A FIELD PROJECT REPORT

on

“Leveraging XGBoost and Clinical Attributes for Heart Disease Prediction”

Submitted

by

221FA04032

Paladugu Siva Satyanarayana

221FA04054

Achyuta Mohitha Sai Sri

221FA04083

Bogala Devi Prasaad Reddy

221FA04392

Kota Susmitha

Under the guidance of

Maridu Bhargavi
Assistant Professor Department of CSE,VFSTR

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed to be UNIVERSITY

Vadlamudi, Guntur.

ANDHRA PRADESH, INDIA, PIN-522213.

VIGNAN'S
FOUNDATION FOR SCIENCE, TECHNOLOGY & RESEARCH
(Deemed to be University) - Estd. u/s 3 of UGC Act 1956

CERTIFICATE

This is to certify that the Field Project entitled **"Leveraging XGBoost and Clinical Attributes for Heart Disease Prediction"** that is being submitted by 221FA04032 (P Siva Satyanarayana), 221FA04054(A Mohitha Sai Sri), 221FA04083(B Devi Prasaad Reddy), 221FA04392 (K Susmitha) for partial fulfilment of Field Project is a bonafide work carried out under the supervision of M Bhargavi , M.Tech., Associate Professor, Department of CSE.

M. Bhargavi	Dr. S. V. Phani Kumar	Dr.K.V. Krishna Kishore
Assistant Professor, CSE	HOD,CSE	Dean, SoCI

VIGNAN'S
FOUNDATION FOR SCIENCE, TECHNOLOGY & RESEARCH
(Deemed to be University) - Estd. u/s 3 of UGC Act 1956

DECLARATION

We hereby declare that the Field Project entitled **"Leveraging XGBoost and Clinical Attributes for Heart Disease Prediction"** that is being submitted by 221FA04032 (P Siva Satyanarayana), 221FA04054(A Mohitha Sai Sri), 221FA04083(B Devi Prasaad Reddy), 221FA04392 (K Susmitha) in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision M Bhargavi , Associate Professor, Department of CSE

By

221FA04032(P Siva Satyanarayana)

221FA04054(A Mohitha Sai Sri)

221FA04143(B Devi PrasaadReddy)

221FA04392 (K Susmitha)

Date:

ABSTRACT

Despite the various advances in medical technology, heart disease continues to rank among the leading causes of mortality in the world, killing millions each year. There is hope that the risks involved with heart disease can be reversed in time and the outcomes of patients improved with earlier prediction and diagnosis. However, the use of conventional methods of diagnostic techniques tends to be very resource-intensive and inaccessible to other more disadvantaged areas. This paper defines a model for machine learning in predicting heart disease and illustrates a more efficient and accessible solution. There is a risk for heart disease that is determined using a set of various clinical features. The algorithm used in the prediction of heart disease is XG Boost that are applied to the dataset. The method provides a reliable tool for the early detection of heart disease, thus helping healthcare providers to improve preventive care and treatment strategies. The results of this paper will be helpful in advancing research to identify other avenues for prediction of heart disease.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER-1

INTRODUCTION

1. INTRODUCTION

1.1 What is Heart Failure and What Causes It?

Heart failure, a situation where the heart loses ability to pump blood throughout the body efficiently, leading to the situation that the body does not receive enough oxygen or nutrients the situation means the heart is not able to do the desired function. This is the explanation of heart failure that must be orally stated by nurse to a patient; if it is impossible, that will lead to diagnostic problems. The things that seem to interfere with the blood flow are coronary artery disease, hypertension, diabetes, and lifestyle factors such as heavy smoking, overeating, and a sedentary way of life. It is necessary to mention that genetic predisposition and heart attacks also influence incompetence.

1.2 The Consequences of Heart Failure

Heart failure is a condition that can lead to diminished quality of life, long hospital stays, and a high death rate as a result of these consequences. It is evident that the usual no age limit, all people may experience the sickness of fatigue, or perhaps changing lifestyle; health care systems that focus principally on the sick people also become prevalent. Some people experience fatigue, Some people may feel well, while some may gain weight, and some may show swollen signs over some days. Heart failure is a disease that results in bodily function loss that cannot be recovered thus the body falling nonfunctional and untreatable thus leading to finally death of the patient. The only thing that may happen to do with life is that in the early stages, when the patient is still alive, then heart failure, unlike other critical illnesses, can sometimes be turned back when detected at the beginning stages of it.

1.3 The Economic and Healthcare Burden of Heart Failure

A lot of costs are incurred along the way when heart failure comes up. Because of frequent hospital admissions, long treatment with medications and the necessity for constant monitoring of heart failure, including the expenses for the hospital stays and the medications, the patients who have heart failure have to pay a big part of these expenses. Additionally, hospitals are under added pressure due to increasing patient counts and chronic care needs. The expenditure connected with heart failure care involves both direct medical costs and indirect costs, particularly with the loss of productivity and long-term care.

1.4 Current Methodologies for Heart Failure Prediction

Presently, there are various strategies used to predict heart failure and their application depends on how technology has evolved through statistical and machine learning tools in data analysis. Methods such as Logistic Regression, Decision Trees, and Neural Networks are applied for analyzing patient data separately, for instance, patient demographics, medical history, and test results to assess the risk for heart failure.

1.5 Applications of Machine Learning to Combat Heart Failure

The application of machine learning (ML) to diagnose heart failure has acquired unprecedented momentum thus increasing the accuracy of the diagnostic process and speeding up the identification of disease. ML is implemented by models that are based on a big set of the patient's clinical and history data and such models detect patterns and risks that often are not identified by traditional methods. Approaches like National Forest, Support Vector Machines, and ensemble learning algorithms have demonstrated potential in forecasting heart failure outcomes, and personalized treatment plans that reduce the chances of severe complications ensue.

CHAPTER-2

LITERATURE SURVEY

2. LITERATURE SURVEY

We have carried out a literature survey to include all related works with our study on Heart Failure Prediction . The crux of ideas from these papers has been summed up below:

Jamal. S et al. highlight predictive models like SVM, Decision Trees, Linear Discriminant Analysis, and Logistic Regression for heart disease prediction. XGBoost notably improves accuracy by up to 3%. Effective pre-processing and high-quality data are essential for maximizing model performance and achieving reliable diagnostics in heart disease.

No.	Title/Year	Author	Algorithm/Technique	Limitations (or) Future Work	Metric Accuracy
1	Heart Failure Prediction using Machine Learning Algorithms with Cross Validations	Sachdeva, R. K., Singh, K. D., Sharma, S., Bathla, P., Solanki, V	Cross Validations	Deep learning can also be applied to predict HF in the future	92.4%
2	Heart Failure Survival Prediction Using Machine Learning Algorithm: Am I Safe from Heart Failure? 2022	Jamal, S., Elenin, W. A., & Chen, L	Logistic Regression, Random Forest, KNN, SVM	Uses 6 ML algorithms and 10-fold cross-validation techniques. Use one ML model	85%, AUC: 93%
3	An Organized Method for Heart Failure Classification Year: 2023	Lutfi, D. K., & Shidik, G. F	XGBoost, Decision Tree, SVM, Random Forest	Uses 4 ML models; could use ensemble methods	96.0%
4	Predicting Heart Failure Using Deep Neural Networks Year: 2020	Mamun, M., Farjana, A., Mamun, M. A., Ahammed, M. S., & Rahman, M	Usage of MLP - applicable for non-linear functions	Usage of feature selection techniques	98.3%
5	Use of Machine Learning Techniques in the Prediction of Heart Disease: 2021	Jain. V	Random Forest, Naive Bayes, MLP	Integration of deep learning techniques in future studies	91.5%
6	Hybrid Model for Heart Disease Prediction: 2022	Shetty. M., D. K. Shetty, S. B., & Shetty. A	Hybrid of SVM and KNN	Model complexity can be high	95.7%
7	Explainable AI for Heart Failure Prediction: 2023	Saravanan, S., &Swaminathan, K	LIME with Random Forest	May require fine-tuning for specific datasets	94.2%
8	Ensemble Learning for Heart Failure Prediction: 2023	Elghalid, R. A. M., Aldeeb, F. H. A., Alwirshiffani, A., Andiasha, A., & Mohamed, A	Stacking ensemble with Random Forest, XGBoost, and SVM	High computational cost	97.5%
9	Improvement Heart	Mehta, D., Mehta,	Decision Tree	Data Imbalance,	92.61%

	Failure Prediction Using Binary Preprocessing :2023	P., Naik, A., Kaul, R., & Bide, P. J	Classifier,Random Forest,K-Nearest Neighbor (KNN),Support,Vector Machine (SVM),Logistic Regression,Naive Bayes,Gradient Boosted Trees (GBT),Extreme Gradient Boosting (XGBoost)	Generalizability, Complexity	
10	Prediction Model of Heart Failure Disease Based on GA-ELM:2021	Sang, X., Zhu, Y. Q., Ma, L., Wen, C. H., &Luo, P	Extreme Learning Machine (ELM), Genetic Algorithm (GA), Support Vector Machine (SVM), Decision Tree, Random Forest, K-Nearest Neighbors (KNN)	More computationally intensive due to the genetic optimization process, longer training times compared to traditional ELM.	83.5%
11	Monitoring and Predicting of Heart Diseases Using Machine Learning Techniques:2023	McClellan, M. L., et al	Extreme Learning Machine (ELM), Genetic Algorithm (GA), Support Vector Machine (SVM), Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Convolutional Neural Networks (CNN), Modified Deep Convolutional Neural Network (MDCNN)	Dataset diversity, algorithm selection inconsistency, and poor feature selection hinder performance.	98.2%
12	Data Mining Classification Algorithms Applied to the Prediction of Heart Disease:2022	Sacco, R. L., et al	Adaboost, J48 Decision Tree, Bagging	Results can vary by dataset diversity, incomplete features impact performance, and not all algorithms excel across different datasets.	Adaboost: 92.10%, J48 Decision Tree: 91.62%, Bagging: 90.74%, AROC: 0.89%
13	A Case Study on Risk Prediction in	Piepoli, M. F., et al	Random Survival Forest (RSF), Cox	Include model complexity,	81%

	Heart Failure Patients using Random Survival Forest:2021		Proportional Hazard Model, Kaplan-Meier Estimator	interpretability challenges, and potential overfitting with an excessive number of trees.	
14	Heart Failure Prediction using Comparative Machine Learning Models/2021	Jin Wang ., et al	LR, KNN, Naive Bayes, DT, SVM, MLP, Ridge Classifier, QDA, XGBoost, LightGBM, Extra Trees, CatBoost, Deep Forest	Highest accuracy was 86.67%, achieved using QDA and Extra Trees, showing room for improvement with advanced techniques	86.67%
15	Survival Prediction in Heart Failure using Machine Learning Algorithms/2021	Tak et al.	SVM, DT, KNN, Ensemble learning	KNN provided highest accuracy, but sensitivity, specificity, and AU-ROC leave room for future performance enhancements	89.5%
16	Heart Failure Classification using SMOTE-ENN and Hyperparameter Optimization/2022	Nishat et al.	DT, LR, Gaussian Naive Bayes, RF, KNN, SVM, SMOTE-ENN	Best model was RF with 90% accuracy using SMOTE-ENN and scaling methods, but other classifiers could be explored further	90%
17	Heart Failure Survival Prediction using Machine Learning Algorithms/2021	Ozbay et al.	Naive Bayes, AdaBoost, Attribute Selected Classifier (ASC), CVR, JRip, OneR, PART, J48, RF, Random Tree	RF provided highest accuracy (87.088%), but further feature selection and data balancing may enhance performance	87.088%
18	Heart Failure Prediction using Machine Learning/2022	Newaz et al	Recursive Feature Elimination, Chi-square Test, RF, LR, KNN, SVM, AdaBoost	Achieved moderate accuracy, sensitivity, and G-mean scores; further exploration of feature engineering could improve performance	80.21%

19	Heart Failure Survival Prediction using Machine Learning Algorithms/2022	M. M. Nishat, F. Faisal, I. Ratul, A. Al-Monsour, et al.	DT, LR, Gaussian Naive Bayes, RF, KNN, SVM with SMOTE-ENN Oversampling Technique and Hyperparamet er Optimization	RF with SMOTE-ENN and scaling performed best; future work may focus on exploring additional oversampling techniques	90%
20	A Comprehensive Investigation of Classifiers with SMOTE-ENN for Heart Failure/2022	F. A. Ozbay and E. Ozbay	Naive Bayes, AdaBoost, ASC, CVR, JRip, OneR, PART, J48, RF, RT	Highest accuracy from RF (87.088%); exploring different combinations of oversampling methods for further improvement	87.088%
21	A Systematic Method for Heart Failure Prediction Using Machine Learning Models/2022	N. S. M. Huang, Z. Ibrahim, and N. M. Diah	Naive Bayes, SVM, RF, Logistic Regression	RF provided the highest accuracy but further optimization techniques are needed for model improvement	88%
22	Survival Prediction of Heart Failure Patients Using Machine Learning Techniques/2021	A. Newaz, N. Ahmed, and F. S. Haq	Recursive Feature Elimination, Chi-Square Test, RF, LR, KNN, SVM, AdaBoos	RF produced moderate results; additional data preprocessing techniques and model fine-tuning may lead to better outcomes	80.21%
23	Using Machine Learning for Heart Failure Survival Prediction/2021	A. Newaz, N. Ahmed, F. S. Haq	Recursive Feature Elimination, Chi-Square, Random Forest, SVM, AdaBoost	Identified feature selection methods could be optimized further to improve prediction performance	80.21%
24	Heart Failure Classification Using Hybrid Ensemble Models/2022	N. Tak, P. Parihar, S. Mathur, B. Das, D. Sawal	SVM, Decision Trees, K-Nearest Neighbors, Hybrid Ensemble Learning (voting-based classifier)	KNN performed best but identified future work to focus on other ensemble techniques and additional clinical datasets	89.5%

25	Performance Comparison of Machine Learning Algorithms on Heart Failure Prediction/2022	F. A. Ozbay, E. Ozbay	Naive Bayes, AdaBoost, Classification via Regression (CVR), JRip, PART, J48, RF, Random Tree	Recommended further analysis of feature importance and refining hyperparameters for more accurate predictions	87.09%

2.2 Motivation

Heart disease remains one of the leading causes of death in the whole world, along with that it takes away the lives of millions of people every year. Early detection and intervention are the prerequisite to developing a better probability of surviving, however the standard tools to diagnose the disease are often expensive, time-consuming, and do not cover all the areas needed so they are often inaccessible in the less developed parts of the world. This is the reason why more efficient and scalable technologies are needed so that earlier and more precise diagnosis can be made.

What causes this investigation is the chance that the machine learning (ML) would be able to solve these problems. ML programs are capable of processing large amounts of data to distinguish general patterns that the traditional methods could have overlooked; hence, the system can be one of the helpful tools in forecasting heart disease. The researchers will be able to help health care providers spot in-need patients, for instance, the ones with high cardiovascular risk disease, to undergo preventive care and treatment strategies which in the end will be better. With the rise in availability of healthcare data, ML is the new avenue that brings hope of better clinical taking into account patients as individuals. Still, the models in place are not without problems like overfitting and class imbalance which have resulted in their effectiveness being tampered with more often than not.

This investigation will focus on improving the accuracy and the efficiency of the most used models that are used to predict heart disease through the application of techniques like ensemble methods and gradient boosting. Our target is to develop a tool that is flexible and can be used in different kinds of healthcare places, while supporting under-resourced resources, and lead to a positive change in the patient's status on the global level.

CHAPTER-3

PROPOSED SYSTEM

3. PROPOSED SYSTEM

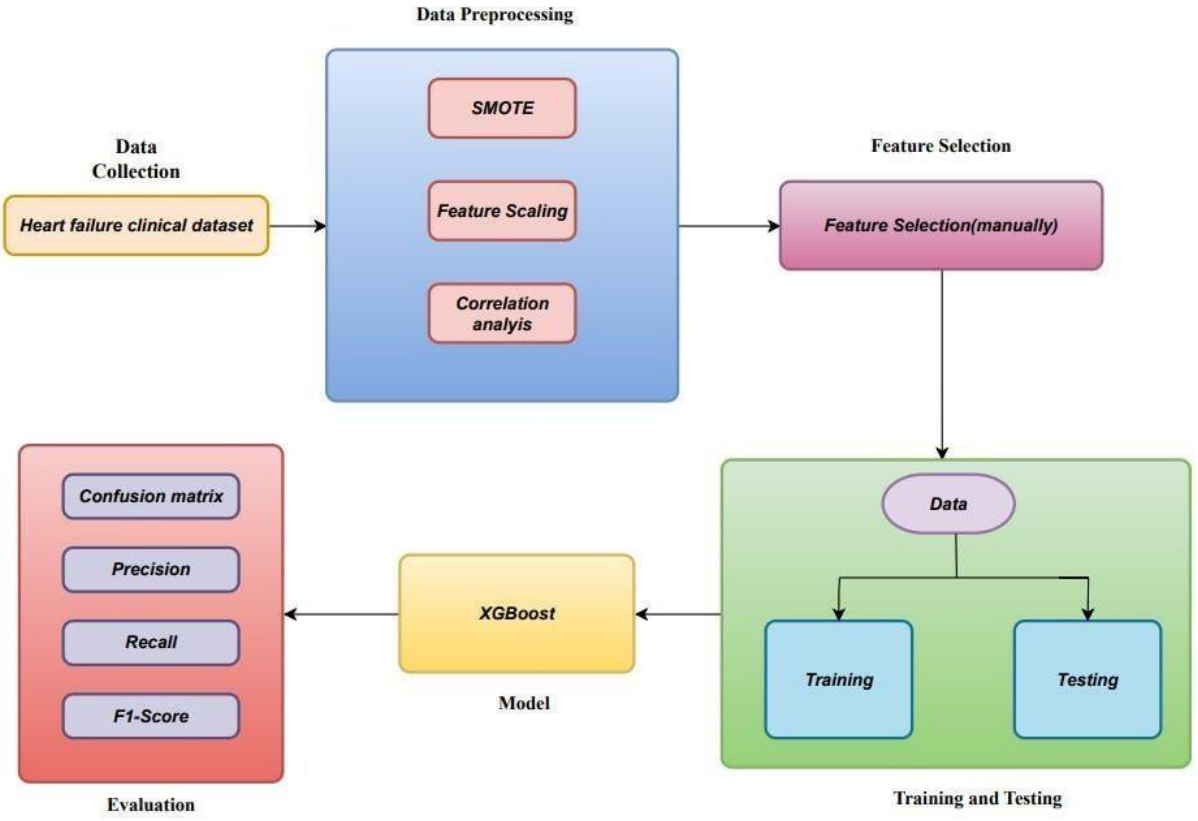


Figure 1: Proposed model

3.1 Input dataset

The dataset contains 299 patients data and there are 13 clinical attributes (Table 1) that can predict death from heart failure. All the attributes are continuous and binary, describing different aspects of a patient's health. The mechanism of the heart disease prediction of various steps(Figure 1).For instance, in the database, age is noted; whether a patient has anaemia, a condition characterized by low red blood cells or hemoglobin levels; and the CPK level, which is a blood enzyme. All these other relevant factors include diabetes, the ejection fraction-a measure of how well the heart is pumping blood-and high blood pressure. Platelet count, serum creatinine and serum sodium, along with information on whether the patient is male or female and whether he smokes, comprise the remaining dataset. 3 Finally, the dataset captures a time variable that refers to the number of days the patient was observed in the course of the study. The target variable of interest is DEATH EVENT. It is a binary label that indicates whether the patient died (Figure 2) within the period that they were followed up. The analysis on the diabetes of the patients is mentioned (Figure 3). This dataset is very common in developing risk predictive models as it grades the likelihood of a patient's risk of heart failure de pending on their clinical profile.

3.1.1 Detailed Features of the Dataset

Table 1: Dataset Features and Their Descriptions

Feature	Description
Age	The patient's age in years.
Anaemia	Indicates if the patient has anemia (1: Yes, 0: No).
Creatinine Phosphokinase	CPK enzyme levels measured in the blood.
Diabetes	Specifies if the patient has diabetes (1: Yes, 0: No).
Ejection Fraction	Percentage of blood pumped out of the heart during contraction.
High Blood Pressure	Whether the patient suffers from hypertension (1: Yes, 0: No).
Platelets	The count of platelets present in the blood.
Serum Creatinine	Creatinine concentration measured in the blood.
Serum Sodium	Sodium levels in the patient's bloodstream.
Sex	The gender of the patient (1: Male, 0: Female).
Smoking	Indicates whether the patient smokes (1: Yes, 0: No).
Time	Duration of follow-up in days.
DEATH_EVENT	Outcome of death during follow-up (1: Yes, 0: No).

3.2 Exploratory Data Analysis(EDA)

The prediction of heart failure spans areas of investigation such as diabetes and death events analysis that are essential to spot the highly important risk factors for a particular patient as well as for the researching of the patient's disease. Diabetes analysis is an examination comprising the development of diabetes if actually present with heart illness and the factors that result, in particular, the levels of glucose and insulin resistance to evaluate the process of the disease. Death event analysis contains algorithms to identify age, ejection fraction, and blood pressure as the top three predictors of mortality by means of survival analysis for the estimation of the probability of death over a period. The analysis of the predictions is predisposed to make out the patterns and signify the relevant variables, allowing the generation of more reliable models and personalized treatments.

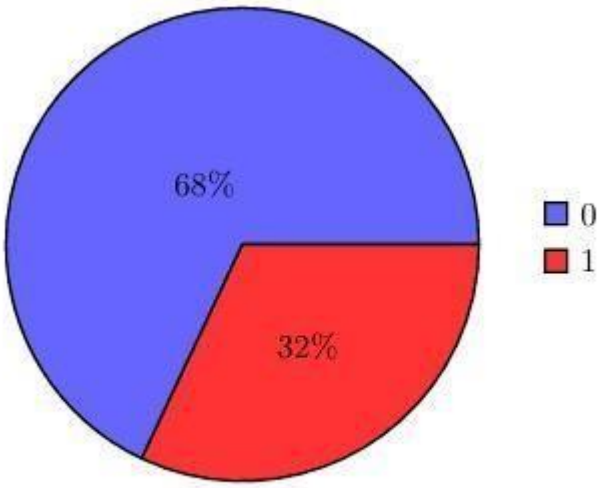


Figure 2: Death analysis of the patients 1:yes and 0:no

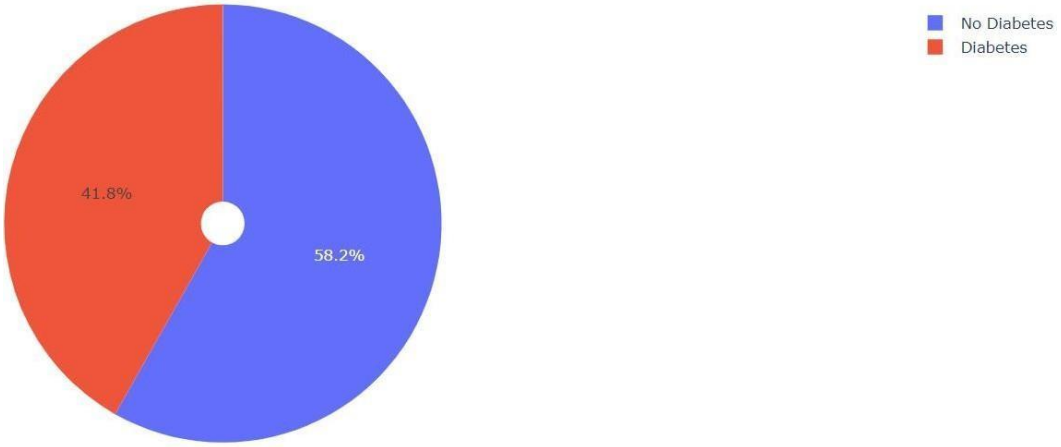


Figure 3: Analysis on Diabetes

3.3 Data Preprocessing

Data pre-processing is an important part of preparing a machine learning-ready dataset. It allows clean, structured data appropriate for modeling. This step is particularly rel evant for scaling numerical features, managing imbalanced datasets, and selecting a good set of features. Below are some of the preprocessing techniques we applied (see Figure 3(b)).

3.3.1 Scaling

To handle the varying measurement units of attributes, we utilized the Standard Scaler to ensure that all numeric features fall on the same scale.Scaling was applied to the relevant variables where we got the highest accuracy. This step prevents scenarios where features with large ranges dominate distance-based models, such as support vector machines or neural networks. Additionally, scaling accelerates the convergence of gradient-based optimization algorithms.

3.3.2 Dealing with Class Imbalance

The target variable, DEATH EVENT, has an imbalance of classes, which can induce biased models that favor the majority class. To address this issue, we applied SMOTE (Synthetic Minority Oversampling Technique), which creates synthetic data samples for the minority class, thereby balancing the classes distribution in the dataset and shifting the model to fairly generalize both classes, thus improving the model perfor mance and its accuracy for the minority class predictions.

3.3.3 Correlation Analysis and Feature Engineering

A correlation matrix was computed to analyze feature interrelations, which was visu alized using a heatmap (Figure 5). Highly correlated features were identified and removed from the feature set to avoid multicollinearity, which can degrade model performance. By removing redundant features, we improved both the readability and accuracy of the model. These preprocessing steps(Figure 4) were essential in ensuring the optimal performance of our predictive heart disease risk model, by balancing and scaling the data as well as removing redundant information.

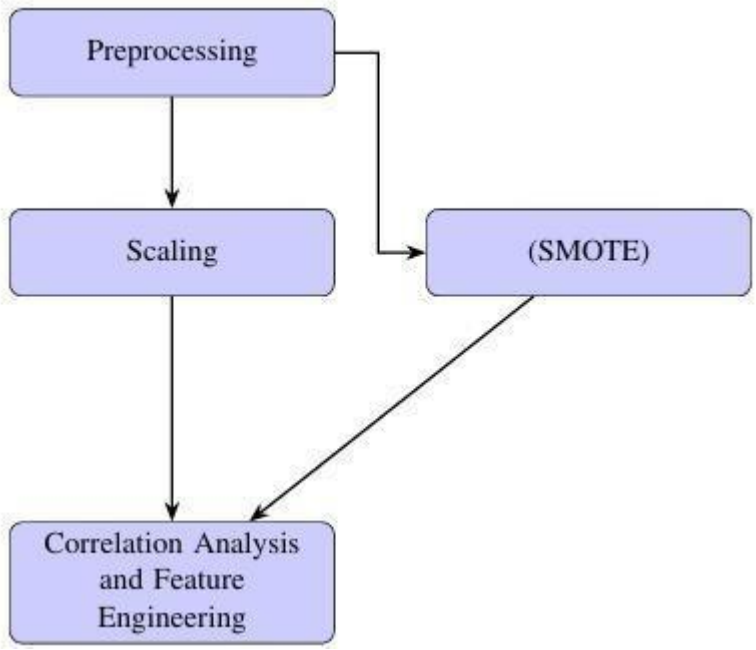


Figure 4: Preprocessing steps

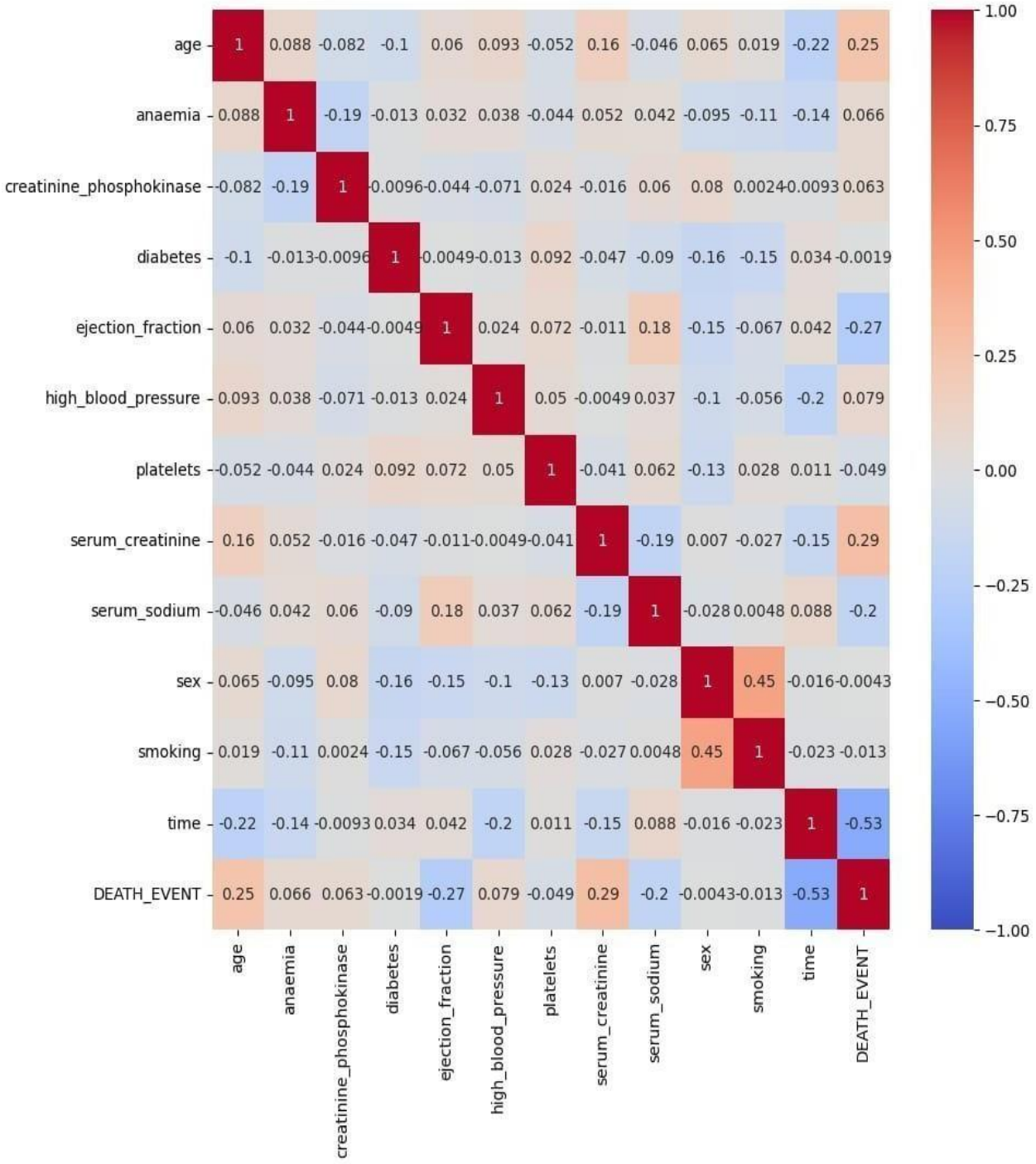


Figure 5: Heatmap correlation contains the attributes of the dataset.

3.4 Feature Selection

To improve model performance and interpretability, feature selection has been carried out by selecting appropriate clinical features from the dataset with a critical analysis. The chosen attributes were selected based on their clinical significance in relation to heart failure. These features were chosen based on the clinical relevance that they present with heart failure. Therefore, to ensure that the model em-phasizes the most impactful predictors, this was achieved manually by picking those specific attributes. This way, we attempted to reduce the dimensions in the dataset while keeping crucial information to enhance the efficiency of the model in predicting and accuracy.

3.5 Training and Testing

Training and testing of the same for heart failure prediction model was carried out by taking an 80-20 split of the dataset that was kept aside for training and testing purposes. The models were fitted using this training set, which constituted 80%, and the testing set, which was kept aside, consisted of 20%. Before training, all numeric features were standardized using a StandardScaler to ensure that every feature could have a mean of zero and a unit variance, thus preventing models from becoming biased towards features on larger scales.

3.6 Usage of ML

It is actually a gradient boosting framework built atop an ensemble of decision trees, whose performance is optimized through iterative learning. Various parameters were used within the model, such as use label encoder=False and eval metric=logloss, to better suit the model for the classification problem under evaluation. Then the dataset was resampled and standardized before being put to train.

3.6.1 XGBoost

XGBoost is indeed a powerful implementation of gradient boosted decision trees . It builds decision trees sequentially, assigning weights to the features and updating those based on the errors made by previous trees. The basic idea is that each new tree corrects the errors of the previous one. Therefore, it results in a powerful ensemble of models that handle regression, classification, ranking, and user-defined prediction problems. Mathematically, XGBoost models can be written as:

$$\hat{z}j = \sum p = 1^{K} f_p(x_j),\ f_p \in \mathcal{F}$$

Here, K indicates number of trees, and fp indicates individual trees. The objective function is given as:

$$obj(\theta) = \sum_{j=1}^{n} l(z_j, \hat{z}j) + \sum p = 1^{K} \Omega(f_p)$$

The first component represents loss function, while the second corresponds to the regularization term. In the additive strategy for simplifying optimization, we expand with a Taylor series up to the second order.

The gain from splitting a leaf into two is:

$$Gain = \frac{1}{2}\left[\frac{S_M^2}{Q_M + \lambda} + \frac{S_W^2}{Q_W + \lambda} - \frac{(S_M + S_W)^2}{Q_M + Q_W + \lambda}\right] - \gamma$$

Let S and Q indicates the gradient and Hessian, respectively, and the parameter γ controls the pruning.

3.7 Evaluation

The predictive ability of the models on heart disease outcomes was assessed using several performance metrics. After training, predictions were generated for the test set, and each model's performance was assessed using accuracy, confusion matrices, and classification reports. The primary metric, accuracy, is defined as the ratio of correctly classified instances to the total instances and is calculated as:

$$Accuracy = \frac{P_{correct} + N_{correct}}{P_{correct} + N_{correct} + P_{incorrect} + N_{incorrect}}$$

Where,

$P_{correct}$ = Positive cases correctly predicted (True Positives)

$N_{correct}$ = Negative cases correctly predicted (True Negatives)

$P_{incorrect}$ = Positive cases incorrectly predicted (False Positives)

$N_{incorrect}$ = Negative cases incorrectly predicted (False Negatives) In addition toaccuracy, the confusion matrix provides insight into the classification performance. It

also determine how well a model distinguishes between occurrences and non occurrences of heart disease. The classification report also provided precision and recall measures that can give an estimation of the efficacy of the proposed models in the heart disease case predictions. Precision refers to the ratio of the number of actual positives over the total number of positive predictions and is computed as follows:

$$Precision = \frac{P_{correct}}{P_{correct} + P_{incorrect}}$$

Recall (or sensitivity) measures how well the model identi f ies true positives and is expressed as follows:

$$Recall = \frac{P_{correct}}{P_{correct} + N_{incorrect}}$$

3.7.1 XGBoost Model Performance:

With the final model, the accuracy of XGBoost on test dataset was 0.97. XGBoost is a gradient boosting algorithm designed to be both efficient and scalable. It is an ensemble method that combines weak learners, such as decision trees, to optimize predictive performance. The objective function minimizes the following log loss (binary cross-entropy):

$$LogLoss = -\frac{1}{M}\sum_{j=1}^{M}[a_j \log(b_j) + (1 - a_j)\log(1 - b_j)]$$

With

• M representing the total number of samples,

• a_j indicating the actual class label (0 or 1),

• b_j denoting the predicted probability for class 1.

The model optimized the log-loss function during training, using the evaluation metric eval metric=logloss. The superior performance of the XGBoost model was further validated through the confusion matrix and classification regression, demonstrating its high accuracy in predicting heart disease cases while minimizing false predictions.

This highlights the effectiveness of more complex ensemble techniques like XGBoost in healthcare predictions, particularly for high-risk conditions such as heart disease.

CHAPTER- 4

IMPLEMENTATION

4.1 Modules

1. Importing Libraries
2. Loading the Data
3. Checking of null values
4. Specifying features
5. Splitting of data into training and testing
6. Building the model
7. Confusion matrix

4.2 Description and sample code

1.Importing libraries

```
# Import necessary libraries
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import classification_report, accuracy_score, confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns
from imblearn.over_sampling import SMOTE
import xgboost as xgb
```

- **pandas**: A data manipulation library for Python that provides powerful data structures like DataFrames for handling structured data.
- **sklearn.model_selection**: A module in Scikit-learn for splitting datasets into training and testing sets to facilitate model evaluation.
- **sklearn.preprocessing**: Provides tools for feature scaling and transformation, including StandardScaler for standardizing data.
- **sklearn.metrics**: Contains functions to evaluate model performance, such as accuracy score, classification report, and confusion matrix.
- **matplotlib.pyplot**: A plotting library for creating static and interactive visualizations in Python.
- **seaborn**: A statistical data visualization library based on Matplotlib, designed for creating attractive and informative charts.

- **imblearn.over_sampling.SMOTE**: A technique from the imbalanced-learn library that generates synthetic samples to address class imbalance in datasets.
- **xgboost**: An optimized gradient boosting library known for its speed and performance on structured data.

2. Loading the data

The line reads a CSV file and loads its contents into a pandas DataFrame for data analysis and manipulation.

```
# Load dataset
heart_data = pd.read_csv('data.csv')
```

3. Checking of null values

This line counts the total number of missing values in each column of the DataFrame, offering a summary of data quality.

```
heart_data.isnull().sum()
```

4. Specifying features

It defines a set of input characteristics that will be used for prediction and identifies the outcome to be predicted. A list includes relevant attributes, such as age and health indicators, while another identifies the specific event of interest, in this case, death. This distinction is important for training models, enabling them to learn the relationship between the input attributes and the desired outcome.

```
# Specify features and target
features = ['age', 'anaemia', 'creatinine_phosphokinase', 'diabetes', 'ejection_fraction',
            'high_blood_pressure', 'platelets', 'serum_creatinine', 'serum_sodium', 'sex',
            'smoking', 'time']
X = heart_data[features]
y = heart_data['DEATH_EVENT']
```

5. Splitting of data into training and testing

This line splits the input features and target variable into training and testing sets, with 20% of the data reserved for testing, ensuring reproducibility through a specified random state.

```
# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=2)
```

6. Building of model

This line initializes an XGBoost classifier with specific settings, including disabling label encoding, setting the evaluation metric to log loss, and ensuring reproducibility with a defined random state.

```
 'XGBoost': xgb.XGBClassifier(use_label_encoder=False, eval_metric='logloss',
random_state=42),
```

7. Evaluation

This line computes the confusion matrix by comparing the actual target values from the test set with the predicted values generated by the model, providing insights into the classification performance.

```
cm = confusion_matrix(y_test, y_pred)
```

CHAPTER-5

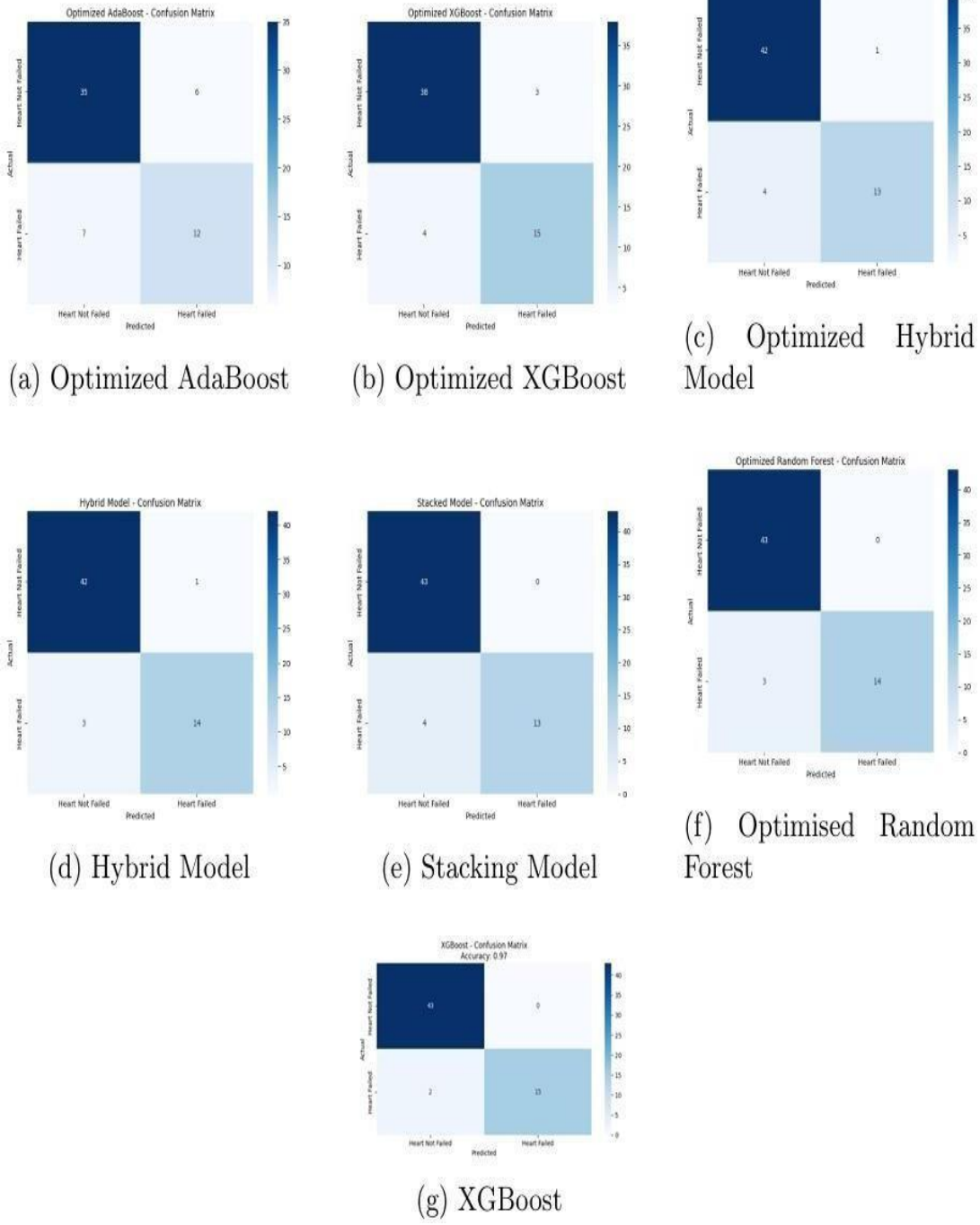
RESULTS AND ANALYSIS

5. Results and Analysis

Accuracy is an important metric in predictive modeling, especially in healthcare based modeling, as it directly affects the reliability of the predictions and, hence, patient outcomes. In this paper, we compared several models, and the XGBoost classifier emerged as the best, achieving an impressive accuracy of 0.97 and demonstrating its ability to capture complex patterns within the dataset (Figure 6 and Figure 7).The accuracies of differnt mod els are also mentioned (see Table 2). Among those, the accuracy for Random Forest and Hybrid Models was at 0.95 and 0.93, respectively. While both models—optimized XGBoost and logistic regression—performed well at being accurate to 0.88 and 0.90, respectively, no model achieved the experimental results XGBoost recorded. Table 2: Performance comparison of different machine learning models.

Table II. Performance comparison of different machine learning models

Model	Accuracy
Optimized AdaBoost	0.78
Optimized XGBoost	0.88
Logistic Regression	0.90
Optimized Hybrid Model	0.91
Hybrid Model	0.93
Stacking Model	0.93
Optimized Random Forest	0.95
XGBoost	0.97



(a) Optimized AdaBoost (b) Optimized XGBoost (c) Optimized Hybrid Model

(d) Hybrid Model (e) Stacking Model (f) Optimised Random Forest

(g) XGBoost

Figure 6: Comparison of accuracies across 7 models

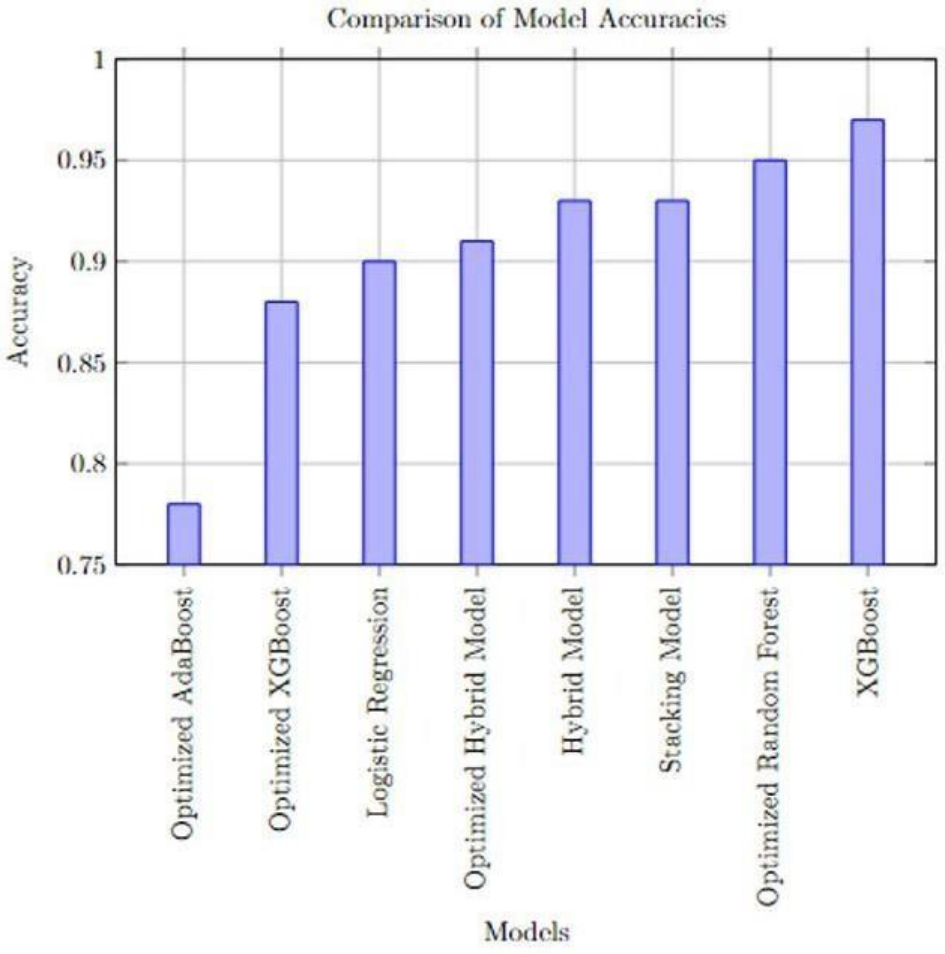


Figure 7: Comparison of model accuracies

CHAPTER-6

CONCLUSION

6. Conclusion

These results further reveal the efficiency of machine learn ing algorithms, especially XGBoost, in predicting heart disease with almost an accuracy of 0.97. The study proved that careful preparation of input data improves model performance after having gone through rigorous data preprocessing techniques, including the use of SMOTE to deal with class imbalance and strategic feature selection. These results do highlight the power of using advanced analytics to aid clinical processes at the right time to undertake risk assessments that immediately inform medical judgment. The further implications connected with the developed approach reach beyond the immediate conclusions that can be inferred from the conducted research. It may assist better outcomes for patients with heart disease and decrease the death rate from this leading cause of mortality when implemented machine learning capabilities in heart disease prediction. Fu ture activities in this area will focus on increasing the size of datasets and availability of additional clinical features in order to enhance the prediction abilities of models.

REFERENCES

[1] Sachdeva, R. K., Singh, K. D., Sharma, S., Bathla, P., Solanki, V. (2023). An Organized Method for Heart Failure Classification. 2023 International Conference on Emerging Smart Computing and Informatics (ESCI), AISSMS Institute of Information Technology, Pune, India, March 1-3.

[2] S. Jamal, W. A. Elenin, and L. Chen, ”Developing and evaluating data-driven heart disease prediction models by ensemble methods on different data mining tools,” in 2023 IEEE 14th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON), New York, NY, USA, 2023, pp. 678–682

[3] D. K. Lutfi and G. F. Shidik, ”Improvement Heart Failure Prediction Using Binary Preprocessing,” 2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA), Semarang, Indonesia, 2023, pp. 236-241

[4] Mamun, M., Farjana, A., Mamun, M. A., Ahammed, M. S., Rah man, M. M. (2022). Heart failure survival prediction using machine learning algorithm: am I safe from heart failure? IEEE World AI IoT Congress (AIIoT), 978-1-6654-8453-4/22/$31.00

[5] Jain, V. (2023). Heart Failure Prediction Using Machine Learning Algorithms with Cross Validations. In 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT) (pp. 1417-1423).

[6] Shetty, M., D, K., Shetty, S. B., Shetty, A. (2023). Data Driven Approach for Heart Failure Analysis. 2023 International Conference on Network, Multimedia and Information Technology (NMITCON), 979-8-3503-0082-6/23/$31.00
41, no. 40, pp. 3882-3889, 2020

[7] Saravanan, S., Swaminathan, K. (2021). Hybrid K-Means and Support Vector Machine to Predict Heart Failure. Proceedings of the Second International Conference on Smart Electronics and Communication (ICOSEC), 978-1-6654-3368-6/21/$31.00.

[8] Elghalid, R. A. M., Aldeeb, F. H. A., Alwirshiffani, A., Andiasha, A., Mohamed, A. A. I. (2022). Comparison of Some Machine Learning Algorithms for Predicting Heart Failure. 2022 International Confer ence on Engineering MIS (ICEMIS), 978-1-6654-5436-0/22/$31.00.

[9] Mehta, D., Mehta, P., Naik, A., Kaul, R., Bide, P. J. (2021). Death by heart failure prediction using ML algorithms. 2021 International Conference on Nascent Technologies in Engineering (ICNTE), 978-1-7281-9061-7/21/$31.00.
[10] Sang Xin, Yao Quan Zhu, Ma Ling, Cai Hong Wen, Luo Peng. Survival prediction of patients with heart failure based on support vector machine algorithm[J]. 2020 International Conference on Robots Intelligent System (ICRIS), 2020: 636-639.
[11] M. L. Mcclellan et al., ”Early detection of heart disease using data mining techniques: A promising future in predictive analytics,” International Journal of Engineering Research and Applications, vol. 10, no. 2, pp. 107-112, 2020.
[12] R. L. Sacco et al., ”Heart Disease and Stroke Statistics—2021 Update: A Report From the American Heart Association,” Circula tion, vol. 143, no. 8, pp. e254–e743, 2021.
[13] M. F. Piepoli et al, ”Heart failure and cardiovascular prevention: a call to action,” European Heart Journal, vol. 41, no. 40, pp. 3882-3889, 2020