**IAES International Journal of Artificial Intelligence (IJ-AI)**
Vol. 15, No. 1, February 2026, pp. 191~212
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i1.pp191-212 

# Explainable hybrid models for cardiovascular disease detection and mortality prediction

**Ali Al-Ataby, Hussain Attia**
Department of Electrical and Electronics Engineering, School of Engineering and Computing, American University of Ras Al Khaimah, Ras Al Khaimah, United Arab Emirates

**Article Info**

***Article history:***

Received Sep 1, 2025
Revised Dec 21, 2025
Accepted Jan 10, 2026

***Keywords:***

Cardiovascular disease
Ensemble learning
Explainable artificial intelligence
Heart failure
Machine learning
Mortality prediction
Shapley additive explanations

**ABSTRACT**

The impact of cardiovascular diseases (CVDs) is devastating, with 20.5 million deaths annually. Early detection and prediction tools exist, but current approaches struggle to balance predictive performance with clinical interpretability. In this work, a two-stage machine learning (ML) framework is proposed for heart disease detection and mortality prediction in heart failure patients. Logistic regression (LR), random forest (RF), and gradient boosting (GB) models were trained using the publicly available heart failure datasets, and their performance was compared, then a stacked ensemble approach was employed to enhance prediction accuracy. Model interpretability was achieved through Shapley additive explanations (SHAP), which provides global feature rankings and specific patient attributes, supporting explainable artificial intelligence (XAI) in clinical practice. The GB model achieved the highest performance in the first stage with a receiver operating characteristic area under the curve (ROC AUC) of 96% and an accuracy of 89% on internal testing, while external validation confirmed strong generalization (ROC AUC of 94%). In the second stage, stacked ensemble model was employed and achieved marginal improvements. Two interactive web applications were developed to enable real-time predictions with SHAP visualizations. The results demonstrate that combining high-performance ML models with interpretable outputs can significantly improve trust in real-world healthcare environments.



***Corresponding Author:***

Ali Al-Ataby
Department of Electrical and Electronics Engineering, School of Engineering and Computing
American University of Ras Al Khaimah
Ras Al Khaimah, United Arab Emirates
Email: ali.ataby@aurak.ac.ae

## 1. INTRODUCTION

With an estimated 20.5 million deaths annually, or nearly 33% of all deaths worldwide, cardiovascular diseases (CVDs) continue to be the leading cause of death [1]. Despite improvements in medical treatments, heart failure is often missed, which leads to a high risk of premature death. Timely treatment planning depends on the early detection of CVDs and the precise prediction of patient outcomes. However, it is still difficult to achieve both high predictive accuracy and clinical interpretability [2]. Traditional statistical models such as logistic regression (LR) provide easy interpretability but lack the predictive power of advanced ensemble-based machine learning (ML) algorithms [3], [4]. Complex models like random forests (RF) and gradient boosting (GB) tend to act as "black boxes", which limits their clinical acceptance. Explainability approaches, such as Shapley additive explanations (SHAP), which provide transparent, patient-specific feature attributions, have been introduced recently in interpretable ML to close

this gap. With these approaches, predictive models can provide clinicians with precise information and useful insights, which result in improving risk assessment and personalized care [4].

This study proposes a two-stage ML framework. The first stage focuses on detecting heart disease, while the second predicts mortality in heart failure patients. A variety of models, including LR, RF, and GB, are compared, and a stacked ensemble is developed to maximize performance. To ensure clinical trust, the framework integrates SHAP analysis for both global and individual level interpretability. Moreover, the best performing models will be deployed into a user-friendly web application to provide clinicians with an interface for real-time risk assessment and decision support. Accordingly, the aim of this work is to deliver a robust, interpretable, and deployable solution that supports both the early diagnosis of heart disease and the proactive management of high-risk patients. The main contributions of this work are:

– The development of a two-stage ML pipeline for heart disease detection and mortality prediction.
– The use a stacked ensemble approach to improve model prediction performance.
– The use of SHAP interpretation to enhance decision transparency.
– The deployment of the best performing models in interactive web applications for real-world usability.

The rest of this paper is organized as follows. Section 2 provides a summary of literature work about the use of ML in CVD detection. Section 3 provides the method followed in this work, including dataset analysis and preprocessing, the developed models, and the performance metrics. Section 4 provides details about the implementation of the models. Section 5 provides the obtained results with a discussion about performance. Finally, section 6 concludes the paper with a summary of the main contributions, limitations, and future work.

## 2. LITERATURE REVIEW

A number of references were used in the review of recent literature related to the use of ML in heart disease detection and mortality prediction. The summary of this review is given in Table 1 (see in Appendix) [2], [3], [5]–[22]. The table provides the work carried out in each reference along with the gap. Recent explainable and hybrid AI studies have excelled at CVD prediction by emphasizing interpretable clinically reliable models that balance predictive accuracy with transparency. Sourov *et al.* [23] introduced an explainable AI-enhanced framework for CVD detection and risk assessment, demonstrating how model explainability can coexist with high performance. Napa *et al.* [24] conducted a comparative analysis of explainable models using SHAP and highlighted how model transparency aids in cardiovascular risk determination and feature interpretation. Similarly, Bilal *et al.* [25] developed an explainable AI system for accurate prediction of CVD and emphasized the importance of explainability for clinical adoption of AI tools in healthcare.

Despite the available literature about the potential of ML models (e.g., RF, XGBoost, support vector machine (SVM), and deep neural network (DNN)) for accurate prediction of CVDs, many gaps exist. Many studies have prioritized accuracy over model interpretability and clinical usability, which are essential for real-world adoption. Also, comparative studies often use inconsistent datasets, lack external validation, and report superior performance on small, imbalanced datasets, which limits generalizability. Furthermore, although explainable AI approaches have been proposed to enhance trust and transparency, the integration into CVDs prediction workflows requires more investigation. This study addresses a number of these gaps by developing a robust and interpretable hybrid ML model for heart disease and mortality prediction by combining classical and ensemble techniques with SHAP explainability. This contributes to predictive accuracy and also to clinical transparency and trust, which are critical factors for ethical and practical deployment in healthcare clinics.

## 3. METHOD

This section provides method followed for developing, training, evaluating, and interpreting ML models for heart disease and mortality prediction. The work consists of dataset selection, preprocessing, and feature engineering. It also includes model development, model evaluation, and model interpretability using SHAP.

### 3.1. Dataset description and data preprocessing

#### 3.1.1. Heart disease prediction dataset

A Kaggle dataset with 12 attributes, including age, sex, cholesterol, chest pain, and ECG findings, was used [26]. This dataset contains 918 entries and 12 columns. The following points summarize important observations from the dataset:

– No missing values were found.
– There are categorical features such as Sex, ChestPainType, RestingECG, ExerciseAngina, and ST_Slope.
– There are numerical anomalies, where RestingBP and Cholesterol have a minimum of 0, which may indicate missing or erroneous values.

– FastingBS can be treated as categorical (0/1).
– Target class (HeartDisease) is fairly balanced (~55% positive cases).

Table 2 shows a summary of the specifications of this dataset. The proposed preprocessing involves handling missing values, outlier removal, encoding categorical variables, and feature scaling. Records with Cholesterol=0 or RestingBP=0 were dropped. Accordingly, the remaining records are 746 samples after cleaning.

Table 2. Heart disease prediction dataset specifications

| Column | Description |
|---|---|
| Age | Age of the patient |
| Sex | Biological sex (M/F) |
| ChestPainType | Type of chest pain (e.g., ATA, ASY, NAP, TA) |
| RestingBP | Resting blood pressure (0–200 mm Hg) |
| Cholesterol | Serum cholesterol in mg/dl (0 - 603) |
| FastingBS | Fasting blood sugar>120 mg/dl (1= true, 0= false) |
| RestingECG | ECG results (Normal, ST, LVH) |
| MaxHR | Maximum heart rate |
| ExerciseAngina | Exercise-induced angina (Y/N) |
| Oldpeak | ST depression induced by exercise |
| ST_Slope | Slope of the peak exercise ST segment |
| HeartDisease | Target variable (1= yes, 0= no) |

### 3.1.2. Mortality prediction dataset

The heart failure clinical records dataset from [27] was used. It consists of 299 records and 13 features, including serum creatinine, ejection fraction, blood pressure, and diabetes history. The following points are important observations from this dataset:

– There are no missing values.
– All features are numerical or binary, so preprocessing will be minimal.
– Class distribution of DEATH_EVENT is imbalanced, with 32% deceased (1) and 68% alive (0). So, there is a class imbalance. This must be handled in model evaluation with stratified splits.

Table 3 shows a summary of the specifications of this dataset. Given the class imbalance of the DEATH_EVENT label (32% positive, 68% negative), imbalance-aware learning strategies were adopted. Specifically, for LR, RF, and GB class_weight='balanced' option was used in scikit-learn to up-weight the minority class during training. In additional experiments, synthetic minority oversampling technique (SMOTE) was evaluated on the training set to generate synthetic minority samples, and it was found that performance trends were consistent, so the class-weight results were reported for simplicity. All train/test splits and cross-validation folds were stratified to preserve the class proportions. Exploratory data analysis (EDA) will be carried out on each dataset before model development. This includes operations such as correlations and feature importance.

Table 3. Mortality prediction dataset specifications

| Feature | Description |
|---|---|
| age | Age of the patient |
| anemia | 1= yes, 0= no |
| creatinine_phosphokinase | Enzyme level |
| diabetes | 1= yes, 0= no |
| ejection_fraction | Percentage of blood leaving the heart |
| high_blood_pressure | 1= yes, 0= no |
| platelets | Platelet count |
| serum_creatinine | Kidney function indicator |
| serum_sodium | Sodium level |
| sex | 1= Male, 0= Female |
| smoking | 1= yes, 0= no |
| time | Duration of follow-up (days) |
| DEATH_EVENT | Target (1= death occurred, 0= survived) |

## 3.2. Model development

To ensure robustness and generalization of this work, a number of ML models were developed and compared, including: LR, SVM, RF, and GB. Additionally, a stacked ensemble model was constructed by combining the best performing classifiers using a meta-classifier (which is LR) trained on their output probabilities. Hyperparameter tuning was carried out using GridSearchCV with stratified 5-fold cross-validation to keep the class distribution in each fold. For each model, a number of hyperparameters

(e.g., number of trees, maximum depth, learning rate, and regularization terms) was tested, and the best configuration was selected based on mean receiver operating characteristic (ROC) area under the curve (AUC) across folds.

For reproducibility, the final hyperparameter settings for heart disease detection and mortality prediction models used in the reported experiments are summarized in Table 4. Each model is then evaluated based on metrics and measurements such as accuracy, precision, recall, and ROC AUC, which are used for binary classification models. To enhance interpretability, SHAP was applied to the best-performing models. SHAP values were computed to identify feature importance and the direction of feature influence for both individual predictions and overall model behavior.

Table 4. Final hyperparameter settings for heart disease detection and mortality prediction models

| Model | Hyperparameter | Value | Description |
|---|---|---|---|
| LR | penalty | L2 | Regularization type |
| | C | 1.0 | Inverse of regularization strength |
| | solver | lbfgs | Optimization algorithm |
| | max_iter | 5000 | Maximum iterations for convergence |
| | random_state | 42 | Seed for reproducibility |
| RF | n_estimators | CVD detection: 200<br>Mortality prediction: 100 | Number of trees in the forest |
| | max_depth | None | Fully grown trees (no limit) |
| | criterion | gini | Split quality criterion |
| | min_samples_split | 2 | Minimum samples required to split |
| | min_samples_leaf | 1 | Minimum samples per leaf |
| | max_features | sqrt | Number of features to consider per split |
| | bootstrap | True | Sampling with replacement |
| | random_state | 42 | Seed for reproducibility |
| GB | n_estimators | CVD detection: 200<br>Mortality prediction: 100 | Number of boosting stages |
| | learning_rate | 0.05 | Shrinkage factor for each stage |
| | max_depth | 3 | Depth of individual weak learners |
| | subsample | 1.0 | Fraction of samples used per iteration |
| | loss | log_loss | Loss function for binary classification |
| | random_state | 42 | Seed for reproducibility |

### 3.3. Temporal feature ablation

The “time” variable in the heart failure dataset, which represents the follow-up duration in days, is highly correlated with the DEATH_EVENT because deceased patients often have shorter observation periods. From one perspective, if the goal is to predict death or survival of a patient, then time should probably not be used as an input to the model. On the other hand, the “time” variable can encode quite useful information by extracting features from it.

To assess the effect of temporal information on mortality prediction, and to minimize potential information leakage, an ablation study was conducted using the “time” variable from the heart failure clinical records dataset (patients who survive longer naturally have higher “time” values). Three options were examined to quantify the influence of temporal features on model performance:

– Without time: models were trained using the clinical features excluding “time”.
– With time: the raw “time” variable was included as an additional feature.
– With derived time features: the raw “time” value is transformed using logarithmic and categorical transformations (log(time)), and bins representing short, medium, and long follow-up durations were incorporated to capture temporal dynamics.

All models were trained using the same preprocessing pipeline, hyperparameters, and RF configuration to ensure comparability. The ROC-AUC was used as the primary evaluation metric.

### 3.4. External validation

To assess generalizability, an independent external dataset was used for validation. The UCI heart disease dataset (n =920) [28] was selected as it contains comparable clinical predictors such as age, sex, cholesterol, resting blood pressure, fasting blood sugar, exercise angina, and maximum heart rate. The target variable was converted to a binary indicator. Feature mappings were aligned with the main Kaggle dataset, and missing values were handled by median imputation. The RestingECG feature was entirely missing in the UCI dataset, hence, a neutral baseline value (0= normal ECG) was assigned to maintain compatibility with the trained models. The data were scaled using the same normalization parameters, which were derived from

the main dataset to ensure consistency. All models were retrained on the main dataset and evaluated without further tuning on the external dataset to measure true performance.

### 3.5. Full pipeline for heart disease and mortality prediction

Figure 1 shows the end-to-end preprocessing and evaluation pipeline for heart disease and mortality prediction. Raw datasets are cleaned, encoded, imputed where necessary, and standardized. Models are trained using stratified train/test splits and subsequently evaluated on both the internal test partition and an independent UCI dataset to assess external generalization.

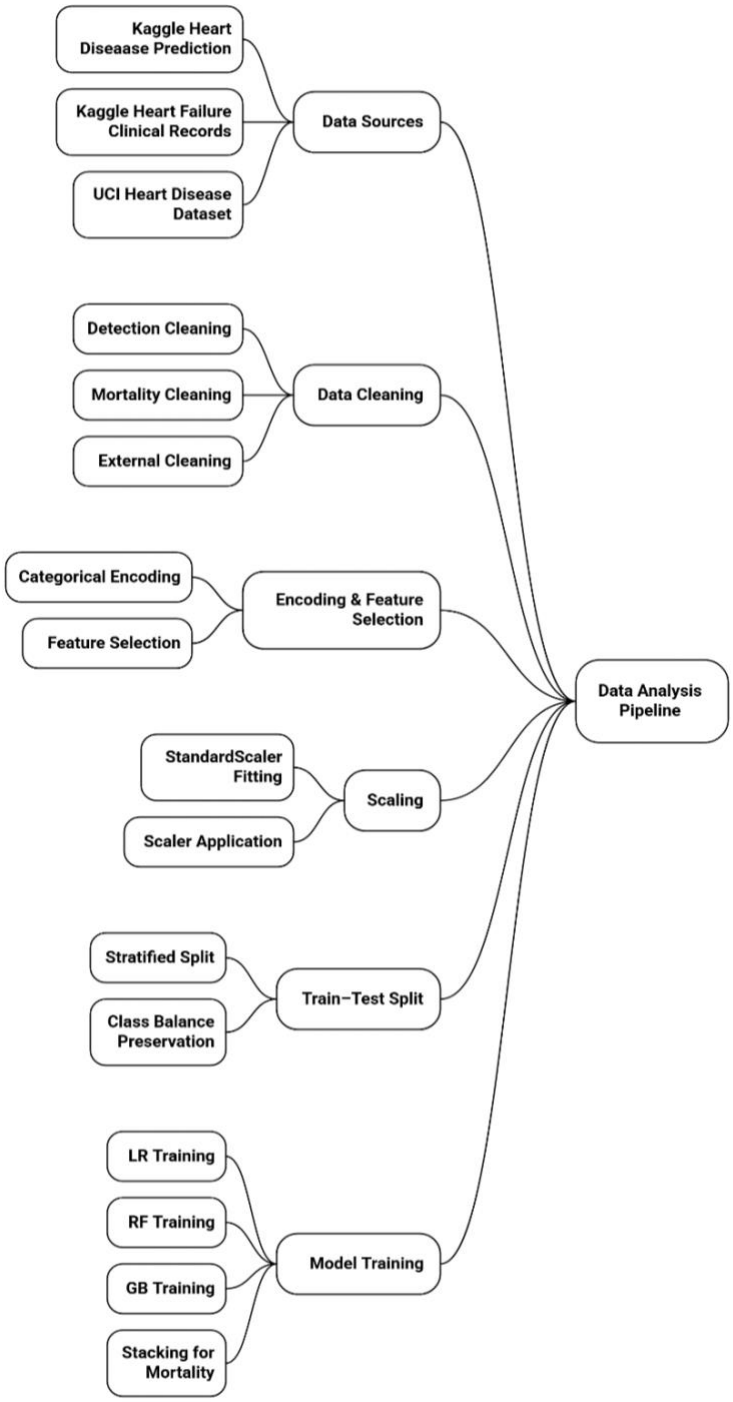



Figure 1. Preprocessing and external validation pipeline for heart disease and mortality prediction

The following steps summarize the end-to-end preprocessing and evaluation pipeline for heart disease and mortality prediction:

i) Data sources
   – Kaggle heart disease prediction (CVD detection) [26].
   – Kaggle heart failure clinical records (mortality) [27].
   – UCI heart disease dataset (external validation) [28].

ii) Data cleaning
   – For the CVD detection dataset, remove RestingBP =0, Cholesterol =0; type checks.
   – For the mortality dataset, range checks, type casting; no rows removed.
   – For the external dataset, handle NaNs: median imputation; RestingECG to "normal".

iii) Encoding and feature selection
   – Encode categorical variables (sex, ChestPainType, RestingECG, ExerciseAngina, ST_Slope).
   – Select shared clinical features (age, sex, BP, cholesterol, MaxHR, Oldpeak, FastingBS, RestingECG, ExerciseAngina).
   – For mortality prediction, retain all 12 predictors.

iv) Scaling
   – Fit StandardScaler on the main training set.
   – Apply the same scaler to internal test and external UCI data.

v) Train/test split: stratified train/test split (80/20) and preserved class balance.

vi) Model training: train LR, RF, GB (and stacking for mortality).

vii) Evaluation
   – Internal evaluation (main dataset): ROC AUC, accuracy, precision, recall, and F1-score.
   – External evaluation (UCI): same metrics, compare generalization.

## 4. IMPLEMENTATION

### 4.1. Dataset exploratory data analysis

EDA was performed on both datasets to get a better idea about the dataset before developing the models. This includes feature correlation heatmap, pairplots, and feature importance analysis. Figure 2 shows feature correlation heatmap for heart disease prediction dataset. From Figure 2, it can be seen that the strongest correlations with HeartDisease feature are: i) ExerciseAngina: positive correlation, ii) Oldpeak: positive correlation, iii) ST_Slope: negative correlation, and iv) ChestPainType: negative correlation. Figure 3 shows pairplot for heart disease prediction dataset. From this figure, the following key observations are obtained: i) age: patients with heart disease appear slightly older on average, ii) cholesterol: no clear visual separation between the classes, iii) MaxHR: generally lower in heart disease patients, iv) oldpeak: clearly higher in patients with heart disease, and v) RestingBP: overlapping distributions, so it may not be useful before feature engineering.

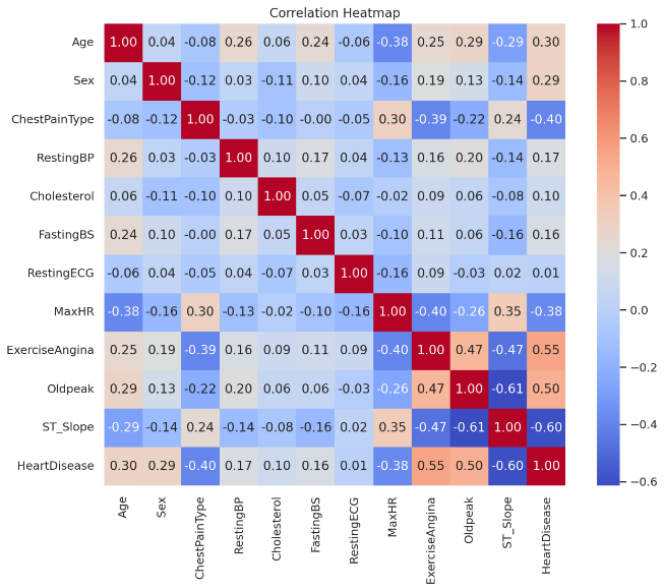



Figure 2. Feature correlation heatmap for heart disease prediction dataset


Figure 3. Pairplot for heart disease prediction dataset

Figure 4 shows feature importance analysis bar chart. From Figure 4, it can be seen that the top 5 most important features for heart disease prediction are as listed in Table 5 with their percentage importance. From Figure 4 and Table 5, it can be concluded that:

- ST_Slope and oldpeak features (related to ST segment of the ECG during exercise) are important indicators.
- Chest pain and angina also play a significant role in heart disease prediction
- Heart rate (MaxHR) is also an important feature, while basic features such as age, cholesterol, or RestingBP contribute less.

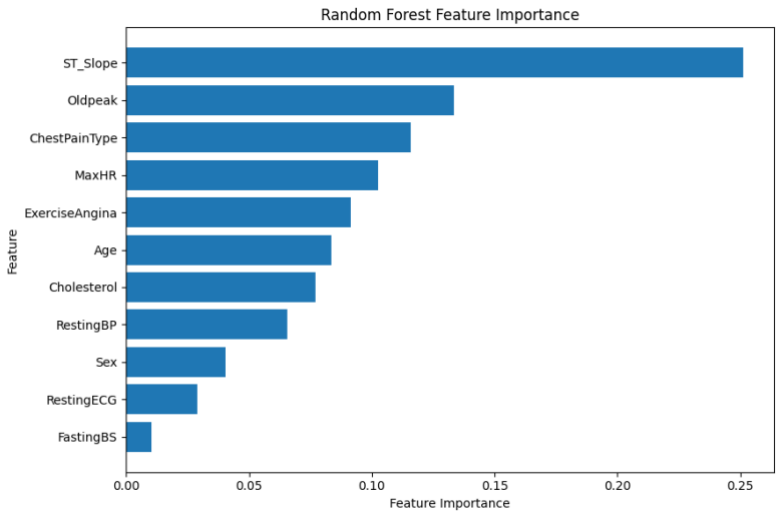



Figure 4. Feature importance analysis for heart disease prediction

Table 5. Top 5 most important features for heart disease prediction dataset

| Rank | Feature | Importance (%) |
|---|---|---|
| 1 | ST_Slope | 25.12 |
| 2 | Oldpeak | 13.33 |
| 3 | ChestPainType | 11.57 |
| 4 | MaxHR | 10.24 |
| 5 | ExerciseAngina | 9.13 |

Figure 5 shows the feature correlation heatmap for mortality prediction from the heart disease dataset. From Figure 5, it can be seen that the strongest correlations with DEATH_EVENT target variable are i) serum_creatinine: positive correlation, ii) age: positive correlation, iii) time: negative correlation, and iv) ejection_fraction: negative correlation. Figure 6 shows the feature importance analysis bar chart. From Figure 6, it can be seen that the top 5 most important features for heart disease mortality prediction are as listed in Table 6 with their percentage importance. From Figure 6 and Table 6, it can be concluded that:

– The time feature dominates the prediction, with patients who survive longer are less likely to die.
– Kidney related features (serum_creatinine, creatinine_phosphokinase) are strong clinical indicators.
– Platelets also influence risk.

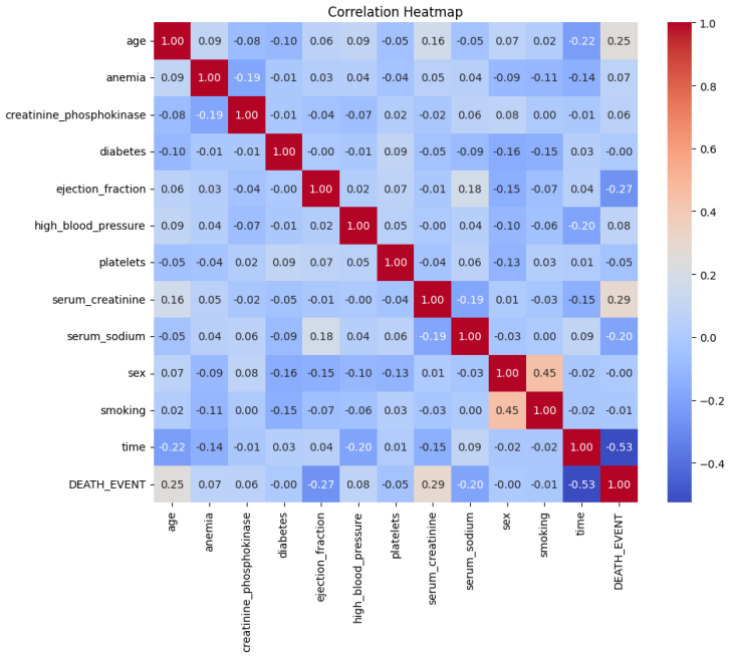



Figure 5. Feature correlation heatmap for heart disease mortality prediction dataset

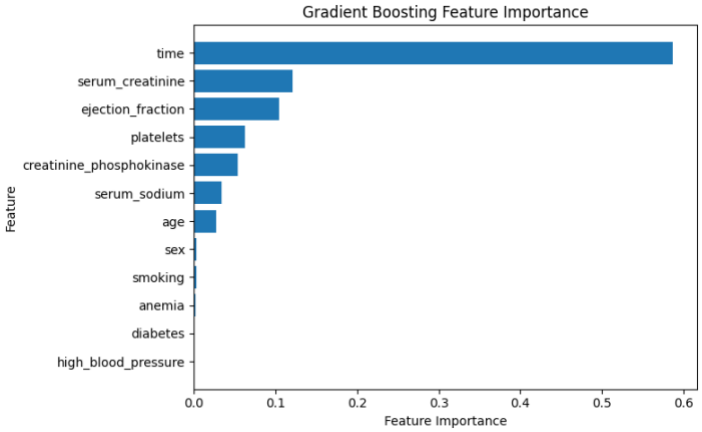



Figure 6. Feature importance analysis for mortality prediction

Table 6. Top 5 most important features for heart mortality prediction

| Rank | Feature | Importance (%) |
|---|---|---|
| 1 | time (follow-up days) | 58.69 |
| 2 | serum_creatinine | 12.13 |
| 3 | ejection_fraction | 10.39 |
| 4 | platelets | 6.23 |
| 5 | creatinine_phosphokinase | 5.41 |

### 4.2. Model implementation

In this work, Python 3 programming language was used in the Google Colab and Jupyter Notebook development environments to implement the suggested models. Moreover, the following Python libraries were utilized in the development:

– Scikit-learn library: used for training and assessing models.
– SHAP library: used to make models interpretable and explainable.
– The pandas and NumPy libraries are used to manipulate and preprocess data.
– The matplotlib and seaborn libraries are used to visualize model performance, SHAP values, and feature importance.

The dataset was split into 80:20 ratio for training and testing sets. RF feature importance was first computed based on Gini impurity for heart disease prediction. Then, gradient boost feature importance was computed for heart disease mortality prediction. To interpret the models and ensure clinical relevance, SHAP analysis was applied across all models to provide a consistent explanation of feature contributions at both global and local (patient-specific) levels. Comparative SHAP plots have shown differences in how LR, RF, and GB utilize clinical features (e.g., ejection fraction, serum creatinine, and age).

### 4.3. Application deployment

The best performing model from each stage was deployed as interactive web applications using Streamlit. The applications allow clinicians to input patient data and receive immediate heart disease and mortality risk predictions. These applications enhance accessibility and provide real-time decision support in healthcare clinics.

## 5. RESULTS AND DISCUSSION

### 5.1. Results of heart disease prediction

Figure 7 shows the ROC curve for the LR model, which was used as the baseline. The obtained AUC was 0.93 in this case. The resulting AUC score indicates excellent discrimination between heart disease and non-disease cases.

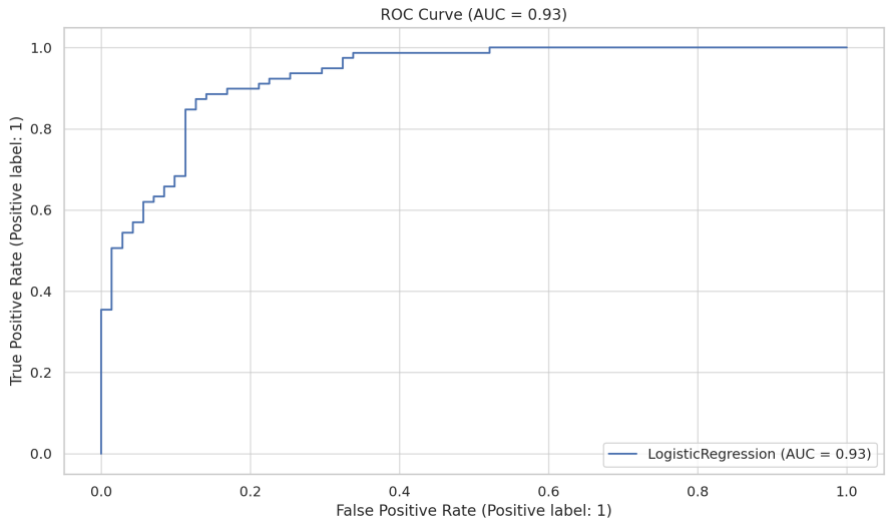



Figure 7. ROC curve for LR model

Table 7 shows the performance of the four models that were developed for this stage. From the table, it can be seen that the GB model performs the best overall. The RF model is also strong and often

easier to interpret. The SVM model underperforms significantly in this setup. Accordingly, the GB model is recommended for deployment or further tuning.

Table 7. Model performance comparison for heart disease prediction

| Model | ROC AUC | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| GB | 0.958 | 0.893 | 0.932 | 0.861 | 0.895 |
| RF | 0.953 | 0.887 | 0.931 | 0.848 | 0.887 |
| LR | 0.930 | 0.867 | 0.893 | 0.848 | 0.870 |
| SVM | 0.721 | 0.673 | 0.727 | 0.608 | 0.662 |

### 5.2. Results of mortality prediction due to heart disease

For this stage, three models were tested. These are GB, LR, and RF. The SVM was excluded in this part because it is likely to underperform similar to the previous part. Table 8 provides a summary of the performance comparison of the three tested models, and Figure 8 shows the ROC curves for the corresponding developed models.

Table 8. Model performance comparison for mortality prediction due to heart disease

| Model | ROC AUC | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| RF | 0.899 | 0.833 | 0.846 | 0.579 | 0.688 |
| LR | 0.855 | 0.800 | 0.733 | 0.579 | 0.647 |
| GB | 0.827 | 0.800 | 0.706 | 0.632 | 0.667 |

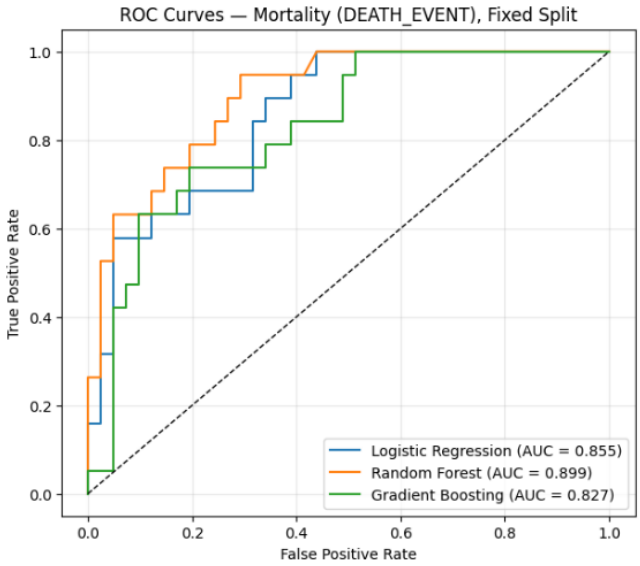



Figure 8. ROC curves for LR, RF, and GB models for mortality prediction

From the Table 8, it can be seen that the RF model has the highest ROC AUC, which indicates it is the best at separating deceased from survivor patients, and achieving the highest accuracy, precision, and F1-score. LR model has shown a good performance, with good ROC AUC and accuracy. GB performs reasonably in terms of ROC AUC and accuracy, with the highest obtained recall value. If interpretability matters most, then LR is recommended, and if precision (avoiding false alarms) matters most, the RF model is recommended. However, if recall (catching all true deaths) matters most, then GB could be adopted.

### 5.3. Model tuning results for mortality prediction

From the previous results of the models predicting mortality due to heart disease, it can be seen that the performance of these models needs to be improved. Since this is a mortality prediction task, recall (i.e., sensitivity) is more critical than precision, so it is better to flag more patients as at risk than to miss a real case. This can be carried out either by model tuning or using a stacked ensemble model.

To push the recall up, probability threshold could be changed from 0.5 to a lower value in the range of 0.1-0.4. A number of tests were carried out on the three models. Table 9 shows the results of the RF model tuning with two threshold values, 0.3 and 0.1. From Table 9, it can be seen that the RF model is very strong at capturing deaths when lowering the threshold (recall increased to 95%), but at the cost of many false positives. A threshold around 0.3 balances the recall and precision better.

Table 9. Threshold tuning for RF model

| Metric | Best F1-score (threshold =0.3) | Best recall (threshold =0.1) |
|---|---|---|
| Accuracy | 0.817 | 0.683 |
| Precision | 0.682 | 0.500 |
| Recall | 0.789 | 0.947 |
| F1-score | 0.732 | 0.655 |

Table 10 shows the results of the LR model threshold tuning with two values, 0.4 and 0.1. From the Table 10, it can be seen that the LR model achieves the highest accuracy at threshold 0.4, with a good balance with precision. At lower thresholds, recall improves significantly, but precision drops.

Table 10. Threshold tuning for LR model

| Metric | Best F1-score (threshold =0.4) | Best recall (threshold =0.1) |
|---|---|---|
| Accuracy | 0.833 | 0.683 |
| Precision | 0.765 | 0.500 |
| Recall | 0.684 | 0.895 |
| F1-score | 0.722 | 0.642 |

Table 11 shows the results of the GB model tuning with two threshold values, 0.2 and 0.1. From Table 11, it can be seen that the GB model provides the most balanced trade off, with both recall and precision around 0.74 at threshold 0.2. It is the most stable model, with the best overall balance (precision ≈ recall ≈ F1-score), so it is the most consistent model. GB gives reliable results compared to the RF, where the recall is highly affected by low thresholds, and the LR, where the performance varies with threshold value. Moreover, GB handles nonlinear relationships and feature interactions well, which is important in medical datasets where risk factors correlate.

Table 11. Threshold tuning for GB model

| Metric | Best F1-score (threshold =0.2) | Best recall (threshold =0.1) |
|---|---|---|
| Accuracy | 0.833 | 0.767 |
| Precision | 0.737 | 0.609 |
| Recall | 0.737 | 0.737 |
| F1-score | 0.737 | 0.667 |

### 5.4. Stacked ensemble results for mortality prediction

To improve prediction accuracy, outputs from the previously used models were combined in a stacked ensemble architecture, where predictions from LR, RF, and GB used as inputs to a meta classifier (which is the LR). This approach improves the complementary strengths of each base model. Stack ensemble predictions were carried out using sklearn's StackingClassifier. The results were evaluated using cross-validation and test metrics. Moreover, soft voting was also used as an additional model, which averages the probability outputs from LR, RF, and GB and then makes the final prediction decision based on the average probabilities. Table 12 provides a summary of the performance comparison of the five models, and Figure 9 shows the corresponding ROC curves.

Table 12. Stacked ensemble results

| Model | ROC AUC | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| RF | 0.899 | 0.833 | 0.846 | 0.579 | 0.688 |
| Stacked ensemble | 0.877 | 0.817 | 0.750 | 0.632 | 0.686 |
| Soft voting | 0.870 | 0.817 | 0.750 | 0.632 | 0.686 |
| LR | 0.855 | 0.800 | 0.733 | 0.579 | 0.647 |
| GB | 0.827 | 0.800 | 0.706 | 0.632 | 0.667 |

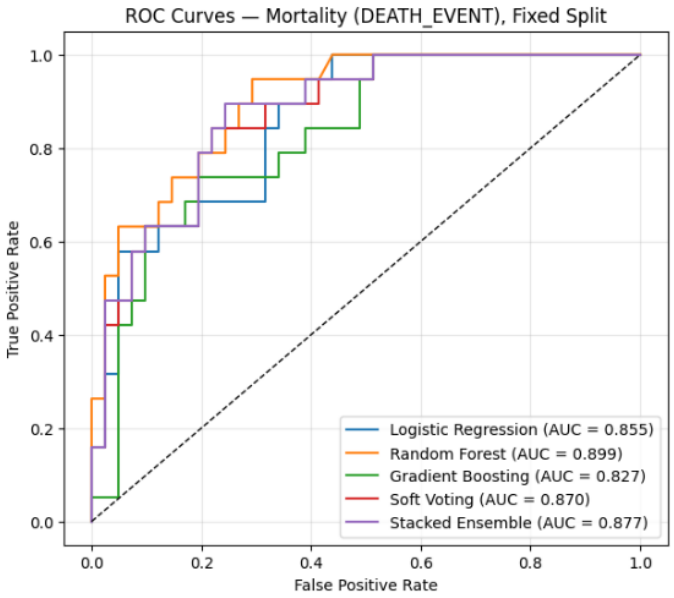



Figure 9. ROC curves for LR, RF, GB, soft voting and stacked ensemble models for mortality prediction

From Table 12, it can be seen that the stacked ensemble model performs well in recall (0.632), which is better than the RF model, but with lower precision (0.750). ROC AUC (0.877) is slightly below RF but still strong, so it sacrifices a little ROC AUC for better recall and slightly more balanced F1. RF still gives the strongest discrimination (highest ROC AUC), but suffers from moderate recall, so it misses some deaths. GB underperforms both, so it is no longer the best model, but it still has the best recall along with the stacked ensemble model. If the goal is best overall classification power (discrimination), then RF is the right choice. While if it is required to capture more true deaths (recall) while keeping decent precision and F1-score, then stacked ensemble model is a better choice. Accordingly, GB should not be the choice for deployment in this case because of its weak ROC AUC and F1-score compared to other models. As a conclusion, the stacked ensemble could be deployed if clinical recall is a priority, but the RF is kept as a benchmark model, especially if threshold tuning could be used to boost recall.

### 5.5. Explainability results

As mentioned before, SHAP was used for explainability. For LR, RF, and GB, SHAP analysis consistently illustrated that time, serum_creatinine, and ejection_fraction are the dominant predictors of mortality, with older age and lower serum sodium also contributing to increased risk. LR exhibited smooth effects consistent with its linear formulation, whereas RF and GB captured more complex, nonlinear interactions and heterogeneous effects across patients. Importantly, the direction of feature influence (e.g., high serum creatinine and low ejection fraction increasing risk) aligned well with established clinical knowledge, reinforcing trust in the models' internal reasoning. Additional SHAP visualizations and comparative analyses are provided as follows.

#### 5.5.1. SHAP analysis details

For mortality prediction, SHAP plot shows how much each feature contributes to the prediction of death (1) or survival (0). Positive SHAP values mean the prediction is moving toward a death event, while negative SHAP values mean the prediction is moving toward survival. The x-axis of the plot (SHAP value) shows the impact of a feature on the model output. Red color in the SHAP plot means high value of the feature, while blue means low value of the feature. Features are sorted from top to bottom by their average contribution to model predictions. SHAP plots were generated for LR, RF, and GB models for the mortality prediction task.

Figure 10 shows the generated SHAP plot for the RF model predictions. From Figure 10, it can be seen that time and serum_creatinine are the most important predictors. On the other hand, sex and smoking are the least influential. The RF model is strongly driven by time since diagnosis, kidney function (serum_creatinine), heart performance (ejection_fraction), and age. These are known clinical predictors of survival in heart failure patients. Features like smoking, sex, anemia are not very predictive in this dataset.

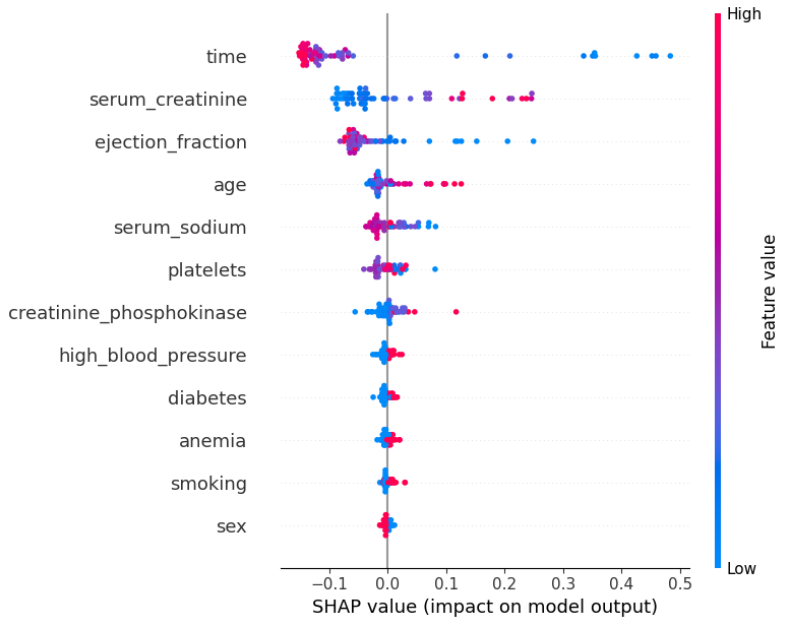



Figure 10. SHAP plot for the RF model

Figure 11 shows the generated SHAP plot for the LR model predictions. In the LR model, SHAP analysis shows that time, ejection_fraction, serum_creatinine, and age were the most important predictors of heart failure mortality. Shorter follow up times (low values of time) and reduced cardiac pumping efficiency (low ejection_fraction) increased predicted death risk, while longer follow up periods and higher ejection fractions are associated with reduced risk. Similarly, older age and elevated serum_creatinine, which is related to problems with kidney function, have led the model to predict higher mortality probability. The remaining variables, such as diabetes, anemia, and smoking, showed comparatively fewer effects, so they have limited predictive contribution in this dataset. Accordingly, the direction and magnitude of these SHAP values align closely with the available clinical evidence, which reinforces model interpretability and reliability for medical decision support.

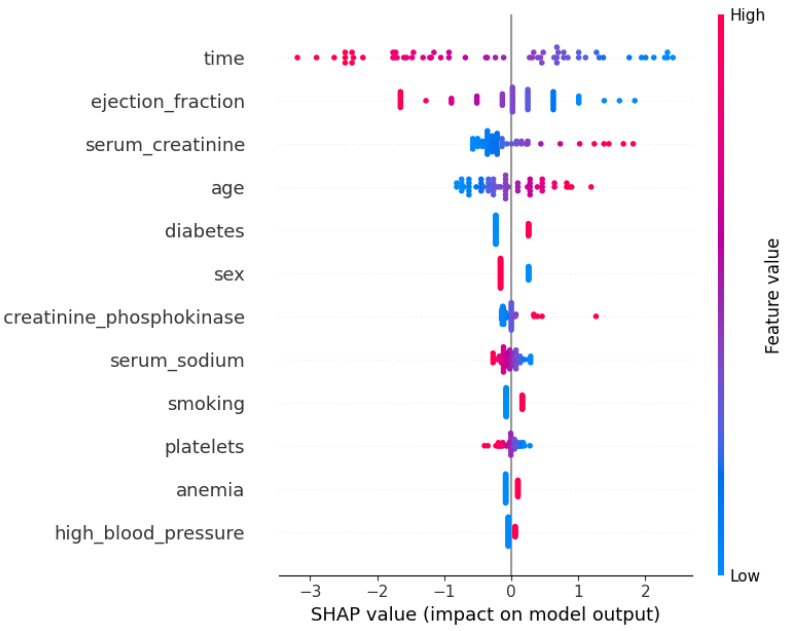



Figure 11. SHAP plot for the LR model

Figure 12 shows the generated SHAP plot for GB model predictions. The SHAP summary plot for the GB model shows that time, serum_creatinine, and ejection_fraction are the most important predictors of mortality in the heart failure dataset. Shorter follow up times (high values in pink) are strongly associated with increased mortality risk, while longer durations (low values in blue) have a protective effect. High serum_creatinine levels, which are linked to issues in kidney function, contribute to higher predicted risk, while lower levels are associated with the opposite. Also, higher ejection_fraction values, which are an indication of better cardiac performance, are linked to reduced mortality risk, while lower values lead to increased risk. The other features, such as creatinine_phosphokinase, age, and platelets, contribute to the prediction but with less effect. The clear separation and distribution of the SHAP values are aligned with the known clinical information of heart failure, which confirms that the GB model effectively captures meaningful relationships between biological features and mortality.

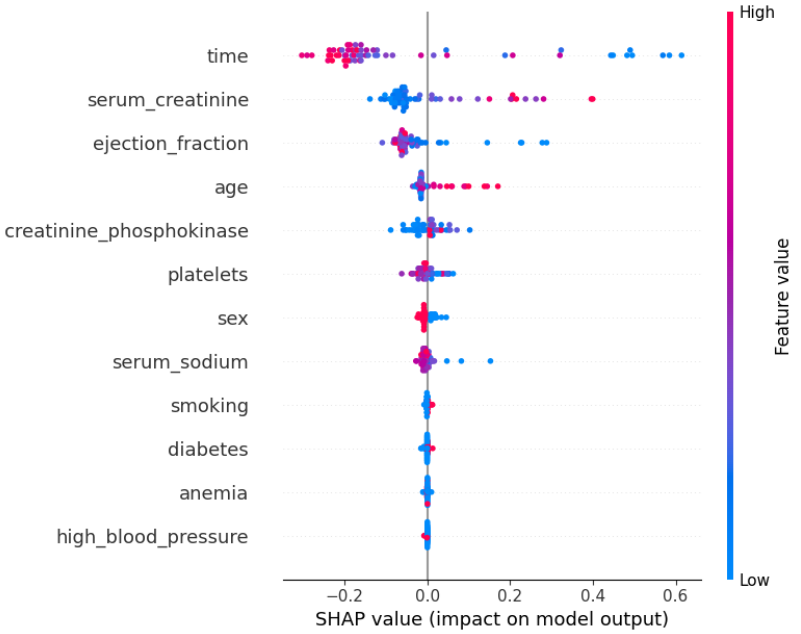



Figure 12. SHAP plot for GB model

A comparative analysis of the SHAP summary plots for RF, LR, and GB reveals a number of model specific differences in how features are prioritized and their directional influence on mortality prediction in heart failure patients. All three models agree on the core importance of time, ejection_fraction, and serum_creatinine, but their relative rankings and the magnitude of SHAP values vary. RF, which is a nonlinear and ensemble model, captures complex interactions, with time, serum_creatinine, and ejection_fraction are the main prediction drivers, but the spread in SHAP values suggests more variability depending on local feature contributions and data partitions, and secondary variables such as age, serum_sodium and platelets have slightly greater influence compared to LR and GB models. However, this comes at the cost of interpretability, as the contribution patterns appear less sharp and more scattered.

The LR SHAP plot shows that the time and ejection_fraction distribute more evenly, with larger and more symmetric SHAP spreads to reflect its linear nature, where feature effects are more constant in the predictor range. LR also shows relatively smoother contributions and importance from secondary predictors, such as age, diabetes, and sex, compared to RF or GB, which enhances interpretability by producing more globally consistent effects. GB produces the most concentrated and sharp importance structure, with time being the strongest and most consistent positive or negative shift in predicted risk, followed by serum_creatinine, while ejection_fraction plays a clear protective role when high. High serum_creatinine and low ejection_fraction are strongly associated with increased mortality risk, while longer survival time has a strong protective effect. Secondary variables such as age, platelets, and sex have more importance in GB than in LR or RF, which suggests that the GB model better captures hidden but consistent effects. Although LR offers simple interpretations and RF spreads importance across variables, GB balances predictive sharpness with interpretability, with strong emphasis on the main three clinical risk drivers.

As a final investigation, a SHAP plot was generated for the stacked ensemble model. This is provided in Figure 13. From the Figure 13, it can be seen that the RF model has the most influence on the output prediction of the stacked ensemble model. It shows the widest SHAP distribution (largest effect on ensemble predictions). LR comes as the next most influential, with consistent and balanced influence. The GB model has an impact but is less weighted compared to both RF and LR. Accordingly, the stacked ensemble model is following the RF model mostly, then LR, with GB as an additional check. This is important to reduce bias from any single model and ensure robustness. Overall, the LR model offers more interpretable and global feature effects, while RF has the ability to model complex interactions but with slightly more spread of importance patterns, and GB captures hidden interactions with high predictive sharpness. This shows that while all models capture similar clinical risk drivers, their internal mechanisms lead to different emphases and interpretability profiles.

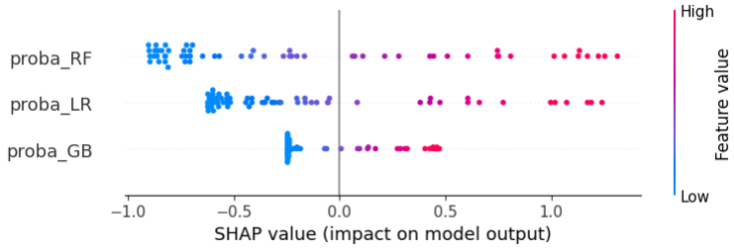



Figure 13. SHAP plot of the stacked ensemble model

### 5.6. External validation results

Table 13 summarizes the external validation performance using the UCI dataset. The RF model achieved the highest external ROC AUC (0.944), followed by GB (0.886) and LR (0.806). In spite of differences in data sources and missing feature patterns, the models retained strong discrimination ability, with only little degradation from internal validation. These results confirm the robustness of the developed hybrid framework and its ability to generalize across independent datasets. The RF model demonstrated superior transfer performance, which makes it the most suitable candidate for deployment in the clinical web application.

The results provided in Table 13 demonstrate strong overall performance of the three models for the internal and external datasets. LR maintained stable but lower recall on the external data, which proves its limited ability to model complex nonlinear dependencies across features. On the other hand, both ensemble models, RF and GB, resulted in excellent generalization capability when evaluated on the independent UCI dataset. The RF model achieved the highest external accuracy (0.864), precision (0.951), and ROC AUC (0.944), which confirms its robustness and discriminative power for heart disease prediction. The slight performance degradation between internal and external evaluations validates the reproducibility and clinical applicability of the proposed modeling framework. Overall, these findings indicate that the ensemble approach provides the best trade-off between predictive accuracy, stability, and generalizability, making it the most suitable candidate for deployment within the explainable hybrid system.

Table 13. Performance comparison between internal (Kaggle heart disease) and external (UCI heart disease) datasets

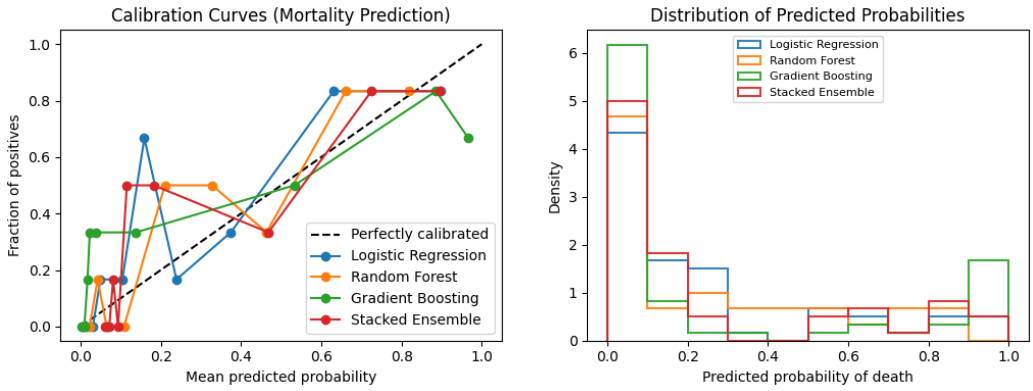
| Model | Dataset | Accuracy | Precision | Recall | F1-score | ROC AUC |
|---|---|---|---|---|---|---|
| LR | Internal (Main) | 0.853 | 0.917 | 0.764 | 0.833 | 0.907 |
| | External (UCI) | 0.725 | 0.830 | 0.633 | 0.718 | 0.806 |
| RF | Internal (Main) | 0.840 | 0.900 | 0.750 | 0.818 | 0.898 |
| | External (UCI) | 0.864 | 0.951 | 0.796 | 0.866 | 0.944 |
| GB | Internal (Main) | 0.847 | 0.889 | 0.778 | 0.830 | 0.899 |
| | External (UCI) | 0.812 | 0.908 | 0.735 | 0.812 | 0.886 |

### 5.7. Ablation study results

To assess the contribution of each component within the hybrid framework, two ablation scenarios were tested: i) models trained without the stacking ensemble layer and ii) models evaluated without SHAP-based interpretability integration. Removing the stacking layer led to a reduction in mean ROC AUC from 0.91 to 0.88, which confirms the ensemble's additive value in improving discrimination ability. Excluding SHAP integration did not affect classification accuracy but eliminated model transparency and feature analysis, which confirms SHAP's essential role in explainability rather than predictive performance. These results indicate that both components enhance the overall system's performance and interpretability.

### 5.8. Calibration plots result to evaluate probability reliability

Figure 14 shows the calibration curves and predicted probability distributions for the four mortality prediction models used in this work. Calibration curves as shown in Figure 14(a) provide analysis to assess how well each model predicted mortality based on the probabilities that correspond to actual outcomes. The diagonal dashed line represents perfect calibration, where predicted and observed event rates are identical. It can be seen that the RF demonstrates the best alignment with the diagonal across all probabilities, which confirms its superior reliability and well-calibrated probability estimates. The LR model shows reasonable calibration but tends to underpredict mortality risk in low-probability regions. GB shows slight overconfidence between 0.2 and 0.4 probability ranges, while the stacked ensemble offers smoother calibration and reduces extreme deviations, which suggests improved overall stability through model aggregation. The probability histograms as shown in Figure 14(b) further show that most patients were assigned low predicted risks (0-0.2), with a small subset in the high-risk range (0.8-1.0). RF and stacked ensemble models produced more polarized distributions, which indicates stronger discriminative ability, while LR resulted in smoother and less extreme outputs with its linear assumptions.




(a) (b)

Figure 14. Evaluation of probability estimates for mortality prediction: (a) calibration curve of mortality prediction and (b) predicted probability distributions of mortality risk

Table 14 provides Brier score comparison. Lower Brier means better calibration and reliability. Accordingly, the RF model achieves the best probability calibration (lowest Brier), which is consistent with its near-diagonal reliability curve above and confirms its suitability for clinical decision support deployment where probability accuracy matters.

Table 14. Brier score comparison

| Model | Brier score | Interpretation |
|---|---|---|
| LR | 0.1448 | Good calibration, moderate discrimination |
| RF | 0.1277 | Best overall calibration and reliability |
| GB | 0.1634 | Slightly overconfident and less reliable |
| Stacked ensemble | 0.1385 | Improved stability over single models |

### 5.9. Effect of the "time" variable

The results of time ablation experiments are presented in Table 15. Incorporating the "time" variable led to a measurable improvement in discrimination, with the ROC-AUC increasing from 0.782 (without "time") to 0.839 (with "time"). The use of the derived time features has improved the ROC-AUC to 0.849, which suggests that transformed or discretized temporal representations can enhance prediction while reducing bias. However, because the "time" variable directly encodes survival duration in such a way that patients who survive longer naturally have higher "time" values, it introduces potential information leakage that could affect performance. To maintain clinical validity and model generalizability, the final deployed hybrid framework excluded the raw "time" variable. Derived temporal features such as log(time) or time bin indicators may be appropriate for future improvements of this work.

Table 15. Effect of the "time" variable on RF model performance

| Configuration | ROC AUC | Comment |
|---|---|---|
| Without time | 0.782 | Baseline model using clinical features only |
| With time | 0.839 | Improved discrimination due to temporal correlation |
| With derived time | 0.849 | Slightly higher performance, reduced leakage |

### 5.10. Web applications for heart disease and mortality predictions

Applications for the two stages of the work (heart disease and mortality predictions) were deployed using Streamlit apps. For the first part, the GB model was identified as the best performing algorithm because of its superior predictive accuracy and clinically meaningful feature explanations compared to LR and RF. Accordingly, a dedicated application was developed to convert the model into a practical tool for clinical decision making. The application offers real-time heart disease prediction for heart patients by allowing healthcare practitioners to input patient data and receive an immediate response. Figure 15 shows two examples of this app execution. Figure 15(a) illustrates an example of the heart disease prediction application where the entered patient profile is classified as "no disease", with a low predicted probability of heart disease. In contrast, Figure 15(b) shows an example for a different patient profile classified as disease present, with an elevated risk score. In both cases, clinicians can inspect the underlying feature inputs and model outputs interactively. The application requires 11 entries, which are the 11 features from the dataset, to make a final prediction about whether the patient has a CVD or not.

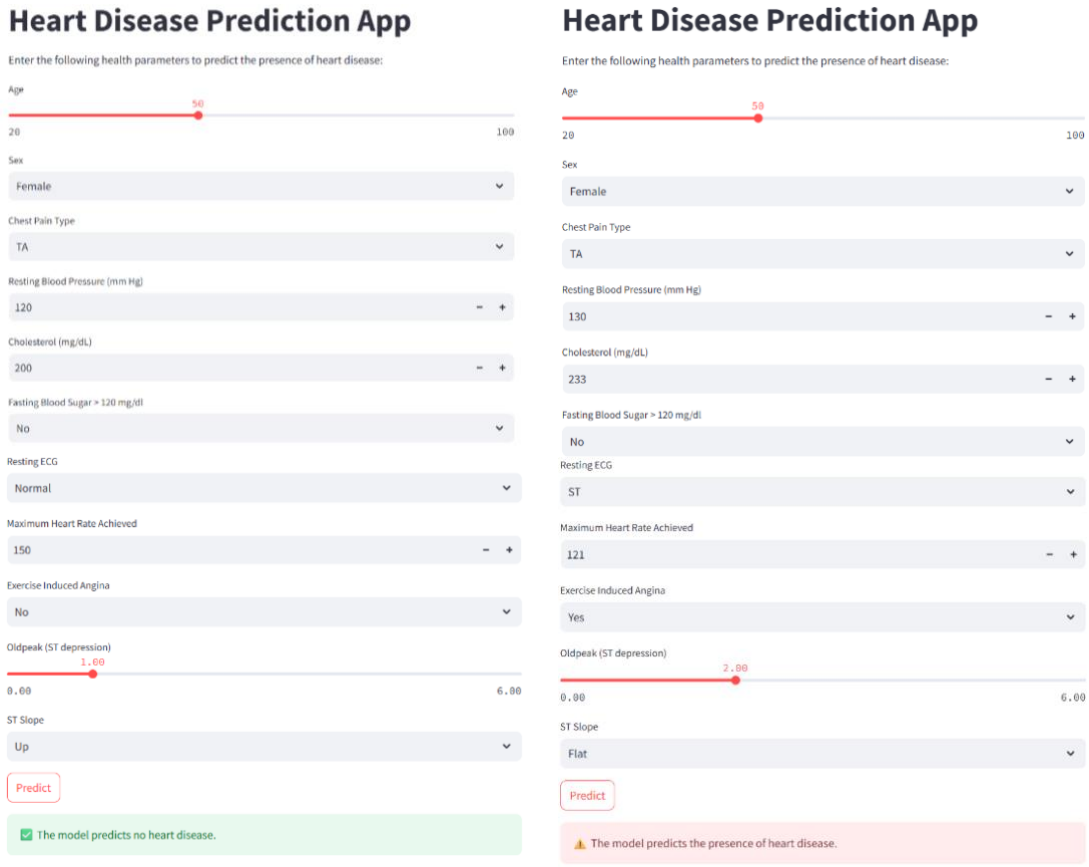
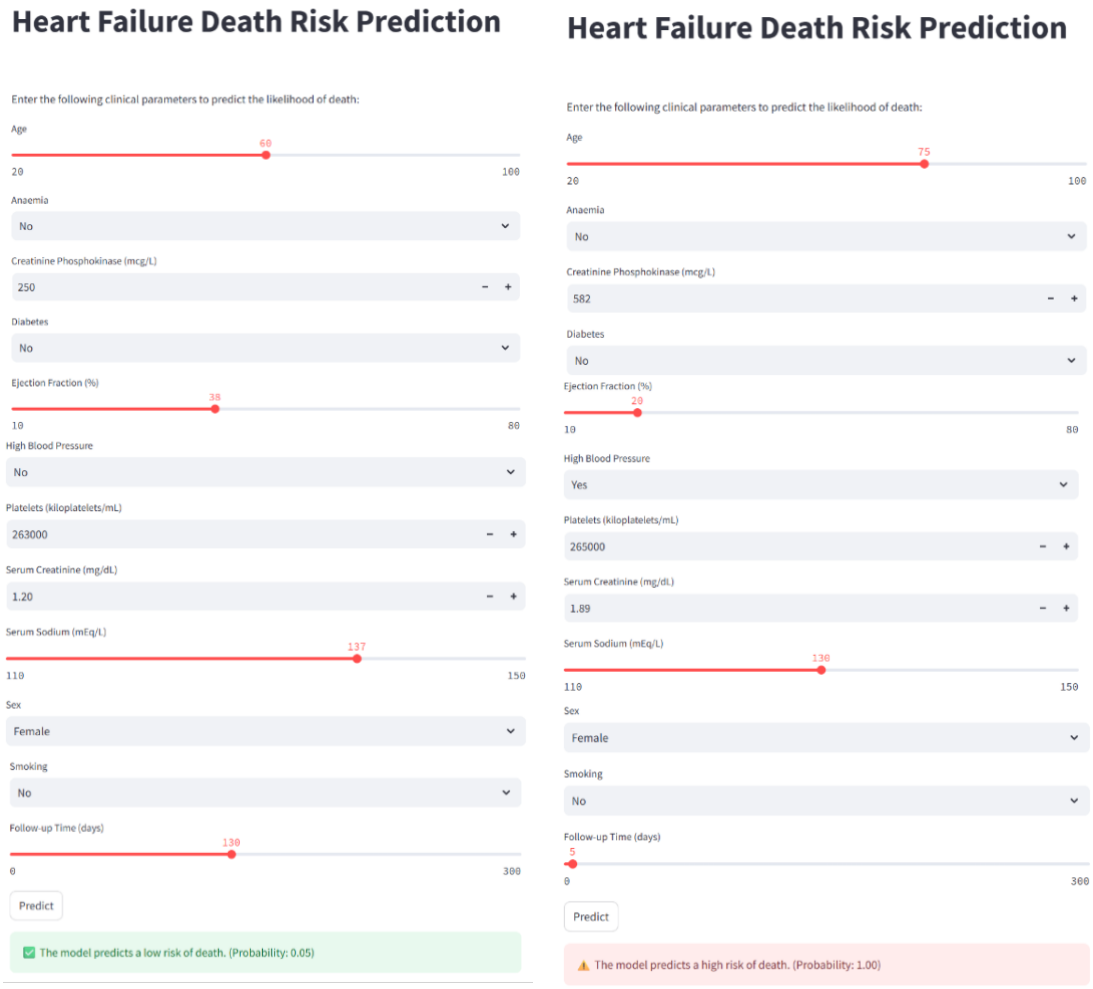



Figure 15. Heart disease prediction app execution examples: (a) "no disease" output prediction and (b) "disease" output prediction

For the second stage, following the model evaluation and SHAP interpretability analysis, and since the task involves predicting mortality, the stacked ensemble model was chosen because it utilizes the strong features of the three models and reduces bias from any single model while ensuring robustness. The application offers real-time mortality risk prediction for heart failure patients, allowing healthcare

practitioners to input patient data and receive an immediate mortality risk probability. Figure 16 shows two examples of mortality prediction application execution. Figure 16(a) shows an example of the application returning a low-risk output for a heart failure patient, whereas Figure 16(b) illustrates a high-risk prediction for a patient with more adverse clinical features. In both examples, the underlying probability estimate is complemented by feature explanations (via SHAP in the deployed application), which allows the clinicians to understand why the model recommends a low or high mortality risk. The application requires 12 entries, which are the 12 features from the dataset, to make a final prediction about whether the patient has a high mortality prediction due to heart disease or not.




(a) (b)

Figure 16. Mortality prediction app execution examples: (a) “low-risk” heart failure prediction output and (b) “high-risk” heart failure prediction output

### 5.11. Discussion

The results of this study demonstrate that ML models, particularly RF and GB, can achieve high predictive performance in both heart disease detection and mortality prediction. As a general observation, the performance of the models in this first stage is higher than that of the second stage. This could be due to the size of the dataset used.

The superior performance of the GB model in the first part over RF and LR models confirms its capacity to model nonlinear interactions and complex feature dependencies, which are often present in medical datasets. In the second part, RF model performed well in all metrics except for the recall value, which is important in clinical tasks where mortality is to be predicted. Threshold tuning improved recall for the three models, but the GB model provided the most balanced trade-off between recall and precision.

Moreover, the findings for the second stage prediction support using ensemble learning for robust CVD mortality risk modelling. Stacked ensemble balances predictive power and interpretability.

The added SHAP explanations increase clinician trust. The stacked ensemble model provides precision and recall advantages in high-risk cases. Combining model outputs with domain knowledge can result in hybrid decision-support systems. The SHAP interpretability analysis revealed serum creatinine and ejection fraction consistently appeared to be dominant predictors in mortality risk, which verifies existing clinical evidence. Model interpretability is important for clinical adoption, as clinicians must understand not only what the prediction is but also why the model reached it. By integrating both global feature rankings and patient attributes, this study bridges the gap between algorithmic decision making and clinical reasoning, which supports explainable artificial intelligence (XAI) in healthcare.

Given the clinical importance of identifying high-risk patients, particular emphasis was on recall (sensitivity) as a key evaluation metric. Among the tested models, GB achieved the highest recall (0.78) on the internal test set and 0.73 on external validation, indicating strong sensitivity to mortality cases. While RF slightly outperformed in precision, GB provided the best balance between sensitivity and specificity, which is crucial for minimizing false negatives in clinical contexts.

Compared to related work in literature, the proposed approach offers three main improvements:

- A two-stage predictive approach to address both disease presence and mortality risk, rather than focusing on a single stage.
- SHAP integration for mortality prediction due to CVDs to ensure model transparency.
- Deployment of interactive web applications based on the best performing models to provide real-time predictions in a user-friendly format for the clinicians.

Moreover, compared with recent studies on heart failure mortality prediction that achieve ROC AUC values between 0.80 and 0.92 using ensemble and deep learning approaches [11], [23]–[25], the proposed models deliver comparable predictive performance while further integrating XAI visualization and a fully deployable clinical decision-support pipeline.

The RF model shows the most consistent and generalizable performance across datasets, with the highest external ROC AUC (0.94) and F1-score (0.87), and confirms its strong transferability to unseen data. The external validation findings show the practical viability of the proposed models beyond a single dataset and strengthen the confidence in deployment within real-world clinical environments. While the results are promising, a number of limitations must be acknowledged. Dataset sizes, especially the one for mortality prediction, remain insufficient for deep learning approaches, which limits model generalizability. Moreover, the current models rely on structured (or numerical) clinical variables only, so incorporating unstructured data such as imaging, ECG waveforms, or clinicians' notes could enhance predictive power further. Future clinical trials are needed to evaluate the practical impact of the deployed applications on decision making and patient outcomes. Finally, in this study, the primary focus was on discrimination metrics (ROC AUC, accuracy, precision, recall, and F1-score). Calibration analysis (e.g., reliability curves and Brier scores) and statistical comparison of models (e.g., DeLong tests or paired comparisons across cross-validation folds) were not included due to time and space constraints but represent important next steps to evaluate the reliability of predicted probabilities and to assess whether observed performance differences are statistically significant. Table 16 summarizes the contribution of this work compared to similar work found in literature.

Table 16. Novel contributions of the proposed work compared to literature

| Aspect | Novel contribution |
|---|---|
| Dual-stage prediction | Few papers combine both heart disease detection and mortality prediction in a single pipeline. |
| Model performance | Solid ROC AUC (~ 0.88 - 0.93) across tasks |
| Stacked ensemble model | Many works compare classifiers, but fewer implement stacking and assess it against base learners. |
| SHAP explainability | The use of SHAP for individual models and ensemble components is a major strength. |
| Deployment | Streamlit apps make the work practical and reproducible, which is highly appreciated in applied ML journals. |

## 6. CONCLUSION

This study presents a two-stage ML framework for heart disease detection and mortality prediction along with model integration with SHAP interpretability to enhance clinical trust. It also introduces a stacked ensemble architecture and a deployment interface for real-world use. The proposed approach achieves high predictive accuracy and delivers transparent and patient-specific explanations with interactive web applications. With the robust performance, interpretability, and deployment, the proposed work offers a step toward implementing AI-driven decision support systems in healthcare. Future work will focus on expanding the dataset for more external validation, including unstructured data inputs, and evaluating the developed tools in real-world scenarios. This research promotes the potential of explainable and high-performance ML

models to improve cardiovascular care to enable earlier interventions and inform decision making to better improve patient outcomes.

## ACKNOWLEDGMENTS

The authors appreciate the support provided by the School of Engineering and Computing, American University of Ras Al Khaimah.

## FUNDING INFORMATION

The authors declare that this research received no external funding.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ali Al-Ataby | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | |
| Hussain Attia | ✓ | ✓ | | | ✓ | | | | ✓ | ✓ | ✓ | ✓ | | |

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

## DATA AVAILABILITY

All data used in this study are publicly available via Kaggle and UCI Machine Learning Repository, with the permission of the provider. The datasets are available at:

– https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
– https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data
– https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

## REFERENCES

[1] M. Di Cesare *et al.*, "The heart of the world," *Global Heart*, vol. 19, no. 1, 2024, doi: 10.5334/gh.1288.
[2] X. Sun, Y. Yin, Q. Yang, and T. Huo, "Artificial intelligence in cardiovascular diseases: diagnostic and therapeutic perspectives," *European Journal of Medical Research*, vol. 28, no. 1, 2023, doi: 10.1186/s40001-023-01065-y.
[3] K. Shahzadi, M. K. Abid, A. Naeem, M. Fuzail, N. Aslam, and R. Sajjad, "Predicting the risk of cardiovascular diseases using deep learning models," *Kashf Journal of Multidisciplinary Research*, vol. 2, no. 6, pp. 1–16, 2025, doi: 10.71146/kjmr475.
[4] S. Lolak, J. Attia, G. J. McKay, and A. Thakkinstian, "Comparing explainable machine learning approaches with traditional statistical methods for evaluating stroke risk models: retrospective cohort study," *JMIR Cardio*, vol. 7, 2023, doi: 10.2196/47736.
[5] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," *SN Computer Science*, vol. 1, 2020, doi: 10.1007/s42979-020-00365-y.
[6] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
[7] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, 2021, doi: 10.1088/1757-899x/1022/1/012072.
[8] K. Vayadande *et al.*, "Heart disease prediction using machine learning techniques: a survey for societal care and information," in *Techno-societal 2022*, Cham, Switzerland: Springer, 2023, pp. 337–347, doi: 10.1007/978-3-031-34644-6_37.
[9] R. Katarya and S. K. Meena, "Machine learning techniques for heart disease prediction: a comparative study and analysis," *Health and Technology*, vol. 11, no. 1, pp. 87–97, 2020, doi: 10.1007/s12553-020-00505-7.
[10] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques," *Algorithms*, vol. 16, no. 2, 2023, doi: 10.3390/a16020088.
[11] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison," *Computers in Biology and Medicine*, vol. 136, 2021, doi: 10.1016/j.compbiomed.2021.104672.

[12] M. I. Al-Janabi, M. H. Qutqut, and M. Hijjawi, "Machine learning classification techniques for heart disease prediction: a review," *International Journal of Engineering & Technology*, vol. 7, no. 4, pp. 5373–5379, 2018, doi: 10.14419/ijet.v7i4.28646.
[13] R. Yilmaz and F. H. Yağin, "Early detection of coronary heart disease based on machine learning methods," *Medical Records*, vol. 4, no. 1, pp. 1–6, 2022, doi: 10.37990/medr.1011924.
[14] N. A. Baghdadi, S. M. F. Abdelaliem, A. Malki, I. Gad, A. Ewis, and E. Atlam, "Advanced machine learning techniques for cardiovascular disease early detection and diagnosis," *Journal of Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00817-1.
[15] N. Nissa, S. Jamwal, and S. Mohammad, "Early detection of cardiovascular disease using machine learning techniques: an experimental study," *International Journal of Recent Technology and Engineering*, vol. 9, no. 3, pp. 635–641, 2020, doi: 10.35940/ijrte.c46570.99320.
[16] U. Nagavelli, D. Samanta, and P. Chakraborty, "Machine learning technology-based heart disease detection models," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–9, 2022, doi: 10.1155/2022/7351061.
[17] P. K. Pande, P. Khobragade, S. N. Ajani, and V. P. Uplanchiwar, "Early detection and prediction of heart disease with machine learning techniques," in *2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET)*, 2024, pp. 1–6, doi: 10.1109/icicet59348.2024.10616294.
[18] P. Chavda, H. Bhavsar, Y. Pithadia, and R. Kotecha, "Early detection of cardiac disease using machine learning," *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3370813.
[19] R. Katarya and P. Srinivas, "Predicting heart disease at early stages using machine learning: a survey," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020, pp. 302–305, doi: 10.1109/icesc48915.2020.9155586.
[20] R. Alizadehsani *et al.*, "Machine learning-based coronary artery disease diagnosis: a comprehensive review," *Computers in Biology and Medicine*, vol. 111, 2019, doi: 10.1016/j.compbiomed.2019.103346.
[21] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.
[22] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1589–1604, 2018, doi: 10.1109/jbhi.2017.2767063.
[23] M. E. A. Sourov, M. S. Hossen, P. Shaha, M. M. Hossain, and M. S. Iqbal, "An explainable AI-enhanced machine learning approach for cardiovascular disease detection and risk assessment," *arXiv:2507.11185*, 2023.
[24] K. K. Napa, R. Govindarajan, S. Sathya, J. S. Murugan, and B. K. P. Vijayammal, "Comparative analysis of explainable machine learning models for cardiovascular risk stratification using clinical data and Shapley additive explanations," *Intelligence-Based Medicine*, vol. 12, 2025, doi: 10.1016/j.ibmed.2025.100286.
[25] A. Bilal, A. Alzahrani, K. Almohammadi, M. Saleem, M. S. Farooq, and R. Sarwar, "Explainable AI-driven intelligent system for precision forecasting in cardiovascular disease," *Frontiers in Medicine*, vol. 12, 2025, doi: 10.3389/fmed.2025.1596335.
[26] F. S. Palacios, "Heart failure prediction dataset," *Kaggle*. Accessed: Jun. 10, 2025. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
[27] A. Maranhão, "Heart failure prediction," *Kaggle*. Accessed: Jun. 28, 2025. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data
[28] R. K. Sony, "UCI heart disease data," *Kaggle*. Accessed: Nov. 30, 2025. [Online]. Available: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

## APPENDIX

Table 1. Literature review summary table

| Reference | Work | Gap |
|---|---|---|
| Sun *et al.* [2] | Comprehensive review of CVD diagnosis using artificial intelligence (AI) including ML and deep learning (DL). | No interpretability, the need for robust AI models for real-world CVD risk assessment and management. No deployment in real-time. |
| Shahzadi *et al.* [3] | CVD prediction with DL convolutional neural networks (CNNs), recurrent neural networks (RNNs), and deep neural networks (DNNs), large dataset, 88.5% accuracy, precision, recall, and F1-scores are all above 85%. | Explainability and clinical applicability. No mortality prediction. |
| Shah *et al.* [5] | CVD prediction with traditional ML algorithms including k-nearest neighbors (KNN), naïve Bayes (NB), and decision tree (DT). | No ensemble learning, interpretability, and deployment in real-time. No mortality prediction. |
| Mohan *et al.* [6] | CVD prediction using a hybrid ML approach, hybrid random forest with a linear model (HRFLM), achieving 88.7% accuracy. | Explainability and clinical applicability. No mortality prediction. |
| Jindal *et al.* [7] | CVD prediction with ML algorithms, including LR and KNN. | Model interpretability or clinical deployment pipelines. No mortality prediction. |
| Vayadande *et al.* [8] | Survey about the critical role of ML in CVDs. | Didn't explore ensemble models or interpretability techniques in depth. No mortality prediction. |
| Katarya and Meena [9] | Comparative study of CVDs using ML techniques. | Model interpretability and real-world deployment. No mortality prediction. |
| Bhatt *et al.* [10] | ML framework for predicting CVD using DTs, RF, extreme gradient boosting (XGBoost), and multilayer perceptron (MLP), large dataset, MLP achieved the highest accuracy (87.28%). | Model explainability and real-time applicability. No mortality prediction. |
| Ali *et al.* [11] | Supervised ML algorithms for early CVD prediction with KNN, DT, and RF. | No external validation, clinical interpretability frameworks, and deployment. No mortality prediction |
| Al-Janabi *et al.* [12] | Comprehensive review of ML classification techniques used for CVDs prediction, DT, SVM, and neural network (NN). | No new modelling strategies. No addressing of practical implementation concerns such as model interpretability or integration. No mortality prediction. |

Table 1. Literature review summary table *(continued)*

| Reference | Work | Gap |
|---|---|---|
| Yilmaz and Yağin [13] | ML algorithms in early CVD detection. RF, SVM, and LR. RF classifier achieved the highest accuracy (92.9%). | No ensemble-based methods. No interpretability. No mortality prediction. |
| Baghdadi *et al.* [14] | ML to improve early detection and diagnosis of CVDs, hospital dataset. CatBoost algorithm demonstrated strong performance with an F1-score of 92.3% and an average accuracy of 90.94%. | No interpretability. No mortality prediction. |
| Nissa *et al.* [15] | Early detection of CVDs using six ML models. DT performed 97.29% accuracy. | Interpretability, generalizability, and deployment. No mortality prediction. |
| Nagavelli *et al.* [16] | ML approaches for early detection of CVDs, including NB, SVM, and XGBoost. | No ensemble learning, model interpretability and real-time deployment. No mortality prediction. |
| Pande *et al.* [17] | Advanced ML framework for early detection and prediction of CVDs. RF, GB, DNN, and AdaBoost | Model transparency and real-time usability. No mortality prediction. |
| Chavda *et al.* [18] | Early detection of CVDs using ML models trained on large dataset. | Interpretability of the models, didn't evaluate deployment feasibility in real-time. No mortality prediction. |
| Katarya and Srinivas [19] | Early prediction of CVDs using ML algorithms including artificial neural networks (ANN), DT, RF, SVM, NB, and KNN. | Model explainability or clinical deployment. |
| Alizadehsani *et al.* [20] | Comprehensive review of ML applications in the diagnosis of coronary artery disease (CAD). | Standardized, explainable, and externally validated ML frameworks. |
| Lundberg *et al.* [21] | Interpretability of tree-based ML models, such as DTs, RF, and GB, by introducing a suite of explainable artificial intelligence (XAI) tools based on game theory. | No ensemble learning and deployment. |
| Shickel *et al.* [22] | Comprehensive survey of DL including RNN and autoencoders in medical record analysis. | Transparent and generalizable deep learning frameworks. |

## BIOGRAPHIES OF AUTHORS

**Ali Al-Ataby** is an associate professor and chair of the Department of Electrical and Electronics Engineering at the American University of Ras Al Khaimah (AURAK), UAE. He holds a Ph.D. in Electrical Engineering with a focus on signal processing and machine learning from the University of Liverpool, Liverpool, UK. He also holds a postgraduate diploma (PgDip) in Higher Education in Learning and Teaching. With professional experience spanning academia and industry since 1997, he joined the University of Liverpool in 2011 as an assistant professor in Signal Processing and Machine Learning and was promoted to associate professor in 2017, before joining AURAK in 2023. He is a chartered engineer (CEng) accredited by the IET, a senior member of IEEE (SMIEEE), and a senior fellow of the UK Higher Education Academy (SFHEA). He has been recognized with a number of prestigious awards, including the sir alastair pilkington award for academic excellence and the Faculty of Science and Engineering Learning and Teaching Award from the University of Liverpool. His research interests include image and signal processing, machine learning and AI, robotics, IoT, technology-enhanced education, and pedagogical research. He can be contacted at email: ali.ataby@aurak.ac.ae.

**Hussain Attia** is an assistant professor at the American University of Ras Al Khaimah (AURAK). He obtained his Ph.D. in power electronics from University Malaya, Kuala Lumpur, Malaysia in 2019, M.Sc. in Electronic Engineering and B.Sc. in Electronics and Communications Engineering from the University of Technology, Baghdad, Iraq in 1999, and 1991 respectively. He served as a technical and organizing member for many IEEE and international conferences such as ICEDSA/2016, ICECTA/2017, ICEWES/2018, ICECTA/2019, and ICECTA/2022. He has published more than 80 articles in international journals and conferences as author or co-author, most of them in the field of multi functions power electronics systems integration, and renewable energy system applications. He won many awards from AURAK including the president's award for faculty – outstanding researcher/AY2019-2020 from president's office of the AURAK in 14th July 2020. His research interests include power electronic systems integration, AC&DC drives, harmonics reduction and power factor correcting techniques, renewable energy systems, and MPPT algorithms. He can be contacted at email: hattia@aurak.ac.ae.