Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 32s (2025)
https://www.jneonatalsurg.com

Leveraging Explainable AI for Improved Heart Failure Survival Prediction Models

Mukesh Kumar Tiwari[1*], Brij Mohan Singh[2]

[1*,2]Department of Computer Science and Engineering, Quantum University, Roorkee, Uttarakhand, India

***Corresponding Author:**

Mukesh Kumar Tiwari

Department of Computer Science & Engineering, Quantum University Roorkee, Haridwar, Uttarakhan, India,

Email ID: mukesh.cse@quantumeducation.in

.Cite this paper as: Mukesh Kumar Tiwari, Brij Mohan Singh, (2025) Leveraging Explainable AI for Improved Heart Failure Survival Prediction Models. *Journal of Neonatal Surgery*, 14 (32s), 2938-2952.

ABSTRACT

Heart failure (HF) is a major cause of illness and death globally, highlighting the need for effective prediction models to identify high-risk patients. Traditional machine learning models are worked as a “black boxes”, offering little understanding of how they make decisions. This study 1) Present a comparative analysis of conventional machine learning models on the HF disease. 2) examines how to incorporate eXplainable Artificial Intelligence (XAI) techniques into heart failure conventional survival prediction models. 3) Analyzes the explainability of heart failure (HF) survival prediction models. This study analyzes a dataset containing 918 patient records with a history of HF. In the first phase of the study, the machine learning model Xtreme Gradient Boosting (XGB) achieves the highest accuracy of 88.59% among all models tested on the HF dataset. The second phase focuses on explainability, emphasizing that cholesterol levels, age, MaxHR, and Oldpeak are crucial features in HF prediction. With the analysis of the experts note that the model performs well because these relevant features significantly contribute to predicting HF and can save the human life.

Keywords: *Artificial Intelligence, Explainable AI, Machine Learning, Heart Failure Prediction, SHAP, LIME*

1. INTRODUCTION

Heart failure is a significant problem for healthcare community that influences millions of lives, highlighting the importance of early detection and intervention [1]. World Health Organization (WHO) identified the HF as a primary global cause of death disease [2]. Several studies, such as [3] and [4], have utilized machine learning and deep learning models on healthcare data to predict heart failure (HF). As a result, they have enhanced the accuracy and reliability of stroke prediction models. However, these traditional models often lack the ability to explain their predictions outcomes. This creates a gap between the model’s results and their practical use in clinical settings [5]. It will offer an opportunity for the research community to explore this area. Coronary heart disease, atrial fibrillation, heart failure, stroke, and vascular dementia are just a few of the many conditions that fall under the broad category of heart failure and circulatory diseases. These conditions can range from genetic disorders to those that develop over time, as shown in Figure 1 of the Global Heart Circulatory Diseases Factsheet [6]. Globally, there are currently about 620 million people who suffer fromcardiac conditions, this number is rising as a result of ageing populations, lifestyle changes, and increased survival rates following heart-related events. According to the Figure 1, around the world, 1 in 13 persons are thought to suffer from a cardiac or circulation condition. In 2019, 290 million women were impacted, more than 260 million males [7]. With 285 million people afflicted, the prevalence of these illnesses has been gradually rising over time.

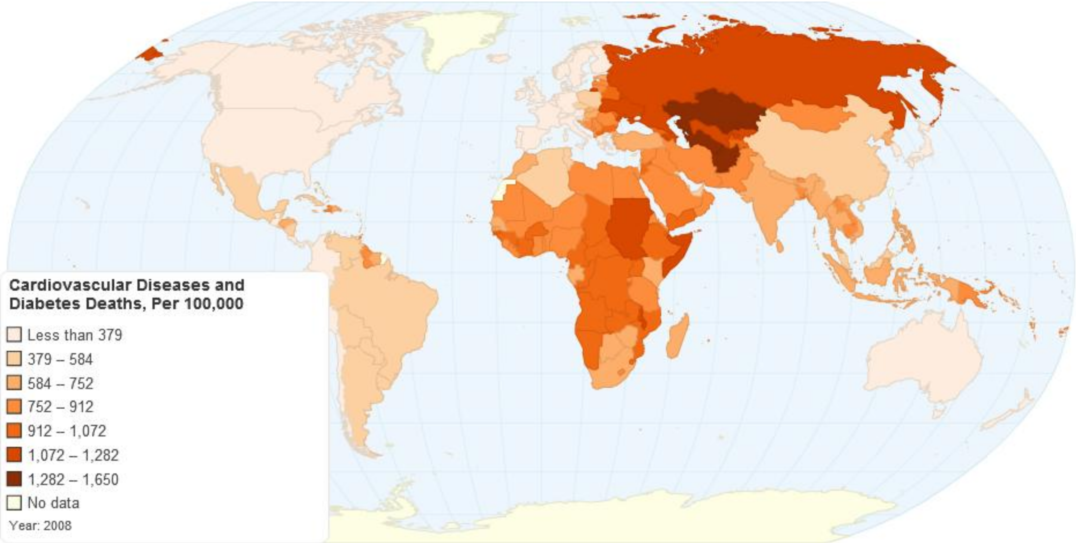

Figure 1. Geographical distribution of heart and circulatory diseases in the world

Throughout many years, clinical cardiac illness has been the subject of substantial research. For instances [8], examined ten machine learning approach, such as Na¨ıve Bayes, Logistic Regression, and discovered that the HRFLM method—which mixes Random Forest with a linear approach—achieved the best accuracy. The HDPM model, which combines DBSCAN, SMOTEENN, and XGB-based MLA, was shown to be the most successful in predicting heart disease by [9]. after they run seven machine learning algorithmson the same dataset. Additionally [10] employed K-Nearest Neighbour and Random Forest techniques, with K-Nearest Neighbour obtaining an accuracy of 86.85%. Traditional methods may include established statistical approaches, while XAI techniques aim to enhance the interpretability of the models [5], allowing healthcare professionals to understand the reasoning behind predictions. This study examines the heart and circulatory diseases by applying various machine learning models, such as support vector machines, random forest and decision trees to enhance model performance and provide recommendations for reducing heart disease risk. The key goals of this study are:

- To analysis the traditional heart failure survival prediction models and explainable artificial intelligence (XAI) techniques.
- To assess the performance and interpretability of traditional approaches.
- To provide insights into the clinical significance of using XAI in heart failure management.

The rest of this paper includes literature review which describe the survival prediction models applied for the heart failure disease on the same dataset as this study, which can be found in Section 2. In Section 3, the propose approach has discussed with the dataset, several machine learning algorithms, feature selection techniques, and metrics used. Section 4 outlines the experiment result after constructing experiment various ML prediction model, the classification and explainability evaluation results, and the significance of the features for both local and global explainability. The work's conclusions are finally presented in Section 5.

2. LITERATURE REVIEW

Heart failure is a crucial health concern which impacts millions of individuals and contributing to high rates of morbidity and mortality [4]. Several predictive machine learning (ML) models including [3][9] in the management of heart failure play a significant role in stratifying patient risk, guiding treatment options, and optimizing resource allocation. However, these advancements often come with a trade-off between model accuracy and interpretability [11]. This literature review presents an overview of recent work focusing on machine learning (ML), artificial intelligence (AI), and explainable AI (XAI) techniques used in heart failure survival prediction models as illustrated in Table 1. Table 1 highlights the need for more focused studies on the application of XAI in heart failure survival prediction models. Addressing these research gaps could lead to the development of more interpretable, clinically relevant and effective AI systems in healthcare.

Table 1: A comparative study of the HF with several predictive models.

Literature	Year	Scope of the existing studies				Dataset	
		AI	**ML**	**DL**	**XAI**	**Public**	**OWN**
Dahri et.al. [12]	2024	χ	✓	χ	χ	✓	χ
Alfredo D et.al. [13]	2021	χ	✓	χ	χ	✓	χ
Sutradhar et.al. [14]	2023	✓	✓	χ	χ	✓	χ
Hassan MM [15]	2022	✓	✓	χ	χ	✓	χ
Yaqoob MM [16]	2023	χ	✓	χ	χ	χ	✓
Cyriac S, [17]	2022	χ	✓	χ	χ	✓	χ
Abbas A [18]	2022	χ	✓	χ	χ	✓	χ
Tambe PM [19]	2025	χ	χ	✓	χ	χ	✓
Acar ZY, [20]	2024	χ	χ	✓	χ	✓	χ
Basak S [21]	2022	χ	χ	✓	χ	✓	χ
Almutairi SA [22]	2023	χ	χ	✓	χ	✓	χ
Afiatuddin N [23]	2024	χ	χ	✓	χ	✓	χ
Barzola-Monteses J [24]	2024	χ	χ	✓	χ	✓	χ

Recent advances in ML and AI have improved predictive accuracy [9]. However, the "black-box" nature of many ML algorithms raises concerns about their application in clinical settings. In addition, AI in healthcare applications requires high levels of interoperability due to the critical nature of clinical decisions [25]. An explainable model such as [26] enhances trust and accountability, making them vital for clinical adoption in dengue disease. Furthermore, healthcare practitioners need to understand the rationale behind AI predictions to make informed decisions about patient care [27]. Many transparent machine learning models, such as linear regression, logisticregression, decision trees, na¨ıve Bayes, and k-nearest neighbors, have been extensively used across multiple clinical fields. These include cardiology, urology, toxicology, endocrinology, neurology, psychiatry, occupational health, breast cancer, Alzheimer's disease severity, knee osteoarthritis, prostate cancer, diabetes, and cardiovascular disease mortality rates.

Explainable AI (XAI) offers methodologies that clarify how predictions are made, thereby increasing trust among healthcare providers and patients [26]. In these various fields, AI is used to perform disease classification and XAI is added to provide explanation to improve the understanding of medical personnel and general users. We refer to these studies to develop a disease prediction model based on machine learning and apply SHapley Additive exPlanations (SHAP) values and a Local Interpretable Model-Agnostic Explanations (LIME), for explainability to visualize the models and provide interpretable rationales for their predictions. These tools allow research community to understand model predictions and the factors influencing them. By applying XAI techniques in the medical field, these methods aim to assist medical professionals in decision making and automate diagnostic processes [28]. In addition, we address the "black box" issue of AI by providing evidence-based insights to healthcare professionals. However, most previous studies [23],[22],[12],[9], focus primarily on improving the performance of AI models through hyper parameter tuning and image pre-processing, offering only a partial representation of the decision-making process rather than fully visualizing it with various XAI techniques.

3. METHODOLOGY

In this study we provide a methodology for developing a heart failure prediction model using explainable AI (XAI) techniques can be divided into three key layers: data collection and pre-processing, model development, and XAI evaluation and visualization. Architecture diagram illustrating the workflow of HF prediction using explainable AI (XAI) technique is shown in Figure 2. First layer, ensures that the input data is clean, representative, and ready for model training. Second layer study the behavior of ML models that not only predicts heart failure outcomes accurately but also generalizes well to unseen data. Third layer integrating XAI techniques to bridges the gap between model predictions and clinical decision-making, ensuring that the model's outputs are interpretable and actionable. Each layer plays a critical role in ensuring the robustness, accuracy, and interpretability of the model.

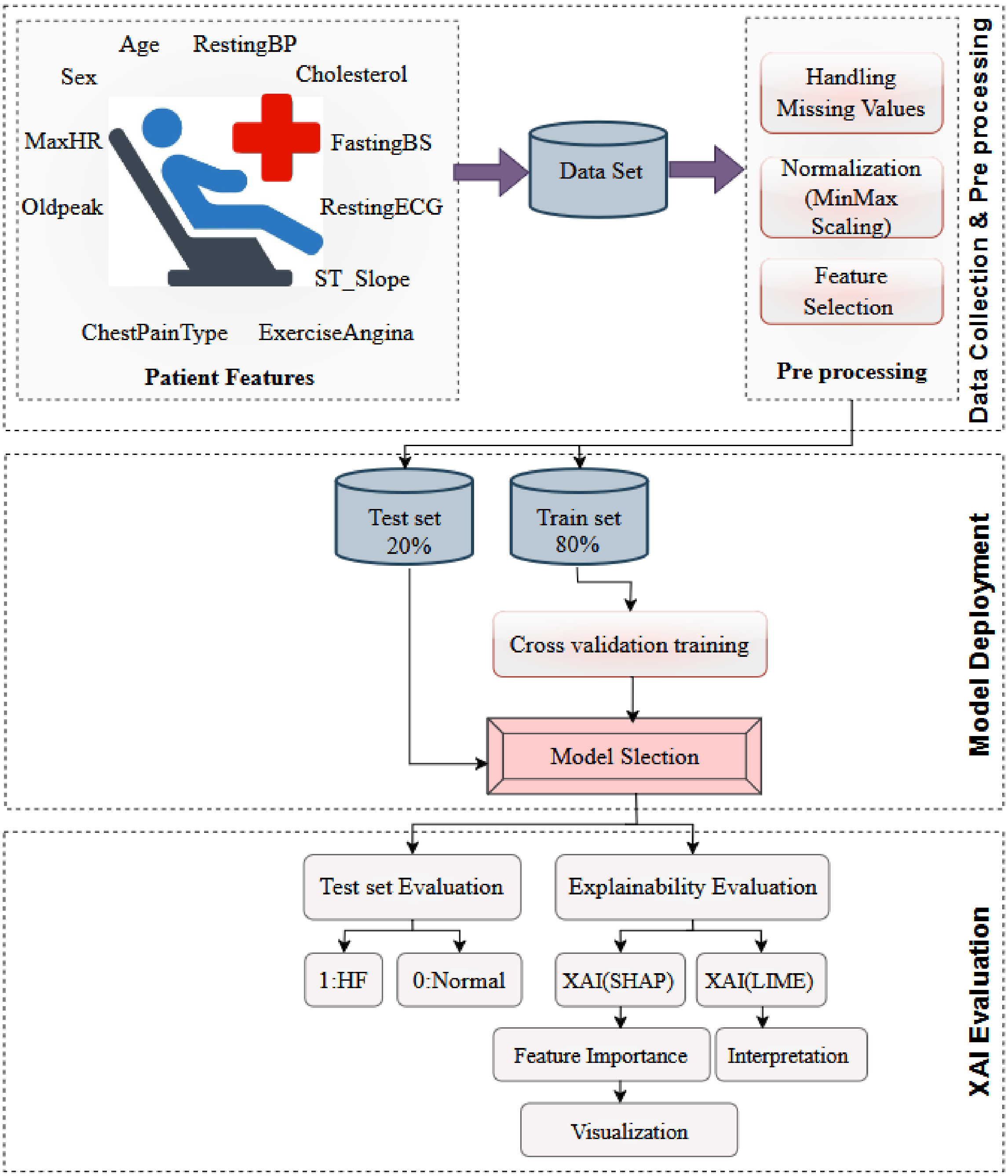

Figure 2. Architecture diagram illustrating the workflow of HF prediction using explainable AI (XAI)

3.1 Data Collection and Pre-processing

According to study [29], this study made use of the Allied Hospital's and Faisalabad Institute of Cardiology of heart failure clinical records dataset. 1190 observations, 272 duplicate observations, and 918 unique observation records make up the dataset. Each record includes 12 clinical features 11 predictive and 1 prediction—that were gathered over a follow-up period. We have list the features in details that can be used to predict a potential heart illness. With the utilizing these characteristics, machine learning models can be quite helpful to recognize and heart failure risks. These models have shown promise in enhancing the precision of diagnoses which will ultimately help to lessen the burden of heart failure illnesses. We have lists all of the attributes' kinds and descriptions, and the study in this paper is predicated on the 12 attributes that are described bellow.

1 **Age**: This feature represents the age of the patient in years. It is a crucial demographic variable that can hold the numeric value influence health outcomes, treatment decisions, and risk assessments with the rang vale of [0-100].

2 **Sex**: This feature indicates the gender of the patient. It is categorical feature like [M: Male; F: Female] and important for understanding gender-specific health issues and can influence the prevalence of certain diseases, treatment responses, and health behaviors.

3 **ChestPainType**: This feature categorizes the type of chest pain experienced by the patient. It helps in diagnosing potential cardiac conditions. Understanding the nature of chest pain data set used the categorical value like [TA: Typical Angina; ASY: Asymptomatic; NAP: Non-Anginal Pain; ATA: Atypical Angina] for determining the urgency and type of medical intervention required.

4 **RestingBP**: This feature measures the patient's blood pressure while at rest. Data set monitoring resting blood pressure in numeric with the range of [70 - 200 mmHg] as an essential feature for assessing overall health and risk for heart disease.

5 **Cholesterol**: This feature indicates the serum cholesterol level in numeric for the patient's blood. Cholesterol levels are critical for evaluating heart failure risk based on the value range [100 - 400 mg/dL], as high levels can lead to heart disease and other health complications.

6 **FastingBS**: This feature indicates the blood sugar levels of patient during fasting. In a data set fasting blood sugar level have binaryvalue [1: above 120 mg/dL, 0: otherwise] suggests potential diabetes.

7 **RestingECG**: This feature reflects the results of the resting electrocardiogram (ECG) test with value of [Normal, ST, LVH]. It helps in assessing the electrical activity of the heart and can indicate various cardiac conditions.

8 **MaxHR**: This feature has numeric value ranges between [60 - 202] to represents the highest heart rate attained by the patient during exercise or stress testing.

9 **ExerciseAngina**: This feature indicates whether the patient experiences angina (chest pain) during exercise with value [N: No, Y: Yes].

10 **Oldpeak**: This feature measures the ST depression induced by exercise, which is an important indicator of cardiac ischemia in range of [-2.6 to 6.2 mm]. The Oldpeak value helps in assessing the severity of heart disease and the effectiveness of treatment.

11 **ST Slope**: This feature describes the slope of the ST segment during peak exercise with possible values of [Flat, Down, Up].

12 **HeartDisease**: This feature indicates the target feature as the presence or absence of heart disease in the patient. It serves as the output binary class [1: heart disease, 0: Normal] for predictive modeling and is crucial for determining treatment strategies and patient management plans.

This dataset was produced by merging five previously unmerged, independently accessible datasets on cardiac disease illustrated in Figure 3. In particular, some feature are all nominal types, meaning they are not numerical, with the exception of Age, Resting Blood Pressure, Cholesterol, Fasting Blood Stream, Old Peak, and Heart Disease. The largest dataset on heart disease available for research reasons was created by integrating 11 common features. The Figure 3 provided a full breakdown of the datasets used for its curation. This final dataset, which has 918 unique observations after duplicates have been removed, is a useful tool for research on heart disease survival prediction.

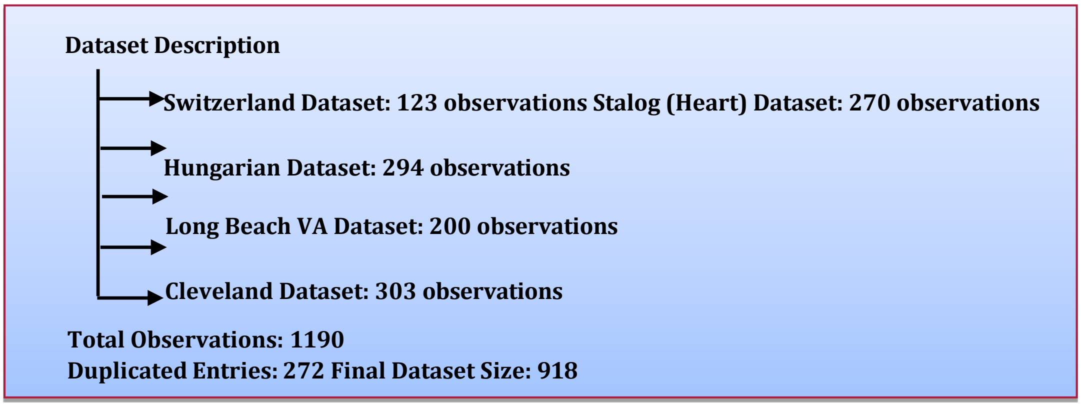
Dataset Description
Switzerland Dataset: 123 observations Stalog (Heart) Dataset: 270 observations
Hungarian Dataset: 294 observations
Long Beach VA Dataset: 200 observations
Cleveland Dataset: 303 observations
Total Observations: 1190
Duplicated Entries: 272 Final Dataset Size: 918

Figure 3. Description of the Dataset

- **Data Cleaning:** Removing missing, inconsistent, or duplicate entries to ensure data quality.
- **Feature Engineering**: Choosing and modifying relevant features to enhance model performance. For instance, normalizing continuous variables like blood pressure or scaling features to a uniform range.
- **Handling Imbalanced Data**: Since heart failure datasets often have an imbalance between positive (heart failure cases) and negative (non- heart failure cases) samples, methods such as oversampling (e.g., SMOTE) or under sampling can be utilized to achieve balance within the dataset.
- **Data Splitting:** To properly assess model performance, split the dataset between training (80%) and testing (20%) groups.

Since five of the features are not quantitative, they must be converted to numeric values in order to further the investigation. Once the quartile range has been calculated, identify the dataset's outliers. 917 observations are now included in the dataset after only one outlier was found and removed. Plotting the data in a distribution map across all characteristics is now possible after the outlier has been removed. Figure 4 shows the potential values for each attribute.

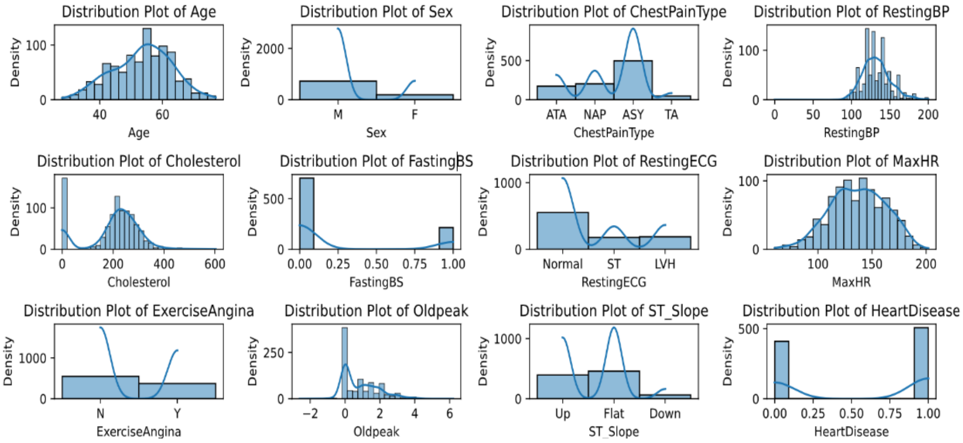

Figure 4. Distribution plot among all features of heart failure data set. The curve line represents Normal distribution

The statistics from the dataset also showed that, of all heart disease cases, 47.1% were women and 52.9% were men. In the data set, the mean year of the onset of heart disease was estimated to be 65.6 years for males and 72.0 years for women. Basically, older men are more likely to develop heart disease than older women. In addition, exercise-induced angina and ST depression can be treated as early warning signs of heart disease. Figure 5 shows a heatmap graphic that shows the correlation matrix between the numerical variables that will be taken into account for the data model. Positive values show that two variables have a positive association, whilst negative values show that two variables have a negative relationship.

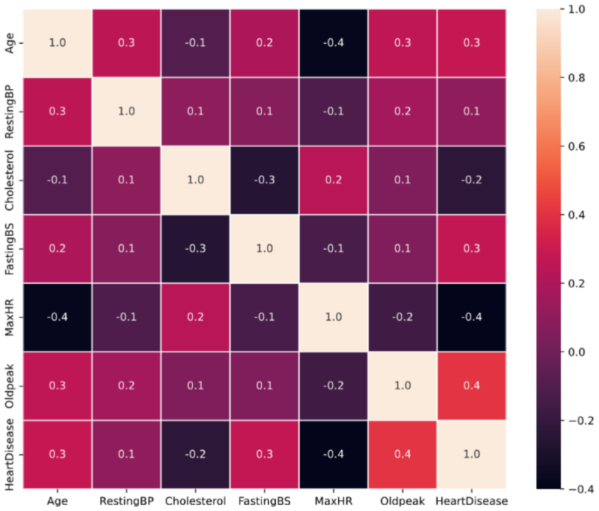

Figure 5. Heatmap plot represented the correlation between the attributes of data set

3.2 Model Development

In this section, we detail the various of heart failure survival prediction models, focusing on the methodologies employed, the rationale for their selection, and the metrics used for evaluation. We have studied the following ML model on collected data set.

- **Logistic Regression (LR)**

LR is a supervised machine learning approach primarily employed for task involving binary in classification where the objective is to estimate the probability of an event occurring, such as success/failure or yes/no outcomes [30]. It achieves this by modeling the relationship to one or any number of distinct (x_i) of ith variables and the probability (PP) of a favourable outcome. The algorithm applies a transformation, typically the sigmoid function, to ensure the predicted probabilities fall within the range of 0 and 1. The mathematical formula used to compute this probability is as follows

$$PP = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m\right)}{1 - \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m\right)}$$

where, β0 is the intercept, and β1, β2, ..., βm are the coefficients for the independent variables (x1, x2, ..., xm).

- **Naive Bayes (NB) Model**

One of the most basic probabilistic classifiers is the Naive Bayes model. Based on the Bayes theorem, it makes the assumption that, given the class label, every feature is conditionally independent. It represents the relationships between variables using a directed acyclic graph (DAG) [31]. In these networks, every node represents a random variable, while the edges connecting the nodes signify conditional dependencies. They are used to model uncertainty and calculate probabilities using Bayesian inference. A Bayesian Network's structure represents a joint probability distribution as a product of conditional probabilities, enabling applications in diagnostics, reasoning, causal modeling, uncertainty-based decision-making, and anomaly detection. The mathematical formula for calculating probabilities is as follows:

$$PP\left(\frac{x}{y}\right) = \frac{PP\left(\frac{x}{y}\right) PP(x)}{PP(y)}$$

where PP (X / Y) represents the subsequent probability, PP (X) represents the class's prior probability, PP (Y) represents the probability prior to the predictor, and PP (Y / X) represents the predictor's probability.

- **K-Nearest Neighbor (KNN)**

A non-parametric, supervised learning technique for classification and regression problems is the K-Nearest Neighbours (KNN) algorithm [32]. Based on a selected distance metric, like Euclidean distance, it determines a data point's closest neighbours and uses that information to classify or predict. Choosing a suitable value for "K" (the number of neighbours) and figuring out the distance metric to gauge proximity are the two main components of the KNN method [33]. The algorithm assigns classifications or makes predictions by clustering data points based on their closeness to others in thedataset. The Euclidean distance formula, is the most commonly used metric for this purpose. In this study, KNN is also applied to identify the optimal model and can be defined as using mathematical formula of Euclidean distance

$$d(x,y) = \sqrt{\sum\nolimits_{i=0}^{n} (y_i - x_i)^2}$$

Where, x and y represent the input features and n is the total number of features.

- **Support Vector Machine (SVM)**

With the ability to handle both classification and regression tasks, the SVM method is one of the most well-known supervised learning strategies [34]. In 1963, Vladimir N. Vapnik created it for linear models. In 1995, he expanded it to handle non-linear data using kernel functions [35]. SVM maximises the margin between data points while identifying the best hyperplane to divide them into classes [34]. For non-linear data, it uses kernel functions to transform the input space into a higher-dimensional space, making it possible to find a separating hyperplane even for complex datasets. This ability to handle both linear and non-linear data makes SVM a powerful tool for predictive modeling and classification tasks. The algorithm generates predictions based on a mathematical function that defines this hyperplane

$$y(x,w) = \sum_{i=1}^{n} w_i K(x, x_i) + w_0$$

Where K(x, xi) is a kernel function (34).

- **Decision Tree (DT)**

The Decision Tree (DT) algorithm was chosen as one of the prediction techniques in this study because of its ease of use and learning efficacy is high. One popular kind of supervised learning technique for classification and regression problems is the decision tree. The way they work is by using a tree-like structure to describe decisions for HF. The process starts at the root node, which stands for the entire dataset. The data is divided into branches according to the results of feature tests conducted at each internal node [36]. This process continues until the data reaches the leaf nodes, which represent the final predictions or classifications. Its ability to visually represent decision-making processes makes it intuitive and easy to interpret. Additionally, the algorithm uses measures like entropy or Gini index to determine the best splits at each node, ensuring that the tree effectively separates the data into meaningful groups. The final output of a Decision Tree is a set of terminal nodes (leaf nodes) that provide the predicted outcomes. This makes Decision Trees a powerful tool for predictive modeling, especially in scenarios requiring straightforward evaluation and interpretation.

$$Entropy\ (S) = \sum_{i=1}^{n} -P_i\, Log_2(P_i)$$

$$Gain(S, A) = \ Entropy\ (S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Where, S_v is the subset of S with A = v, where S is a set of instances, A is an attribute, and Values (A) is the set of all possible values of A. The Gini value is then computed for the entire dataset D as

$$Gini(D) = 1 - \sum\nolimits_{i=1}^{n} PP_i^2$$

Where PP is the probability.

- **Random Forest (RF)**

An ensemble learning method called the Random Forest (RF) algorithm creates base learners in simultaneously. Several decision trees are constructed separately using this method, and each tree is trained using a different subset of the data [37]. A majority voting system (for classification) or average (for regression) determines the final outcome, while each tree generates its own prediction. This aggregation of predictions from multiple trees helps improve the accuracy and robustness of the model. The decision trees serve as the foundational models in the Random Forest framework. By leveraging the independence of these base learners, Random Forest significantly reduces errors, particularly variance, through the averaging of predictions. This parallel ensemble approach ensures that the model remains stable and reliable, even when dealing with noisy or complex datasets. It can be measure by Gini impurity for the D using the equation of Gini.

- **Gradient Boost (GB)**

By comparing the current model with earlier iterations, GB, a machine learning algorithm, is founded on the idea that prediction error can be reduced. Determining the desired results for the subsequent model in the series is the basic idea behind lowering the error. The following formula serves as a mathematical representation of this process.

$$g_t(x) = E_y\left[\frac{\partial\varphi(y, f(x))}{\partial f(x)}\middle| x\right] f(x) = \widehat{f^{t-1}}\,(x)$$

- **AdaBoost (AB)**

In order to dynamically reward the relative importance of a limited set of training characteristics, this machine learning technique prioritises complex samples, linearly integrates the classifier’s components, and generates a single-ended hypothesis. It is carried out for a total of F iterations (Size of ensembling).

$$g(p) = sign(\sum\nolimits_{f=1}^{F} a_f l(p))$$

- **Xtreme Gradient Boosting (XGBoost)**

The sum of tree branches (K-trees), or the overall score derived from the projected value’s leaves, is what makes this approach a form of continuous optimization. With the decision tree serving as the primary method and the loss function regulating the tree’s involvement, XGBoost is an additive extension to the target task design by minimizing the loss function

$$\emptyset\tau = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_x)$$

- **Performance Evaluation Metrics**

This study assessed the effectiveness of the proposed approach using advanced performance evaluation metrics, including accuracy, precision, recall, F1-score, AUC-ROC score, and explainable matrix evaluation. The evaluation results are summarized in Section 4.

- **Accuracy**: Accuracy is used to evaluates the level of precision of the machine learning classifier by dividing the number of correct predictions by the total predictions. Mathematically, it is represented in Equation.

$$Acc = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

where, T_P = true positives, T_N = true negatives, F_P= false positive, and F_N = false negatives.

- **Recall**: It evaluates the model's ability to identify all true positive cases. It is calculated as the ratio of true positives (TP) to the sum of true positives and false negatives (FN). This metric is particularly useful when minimizing false negatives is critical, and its formula is provided in Equation

$$Rec = \frac{T_p}{T_p + F_n}$$

- **Precision:** It focuses on the proportion of correctly predicted positive instances out of all predicted positives. It is especially important in scenarios where false positives need to be minimized. The mathematical representation of precision is shown in Equation.

$$Pre = \frac{T_p}{T_p + F_p}$$

- **F1-score:** The F1-score provides a balanced assessment of the precision and recall by describing the harmonic mean of these two. It is especially valuable in situations with imbalanced class distributions. The formula for calculating the F1-score is presented in Equation.

$$F_1 = 2 * \frac{Pre + Rec}{Pre * Rec}$$

- **AUC-ROC score**: AUC-ROC (Area Under the Receiver Operating Characteristic Curve) illustrates the balance between the true positive rate (TPR) and the false positive rate (FPR). It offers a thorough assessment of the model's effectiveness in differentiating between classes. This metric is especially useful for assessing binary classification models.

- **Explainable** Matrix Evaluation: To evaluate the explainability of the proposed approach, we utilized two popular interpretability tech- niques, LIME and SHAP, to compute feature importances based on the model's predictions for specific instances [26]. Since it is computa- tionally impractical to apply these methods to number of instances, we implemented a uniform sampling strategy to select k representative pa- tients sample, ensuring a balanced distribution between the two classes (heart failure and non heart failure).

3.3 Proposed Algorithm

The pseudo-code details the workflow for developing and testing an ex- plainable AI model designed to predict heart failure survival. It covers key stages such as data pre-processing, feature selection, model training, pruning methods, and performance assessment to balance accuracy with interpretabil- ity. Algorithm 1 presents the complete pseudo-code for this approach.

Algorithm 1 Pseudo-code of Proposed Explainable AI for Improved Heart Failure Model

```
1: Begin
2: Pre-process heart datasets D to clean and prepare the data for analysis
3: Apply oversampling algorithms (e.g., SMOTE) to handle class imbalance in the datasets
4: Conduct attribute pruning to identify the most significant features.
5: Split dataset into Train and Test.
6: Train the selected features of the dataset using ML models.
7: Apply the optimized parameters to the model, then conduct post-pruning to minimize complexity.
8: Cross-validate the Explainable AI for Improved Heart Failure model.
9: Assess the performance of models using the test dataset.
10: Calculate performance evaluationmetrics, including accuracy, precision, recall, F1-score, and AUC-ROC, to evaluate the model's effectiveness.
```

11: End

4. EXPERIMENT RESULTS

To achieve higher accuracy, the data must first undergo standardization. This process ensures that the data is transformed into a consistent format, making it suitable for machine learning algorithms. After standardization, the dataset is split into four variables: x train, x test, y train, and y test. This split allocates 80% of the data for training and 20% for testing purposes. Once the pre-processing steps are complete, these variables are ready to be utilized in the selected nine machine learning algorithms. This study assessed the effectiveness of the proposed approach using two advanced metrics, including 1) performance evaluation, and 2) explainable matrix evaluation. This study discusses each result with detailed in further sub section.

4.1 Performance Evaluation Metrics

The experiment was conducted to evaluate the performance based on the metrics discussed in the Section 3.2.10 to evaluate the results. We have ana- lyzed the nine traditional ML models given in the Section 3.2 in detailed. The first algorithm applied is the LR. Here, the number of features considered for the optimal split (max features) is set to 10, while all other parameters remain at their default values. For each of the machine learning algorithms mentioned, the classification results are calculated and illustrated in the Table 2, and the prediction outcomes are visualized using a confusion matrix.The evaluation of nine machine learning models reveals that ensemble methods like Random Forest and XGBoost outperform others, achieving the highest accuracy and balanced metrics across both classes. XGBoost, in particular, stands out with the best test accuracy (88.59%) and F1-scores, making it the most effective model for this dataset. Random Forest also demonstrates strong generalization with high precision and recall. Logistic Regression, KNN, Gradient Boost, and AdaBoost perform moderately well, offering balanced results but slightly lower accuracy compared to the top models. Naive Bayes and Decision Tree show acceptable performance but are less robust, while SVM struggles with over-fitting and fails to generalize effectively. Overall, the results highlight the superiority of ensemble methods for reliable classification in this context.

Table 2: Performance Metrics of Various Models for Heart Failure Prediction

	Model	**Accuracy (%)**		**Class**	**Precision**	**Recall**	**F1-Score**
S.No		Train	Test				
1	Logistic Regression (LR)	85.42	81.52	(Normal)	0.87	0.74	0.8
				(HeartFailure)	0.78	0.89	0.83
2	Naive Bayes (NB)	86.1	79.98	(Normal)	0.86	0.72	0.78
				(HeartFailure)	0.76	0.88	0.81
3	K-Nearest Neighbor (KNN)	81.34	82.7	(Normal)	0.87	0.74	0.8
				(HeartFailure)	0.79	0.89	0.84
4	Support Vector Machine (SVM)	99	59	(Normal)	0.05	0.67	0.1
				(HeartFailure)	0.98	0.59	0.74
5	Decision Tree (DT)	80	79.89	(Normal)	0.87	0.71	0.78
				(HeartFailure)	0.75	0.89	0.81
6	Random Forest (RF)	88	87.5	(Normal)	0.9	0.82	0.86
				(HeartFailure)	0.86	0.92	0.89
7	Gradient Boost (GB)	90.33	82.07	(Normal)	0.87	0.74	0.8
				(HeartFailure)	0.79	0.89	0.84
8	AdaBoost (AB)	89.37	82.07	(Normal)	0.87	0.74	0.80
				(HeartFailure)	0.79	0.89	0.84
9	Xtreme Gradient Boosting (XGB)	93.6	88.59	(Normal)	0.91	0.83	0.87
				(HeartFailure)	0.87	0.93	0.90

Random Forest (RF) and XGBoost stand out as the top-performing models, with high accuracy, precision, recall, and F1-scores for both classes. The ROC curves and AUC (Area Under the Curve) values for the nine models provide a comprehensive evaluation of their performance in distinguishing between the two classes (e.g., Normal vs. Heart Failure). Figure 6 shows de- tailed explanation of the results and a summary of the models' performance. LR achieved a moderate

AUC score, indicating that it performs reasonably well in distinguishing between the two classes. NB showed a slightly lower AUC compared to Logistic Regression, suggesting that it may not handle the data distribution as effectively. AdaBoost and XGBoost are ensemble method that focuses on misclassified samples, improving performance iteratively. It works well with weak learners like decision trees.

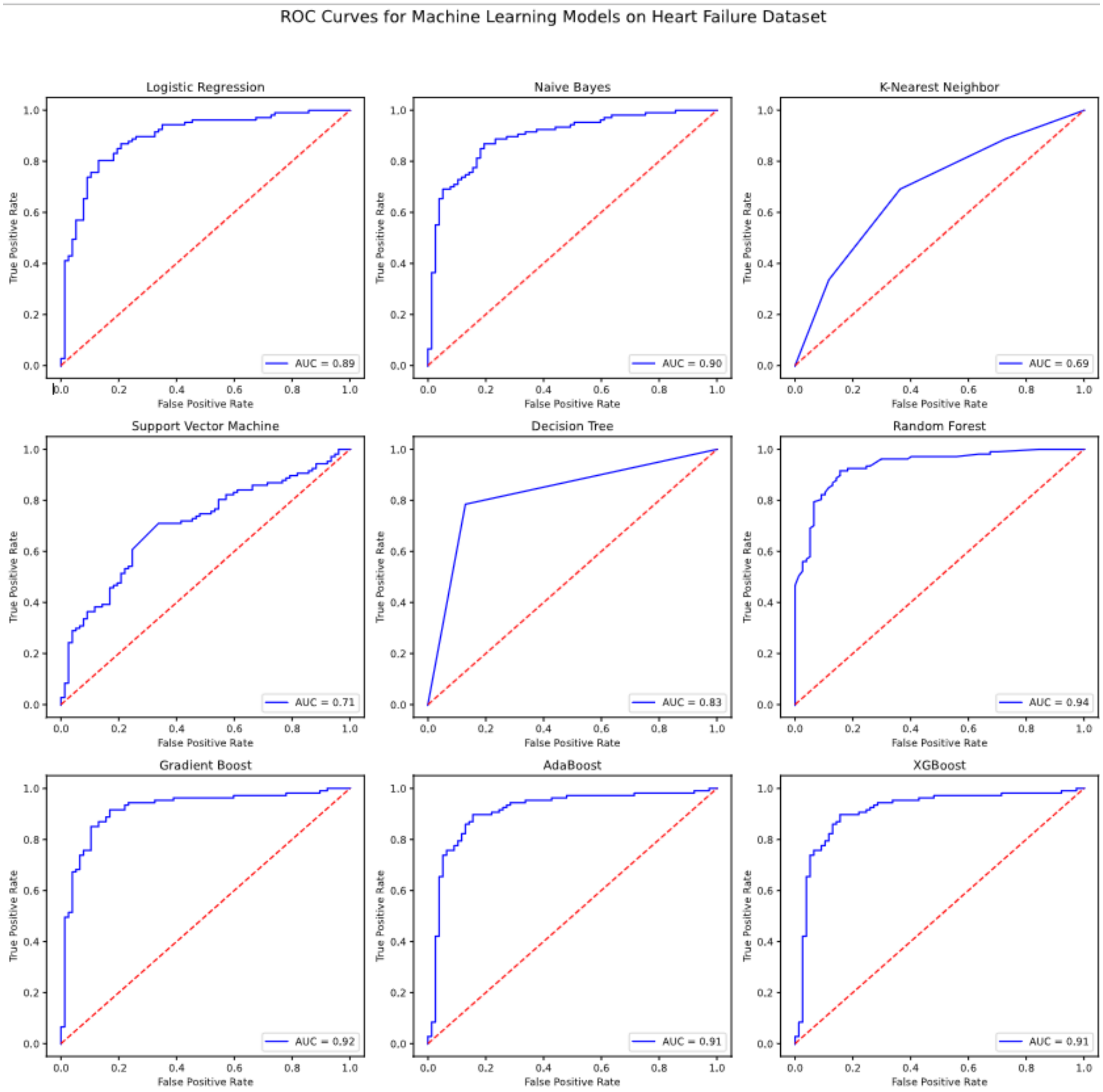

Figure 6: ROC Curves for Machine Learning Models on Heart Failure Dataset.

4.2 Explainable Evaluation Metrics

The evaluation of explainable machine learning models on the heart fail- ure dataset highlights the importance of balancing predictive accuracy with interpretability to support clinical decision-making using LIME (38) and SHAPE (39) model. The feature importance of applied machine learning models is illustrated in the Figure 8 and 7. This study observe that XG- Boost demonstrated superior performance, achieving the highest test accuracy (up to 88.59%) and balanced metrics with the highest importance of cholesterol feature in the heart failure surveillance. Where as ensemble methods like Random Forest demonstrated good performance, achieving the highest test accuracy (up to 87.50%) and balanced metrics such as precision, recall, and F1-scores, making them reliable for identifying both normal and heart failure cases. These models excel in capturing complex patterns in the data and show that ST Slope feature have the high importance in surveillance HF, which is critical for accurate predictions in medical contexts. However, simpler models like Logistic Regression and KNN also performed well, offering competitive feature importance of there prediction accuracy and interpretability, which are essential for building trust in clinical applications. On the other hand, models like SVM struggled with overfitting, highlighting the need for careful parameter tuning and validation. Explainable evaluation metrics, such as precision and recall, provide deeper insights into the models' decision-making processes, ensuring that predictions align with clinical priorities, such as minimizing false negatives in heart failure detection. Overall, the results emphasize the potential of machine learning in improving heart failure diagnosis while underscoring the need for explainability to foster trust and usability in healthcare settings.

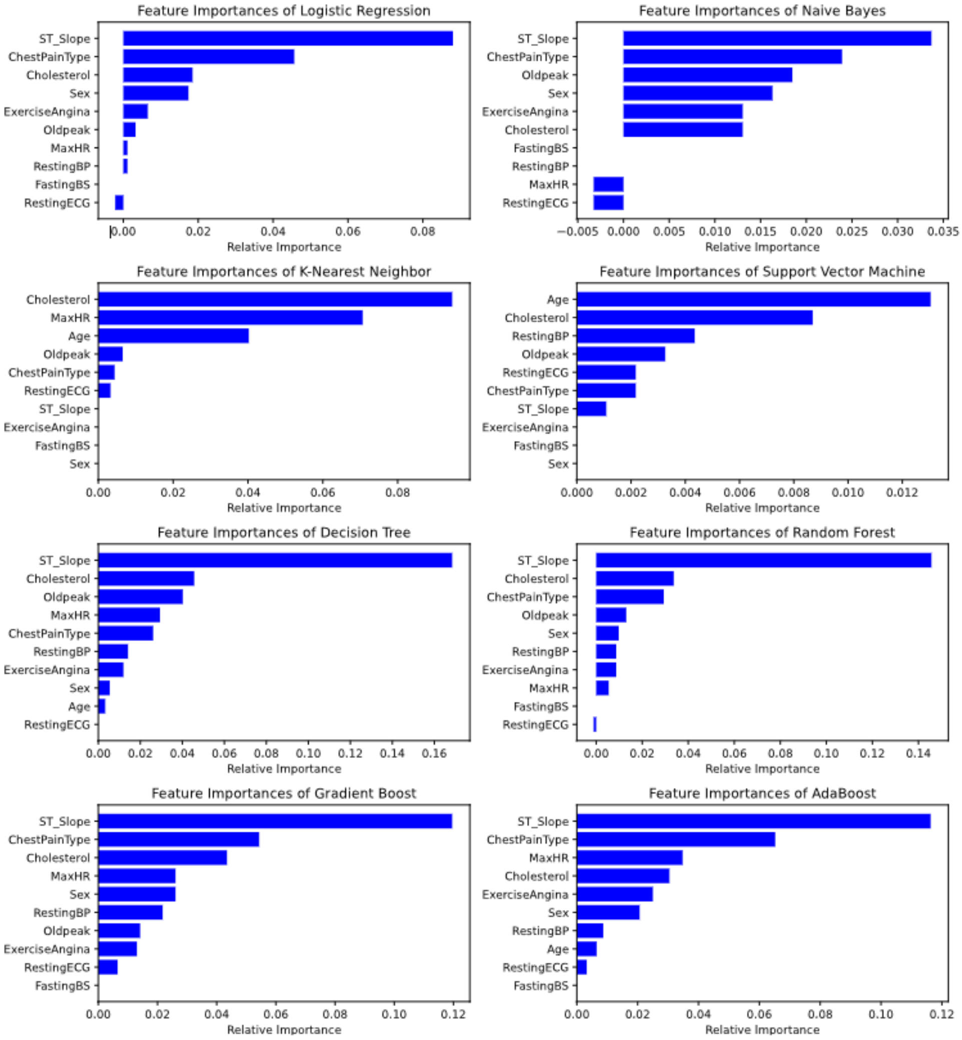

Figure 7: Feature Importance of Different ML Models.

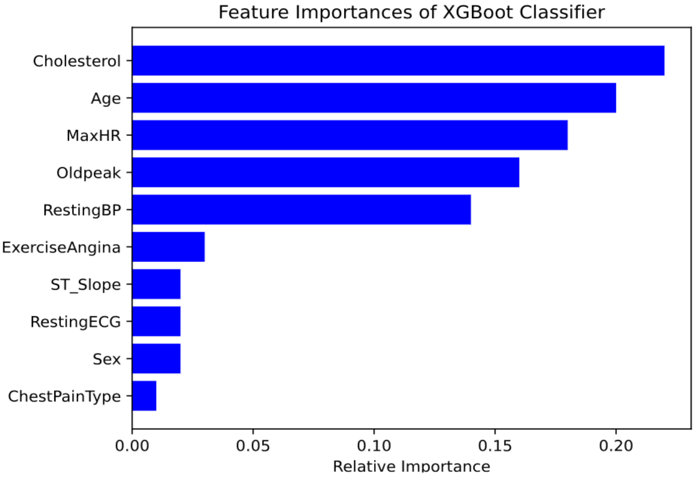

Figure 8: Feature Importance of XGBoots Models.

5. CONCLUSION

This study underscores the transformative role of machine learning (ML) and explainable artificial intelligence (XAI) in heart failure (HF) survival prediction. By comparing traditional ML models with XAI-enhanced ap- proaches, it highlights the superior predictive accuracy and balanced perfor- mance metrics of nine ML model. This study shows that ensemble methods like Random Forest and XGBoost perform well and highly effective in fore- casting HF patient mortality. Further, the integration of XAI techniques ad- dresses the interpretability challenges of conventional "black-box" nature of traditional models and enabling clinicians to better understand the decision- making process in heart failure. This interpretability is essential for fostering trust and facilitating the adoption of these models in clinical practice. Ad- ditionally, the study underscores the importance of explainable evaluation metrics, such as precision, recall, and F1-scores, which provide deeper in- sights into model performance and ensure alignment with clinical priorities, such as minimizing false negatives in HF detection. Overall, the findings emphasize the potential of XAI-enhanced models to not only improve pre- dictive accuracy but also enhance clinical utility by offering interpretable and actionable insights, paving the way for more personalized and effective HF management strategies.

The result of the ensemble models like Random Forest and XGBoost stand out, achieving the highest test accuracy (87.5% and 88.59%, respec- tively) and balanced metrics, making them the most effective for both classes. XGBoost, in particular, achieves the best F1-scores (0.87 for Normal and 0.9 for HeartFailure), indicating its robustness. Simpler models like Logistic Regression and KNN also perform well, offering competitive accuracy and interpretability. Gradient Boost and AdaBoost show similar performance to KNN but slightly lower accuracy than Random Forest and XGBoost. SVM, however, struggles with overfitting, achieving high training accuracy (99%) but poor test accuracy (59%), and fails to generalize effectively. Overall, the results highlight the superiority of ensemble methods for accurate and reli- able heart failure prediction, while simpler models provide a balance between performance and interpretability.

ACKNOWLEDGMENTS

Author thaks to the Quantum University Roorkee for there moral and technical support.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT *(mandatory)* (10 PT)

All the author in this work have equal contribution. The contributing author work under the supervison of second author with following 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
Mukesh Kumar Tiwari	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Brij Mohan Singh		✓				✓		✓	✓	✓	✓	✓		

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

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

This study utilizes the Fedesoriano: Heart failure prediction dataset. This dataset is publicaly available. As part of our commitment to supporting open research, we provide a data availability statement in order to be accepted for publication. Examples:

The data that support the findings of this study are openly available in, kaggle, sept. 2021. [On- line]. Available: https://www.kaggle.com/datasets/fedesoriano/heart- failure-prediction.

REFERENCES

[1] G. A. Roth, G. A. Mensah, C. O. Johnson, G. Addolorato, E. Ammirati,L. M. Baddour, N. C. Barengo, A. Z. Beaton, E. J. Benjamin, C. P. Ben- ziger et al., “Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the gbd 2019 study,” Journal of the American college of cardiology, vol. 76, no. 25, pp. 2982–3021, 2020.

[2] A. Daza, J. Bobadilla, J. C. Herrera, A. Medina, N. Saboya, K. Zavaleta, and S. Siguenas, “Stacking ensemble based hyperparameters to diagnos- ing of heart disease: Future works,” Results in Engineering, vol. 21, p. 101894, 2024.

[3] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, “Prediction of heart disease using machine learning,” in 2018 second interna- tional conference on electronics, communication and aerospace technol- ogy (ICECA). IEEE, 2018, pp. 1275–1278.

[4] D. Paikaray and A. K. Mehta, “An extensive approach towards heart stroke prediction using machine learning with ensemble classifier,” in Proceedings of the International Conference on Paradigms of Commu- nication, Computing and Data Sciences: PCCDS 2021. Springer, 2022,pp. 767–777.

[5] [P. A. Moreno-Sanchez, “Improvement of a prediction model for heart failure survival through explainable artificial intelligence,” Frontiers in Cardiovascular Medicine, vol. 10, p. 1219586, 2023.

[6] C. W. Tsao, A. W. Aday, Z. I. Almarzooq, C. A. Anderson, P. Arora,C. L. Avery, C. M. Baker-Smith, A. Z. Beaton, A. K. Boehme, A. E. Buxton et al., “Heart disease and stroke statistics—2023 update: a re- port from the american heart association,” Circulation, vol. 147, no. 8,pp. e93–e621, 2023.

[7] A. F. Members, J. J. McMurray, S. Adamopoulos, S. D. Anker, A. Auric- chio, M. B¨ohm, K. Dickstein, V. Falk, G. Filippatos, C. Fonseca et al., “Esc guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The task force for the diagnosis and treatment of acute and chronic heart failure 2012 of the european society of cardiol- ogy. developed in collaboration with the heart failure association (hfa) of the esc,” European heart journal, vol. 33, no. 14, pp. 1787–1847, 2012.

[8] S. Mohan, C. Thirumalai, and G. Srivastava, “Effective heart disease prediction using hybrid machine learning techniques,” IEEE access, vol. 7, pp. 81 542–81 554, 2019.

[9] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. Al Ma- mun, and M. S. Kaiser, “Performance analysis of machine learning ap- proaches in stroke prediction,” in 2020 4th international conference on electronics, communication and aerospace technology (ICECA). IEEE, 2020, pp. 1464–1469.

[10] M. Zeng, “The prediction of heart failure based on four machine learning algorithms,” Highlights in Science, Engineering and Technology, vol. 39,pp. 1377–1382, 2023.

[11] S. Sharma and S. Jain, “The coronavirus disease ontology (covido),” in Semantic Intelligence: Select Proceedings of ISIC 2022. Springer Nature Singapore Singapore, 2023, pp. 89–103.

[12] F. H. Dahri, A. A. Laghari, D. K. Sajnani, A. Shazia, and T. Kumar, “Heart failure prediction: a comparative analysis of machine learning algorithms,” in International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), vol. 13395. SPIE, 2024,pp. 634–640.

[13] A. Daza Vergaray, J. Bobadilla Cornelio, J. C. Herrera Miranda,A. Medina Ch´avez, N. Saboya Rios, K. Zavaleta Ramos, and S. M.Sigu¨en˜as Gonzales, “Stacking ensemble based hyperparameters for the prediction of heart disease,” Available at SSRN 4553047.

[14] A. Sutradhar, M. Rafi, M. J. Alam, and S. Islam, “An early warning system of heart failure mortality with combined machine learning meth- ods,” Indones. J. Electr. Eng. Comput. Sci, vol. 32, pp. 1115–1122, 2023.

[15] M. M. Hassan, M. Raihan, M. H. B. Khan, T. Dhali, M. M. Rahman, andZ. H. Sneha, “A hybrid machine learning approach to predict the risk of having stroke,” in Proceedings of the 2nd International Conference on Computing Advancements, 2022, pp. 460–465.

[16] M. M. Yaqoob, M. Nazir, M. A. Khan, S. Qureshi, and A. Al-Rasheed, “Hybrid classifier-based federated learning in health service providers for cardiovascular disease prediction,” Applied Sciences, vol. 13, no. 3,p. 1911, 2023.

[17] S. Cyriac, R. Sivakumar, N. Raju, and Y. W. Kim, “Heart disease prediction using ensemble voting methods in

machine learning," in 2022 13th International Conference on Information and Communica- tion Technology Convergence (ICTC). IEEE, 2022, pp. 1326–1331.

[18] A. Abbas, A. Imran, A. A. N. Al-Aloosy, S. Fahim, A. Alzahrani, andS. K. Muzaffar, "Heart failure prediction using machine learning ap- proaches," in 2022 Mohammad Ali Jinnah University International Con- ference on Computing (MAJICC). IEEE, 2022, pp. 1–7.

[19] P. M. Tambe and M. Shrivastava, "Hybrid brave-hunting optimisation for heart disease detection model with svm coupled deep cnn," Interna- tional Journal of Intelligent Information and Database Systems, vol. 17, no. 1, pp. 92–123, 2025.

[20] Z. Y. Acar and U¨ . Tok, "Combining lstm-enhanced features with ma- chine learning algorithms for improved heart failure prediction," Selcuk University Journal of Engineering Sciences, vol. 23, no. 2, pp. 48–53, 2024.

[21] S. Basak and K. Chatterjee, "Smart healthcare surveillance system using iot and machine learning approaches for heart disease," in International Conference on Advancements in Smart Computing and Information Se- curity. Springer, 2022, pp. 304–313.

[22] S. A. Almutairi, "An optimized feature selection and hyperparameter tuning framework for automated heart disease diagnosis." Computer Systems Science & Engineering, vol. 47, no. 2, 2023.

[23] N. Afiatuddin, R. Rahmaddeni, F. Pratiwi, R. Septia, and H. Hen- drawan, "Evaluation of data mining in heart failure disease classfica- tion," CogITo Smart Journal, vol. 10, no. 2, pp. 460–473, 2024.

[24] J. Barzola-Monteses, R. Caicedo-Quiroz, F. Parrales-Bravo, C. Medina- Suarez, and W. Yanez-Pazmino, "Convolutional neural networks applied to the diagnosis of cardiovascular disease," in 2024 43rd International Conference of the Chilean Computer Science Society (SCCC). IEEE, 2024, pp. 1–8.

[25] S. Sharma and S. Jain, "Comprehensive study of semantic annotation: Variant and praxis," Advances in Computational Intelligence, its Con- cepts Applications (ACI 2021), vol. 2823, pp. 102–116, 2021.

[26] S. Sharma and S. Jain "Ontoxai: a semantic web rule language approach for dengue fever classification using explainable ai and ontology," Available at SSRN 4726837, 2024.

[27] S. Jain, S. Sharma, J. M. Natterbrede, and M. Hamada, "Rule-based actionable intelligence for disaster situation management," International Journal of Knowledge and Systems Science (IJKSS), vol. 11, no. 3, pp. 17–32, 2020.

[28] S. Sharma, S. Jain, M. K. Tiwari, and S. lal, "Classifying the state of knowledge-based question answering: patterns, progress, and prospects," International Journal of Computers and Applications, vol. 47, no. 1, pp. 93–105, 2025.

[29] Fedesoriano: Heart failure prediction dataset, kaggle, sept. 2021. [On- line]. Available: https://www.kaggle.com/datasets/fedesoriano/heart- failure-prediction.

[30] H.-A. Park, "An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain," Journal of Korean academy of nursing, vol. 43, no. 2, pp. 154–164, 2013.

[31] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classi- fiers," Machine learning, vol. 29, pp. 131–163, 1997.

[32] Y. Liao and V. R. Vemuri, "Use of k-nearest neighbor classifier for in- trusion detection," Computers & security, vol. 21, no. 5, pp. 439–448, 2002.

[33] L. E. Peterson, "K-nearest neighbor," Scholarpedia, vol. 4, no. 2, p. 1883, 2009.

[34] D. A. Pisner and D. M. Schnyer, "Support vector machine," in Machine learning. Elsevier, 2020, pp. 101–121.

[35] V. N. Vapnik, "Pattern recognition using generalized portrait method,"Automation and remote control, vol. 24, no. 6, pp. 774–780, 1963.

[36] A. J. Albert, R. Murugan, and T. Sripriya, "Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology," Research on Biomedical Engineering, vol. 39, no. 1, pp. 99–113, 2023.

[37] U. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," Phys. Rev E., vol. 76, p. 036106, 2007.

[38] D. Garreau and U. Luxburg, "Explaining the explainer: A first theoreti- cal analysis of lime," in International conference on artificial intelligence and statistics. PMLR, 2020, pp. 1287–1296.

[39] C. Biffi, J. J. Cerrolaza, G. Tarroni, W. Bai, A. De Marvao, O. Oktay,C. Ledig, L. Le Folgoc, K. Kamnitsas, G. Doumou et al., "Explainable anatomical shape analysis through deep hierarchical generative models," IEEE transactions on medical imaging, vol. 39, no. 6, pp. 2088–2099, 2020.