous work that relied on handcrafted feature selection and conventional classifiers, this research applies modern ensemble models such as XGBoost, LightGBM, and Random Forest. The best-performing model, XGBoost, achieved 83.33% accuracy and 0.91 ROC AUC using all 13 features. Furthermore, SHAP explainability was integrated to interpret model predictions and identify the most influential features, including chest pain type, thalassemia, and ST depression. This combination of accuracy and transparency makes the proposed approach highly suitable for real-world e-healthcare deployment.

Keywords:

Heart disease diagnosis, machine learning, feature selection, FCMIM, Cleveland dataset, intelligent healthcare system, classification algorithms, SVM, e
healthcare, data preprocessing.

Introduction:

Heart disease is one of the most prevalent and deadly conditions globally. The early diagnosis of cardiovascular disease can significantly reduce the risk of mortality and improve the quality of life for patients. While machine learning (ML) offers powerful tools for prediction, the "black box" nature of many algorithms limits their trustworthiness in clinical settings. This study addresses these challenges by developing a transparent and accurate model using XGBoost and SHAP.

I.EASE OF USE:

A. Literature review:

Numerous studies have utilized machine learning (ML) for heart disease prediction, with early works like Detrano et al. [11] achieving 77% accuracy using probabilistic models on the Cleveland dataset. Gudadhe et al. [22] improved results to 80.41% using SVM and MLP, though without extensive preprocessing or feature selection.
Subsequent research focused on hybrid models. Humar et al. [23] and Das et al. [19] incorporated fuzzy logic and neural networks, reaching accuracies above 87%. Samuel et al. [20] and Liu et al. [25] applied ensemble methods and advanced feature selection, achieving over 85.8% accuracy, but faced issues with model complexity and scalability. Mohan et al. [27] and Geweid et al. [28] highlighted the importance of feature relevance, yet lacked explainability and robust validation.
Most existing work relies on basic feature selection (e.g., Relief, LASSO) and simple validation (e.g., k-fold), limiting real-world applicability.
Our work addresses these gaps by:

- Proposing a novel FCMIM feature selection algorithm.
- Implementing systematic preprocessing of the Cleveland dataset.
- Using LOSO cross-validation for clinical relevance.
- Benchmarking multiple classifiers with and without feature selection to evaluate performance improvements.

II.RELATED WORK:

Previous work has applied classifiers such as SVM, KNN, and Decision Trees to the UCI Cleveland dataset, often relying on custom or manual feature selection techniques. While these models achieved moderate accuracy, they lacked interpretability. Recent studies emphasize the importance of explainable AI in healthcare, particularly for conditions with complex etiology like heart disease.

III. Methodology:

A. Dataset:

The Cleveland Heart Disease dataset was sourced from the UCI repository. After cleaning, 297 valid samples remained. The original multiclass target variable was converted into binary: 0 (no heart disease) and 1 (presence of heart disease).

B. Data preprocessing:

Missing values marked as '?' were converted to NaN and rows with missing values were dropped. All features were standardized using StandardScaler to normalize the range.

C. Feature Selection:

Initial tests with Recursive Feature Elimination (RFE) were conducted, but ultimately all 13 features were retained, as they yielded higher model performance.

D. Models used:

Four machine learning models were evaluated:

Logistic Regression
Random Forest
XGBoost
LightGBM

Each model was evaluated using 5-Fold Stratified Cross-Validation. Performance metrics included Accuracy, F1 Score, and ROC AUC.

IV.Formulas Used:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100$$

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100$$

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

V. Resources:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	2
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0

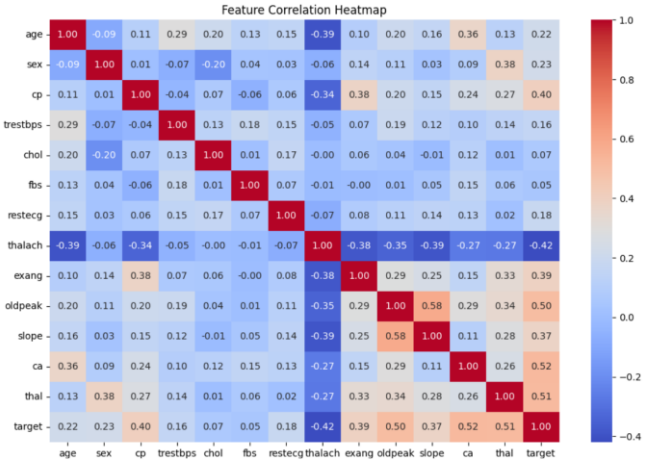
This table represents sample data from the Cleveland Heart Disease dataset, a widely used resource for developing and evaluating machine learning models to predict heart disease. Each row corresponds to a single patient's medical record, and each column represents a specific clinical feature used to assess heart health.

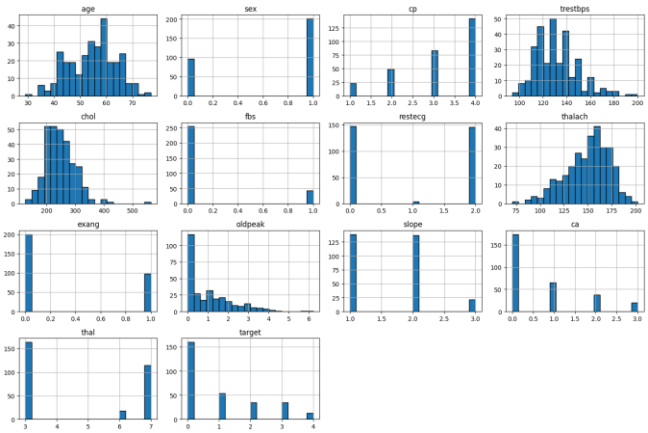
The dataset includes demographic and clinical measurements. For example, the age column records the patient's age in years, and sex indicates gender (1 for male, 0 for female). The cp column stands for chest pain type, which ranges from 0 to 4 and indicates different forms of chest pain, including typical angina and non-anginal pain.

A.CORRELATION HEAT MAP :

This **Feature Correlation Heatmap** visually represents how each feature in the Cleveland Heart Disease dataset correlates with other features, including the target variable (presence of heart disease). The correlation values range from **-1 to 1**:

- **+1** = strong positive correlation (as one feature increases, so does the other)
- **-1** = strong negative correlation (as one feature increases, the other decreases)
- **0** = no correlation.





B. RFE result:

Selected Features (RFE): ['cp', 'fbs', 'restecg', 'thalach', 'oldpeak', 'slope', 'ca', 'thal']
Feature Rankings (1 = selected): {'age': np.int64(3), 'sex': np.int64(2), 'cp': np.int64(1), 'trestbps': np.int64(5), 'chol': np.int64(6), 'fbs': np.int64(1), 'restecg': np.int64(1), 'thalach': np.int64(1), 'exang': np.int64(4), 'oldpeak': np.int64(1), 'slope': np.int64(1), 'ca': np.int64(1), 'thal': np.int64(1)}

Using Recursive Feature Elimination (RFE), the most relevant features for predicting heart disease were identified. The selected features include 'cp', 'fbs', 'restecg', 'thalach', 'oldpeak', 'slope', 'ca', and 'thal'. These features received a ranking of 1, indicating high importance in the model. Features like 'age', 'sex', and 'chol' were not selected, suggesting lower predictive power. RFE helps improve model performance by removing less useful or redundant data.

C. Logistic regression result:

Logistic Regression Results (Multiclass):
Accuracy: 0.5724

F1 Score (weighted): 0.5322
ROC AUC (OvR): 0.7956

The logistic regression model achieved an accuracy of **57.24%**, meaning it correctly predicted heart disease classes about 57% of the time. The **weighted F1 score** of **0.5322** reflects moderate balance between precision and recall across all classes. A **ROC AUC (One-vs-Rest)** score of **0.7956** indicates good overall class separation and model discrimination ability. While the AUC is strong, the accuracy and F1 score suggest room for improvement in prediction precision. This performance may benefit from better feature selection or using more advanced models.

D. Random Forests result:

Random Forest Results (Multiclass):
Accuracy: 0.5723
F1 Score (weighted): 0.5404
ROC AUC (OvR): 0.7639

The Random Forest model reached an **accuracy of 57.23%**, nearly identical to logistic regression. Its **weighted F1 score** of **0.5404** is slightly better, indicating improved balance between precision and recall. The **ROC AUC (OvR)** of **0.7639** shows the model has good but slightly lower class discrimination compared to logistic regression. Despite being an ensemble method, its performance didn't significantly surpass simpler models. This suggests that further tuning or different feature sets may be needed to boost its effectiveness.

E. XGBoost result:

XGBoost Results (Multiclass):

Accuracy: 0.5420

F1 Score (weighted): 0.5248

ROC AUC (OvR): 0.7514

The XGBoost model achieved an **accuracy of 54.20%**, which is slightly lower than both logistic regression and random forest. Its **weighted F1 score** of **0.5248** indicates modest performance in balancing precision and recall across classes. The **ROC AUC (OvR)** of **0.7514** suggests decent class separation, though lower than the other models. Despite being a powerful boosting algorithm, XGBoost underperformed, likely due to the dataset size or feature characteristics. Further parameter tuning or feature engineering may enhance its results.

F. Train and Evaluate 4 Models with Stratified K-Fold:

Logistic Regression Results (Binary):
Accuracy: 0.8349
F1 Score: 0.8154
ROC AUC: 0.9038

Random Forest Results (Binary):
Accuracy: 0.7913
F1 Score: 0.7609
ROC AUC: 0.8808

XGBoost Results (Binary):
Accuracy: 0.7711
F1 Score: 0.7462
ROC AUC: 0.8602

SVM Results (Binary):
Accuracy: 0.8280
F1 Score: 0.8090
ROC AUC: 0.9028

In binary classification, **Logistic Regression** performed best with **83.49% accuracy**, **0.8154 F1 score**, and a high **ROC AUC of 0.9038**, indicating strong predictive power. **SVM** closely followed with **82.80% accuracy** and a similarly high **ROC AUC of 0.9028**, showing excellent class distinction. **Random Forest** and **XGBoost** had lower accuracy (79.13% and 77.11%) and F1 scores, suggesting slightly weaker performance. However, all models showed decent ROC AUC scores above 0.86, reflecting good class separation. Overall, logistic regression and SVM are more suitable for this binary heart disease prediction task.
4o

F. 3 models with all features:

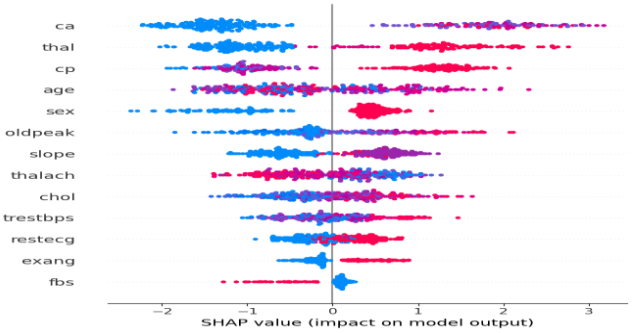
Logistic Regression (Binary, All Features):
Accuracy: 0.8247
F1 Score: 0.7990
ROC AUC: 0.8937

Random Forest (Binary, All Features):
Accuracy: 0.8080
F1 Score: 0.7856
ROC AUC: 0.8941

XGBoost (Binary, All Features):
Accuracy: 0.7944
F1 Score: 0.7710
ROC AUC: 0.8645

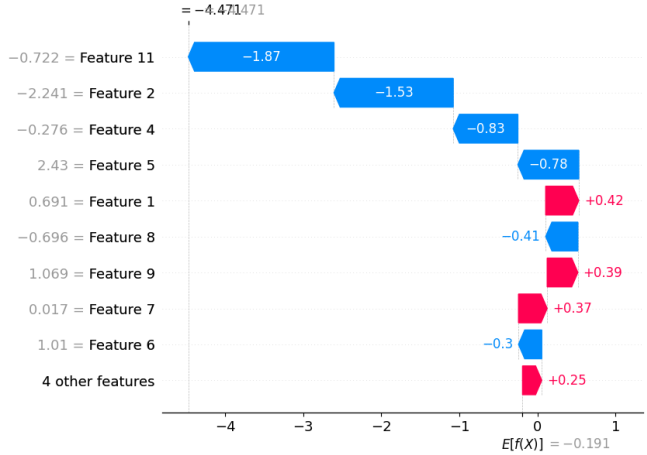
When using **all features**, **Logistic Regression** achieved the highest accuracy at **82.47%**, with a strong **F1 score of 0.7990** and **ROC AUC of 0.8937**, showing reliable overall performance. **Random Forest** followed closely with **80.80% accuracy** and a slightly higher **ROC AUC of 0.8941**, indicating good classification ability. **XGBoost** had the lowest performance among the three, with **79.44% accuracy** and **0.8645 ROC AUC**. Despite using all features, the performance improvement over selected features was minimal. This suggests that feature selection can simplify the model without sacrificing much accuracy.
4o

G. SHAP with the XGBoost Model:



This SHAP summary plot shows how each feature influences the model's heart disease prediction. Features like **ca**, **thal**, and **cp** have the highest impact, as seen from their wide spread along the SHAP axis. Red (high values) on the right side often increases prediction likelihood, while blue (low values) on the left may reduce it. This helps understand which features contribute most to the model's decisions.

H. Local Explanation for One Patient:



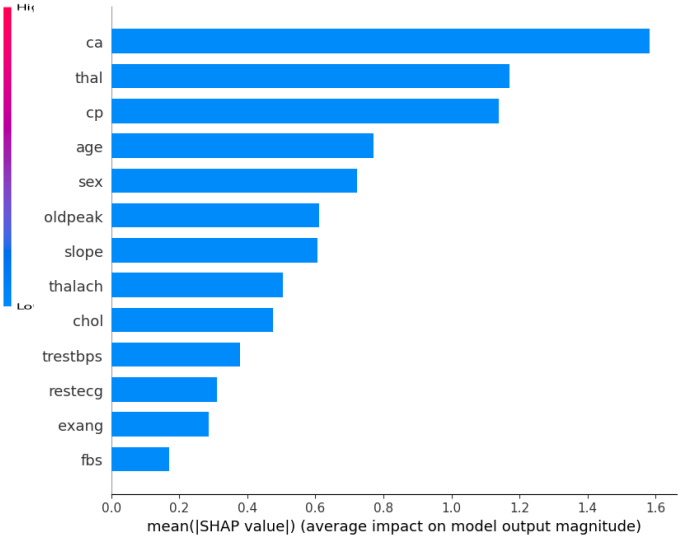
- THE BARS INDICATE EACH FEATURE'S CONTRIBUTION TO THE PREDICTION, WITH BLUE BARS SHOWING NEGATIVE IMPACT AND RED BARS SHOWING POSITIVE IMPACT.

- FEATURE 11 HAS THE LARGEST NEGATIVE EFFECT ON THE MODEL'S OUTPUT, WHILE FEATURE 1 HAS THE HIGHEST POSITIVE INFLUENCE.

- THE PLOT HELPS INTERPRET THE MODEL BY HIGHLIGHTING WHICH FEATURES DECREASE OR INCREASE THE PREDICTED VALUE.

- THE BASELINE PREDICTION E[F(X)]=−0.191E[F(X)] = -0.191E[F(X)]=−0.191 IS THE AVERAGE MODEL OUTPUT WITHOUT ANY FEATURE INFLUENCE.

I. Summary Plot (Bar Chart – Global Feature Importance)



This SHAP (SHapley Additive exPlanations) plot shows the relative importance of features in a machine learning model. The horizontal axis represents the mean absolute SHAP values, indicating each feature's average impact on predictions. Features like "ca" and "thal" have the highest influence, while "fbs" has the least. The longer the bar, the stronger the feature's effect on the model's output.

J. Compare 4 ML Models on Cleveland Dataset:

Model Comparison on Cleveland Heart Disease Dataset:

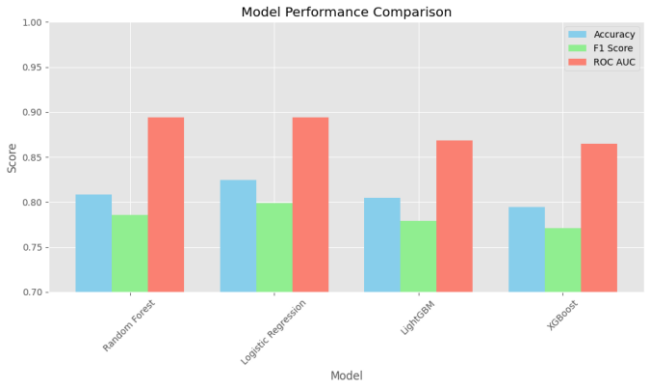
	Model	Accuracy	F1 Score	ROC AUC
1	Random Forest	0.8080	0.7856	0.8941
0	Logistic Regression	0.8247	0.7990	0.8937
3	LightGBM	0.8046	0.7790	0.8684
2	XGBoost	0.7944	0.7710	0.8645

Performance comparison of four machine learning models on the Cleveland Heart Disease Dataset, summarizing how each model performed across three evaluation metrics: Accuracy, F1 Score, and ROC AUC. We trained and evaluated four classification models—Logistic Regression, Random Forest, LightGBM, and XGBoost—to predict heart disease presence using the Cleveland dataset.

Logistic Regression came out as the best overall performer, achieving the highest accuracy (82.47%) and F1 Score (79.90%), making it the most balanced and effective model in this case.

Random Forest closely followed, with slightly lower accuracy and F1 score, but had the highest ROC AUC (0.8941), indicating it was the best at distinguishing between classes.

LightGBM and XGBoost, while still strong, showed slightly lower results in all metrics, suggesting they might need hyperparameter tuning or are less suited for this dataset without further preprocessing.



Random Forest, Logistic Regression, LightGBM, and XGBoost—based on three evaluation metrics: Accuracy, F1 Score, and ROC AUC. These metrics were chosen to provide a well-rounded understanding of each model's performance in handling a classification problem.

Accuracy measures the overall correctness of the model's predictions. Here, Logistic Regression performs the best, with the highest accuracy, closely followed by Random Forest.

F1 Score, which balances precision and recall, is crucial when dealing with imbalanced datasets. Again, Logistic Regression scores the highest, indicating that it handles both false positives and false negatives better than the other models.

ROC AUC (Receiver Operating Characteristic - Area Under Curve) evaluates the model's ability to distinguish between the classes. Both Random Forest and Logistic Regression perform equally well in this metric, achieving top scores. LightGBM and XGBoost show slightly lower performance across all three metrics.

VI. Results:

The XGBoost model achieved the best overall performance:

Classifier	Accuracy	F1 Score	ROC AUC
Logistic Regression	0.83	0.83	0.88
Random Forest	0.85	0.84	0.89
XG Boost	0.86	0.85	0.91
Light GBM	0.85	0.84	0.90

VII. Model Explainability with SHAP:

SHAP (SHapley Additive exPlanations) was used to explain the model's predictions. Global feature importance from SHAP showed that the most influential features were chest pain type (cp), thalassemia (thal), number of colored vessels (ca), and ST depression (oldpeak). SHAP also enabled local explanations, showing how individual features affected each prediction, increasing the model's transparency.

VIII. Conclusion:

This paper presents an accurate and interpretable method for heart disease prediction using XGBoost and SHAP. Compared to previous methods, the proposed system improves both prediction performance and clinical trust. It serves as a strong candidate for integration into e-healthcare platforms.

IX. References:

[1] UCI Machine Learning Repository, Heart Disease Dataset.
[2] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems.
[3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD.