Design and Development of an Efficient Explainable AI Framework for Heart Disease Prediction

Deepika Tenepalli, Navamani T M*
SCOPE, VIT Vellore, Tamil Nadu, 632014, India

***Abstract*—Heart disease remains a global health concern, demanding early and accurate prediction for improved patient outcomes. Machine learning offers promising tools, but existing methods face accuracy, class imbalance, and overfitting issues. In this work, we propose an efficient Explainable Recursive Feature Elimination with eXtreme Gradient Boosting (ERFEX) Framework for heart disease prediction. ERFEX leverages Explainable AI techniques to identify crucial features while addressing class imbalance issues. We implemented various machine learning algorithms within the ERFEX framework, utilizing Support Vector Machine-based Synthetic Minority Over-sampling Technique (SVMSMOTE) and SHapley Additive exPlanations (SHAP) for imbalanced class handling and feature selection with explainability. Among these models, Random Forest and XGBoost classifiers within the ERFEX framework achieved 100% training accuracy and 98.23% testing accuracy. Furthermore, SHAP analysis provided interpretable insights into feature importance, improving model trustworthiness. Thus, the findings of this work demonstrate the potential of ERFEX for accurate and explainable heart disease prediction, paving the way for improved clinical decision-making.**

***Keywords*—Machine learning; heart disease; explainable AI; XGBoost; SHAP**

I. Introduction

The heart is a very important organ in the human body. It transports blood, oxygen, and other materials to the body's organs via the circulatory system's blood veins. While an artery in the chest is partially or fully clogged by cholesterol or a blood clot, blood supply to the heart tissue is decreased or stopped. This may damage or destroy heart muscle cells, resulting in a heart attack [1]. Cardiovascular diseases (CVDs), encompassing a range of heart and blood vessel disorders, pose the deadliest hazards in the world. Statistics demonstrate that over 17 million lives are lost tragically to CVDs each year. The World Health Organization (WHO) anticipates this figure will grow drastically to 23 million by 2030 [2]. This concerning trend suggests an ominous future unless significant advances in early identification and prevention measures are implemented. Early detection is critical for efficiently managing cardiac disease and improving patient outcomes. Healthcare providers can avoid heart attacks, strokes, and other serious problems by identifying high-risk individuals early on [2]. The main causes of heart disease are unhealthy habits like smoking, excessive alcohol drinking, unhealthy diet, being physically inactive, diabetes, obesity, stress, high cholesterol, high blood pressure, age, gender, genetics, etc. [2]. Heart disease is one of the most serious diseases that can be recognized by monitoring symptoms and receiving alerts from the devices before an attack happens. The symptoms found are chest pain, discomfort in body parts like the back pain, abdomen, or jaw, left arm, and breathing difficulty [3]. Early detection of the disease will help the patients from the extreme damage. This leads to the demand for advancements in early diagnosis and accurate prediction. Although traditional diagnostic techniques have long been used to detect cardiac disease, they frequently have limitations that hinder prompt and accurate diagnosis. These methods often rely on tracking symptoms such as chest discomfort, shortness of breath, and exhaustion. However, these symptoms may not always be present or obvious, especially in the early stages of the disease. Furthermore, traditional approaches frequently include intrusive procedures such as stress tests and angiograms, which can be costly, time-consuming, and even risky for patients [4].

In recent years, Artificial intelligence (AI) and Machine learning (ML) have been revolutionary technologies with significant impact on healthcare and personalized clinical support [5]. Early identification is critical for successfully managing heart disease. Machine learning algorithms can evaluate vast amounts of patient data, revealing hidden patterns and risk factors that older methods may overlook. This enables healthcare practitioners to identify high-risk patients earlier, allowing them to implement preventative tactics and therapies. Consider an era in which routine checks include an AI-powered system that analyzes medical data and identifies potential risks for heart disease before symptoms occur. Early identification, combined with preventative interventions, can greatly improve patient outcomes and perhaps save lives [6]. Machine learning and Artificial intelligence have been used for the prediction and diagnosis of Chronic diseases like Heart Disease, Cervical Cancer, Breast Cancer, Lung Cancer, etc. Machine Learning Techniques are also used to improve the prediction accuracy for heart disease predictions [2]. ML is a technique in which a machine is trained automatically rather than explicitly personalized [7]. ML has been utilized for disease prediction or to find the Risk level or survival of the patients. Logistic Regression (LR), Support Vector Machines (SVM), K-Nearest Nighbor (KNN), Decision Trees (DT), Random Forest (RF), Naive Bayes (NB), ADA Boosting, Gradient Boosting (GB), etc. are the most used algorithms in disease prediction. Heart disease prediction is being greatly aided by deep-learning neural networks as well as machine learning algorithms. These models aid in analyzing the vast amount of data that doctors have at their disposal, look for patterns in diagnosis, streamline the process, and integrate patient records to reduce errors [8]. It is generally known that early detection of subsequent heart attacks is crucial for both providing emergency care and preventing deadly consequences [9].

*Corresponding authors

Today's healthcare system has challenges in providing high-quality, efficient, and effective services. Heart disease is the main cause of death globally. The ability to handle an illness accurately depends on its detection time [10]. Machine learning allows for the discovery of hidden patterns in data. Mondal, S., et al. [11] proposed a model for cardiovascular diseases using a two-stage stacking approach with machine learning algorithms. It shows a potential improvement in risk prediction. It is observed that feature selection can be improved better, and a generalized model can be considered. Subathra, R., & Sumathy, V. [12] discussed the heart disease prediction model by utilizing a swarm optimization technique with ensemble learning. However, the issues like early detection, versatility, and accuracy can be improved. Rani, P. et al. [13] presented a survey on heart disease classification and predictions using machine learning and deep learning techniques. In that study, the main challenges faced in heart disease were missing values in the dataset, unbalanced datasets, irrelevant features, and different types of attributes. Manikandan, G., et al. [14] implemented a prediction model using machine learning algorithms such as logistic regression, Decision trees, and Support Vector Machine, along with Boruta feature selection to predict heart disease. This model provides improved performance and feature selection but suffers potential performance reduction for certain algorithms such as Random Forest and XGBoost after feature selection. It is observed that most of the prediction models suffer from less accuracy, class imbalance, feature selection, and overfitting issues [15] [16].

To address these issues, we propose an Explainable Recursive Feature Elimination with the eXtreme Gradient Boosting (ERFEX) framework for heart disease prediction. To handle class imbalance, overfitting, and better feature selections, the Support Vector Machine-based SMOTE (SVMSMOTE) technique and Recursive Feature Elimination (RFE) are employed. Explainable AI technique SHapley Additive exPlanations (SHAP) is utilized to enhance the trustworthiness of our prediction model. Using this model, we examined different machine learning algorithms like Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), Logistic Regression (LR), and Ada Boosting Classifier, eXtreme Gradient Boosting (XGB) Classifier, and Gaussian Naive Bayes (GNB) to classify and predict the heart disease. Since maintaining class balance is essential for developing effective heart disease prediction algorithms, the Support Vector Machine-based Synthetic Minority Over-sampling Technique (SMOTE) was employed.

The main contributions of our work are as follows:

- We propose an Efficient Explainable Recursive Feature Elimination with eXtreme Gradient Boosting (ERFEX) framework for heart disease prediction: This framework combines RFE with XGB for feature selection and prediction, potentially improving accuracy and interpretability.
- Systematic comparative analysis of various machine learning algorithms such as KNN, SVM, RF, DT, LR, MLP, XGB, AdaBoost, and GNB is performed, and also identified the suitable algorithm for better prediction of heart diseases.
- Addressing class imbalance and feature selection: We incorporate SVMSMOTE for handling imbalanced classes and RFE for selecting the most relevant features. This helps to improve the effectiveness of the chosen models and potentially reduces overfitting.

The remaining Sections of this work is organized as Related Work in Section 2, Materials and Methodology in Section 3, and Result Analysis and Discussion in Section 4. Finally, the conclusion and future work are discussed in Section 5.

II. Related Work

Recently researchers have been interested in developing technologies and approaches to monitor and forecast diseases that significantly impact people's health. Here, Heart disease prediction and classification works are discussed.

Paudel, P. et al. [1] focused on early heart attack detection using machine learning and explainable AI (XAI) techniques. The study compares the performance of different classification algorithms, including AdaBoost, Random Forest, Gradient Boosting Classifier, and Light Gradient-Boosting Machine (LGBM), in predicting heart diseases. LGBM algorithm showed better performance in terms of accuracy. Jafar, A., & Lee, M., [2] presented the development of the HypGB model, a high-accuracy heart disease prediction system that utilizes Gradient Boosting (GB) modeling and the LASSO feature selection technique. The model needs to enhance the heart disease prediction accuracy. Tn, K. et al. [3] presented a comprehensive study on the development of a data-driven prediction model for the early detection of heart disease. By evaluating multiple machine learning algorithms, such as SVM, Random Forest, KNN, and others, the research demonstrates that Random Forest outperforms other models. The model needs to be expanded and improved to predict all the similar types of cardiovascular diseases. Javaid, M. et al. [5] discussed the growing impact of machine learning (ML) applications in healthcare, emphasizing its potential to enhance the speed and accuracy of physicians' work and alleviate the challenges posed by overburdened healthcare systems and shortages of skilled physicians.

In our previous work [7], we presented a review on the integration of machine learning, blockchain technology, and cloud computing in the e-healthcare domain. It emphasizes the advantages of cloud services in providing flexible and affordable access to patients' Electronic Health Records (EHR) while highlighting the crucial concerns regarding EHR security and privacy. Also, we addressed the need for prediction techniques for chronic diseases in their early stages, emphasizing the potential of machine learning and blockchain technology in improving diagnosis and prognosis. Amato, F., et al. [8] provided a comprehensive overview of the application of Artificial Neural Networks (ANNs) in medical diagnosis, highlighting their potential to streamline the diagnostic process and prevent misdiagnosis. ANNs, as a form of artificial intelligence, are adopted for handling diverse medical data, including clinical symptoms, biochemical information, and imaging outputs, and integrating them into categorized outputs. The review discusses the capabilities and limitations of ANNs through selected examples, showcasing their use in diagnosing conditions such as cancer, cardiovascular diseases, and diabetes. Dev, S., et al.

[10] presented a comprehensive approach to stroke prediction using machine learning and neural networks. However, there exists potential overfitting, reliance on a single dataset, and lack of consideration for all confounding variables in stroke prediction.

Mishra, I., & Mohapatra, S. [15] presented an improved method for evaluating the effectiveness of cardiac stroke prediction using machine learning approaches. It emphasizes the significance of early detection of strokes and their potentially distressing effects, highlighting the role of rapidly evolving AI/ML models in uncovering significant risk factors and estimating stroke probability. The potential drawback of the model is overfitting. Sharma, C., et al. [17] worked on various machine learning classification algorithms such as Naïve Bayes (NB), Random Forest (RF), Decision Trees (DT), Multilayer Perceptron (MLP), and JRip Algorithm. The results showed that the Random Forest Classifier got the highest accuracy. Venkata MahaLakshmi, N., & Rout, R. K. [18] presented an intelligent health monitoring framework for heart disease prediction, utilizing deep learning models and function fusion to enhance diagnostic accuracy. This approach incorporates evolutionary search, optimized ensemble classifier, and integrated filter-based feature selection for accurate diagnosis of heart disease.

Rimal, Y., et al. [16] compared the performance and accuracy of different machine learning models for predicting heart disease, with a focus on ensemble learning and AutoML. The researchers analyzed the correlation between variables using a cluster map and split the data into training and test sets. They compare 18 different models, including eight individually trained models and 10 from AutoML, using boosting, bagging, and voting algorithms. However, it must be improved in terms of accuracy and overfitting issues. Hera, S. Y., et al. [19] proposed a multi-tier ensemble model for improved diagnosis of heart disease using machine learning techniques. The selection of 3-tier can be made automated and needs to be implemented for all machine learning algorithms. Asif, S., et al. [20] proposed an ensemble machine learning approach to improve the accuracy of detecting and predicting coronary heart disease utilizing Random Forest, XGBoost, and Gradient Boosting. Still, it requires accuracy needs to be improved.

Uma Maheswari, K., & Valarmathi, A. [21] presented an approach for predicting heart disease using a Support Vector Machine (SVM) classification with an Optimized Deep Belief Network (DBN). The authors emphasize the importance of accurate diagnosis and treatment of heart disease, highlighting the limitations of conventional diagnostics. However, the technique can be improved better by utilizing hybrid techniques and real-time medical datasets for development. Isik, I. [22] discussed the efficiency and precision of various algorithms and techniques in medical diagnostics, particularly in the context of heart disease detection. This has highlighted the effectiveness of approaches such as random forest, KNN, and SVM in achieving high accuracy rates for heart sound classification.

In summary, recent advancements in heart disease prediction using machine learning and AI have shown significant progress and innovation. Techniques such as Light Gradient-Boosting Machine (LGBM) [1], HypGB model [2], and Random Forest [3] have achieved high accuracy in specific contexts, though they also underscore the need for broader applicability and further accuracy enhancements. Research has highlighted the potential of machine learning to improve healthcare efficiency and the integration of secure, accessible Electronic Health Records (EHR) through blockchain technology [7]. While Artificial Neural Networks (ANNs) and deep learning frameworks hold promise for diagnostics, they encounter challenges like overfitting and the necessity for real-time data integration [8] [10] [18]. Thus, despite these substantial strides, future research must address the ongoing issues of prediction accuracy, model generalization, and comprehensive data utilization to fully harness the potential of these technologies in clinical settings. In our work, we propose the ERFEX framework for heart disease prediction to enhance the prediction accuracy and reduce the overfitting problem.

III. Materials and Methodology

Heart disease is becoming one of the most common diseases that occur due to several reasons. Early prediction of heart disease is essential to begin the treatment to avoid great losses. In our work, we implemented an ERFEX framework to predict heart diseases at their early stages. In this section, we first discuss the Dataset used in our work. Then we describe the methodologies used in this heart disease prediction model.

A. Dataset Description

In this work, the publicly available Heart Attack Dataset [23] is utilized with eight features including 'age', 'gender', 'impulse', 'pressurehight', 'pressurelow', 'glucose', 'kcm', 'troponin' and a target class as 'Class' with 'Positive' and 'Negative' samples are presented in Table I.

Cardiovascular Diseases (CVDs) are the primary cause of death worldwide. Heart and blood vessel problems collectively known as CVDs include conditions like rheumatoid heart disease, coronary heart disease, and cerebrovascular illness. A thorough database of the elements that lead to a heart attack has been created [23].

TABLE I. Dataset Attributes

Attribute	Description	Type
Age	Age in number, minimum age is 14 and maximum age is 103	Int64
Gender	0 for female, 1 for male	Int64
Impulse	Heart Rate	Int64
Pressure High	Systolic BP	Int64
Pressure Low	Diastolic BP	Int64
Glucose	blood sugar	Float64
Kcm	Creatine kinase Myocardial Band	Float64
Troponin	troponin is a protein complex found in the heart muscle cells	Float64
Class	Output class Heart Attack Presence Positive or Negative	object

B. Methodology

Fig. 1 shows the sequence of processes involved in the proposed architecture model. Initially, Data pre-processing is performed by identifying null and missing values. Here the selected data set does not have any null or missing values. Data is encoded to convert all features into the same type to process further and create a new target value. The collected patient data encompasses medical history, demographics, and lifestyle factors. This data is then pre-processed to address missing values and inconsistencies. Techniques like filling in

missing entries, and encoding categorical data are employed. Normalization is performed using the StandardScalar to ensure all features are on a similar scale, preventing features with larger ranges from dominating the model during training.

An exploration of data distribution, possible correlations between variables, and outlier detection are the goals of exploratory data analysis. This helps in comprehending the data and selecting the most relevant features for model building. Not all collected data is equally important; feature selection techniques are used to identify the most relevant features that best correlate with heart disease risk. In our work, we utilized the RFE technique for extracting the important features in our prediction model. The important features identified are troponin, kcm, gender, and so on.

SVM-SMOTE is a technique for dealing with imbalanced datasets in machine learning. Imbalanced datasets, where one class has significantly more data points than another, can confuse algorithms. SVM-SMOTE tackles this by focusing on the minority class. It first trains a model called a Support Vector Machine (SVM) to identify the decision boundary, the line that separates the classes. Then, it creates new synthetic data points for the minority class, specifically around this decision boundary. This helps the model to understand the important areas for classification better and improve its accuracy for the minority class.

Several machine learning algorithms are then explored for model building. These algorithms include SVM, DT, LR, XGBoost, RF, KNN, AdaBoost, GNB, and MLP. Each algorithm has its own strengths and weaknesses, and experimentation analysis identifies the model with the better performance on the given dataset. Our research found that XGBoost and Random Forest classifiers with RFE outperformed in all aspects. The model development utilized the different parameters of each model. In KNN, parameters considered are n_neighbors, weights, algorithm, metric, and leaf_size. Likewise, XGBoost parameters are maximum depth, alpha, learning rate, and number of estimators.

The model's performance is assessed by a range of criteria, including F1 Score, accuracy, precision, and recall. These measures shed light on how well the model predicts the heart disease risk. After evaluating different models, XGBoost and Random Forest were chosen due to their superior performance in all aspects. The data is split into two mutually exclusive subsets: a training set and a testing set. The training set (comprising 70% of the data) was used to develop the model. The testing set (comprising 30% of the data) served to evaluate the model's generalizability to unseen data.

SHAP (SHapley Additive exPlanations) is used to explain the model's predictions. SHAP highlights the features that most influence a specific prediction, making the model's reasoning more transparent. In our work, troponin and kcm were identified as having the highest importance compared to other features.

C. Feature Selection

Feature selection techniques are used to identify the most relevant features that best correlate with heart disease risk. In our work, Recursive Feature Elimination (RFE) is used to simplify the model by separating the most significant features from the least significant ones. Building a pipeline of classification models, oversampling the data to adjust for class imbalance, and using cross-validation to assess model performance are all part of the model training process. The evaluation procedure involves comparing the models using various measures such as precision, F1-score, accuracy, and recall. Feature selection is essential for reducing the dimensionality of the input, improving model performance by using pertinent data, and enabling the model to represent the deeper trends more accurately. The feature importance of the attributes is determined using the XGBoost Classifier [24] [25]. The XGBoost Classifier is an embeddable algorithm that maximizes model performance by gradient boosting, utilizing tree-based techniques.

Feature selection techniques are used to identify the most relevant features that best correlate with heart disease risk. In our work, Recursive Feature Elimination (RFE) is used to simplify the model by separating the most significant features from the least significant ones. Building a pipeline of classification models, oversampling the data to adjust for class imbalance, and using cross-validation to assess model performance are all part of the model training process. The evaluation procedure involves comparing the models using various measures such as precision, F1-score, accuracy, and recall. Feature selection is essential for reducing the dimensionality of the input, improving model performance by using pertinent data, and enabling the model to represent the deeper trends more accurately. The feature importance of the attributes is determined using the XGBoost Classifier [24] [25]. The XGBoost Classifier is an embeddable algorithm that maximizes model performance by gradient boosting, utilizing tree-based techniques.

D. Model Training

In model training, the model is fed with the data to learn from experience then new or unknown data is fed to the model for test. In our work, 70% of the data is used for training the model, and 30% of the data is used for testing purposes. Upon evaluating each feature's significance, the most crucial characteristics were selected, and the other features were eliminated from the dataset to enhance the model's overall performance and data quality. To deliver an accurate prediction and risk percentage, we employ a pipeline of models to determine which model is best suited for training with a high degree of precision.

A binary classification problem is said to have class imbalance when one class contains substantially fewer samples than the other. Because it will be biased in favor of the dominant class in these situations, a model's performance can be inferior. Oversampling requires the production of artificial minority samples to balance the class distribution. These synthetic samples can be produced using methods like Random Oversampling, SMOTE, or ADASYN [26] [27]. The model's accuracy might not be up to par because the dataset only has 1319 samples and over 9 characteristics. We developed a pipeline that included popular machine learning classification models such as Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Multilayer Perceptron (MLP), Ada Boosting Classifier, eXtreme Gradient Boosting (XGB) Classifier and, Gaussian Naive Bayes (GNB). The models' performance was then enhanced by fine-tuning them using hyperparameters to

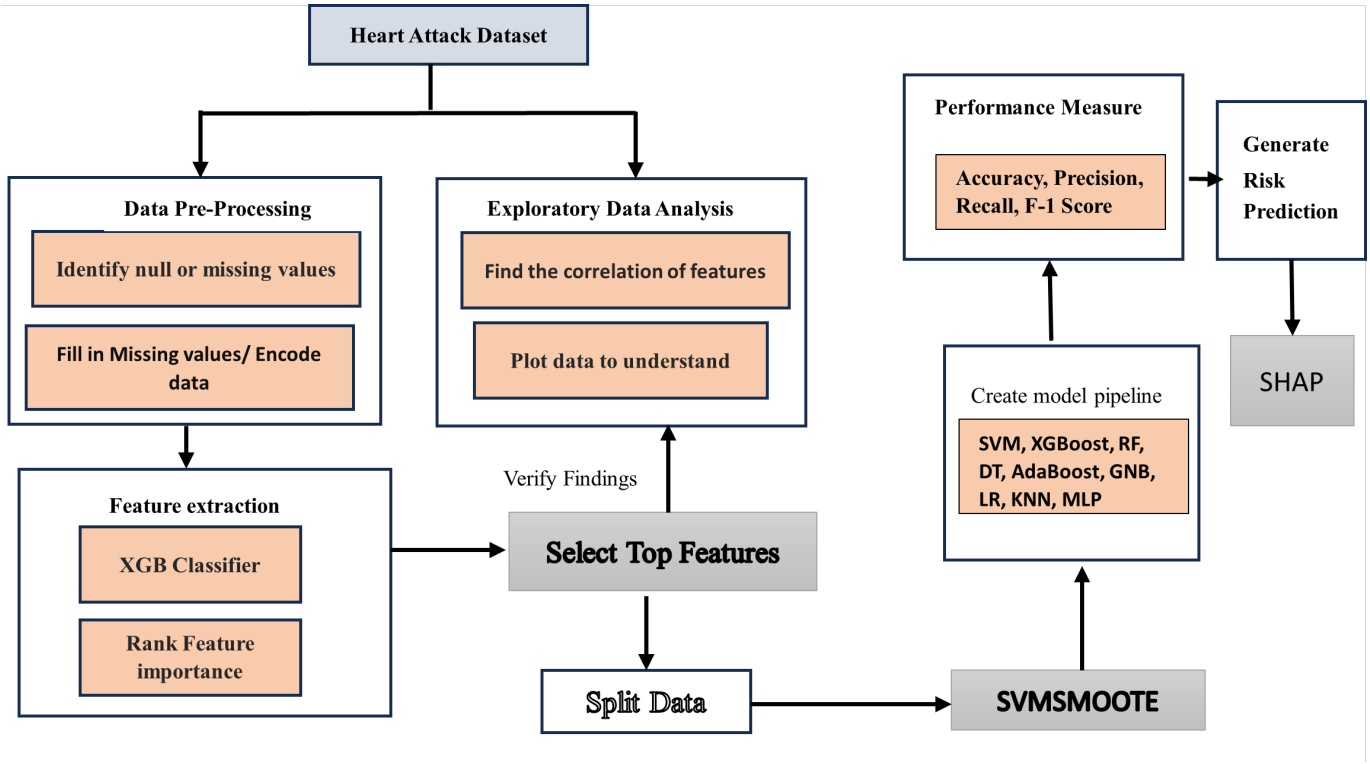

Fig. 1. Architecture of the proposed framework.

increase the model's sensitivity and adaptability to the features of the data.

E. Model Evaluation

The process of model evaluation assesses a model's effectiveness in accomplishing heart disease prediction. Performance metrics like Training Accuracy, Testing Accuracy, F1-Score, precision, and recall, are used to compare the models. Eq. (1) to (5) are used for performance analysis. The performance of a machine learning model on the data it was trained on is referred to as training accuracy [27]. It calculates the model's accuracy percentage based on training data predictions.

$$Training\ Accuracy = \frac{P}{Q} \tag{1}$$

where,

P: Number of correctly predicted training examples

Q: Total number of training examples

The performance of a machine learning model on data that it hasn't seen during training is referred to as testing accuracy [27]. It calculates the model's accuracy percentage in predicting the test set of data.

$$Testing\ Accuracy = \frac{R}{S} \tag{2}$$

where,

R: Number of correctly predicted test examples

S: Total number of test examples

Precision [27] assesses the percentage of accurate predictions a model produces. It explains how the model can prevent false positives. A high precision score means that the model predicts few false positives.

$$Precision = \frac{True\ Positive}{True\ positive + False\ Positive} \tag{3}$$

Recall [27] assesses the percentage of accurate forecasts that turn out to be favorable about all the real cases of positive data. It explains how the model may identify positive examples. A high recall score means that most positive data items are accurately identified by the model.

$$Recall = \frac{True\ Positive}{True\ positive + False\ Negative} \tag{4}$$

The F1-Score is a metric that balances precision and recall by calculating their harmonic mean [27]. It provides a single figure that sums up the precision and recall capabilities of the model. A model with a high F1 score demonstrated excellent precision and recall performance.

$$F1 - Score = \frac{2 * Precision * Recall}{(Precision + Recall)} \tag{5}$$

IV. Result Analysis and Discussion

The outcomes of the feature extraction, oversampling, and model evaluation are presented here. Feature selection was performed using Pearson's correlation coefficient, and Fig. 2 illustrates the correlation values between feature pairs in the dataset. Here the value varies from 1 to -1, where 0 indicates

no impact, -1 indicates negative impact and 1 indicates the positive impact. From Fig. 2, it is observed that there is a positive correlation between the pairs 'age and troponin', 'gender and kcm'. It was so determined that each of these elements ought to be incorporated into the model.

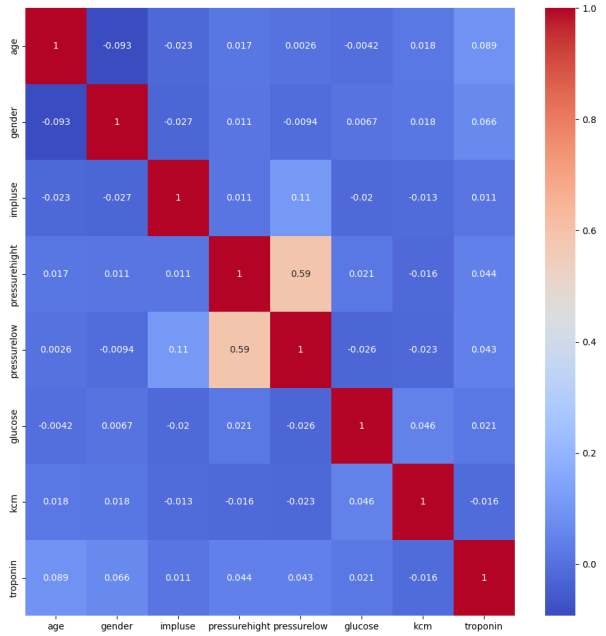

Fig. 2. Correlation coefficients.

As illustrated in Fig. 3, the significant features are arranged in order of relevance. A feature that is noticeably more substantial than the rest is depicted in Fig. 3. Troponin has the highest importance, then kcm, gender, and so on. Important Features are given to the models for further processing to enhance the model accuracy. Important features are extracted using the Recurrence Feature Elimination (RFE) technique with the XGB classifier.

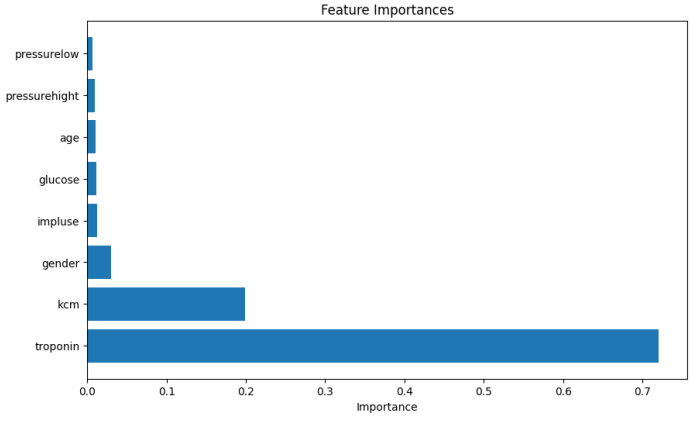

Fig. 3. Feature importance by RFE based XGBoost classifier.

Table II shows the performance of the Classification models. The models were trained on a training dataset, and then their performance was evaluated on a test dataset. From Table II, it is observed that the Random Forest Classifier, XGB Classifier, Decision Tree Classifier, and Ada Boost Classifier showed 100% training accuracy. Random Forest Classifier and XGB Classifier provide 98.23% testing accuracy over the remaining models. Random Forest Classifier outperformed in all aspects than all other classifiers. The Decision Tree Classifier and XGB classifier also performed well compared to other Classifiers. K Nearest Neighbor Classifier provided the least testing accuracy compared to others. The prediction results are also shown graphically as in Fig. 4. Here Y-axis shows the performance measures and the X-axis represents the prediction models.

Fig. 5 shows the confusion matrix which illustrates that among 396 samples 5 normal samples were misclassified as diseased and 3 diseased samples were classified as normal using the XGBoost classifier. Here 0 represents Normal and 1 represents heart disease.

Explainable AI can be used to provide a global view of the predictions. To gain insights into feature importance, we employed SHAP (SHapley Additive exPlanations), a technique within Explainable AI (XAI). SHAP assigns scores to each feature, reflecting its contribution to the model's output. Fig. 6 presents the SHAP feature importance plot, where the x-axis represents the impact on the predicted outcome, and the y-axis shows the features. As evident from the figure, troponin, and kcm have the highest impact on the model's predictions compared to other features.

Fig. 7 shows the model output value analysis of a machine learning model using SHAP. The x-axis shows input features, while the y-axis shows the corresponding model output value. Each data point is represented, and the line shows the overall trend of the data. The model output value ranges from 0.3 to 1.0. There is a positive correlation between the model output value and the input features. This means that as the values of the input features increase, the model output value also tends to increase. The most important input features for the model are "troponin", "kcm", "pressurehight", and "impulse". This is because these features have the largest effect on the model output value. The least important input features for the model are "gender", "glucose", and "age". This is because these features have the smallest effect on the model output value.

Fig. 8 shows the SHAP value for the feature troponin. This technique helps us to understand the inner workings of a machine learning model by analyzing how each input feature influences the final prediction. The x-axis of the plot depicts the SHAP value for each input feature. SHAP values can be positive or negative, and they indicate how much a particular feature increases or decreases the model's prediction. The y-axis shows the possible values of the troponin level. Each dot on the plot represents a different data point, and the color of the dot indicates the model's prediction for that data point. The darker the color, the higher the predicted troponin level. Fig. 9 shows the SHAP value for the feature kcm. The x-axis of the plot shows the SHAP value for each input feature. SHAP values can be positive or negative, and they indicate how much a particular feature increases or decreases the model's prediction. The y-axis shows the possible values of the kcm level.

TABLE II. PERFORMANCE OF ML MODELS

Model	Cross Validation Score	Training Accuracy	Testing Accuracy	Precision Score	Recall Score	F-1 Score
KNeighbors Classifier	71.89	79.61	62.88	72.82	62.24	67.11
SVC	74.35	74.96	65.15	77.54	60.17	67.76
RandomForestClassifier	98.06	100	98.23	98.75	98.34	98.54
LogisticRegression	74.88	76.19	69.19	75.54	73.03	74.26
DecisionTreeClassifier	97.97	100	97.98	98.34	98.34	98.34
AdaBoostClassifier	97.71	100	96.97	96.73	98.34	97.53
MLPClassifier	81.20	83.48	76.52	82.46	78.01	80.17
XGBClassifier	97.88	100	98.23	98.35	98.76	98.55
GaussianNB	92.53	92.53	90.40	99.03	85.06	91.52

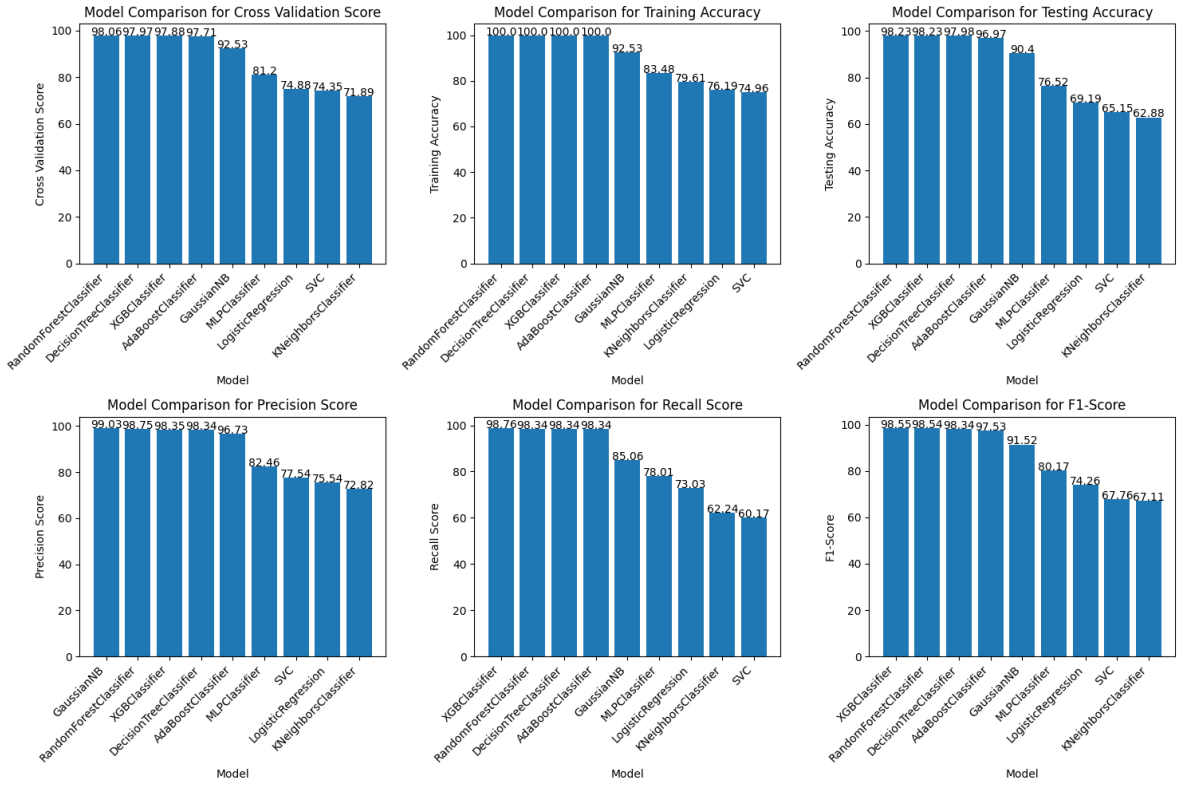

Fig. 4. Performance analysis of ML models.

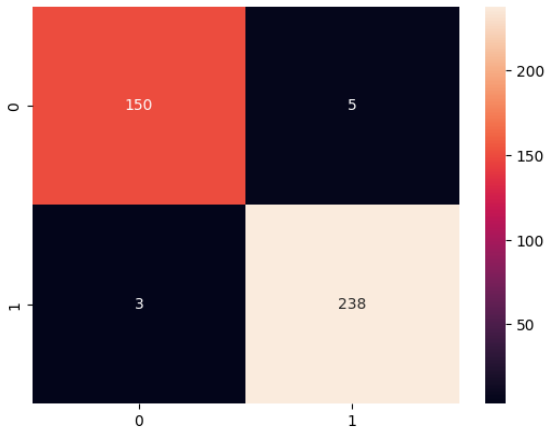

Fig. 5. Confusion matrix of XGBoost classifier.

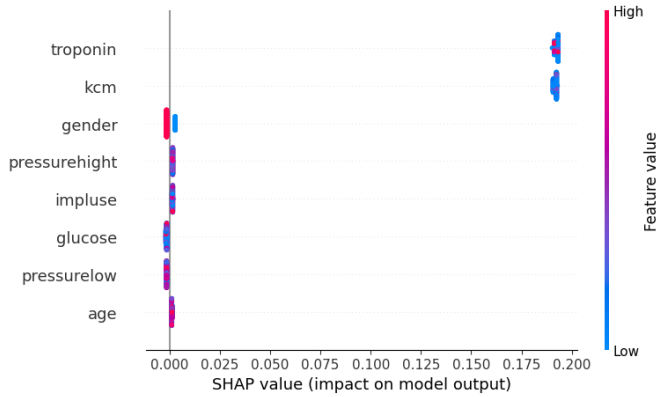

Fig. 6. Feature importance using SHAP.

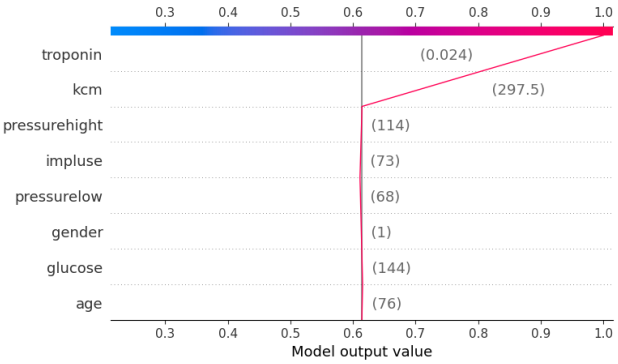

Fig. 7. Model output value analysis using SHAP.

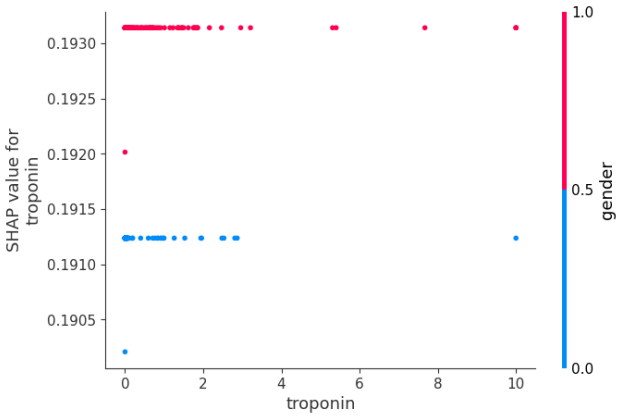

Fig. 8. SHAP value for Troponin.

From the obtained results, it is observed that a positive correlation between age and troponin indicates a potential rise in troponin levels (a marker for heart damage) concerning age. This aligns with the increased risk of heart disease in older populations. The association between gender and kcm (potassium level) warrants further investigation. While potassium imbalances can affect heart rhythm, understanding the gender-specific link could be crucial for personalized medicine approaches. The proposed ERFEX framework has shown better performance for XGBoost and Random Forest classifiers. This is accomplished through a combination of Recursive Feature Elimination for feature importance, SVM-SMOTE to address class imbalance and SHAP for explaining the model's predictions.

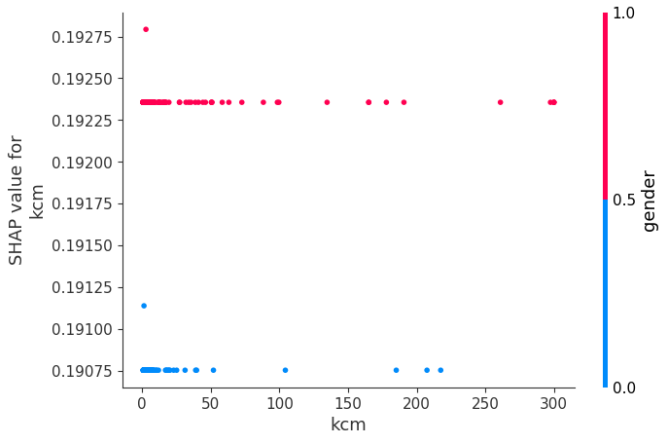

Fig. 9. SHAP value for kcm.

A. Comparative Analysis

Table III shows the performance variation of several algorithms that were assessed on various datasets related to heart disease. Tn, K., et al. [3] developed a comparative analysis of machine learning algorithms for heart disease prediction. The result of that analysis showed that Random Forest performed well compared to other algorithms. Guleria, P., et al. [28] proposed a framework to classify cardiovascular diseases using SVM, KNN, LR, Gaussian NB, AdaBoost, and Bagged Tree. In this model, the SVM classifier got 82.5% accuracy. Ali, F., et al. [29] implemented the Ensemble Deep Learning (DL) model and got 98.5% for the Cleveland dataset and not use any Explainable AI techniques. Paudel, P., et al. [1] developed a prediction model for early detection of heart attack with incorporation of LIME. The model got 99.33% of testing accuracy for the LGBM. From Table III, it is observed that the proposed model outperformed the XGBoost and Random Forest classifiers with a training accuracy of 100% and a testing accuracy of 98.23% than the remaining existing works.

B. Comparison of Performance Metrics for ML Models

Fig. 10 shows the comparison of performance metrics of ML models. It is observed that XGBoost, Decision Tree, Random Forest, and AdaBoost classifiers provided 100% training accuracy and performed well in all other aspects when compared to other models. K-Nearest Neighbor classifier performed least in all aspects except Training accuracy compared with other models. Gaussian Naïve Bayes (GNB) provided the highest precision value compared to other models and K-Nearest Neighbor provided the lowest precision value. XGBoost and Random Forest classifier provided the highest 98.23% testing accuracy and K-Nearest Neighbor provided the lowest accuracy of 62.88%. XGBoost and Random Forest classifiers provided a better recall value of 98% compared to the remaining models and the Support Vector Machine (SVM) classifier provided the lowest recall of 60%. XGBoost, Decision Tree, and Random Forest classifiers provided a better F1- Score value of 98% and K-Nearest Neighbor provided the lowest F1-Score of 67%. It is observed from Fig. 10 that XGBoost and Random Forest Classifiers outperformed well in all aspects compared to other models. Even though the proposed ERFEX Framework works better in the prediction of heart disease, it needs improvement in terms of testing accuracy, overfitting issues, and datasets.

V. Conclusion and Future Work

This work introduces ERFEX, a framework that combines Explainable Recursive Feature Elimination (ERFE) with eXtreme Gradient Boosting (XGBoost) for achieving accurate heart disease prediction. While many machine learning algorithms can achieve high accuracy during training, they often suffer from a lack of transparency in their decision-making process. ERFEX addresses this by incorporating a

TABLE III. COMPARATIVE ANALYSIS OF RELATED WORKS WITH THE PROPOSED MODEL

SNo	Related Work	Algorithms Implemented	Highest Training Accuracy	Highest Testing accuracy	XAI
1	[28]	SVM, KNN, LR, Gaussian NB, AdaBoost, Bagged Tree Dataset: Heart Disease	-	SVM -82.5%	SHAP
2	[30]	DT, RF, XGBoost, NB, KNN, SVM, LR, AdaBoost Dataset: UCI Vascular heart disease	-	DT and RF: 99%	No
3	[3]	LR, KNN, RF, SVM, Polynomial SVM, Gaussian SVM, Sigmoid SVM, bagging Classifier, AdaBoost, Gradient Boost, XGBoost, DT, Naïve Bayes Dataset: UCI Repository	-	RF-87.9%	No
4	[29]	Ensemble Deep learning Dataset: Cleveland	-	98.5%	No
5	[19]	Multi-Tier Ensemble (MTE) with RF feature selection Dataset: Cleveland and Statlog datasets	-	93.76%	No
6	[1]	AdaBoost, RF, GB, LGBM Dataset: Heart disease classification	LGBM- 99.33%	-	LIME
7	Proposed work (**ERFEX**)	RF, DT, XGBoost, Ada Boost, SVM, LR, KNN, MLP Dataset: Heart disease classification	XGBoost, Ada Boost, RF, DT-100%	RF and XGBoost-98.23%	SHAP

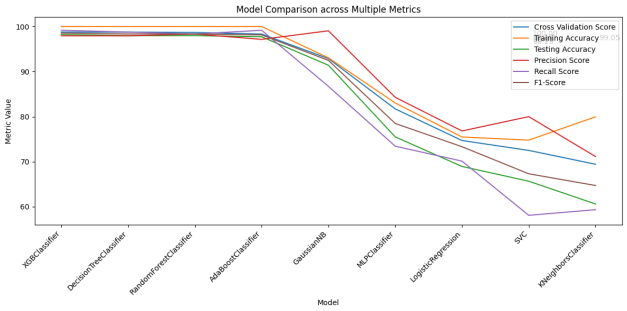

Fig. 10. Comparison of performance metrics of ML models.

technique that iteratively removes the least important features, ultimately leading to a more focused and interpretable model. This model is then built using XGBoost, a powerful machine learning algorithm known for its accuracy and efficiency. In this work, Random Forest and XGBoost classifiers using the ERFEX framework achieved an exceptional testing accuracy of 98.23%. This signifies the model's effectiveness in correctly identifying heart disease cases from the test data. Furthermore, to strengthen the trustworthiness of these predictions, Explainable AI techniques like SHAP were employed. SHAP helps us to understand which features in a patient's data most significantly influence the model's prediction of heart disease. This level of explainability is crucial in healthcare settings, as it allows healthcare professionals to not only rely on the prediction but also understand the reasoning behind it. Overall, these results suggest that the ERFEX framework has the potential to significantly improve heart disease prediction. By providing accurate and interpretable results, ERFEX can potentially aid healthcare professionals in the early detection and intervention of heart disease, leading to better patient outcomes. Future research works will be focused on validating ERFEX's efficacy on even larger and more diverse datasets. This will ensure that the work accuracy and generalizability hold across broader populations, making it a more robust tool for real-world application.

REFERENCES

[1] P. Paudel, S. K. Karna, R. Saud, L. Regmi, T. B. Thapa, and M. Bhandari, "Unveiling key predictors for early heart attack detection using machine learning and explainable ai technique with lime," in *Proceedings of the 10th International Conference on Networking, Systems and Security*, 2023, pp. 69–78.

[2] A. Jafar and M. Lee, "Hypgb: High accuracy gb classifier for predicting heart disease with hyperopt hpo framework and lasso fs method," *IEEE Access*, 2023.

[3] K. Tn, S. Meghana, A. Kodipalli, T. Rao, S. Kamal *et al.*, "Prediction of early heart attack possibility using machine learning," in *2023 2nd International Conference for Innovation in Technology (INOCON)*. IEEE, 2023, pp. 1–5.

[4] Y. Muhammad, M. Tahir, M. Hayat, and K. T. Chong, "Early and accurate detection and diagnosis of heart disease using intelligent computational model," *Scientific reports*, vol. 10, no. 1, p. 19747, 2020.

[5] M. Javaid, A. Haleem, R. P. Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," *International Journal of Intelligent Networks*, vol. 3, pp. 58–73, 2022.

[6] L. B. Elvas, M. Nunes, J. C. Ferreira, M. S. Dias, and L. B. Rosário, "Ai-driven decision support for early detection of cardiac events: Unveiling patterns and predicting myocardial ischemia," *Journal of Personalized Medicine*, vol. 13, no. 9, p. 1421, 2023.

[7] D. Tenepalli and N. TM, "A systematic review on iot and machine learning algorithms in e-healthcare," *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 1–14, 2024.

[8] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, A. Hampl, and J. Havel, "Artificial neural networks in medical diagnosis," pp. 47–58, 2013.

[9] S. Safdar, S. Zafar, N. Zafar, and N. F. Khan, "Machine learning based decision support systems (dss) for heart disease diagnosis: a review," *Artificial Intelligence Review*, vol. 50, no. 4, pp. 597–623, 2018.

[10] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, "A predictive analytics approach for stroke prediction using machine learning and neural networks," *Healthcare Analytics*, vol. 2, p. 100032, 2022.

[11] S. Mondal, R. Maity, Y. Omo, S. Ghosh, and A. Nag, "An efficient computational risk prediction model of heart diseases based on dual-stage stacked machine learning approaches," *IEEE Access*, 2024.

[12] R. Subathra and V. Sumathy, "An offbeat bolstered swarm integrated ensemble learning (bsel) model for heart disease diagnosis and classification," *Applied Soft Computing*, vol. 154, p. 111273, 2024.

[13] P. Rani, R. Kumar, A. Jain, R. Lamba, R. K. Sachdeva, K. Kumar, and M. Kumar, "An extensive review of machine learning and deep learning techniques on heart disease classification and prediction," *Archives of Computational Methods in Engineering*, pp. 1–19, 2024.

[14] G. Manikandan, B. Pragadeesh, V. Manojkumar, A. Karthikeyan, R. Manikandan, and A. H. Gandomi, "Classification models combined with boruta feature selection for heart disease prediction," *Informatics in Medicine Unlocked*, vol. 44, p. 101442, 2024.

[15] I. Mishra and S. Mohapatra, "An enhanced approach for analyzing the performance of heart stroke prediction with machine learning techniques," *International Journal of Information Technology*, vol. 15, no. 6, pp. 3257–3270, 2023.

[16] Y. Rimal, S. Paudel, N. Sharma, and A. Alsadoon, "Machine learning model matters its accuracy: a comparative study of ensemble learning and automl using heart disease prediction," *Multimedia Tools and Applications*, pp. 1–18, 2023.

[17] C. Sharma, S. Sharma, M. Sharma, and A. Sodhi, "Early stroke prediction using machine learning," 03 2022.

[18] N. Venkata MahaLakshmi and R. K. Rout, "An intelligence method for heart disease prediction using integrated filter-evolutionary search based feature selection and optimized ensemble classifier," *Multimedia Tools and Applications*, pp. 1–25, 2023.

[19] S. Y. Hera, M. Amjad, and M. K. Saba, "Improving heart disease prediction using multi-tier ensemble model," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 11, no. 1, p. 41, 2022.

[20] S. Asif, Y. Wenhui, Q. ul Ain, Y. Yueyang, and S. Jinhai, "Improving the accuracy of diagnosing and predicting coronary heart disease using ensemble method and feature selection techniques," *Cluster Computing*, pp. 1–20, 2023.

[21] K. Uma Maheswari and A. Valarmathi, "A novel mechanism to recognize heart disease by optimised deep belief network with svm classification," *Journal of Intelligent & Fuzzy Systems*, vol. 44, no. 1, pp. 167–184, 2023.

[22] I. Isik, "Heart disease prediction with feature selection based on meta-heuristic optimization algorithms and electronic filter model," *Arabian Journal for Science and Engineering*, pp. 1–14, 2023.

[23] Bharath011, "Heart disease classification dataset," aug 2023. [Online]. Available: https://www.kaggle.com/datasets/bharath011/heart-disease-classification-dataset

[24] T. Maguire, L. Manuel, R. Smedinga, and M. Biehl, "A review of feature selection and ranking methods," *19th SC@ RUG 2021-2022*, p. 15, 2022.

[25] X. Shi, Y. D. Wong, M. Z.-F. Li, C. Palanisamy, and C. Chai, "A feature learning approach based on xgboost for driving assessment and risk prediction," *Accident Analysis & Prevention*, vol. 129, pp. 170–179, 2019.

[26] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "Smote for handling imbalanced data problem: A review," in *2021 sixth international conference on informatics and computing (ICIC)*. IEEE, 2021, pp. 1–8.

[27] R. Hariprasad, T. Navamani, T. R. Rote, and I. Chauhan, "Design and development of an efficient risk prediction model for cervical cancer," *IEEE Access*, 2023.

[28] P. Guleria, P. Naga Srinivasu, S. Ahmed, N. Almusallam, and F. Alarfaj, "Xai framework for cardiovascular disease prediction using classification techniques. electronics 2022, 11, 4086," 2022.

[29] F. Ali, S. El-Sappagh, S. R. Islam, D. Kwak, A. Ali, M. Imran, and K.-S. Kwak, "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," *Information Fusion*, vol. 63, pp. 208–222, 2020.

[30] M. M. Rahman, "A web-based heart disease prediction system using machine learning algorithms," *Network Biology*, vol. 12, no. 2, p. 64, 2022.