PERFORMANCE MEASUREMENT IN THE DIAGNOSIS OF HEART DISEASE USING FISHER SCORE

INTERNATIONAL ISLAMIC UNIVERSITY CHITTAGONG (IIUC)
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Nusrat Enam (C181225)

Jannatul Maowa (C181239)

Amreen Khan (C181260)

SUPERVISOR
Md. Mahiuddin
Associate Professor
Department of Computer Science and Engineering

THESIS SUBMITTED IN FULFILLMENT FOR THE DEGREE OF
B.Sc IN COMPUTER SCIENCE AND ENGINEERING

PERFORMANCE MEASUREMENT IN THE DIAGNOSIS OF HEART DISEASE USING FISHER SCORE

Nusrat Enam (C181225)

Jannatul Maowa (C181239)

Amreen Khan (C181260)

INTERNATIONAL ISLAMIC UNIVERSITY CHITTAGONG (IIUC)
CHITTAGONG, BANGLADESH

PERFORMANCE MEASUREMENT IN THE DIAGNOSIS OF HEART DISEASE USING FISHER SCORE

Nusrat Enam (C181225)

Jannatul Maowa (C181239)

Amreen Khan (C181260)

THESIS SUBMITTED IN FULFILLMENT FOR THE DEGREE OF B. SC. IN COMPUTER SCIENCE AND ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (CSE)

INTERNATIONAL ISLAMIC UNIVERSITY CHITTAGONG (IIUC)

CHITTAGONG, BANGLADESH

DECLARATION

We hereby declare that the work in this thesis is my/our own except for quotations and summaries which have been duly acknowledged.

Nusrat Enam-C181225
Jannatul Maowa-C181239
Amreen Khan-C181260

SUPERVISOR'S DECLARATION

We hereby declare that we have read this thesis and in our opinion, this thesis is sufficient in terms of scope and quality for the award of the degree of B. Sc. in Computer Science & Engineering.

Md. Mahiuddin
Associate Professor
Department of Computer Science and Engineering

DECLARATION OF THESIS REPORT AND COPYRIGHT

THESIS REPORT TITLE

AUTHORS:

SL NO.	AUTHOR'S NAME	STUDENT ID	SIGNATURE
1			
2			
3			
NAME OF SUPERVISOR: Md Mahiuddin **SIGNATURE OF SUPERVISOR:**			

we declare

1. Our thesis is to be published as online open access (full text) at the IIUC database or archive.

2. Dept. of CSE, IIUC reserves the right as follows:

i. The thesis is the property of the Dept. of CSE, IIUC

ii. The Library of IIUC has the right to make copies for the purpose of research only.

iii. The Library has the right to make copies of the thesis for academic exchange.

ACKNOWLEDGEMENT

The greatest thanks go to Almighty Allah for all of his favors, including the patience and good health he provided us with while we were working on this thesis. We are extremely fortunate to have Md Mahiuddin Sir as our supervisor. We appreciate the help from all of our friends and classmates. Our parents deserve our thankfulness as well. We want to thank International Islamic University Chittagong for giving us all of the comforts.

ABSTRACT

One of the most prevalent and significant diseases affecting people's health is cardiovascular disease (CVD). Early diagnosis may allow for CVD mitigation or prevention, which could lower mortality rates. A viable strategy is to find risk factors using machine learning algorithms. We would like to suggest a model that combines various approaches to obtain accurate cardiac disease prediction. Our dataset was appropriately organized, and we updated it with the required imputations. Five models are trained and tested to obtain the model with higher accuracy. We have combined two datasets, the UCI heart disease, and the Framingham heart disease dataset. To select appropriate features, the fisher score is used two times and the chi-square is used three times. MICE imputation, Simple imputer, and Median value are used for missing value imputation during data pre-processing. Using selected features, models are trained and further tested using different classifiers such as Gradient Boosting, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, XgBoost, AdaBoost, and Naive Bayes and compared accuracy for each model. For each of the five models, most of the classifiers performed significantly better than the existing works. Although errors can be pretty expensive, maximizing model accuracy helps to reduce such costs. Using feature selection approaches, the performance is noticeably enhanced. The accuracy of the ensemble algorithms was enhanced by the feature selection strategies. Therefore, our results demonstrated a significant improvement in prediction accuracy, enabling the application of machine learning models to make sound business decisions. Better choices are the result of more accurate model outcomes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AdaBoost	Adaptive Boosting
AI-based	Artificial intelligence
AUC	Area under Curve
CAD	Coronary Artery Disease
CHD	Coronary Heart Disease
CSV	Comma-Separated Values
CVD	Cardiovascular Disease
GB	Gradient Boosting
K-NN	K-nearest neighbor
LASSO	Least Absolute Shrinkage and Selection Operator
LightGBM	Light Gradient-Boosting Machine
LR	Logistic Regression
MAR	Missing At Random
MICE	Multivariate imputation by chained equations
ML	Machine Learning
MLP	Multi-Layer Perceptron
ROC	Receiver Operating Characteristics
SVM	Support Vector Machine
UCI	University of California Irvine machine learning repository
XGBoost	Extreme Gradient Boost

CHAPTER I

INTRODUCTION

1.1 RESEARCH BACKGROUND

Blood arteries and the heart make up the cardiovascular system. Any severe, abnormal heart or blood vessel ailment is referred to as cardiovascular disease (CVD) (arteries, veins). Cardiovascular disease (CVD), a form of heart disease that continues to be a leading factor in death worldwide, is responsible for more than 30% of all fatalities [1]. Many medical conditions can be identified, detected, and predicted using machine learning. This study's primary purpose is to spot cardiac issues at an early stage.

In order to comprehend and further explore future research in this field of heart disease prediction, this study will investigate the supervised machine learning approach. Our contribution to this paper consists of developing machine learning models accurately and comparing them.

Many studies have been conducted on heart disease forecasting systems in medical facilities using various data mining techniques and machine learning algorithms, but there is still much potential for improvements to enhance the efficiency of this prediction model using modern approaches to machine learning. Deep Learning and neural networks can be used for further research on the topic of heart disease prediction.

1.2 PROBLEM STATEMENT

A heart disease prediction algorithm can assist in the early diagnosis of cardiac problems, even though supervised learning techniques have shown promise in text categorization, it was still not very effective. We will develop models based on supervised learning and compare models in this work to forecast cardiac illness earlier and to a larger extent.

In light of this, we create models utilizing the often-used feature selection strategy, chi-square, as well as the rarely-used feature selection technique, fisher score. Once more, we assess each model using seven machine-learning classifiers. Finally, using the corresponding accuracy score, we compare each model. As far as we are aware, this is the first time that models for cardiovascular disease prediction have been developed using a unified dataset. (Framingham dataset & UCI heart disease dataset).

1.3 MOTIVATION

1. Cardiovascular disease is a major killer worldwide (CVD). According to a prediction, 17.9 million individuals died from cardiovascular fatal diseases in 2019—32% of all deaths. Heart attacks and strokes were the main causes of 85 percent of these fatalities.[3]
2. Throughout the United States, heart disease was the culprit of 1 in 5 fatalities in 2020,

claiming the lives of nearly 697,000 people. [4]

3. By addressing behavioral hazard elements like smoking, and bad eating habits. Most cardiovascular problems can be prevented by losing weight, getting more exercise, and cutting back on alcohol usage.

4. In order to begin counseling and antidepressant treatment for cardiovascular disease, early identification is essential.

1.4 OBJECTIVE OF RESEARCH

The following can be used to summarize the efforts of our study:

- The primary objective is to enhance the prediction accuracy of five models for diagnosing cardiac disease.
- On a combined dataset (Framingham and UCI heart disease dataset), the proposed five models make use of the most commonly used chi-square, the fisher score, and a rarely used feature selection methodology. Different classifiers are used to evaluate the models and further, we compared the prediction accuracy for each model.

1.5 ORGANIZATION OF THE THESIS

This thesis is organized as follows.
A background on supervised learning is provided in Section 2. Section 3 describes the methodology of the proposed model. Performance evaluation is included in Section 4. Finally, Section 5 concludes with some concluding remarks on this work and suggests areas for future research.

1.6 SUMMARY

We introduced this thesis's work in the first chapter. Additionally, the context of our work, the problem statement, the objective of the study, and contributions are discussed.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

There has been a lot of excitement in applying machine learning, deep learning, and data mining techniques and tools to predict cardiac illness.Researchers have employed a variety of datasets, algorithms, and procedures; the results they have discovered so far and future work will be used to figure out the most efficient techniques for predicting cardiovascular disease. Following are some of the main strategies for predicting cardiovascular diseases:

2.2 AN OVERVIEW OF SUPERVISED LEARNING

2.2.1 Simple-Imputer

An estimator called the imputer is used to fill in the blanks in datasets. It uses mean, median, and constant in relation to numerical values. The most common and constant value is selected for categorical values

A dataset's missing values can be replaced using a variety of input methodologies by the SimpleImputer class found in the sklearn.impute module. Although it was created to work with numerical data, SimpleImputer can also handle categorical data that is supplied as strings. A scikit-learn pipeline may include SimpleImputer. With the "mean" default method, missing values are replaced with the column's median value. Other choices include "most frequent," which substitutes missing values with the column's most prevalent value, and "constant" (which replaces missing values with a constant value). By providing a list of column names, SimpleImputer can also be used to simultaneously impute several columns. The missing values in each of the designated columns will subsequently be replaced by SimpleImputer.Then, SimpleImputer will add missing values to each of the designated columns.

2.2.2 MICE

With the help of several models, partial attributes are imputed in the MICE approach, which is according to fully implicit specifications. Therefore, by applying a distinct model for each feature, datasets containing continuous, binary, and categorical attributes may have missing values imputed using MICE. Every property is therefore modeled in accordance with how frequently it occurs; for instance, utilizing logistic regression model binary or categorical variables, while linear regression is performed to model continuous variables. The modeled attribute serves as the variable that is dependent in regression models, while the other qualities serve as the independent variables. Louridi et al. [6] used the MICE imputation mechanism for imputing missing values. Here is a description of the MICE algorithm:

1. Create a simple imputation for each dataset's missing values.

2. Replacing one feature's missing values (Fx).

3. Amount of data points of Fx is utilized to train an independent forecasting system in which Fx is the dependent feature.

4. The predictions generated by the model created in step three are used to replace the missing values for Fx.

5. Steps 2-4 are repeated for each feature that lacks values. One cycle or iteration of a prediction model is complete once all features with missing values have been imputed.

6. The imputations are updated after each cycle as steps 2 through 5 are repeated for n cycles. The goal is to produce a stable imputation by using the number of iterations. In the most recent iteration, the imputed dataset was acquired.

2.2.3 Logistic regression

Logistic regression is used for supervised learning. It serves to ascertain or forecast the likelihood that a binary (yes/no) occurrence will take place. Linear regression produces continuous, unbounded data. However, in the case of logistic regression, the projected outcome is discrete and restricted to a small range of values. Binary logistic regression, where the result is binary, is the most typical use of this for classification issues (yes or no).In the real world, many different industries and areas use logistic regression. A method for diagnosing a cardiovascular disease that aids a patient based on clinical data about a past heart problem they have been diagnosed with was offered by Folsom A R et al. [5]. They used Logistic Regression in their model.

- The likelihood that a tumor will be benign or malignant can be determined using logistic regression in the medical field.
- This classifier can be used to determine if a transaction is fraudulent or not in the financial sector.
- It can be used to forecast whether a specific audience will respond or not in marketing.

2.2.3.1 Logistic function

The strategy's core component, the logistic function, served as the basis for the movement's name, logistic regression. Using this S-shaped curve, any number with a real value can be changed to a value ranging from 0 to 1, but never precisely at those values.

$$1/(1 + e \text{ ^-value})$$

Where value is the precise number we want to change and e (Euler's number or the EXP) is the base of the equations ().

2.2.3.2 Logistic regression's representation

In this classifier, the representation is a mathematical equation. Linear weights or coefficient values are added to input values (x) (referred to as the Greek capital letter Beta)to forecast an outcome value (y). Unlike linear regression, which results in a value the output modeled is a binary value (0 or 1) rather than a numeric number.
Here is an illustration of the formula:

$$y=e^{(b_0+b_1*x)}/(1+e^{(b_0+b_1*x)})$$

When y is the projected outcome, the single input value's coefficient is represented by b1, and b0 is the bias or intercept term (x). For each column in your input data, you must learn the corresponding b coefficients (constant real values) from your training set.
The real visual depiction of the model you would maintain in memory or a file is the equation's coefficients, also referred to as the beta value or b's.

2.2.4 Support Vector Machine

Support Vector Machine, one of the most well-liked supervised learning techniques, is used to address classification and regression problems. The SVM technique seeks the best range or decision boundary that can categorize n-dimensional space into classes in order to swiftly classify fresh data points in the future. A hyper-plane is the name for this perfect decision boundary. It is employed to pick the critical points and vectors that make up the hyper-plane. The backbone of the SVM classifier is a pair of support vectors, which are exploited to represent such specific cases. Saboor et al. [7] suggested a solution to grow accuracy and diminish the cost of cardiac illness prediction while employing machine learning techniques. They have used a variety of machine learning classifiers to characterize the prediction of heart disease, with SVM achieving a 96.72% accuracy rate.

Example:

Now let us say we want a prototype that can perfectly distinguish a cat from a dog and we come across a rare cat that also resembles a dog. Such a prototype can be built using the SVM approach. Before testing it with this weird species, we will initially train our approach on a number of photos of cats and dogs to get used to the variety of traits that cats and dogs have. Therefore, the support vector will be able to recognize both the extreme cases of cats and dogs when it builds a judgment border between these two groups of data (cat and dog). The support vectors will identify it as a cat.

2.2.4.1 Types of SVM

Linear SVM

Data that can be divided into two groups along a single straight line are known as linearly separable data. It is used to categorize such data, and the algorithm employed is referred to as the Linear SVM classifier.

Non-Linear SVM

If a dataset cannot be characterized using a straight line, the classification algorithm is known as a non-linear SVM classifier.

2.2.5 K-Nearest Neighbor(KNN)

Regression analysis, as well as classification, can be accomplished using K-nearest neighbors (KNN). By measuring the separation between the test data and all of the training points, KNN seeks to identify the proper category for the test data. Then, select the K places with the largest test data correlation. The test data will be divided into one of the "K" training data classes by the KNN method, and The class with the greatest chance will be found. The average of the "K" defined training point serves as the value in a regression scenario. Jindal et al. [8] proposed a model that tells that LR and KNN outperform Random Forest Classifiers in order to forecast which clients may develop cardiovascular disease. This demonstrates that KNN and Logistic Regression are superior for the diagnosis of coronary disease.

2.2.5.1 KNN working principle

Step 1: Establish the K-numbers for the neighbors.
Step 2:It should be known what the Euclidean distance is between K neighbors.
Step 3: Based on the measured Euclidean distance, the K nearest neighbors should be selected.
Step 4: The total of each department's data points across these k neighbors should be totaled.
Step 5: The department with the largest neighbor count should receive the new data points.
Step six: We have finished our model.

2.2.6 Gradient boosting algorithm

One well-liked boosting technique is gradient boosting. Each forecast in gradient boosting adjusts the error of its predecessor. Instead of altering the training instance weights like Adaboost does, each predictor in this method is trained using the predecessor's residual blunders as labels. Gradient Boosted Trees uses the CART learner as its base learner (Classification and Regression Trees). Ch. Anwar ul Hassan et al. [13] introduced that the presence of heart issues is predicted

using ML classifiers. The use of ML models for prediction is then employed. These eleven used ML methods' potential to forecast heart illness was evaluated. In Random forest and gradient boosting, they achieved good accuracy.

2.2.6.1 Method

There are three components to the gradient boosting algorithm. The loss function varies depending on the task at hand, the additive model, which adds trees using a gradient descent method, and weak learners who are employed to make predictions. By merging the upcoming model with the ones that came before it, the approach minimizes error while predicting the best model feasible.

2.2.6.2 Time of using gradient boosting

When processing vast and complicated amounts of data, gradient boosting is frequently employed to lower the possibility of inaccuracy. Additionally, it is employed in regression and classification techniques to get the most accurate predictions.

2.2.7 AdaBoost Algorithm

The name AdaBoost stands for adaptive boosting, a machine learning method used in ensemble learning. Decision trees with one level are the AdaBoost algorithm that is most frequently employed. This suggests that there is only one divide in the AdaBoost decision. These decision trees are referred to as Decision Stumps in the AdaBoost algorithm. Sai Bhavan Gubbala [9] utilized Adaboost, an ensemble technique to evaluate the model and got an accuracy level of 78.59%.

2.2.7.1 How Adaboost works

The main goal of this approach is to draw attention to the observations that were improperly classified. The word "Adaptive Boosting" (AdaBoost) refers to a meta-learner that adapts to the performance of dim classifiers by forcing big weight to the incorrectly classified examinations of the previous weak learner (typically a decision stump: one-level decision trees, which signify that the tree is constructed using just one decision variable and that all the child's directly connected to the root).

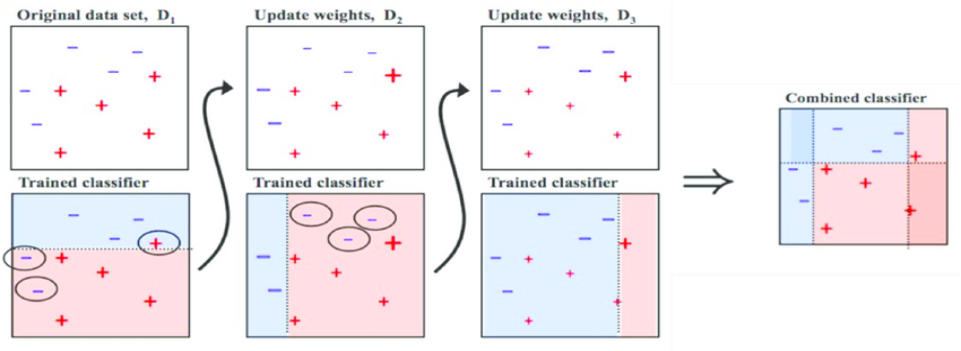

Figure 2. 1 Working Principle of Adaboost

The plus and minus symbols are circled. If incorrectly classified, they expand; alternatively, if flawlessly classified, they contract.
Here is a generalized meta-learner (weighted sum) form:

$$F(x)=\text{sign}\left(\sum_{m=1}^{M} \theta_m f_m(x)\right),$$

Where theta denotes the weight corresponding to the particular weak classifier, and f represents the weak learner. Given the presumptions, the implementation of the vanilla AdaBoost algorithm includes the following stages:
Assumptions: Considering an m-sample training set in the form of data sets (x,y), where the set X includes x (the set of potential values for the dependent variables) and y is a member of set -1,1 indicating that the problem being studied is a classification problem (vanilla AdaBoost can only be used for classification issues).
1. The weight distributions for the observations can then be added up, starting with a weight distribution that is inversely proportional to the number of findings (1/m, compatibility between observations, which indicates that it is a standardized distribution and the sum of the distributions for the m observations is 1). This weight distribution will be used to mimic the decisions made by the weak learner.
2. For T iterations, iteratively:

a)We use the probability distribution presented above to train the weak classifier.

b)The weak learners forecast the grouping of data, and the predictions are then used to compute an error metric, which is equal to the sum of the weak learner values of the incorrectly labeled forecasts (indicating that they are out of sync in conflict with the real value y).

c)We assess the weight given to a particular weak learner based on the error; the higher the error, the more weight is given.

d)The dim learner's weight is then used to adjust the dim classifier's initial distribution for the iteration cycle: We assign misclassified observations with less weight than correctly classified ones. A normalizing factor helps the distribution keep its characteristic of normalcy.

3. Then, based on T-products between the total of the dim learner's predictions and the weights assigned to the specific dim learner, we output the resultant value associated with any hypothesis we choose (especially a not-tested one, testing set). The sign function is a function that outputs values of -1 or 1, depending on the sign of the function's argument.

4. The algorithm appears to operate iteratively for each data observation, changing the weight of the data after each observation in accordance with whether or not the prediction was correct, at least according to the implementation given above.

5. The output, which delivers a final hypothesis, is supplied by an approximation (or, more precisely, activation in this case).

2.2.8 Naive Bayes

It is a categorization procedure based on predictor independence and the Bayes Theorem. To put it simply, a Naive Bayes classifier thinks that the existence of one feature in a class has no influence on the existence of any other characteristics.

The naive Bayes model is easy to build and especially useful for very large data sets. In addition to being simple, Naive Bayes is known to outperform even the most sophisticated classification methods.A system for expert medical diagnosis was put up by S. Palaniappan and R. Awang [10] for the identification of HD. 86.12% accuracy was attained via Naive Bayes (NB), a machine learning predictive model, during the system's development.

For instance, if a fruit is red, spherical, and around 3 inches in diameter, it might be classified as an apple. Even though these attributes are interdependent or dependent on the existence of other characters, taken individually, each of these traits raises the possibility that this fruit is an apple, hence the name "Naive."

2.2.9 Xgboost

A popular method used in ensemble learning is the extreme gradient boosting algorithm, or XGBoost. Machine learning techniques known as "ensemble learning" use a variety of models to jointly predict outcomes. By transforming a series of initially weak models into increasingly powerful models, boosting algorithms set themselves apart from other ensemble learning techniques. Gradient boosting methods make choices on how to enhance a model's performance relying on the gradient of a performance-measuring loss function.

Decision trees are graph-building models that examine the input under various "if" statements and they are used by the XGBoost algorithm (vertices in the graph). The next "if" condition and

ultimate prediction depend on whether the "if" condition is met. The XGBoost algorithm continuously adds more and more "if" conditions to the decision tree in order to produce a more reliable model.Bharti et al. [2] utilized Xgboost in their approach where they used a feature selection technique and outliers are also detected getting an accuracy level of 71.4%.

2.2.10 Fisher Score

An exceptionally convincing feature selection technique is the Fisher score. It picks every attribute uniquely based on how well it performs under the Fisher criterion, which results in a less-than-ideal collection of features. To jointly pick the features for this investigation, we provide a modified Fisher score.

The Fisher score's overall goal is to identify a subset of features such that the distances between data points in the data space covered by the selected features are as great as possible, while the distances between data points in the same class are as manageable.More specifically, the input data matrix $X\varepsilon R^{d\times n}$ given the chosen m characteristics decreases to $\mathbf{Z} \in \mathbf{R}^{m\times n}$**.** Afterward, the Fisher Score is determined as follows:

$$F(\mathbf{Z}) = \mathrm{tr}\{(\underline{S}_b)(\underline{S}_t+\gamma\mathbf{I})^{-1}\},$$

Where γ is a positive regularization parameter, $\underline{S}_b$ is called between-class scatter matrix, and

$\underline{S}_t$ is called total scatter matrix, which is defined as

$$\underline{S}_b = \sum_{k=1}^{c}(\underline{\mu}_k - \underline{\mu})(\underline{\mu}_k - \mu)^T$$

$$\underline{S}_b = \sum_{i=1}^{n}(\mathbf{z}_i - \underline{\mu})(\mathbf{z}_i - \underline{\mu})^T,$$

Where $\underline{\mu}_k$correspondingly the k-th class's mean vector and size in the compressed data space, i.e.Z,$\underline{\mu} = \sum_{k=1}^{c}\eta_k\underline{\mu}_k$is the reduced data's total mean vector. Since $\underline{S}_t$is typically singular, we add a disruption term $\gamma^I$to acquire it a semi-definitely positive. Because there are $\binom{d}{m}$ prospect The feature selection problem is a challenging combinatorial optimization problem that selects Z's out of X. As a means of the challenge, a common heuristic approach is to independently calculate a score for each feature in accordance with criterion F. That is to say, it

merely takes into $x^{j} \epsilon R^{1*n}$ In this case, there are only $\binom{d}{1} = d$ candidates. Mainly, let μ_k^j and σ_k^j be the mean and standard deviation of k-th class, corresponding to the j-th feature. Let corresponding to the j-th feature, denote the mean and standard deviation of the entire data set. After that, the j-th feature's Fisher score is calculated below,

$$F(x^j) = \sum_{k=1}^{c} \eta_k (\mu_k^j - \mu^j)^2 / (\sigma^j)^2$$

where $(\sigma^j)^2 = \sum_{k=1}^{c} \eta_k (\sigma_k^j)^2$. It chooses features with high rankings in the top m after computing the f score for each feature. The heuristic algorithm's chosen features are not the best because each feature's score is calculated separately. More importantly, as we have already indicated, the heuristic algorithm is unable to select features that have relatively low individual scores but extremely high aggregate scores. It also cannot reduce superfluous features. We are inspired to suggest a generic Fisher score to address these issues because of this. Saqlain et al. [23] used fisher score, a feature selection technique with two of other feature selection techniques namely forward and backward feature selection algorithms for heart disease prediction.

2.2.11 Chi-square

A critical barrier in machine learning is determining which features to utilize in the model-building process when there are many features available. The chi-square test assists in solving the feature selection problem by looking at the relationships between the features. In their study, Jabbar et al. [12] advised utilizing the measure of feature selection that uses chi-square to examine the interactions between variables and determine whether or not they are associated. In this study, the data recommend a classification model to predict cardiovascular disease employing chi-square and genetic algorithm as attribute selection measures and random forest as a classifier.

2.2.11.1 Feature extraction Chi-square test

We frequently ponder the Chi-Square test's use in machine learning and how it affects results. When there are many features in line, choosing the best ones to use in the process of modeling is a significant difficulty in machine learning. By investigating the relationships between the features, the chi-square test assists in feature selection and resolving issues. The features with the greatest Chi-square scores are chosen after measuring the correlation between each feature and

the desired result. The chi-square score is calculated using:

$$x^2 = (\text{Observed frequency} - \text{Expected frequency})^2 / \text{Expected frequency}$$

we can write,
Observed frequency = No. of observations of class
Expected frequency = No. of expected observations of class if there was no relationship between the feature and the target.

2.3 LITERATURE REVIEW

2.3.1 Existing work on Heart Disease Prediction:

Bharti et al used Deep learning and Machine learning algorithms to compare the result. Isolation Forest is utilized to manage datasets with insignificant features, and data is also sanitized to yield better results. [2]

Folsom A R et al proposed the findings in this paper were supported by multiple logistic regression. These findings lend credence to the idea that there is a causal link between the distribution of body fat and the development of cardiovascular disease in older women.[5]

Louridi et al developed a system by which they solved the missing data problem by using the MICE model. They used many machine learning techniques but got good accuracy in the stacking algorithm. Here two different datasets are used.[6]

Saboor et al suggested a solution to grow accuracy and diminish the cost of cardiac illness prediction while employing machine learning techniques. They have used a variety of machine learning classifiers to characterize the prediction of heart disease, with SVM obtaining an accuracy of 96.72%.[7]

Harshit Jindale et al presented a technique. This procedure includes a dataset of patient health records, including information on chest discomfort, blood sugar levels, blood pressure, and other conditions, to forecast individuals who may develop cardiovascular disease.By analyzing the past health cardiac disease diagnosis on the basis of clinical data, this coronary heart detection system aids the patient.[8]

Gubbala proposed this paper's comparative analysis of machine learning algorithms' accuracy is its main objective. In order to anticipate models with the highest degree of accuracy, this research will be expanded.[9]

S. Palaniappan et al proposed a prototype Intelligent Heart Disease Prediction System (IHDPS) that was created by combining Decision Trees, Naive Bayes, and Neural Network data mining

approaches.[10]

Shrivastava et al suggested the research analysis demonstrated that the suggested model is a highly optimized version of the machine learning classification methods that are already in practice.[11]

M.A. Jabbar et al. constructed a useful technique for anticipating heart illness using the random forest.Data mining has significant benefits for the prognosis of heart disease. We employed feature selection based on chi-square and genetic algorithm measures to categorize cardiovascular diseases. The recommended approach (Random forest and Chi-square) has an accuracy of 83.70% for the Cardiac Stalog data set. The accuracy of coronary disease prediction has risen with the usage of random forests. In this work, systematic testing with 10-fold cross-validation is used to identify the best effective strategy.
.[12]

Ch. Anwar ul Hassan et al introduced that the presence of heart issues is predicted using ML classifiers. The UCI repository was used to obtain the dataset. Preprocessing and data cleansing are done using the acquired data. The use of ML models for prediction is then employed. These eleven used ML methods' potential to forecast heart illness was evaluated. In Random forest and gradient boosting, they achieved good accuracy.[13]

K. Karthick et al suggested in this study, a forecasting model for cardiovascular disease was created utilizing the Cleveland HD dataset from the UCI ML repository and the six machine learning (ML) classification methods SVM with RBF kernel, Gaussian Naive Bayes, logistic regression, LightGBM, XGBoost, and random forest. The 13 qualities chosen using the chi-square distribution are most accurately predicted by the random forest method (88.5%), followed by SVM and logistic regression (80.32%).[14]

Enhancing categorization prediction performance requires tight control over missing value and feature selection, as Md. Julker Nayeem et al. have demonstrated. To discover which classification model is the best, they have compared them. Applying mean imputation and the info gain feature selection method to our dataset, each classification algorithm exhibits excellent performance when dealing with observations from the dataset that contain missing values. The low classification accuracy in the dataset could be attributed to null values and low contribution features. The Random Forest classifier performed the best out of the three.[15]

Chen et al proposed Pooled area curve is constructed in a machine learning algorithm to predict CAD and also used a noisy dataset for better clarification of classifiers. This paper used adaptive image-based classification techniques. [16]

Yahaya et al suggest in this paper gave us an idea about many papers. It combined many papers and helped us to know about many techniques [17]

Nissa et al in this paper improved the quality of the dataset by using preprocessing techniques namely removing outliers. They mainly gave attention to handling corrupted and missing values. They used three machine learning techniques. [18]

Salhi et al proposed data analytics to detect and predict heart disease by using a confusion matrix. To compare accuracy, three data analytics methods—Neural networks, Support Vector Machine, and K - nearest neighbors used on various datasets. To choose pertinent features, the correlation matrix is used.[19]

Ghosh et al have proposed efficient data gathering, data cleaning, and data revolution methods to predict CVD using a combined dataset. Feature selection techniques such as Relief and Least absolute shrinkage and selection operator LASSO. Hybrid classifiers like the Decision tree bagging method, RFBM, KNNBM, etc.[20]

Shetgaonkar et al Three AI-based techniques—Decision Tree, Naive Bayes, and Neural Network—are used by this paper's researchers to predict cardiovascular or heart disease. Each of these approaches will be tested according to a wide range of particular standards, with improvements made for greater accuracy. Experimented with different hidden layers, learning rates, and attribute adjustments to increase accuracy in a neural network. As a result, accuracy levels have varied.[21]

Pal et al suggested using the K-nearest neighbor (K-NN) and multi-layer perceptron (MLP) machine learning techniques. Outliers and characteristics with null values should be removed to enhance model performance. MLP achieved the highest accuracy in this instance.[22]

Saqlain et al. [23] used the fisher score, a feature selection technique with two other feature selection techniques namely forward and backward feature selection algorithms for heart disease prediction.

Armin Yazdani et al achieved the highest confidence score when employing central elements in WARM to forecast cardiovascular illness. It has been demonstrated that allocating suitable weight scores enhances the prediction's confidence level performance. To predict cardiovascular events, a series of characteristics with varying scores to represent the strength of each characteristic were included.[24]

Mohammad Alsaffar et al present a particular decision support tool for the diagnosis of ischemic heart disease with the ability to be modified and learned so that it may be used to diagnose different diseases. This tool is offered as an alternative to current options, applying a hybrid approach that combines elements of machine learning and evolutionary computing (Genetic

Algorithms) (Case-Based Reasoning and Artificial Neural Networks). As a result, the system was able to analyze EKG signal images in addition to predicting diagnoses through the study of clinical data, giving medical practitioners more evidence to back their choice.[25]

BhaveshDhande et al presented a method by which, Diabetes and heart disease can both be combined to influence a patient's judgments. In order to comprehend the necessity of these predictions utilizing the ensemble technique, sufficient exploratory analysis and pre-analysis of normalized models have been conducted in this research. Using machine learning principles, the system promises to manage and link both cardiac and diabetes events to enable faster prediction. The Voting Classifier of Decision Tree, Sigmoid SVC, and Adaboost is shown to have a maximum accuracy of 88.57 % for heart disease and an accuracy of 80.95% for diabetes.[26]

Pooja Anbuselvan et al presented the overarching aim to define several data mining methods that can be effectively used to forecast cardiac disease. The purpose of this dissertation is to create prediction techniques that are successful and efficient while utilizing fewer features and tests.The model employed pre-processed data that had been previously altered. The most effective algorithms are XGBoost with 78.69% and Random Forest with 86.89%.The least accurate algorithm, K-Nearest Neighbor, has a performance of 57.83%.[27]

Sonam Nikhar et al explained the system used in this work for predicting coronary heart disease along with different classifier methods. Naive Bayes classifier and decision tree classifier are the two methodologies; according to our analysis, the decision tree is more accurate.[28]

J.Vijayashree et al suggested a method for the diagnosis of cardiac sickness, 11 attributes, the Waikato Environment for Knowledge Analysis (WEKA) tool, and five data mining algorithms are utilized: J48, Bayes Net, Naive Bayes, Simple Cart, and REPTREE. According to research, J48, REPTREE, and SIMPLE CART are the three best algorithms.[29]

2.4 SUMMARY

We have discussed our work's detailed literature review in this chapter. We carefully considered the precise phrases and used them in our work. We also reviewed some relevant articles that served as inspiration for and motivation for us to complete our method.

CHAPTER III

METHODOLOGY

3.1 INTRODUCTION

There are a few main sections to our research. Together, these components make up our research models. Following is the procedure we followed:

1. Data collection
2. Data preprocessing
3. Feature Selection
4. Feature Scaling
5. Data splitting
6. Training and testing of the algorithms
7. Performance evaluation of algorithms

3.2 DATA COLLECTION

3.2.1 UCI heart disease dataset

The UCI Heart condition Dataset, which includes data from four organizations, is gained from the UCI machine learning library. Foundation for Cleveland Clinic, Hungarian Institute of Cardiology in Zurich, Switzerland; University Hospital; Long Beach, California; and Budapest, Hungary.

Only fourteen of the 76 variables in the open dataset UCI heart disease were selected since they were determined to be the most crucial for diagnosing heart disease in the literature.

3.2.2 Framingham dataset

The dataset, which can be downloaded for free from the Kaggle website, came from a cardiovascular study that involved people from Framingham, Massachusetts. The categorization is used to estimate a patient's possibility of developing coronary heart disease within the next 10 years (CHD). The dataset comprises a total of 4,239 records, and 16 attributes, and contains patient information. Every trait might be a risk factor. Risk factors include worries about demographic, behavioral, and medical traits.

These two datasets have been combined into one. The total number of features is 29 and instances are 122132 as a result.

3.3 DATA PREPROCESSING

Pre-processing the data was the second critical stage in our workflow. In order for the machine to comprehend the data's content, this is done. Finding, eliminating, or changing data that are inaccurate, incomplete, or unrelated to the model are typically part of this process. However, in our instance, this stage also included converting all data into the appropriate unit system, locating any values or entries that were lacking, styling inputs, eliminating unwanted values, reviewing, and sanitizing.

Many missing values were found after combining the dataset. Therefore, we preprocessed the dataset to increase its usefulness. We employ three alternative methods to address the missing value issue. For Missing value imputation, we employed the Mice Imputation Technique, Simple Imputer, and Median. We used these three methods individually.

3.4 FEATURE SELECTION

The method of feature extraction is used to minimize the number of attributes that define a large data collection in order to decrease computing complexity. Numerous variables can cause a data set to overfit training samples and have low generalizability for new samples. This was unnecessary, though, as the final data set was generated manually and as a result did not start out with any redundant variables.

On the data, we obtained after using the Mice Imputation Technique, the Simple Imputer, and the Median, we employed two types of feature selection techniques named Fisher score and the Chi-square technique. The data obtained using the mouse imputation technique was subjected to the Fisher score feature selection technique. Furthermore, chi-square was used to analyze the data from the Simple Imputer and Median.

3.5 FEATURE SCALING

This technique involves standardizing sets of independent variables or attributes. The data processing stage took care of this, sometimes referred to as data normalization. To scale our data, we used the MinMax Scalar library, which is a component of the Scikit Learn package.

3.6 DATA SPLITTING

The collected data were initially divided into two portions. The division was created for system testing and training. Any supervised machine learning or data science application is required to attain the method. Because the precision of the machine's or model's result will have a significant impact on the final result.

Then we are going to split the data then test and train them in a ratio of 60:40.

3.7 TRAINING AND TESTING OF THE ALGORITHMS

The whole investigation's primary aim was to assess the threat of coronary heart disease. We will try various classification techniques in order to do this. The most accurate accomplishment in terms of metrics is displayed in this part, which also provides an overview of all study data. We selected a number of algorithms that are frequently used to resolve supervised learning issues in classification techniques.

3.8 PERFORMANCE EVALUATION OF ALGORITHMS

Some matrices are taken into account when assessing the effectiveness of any algorithm such as Confusion Matrix, Classification Accuracy, Area under Curve(AUC). The confusion matrix, which assesses how well classification models perform when they make judgments based on test data, demonstrates the reliability of our classification model. Using the confusion matrix, we can find out the various model parameters, such as accuracy, precision, etc. When we use the term "accuracy," we often mean classification accuracy. By dividing the percentage of correct forecasts by the total number of input Samples, this is calculated. It works perfectly if there are a similar number of samples for each class. The likelihood that a classifier will select a positive example at random and give it a maximum ranking than a negative example is indicated by the classifier's AUC. AUC is a curve with a [0, 1] range that is plotted between the False Positive Rate Vs True Positive Rate at all various data points. The model performs better when the AUC value is greater.

Table 3.1 Proposed Approaches

Approaches	Missing Value Imputation	Feature Selection
1st Approach	Median	Chi-square
2nd Approach	Simple Imputer	Chi-square
3rd Approach	MICE	Fisher Score
4th Approach	Simple Imputer	Fisher Score
5th Approach	MICE	Chi-square

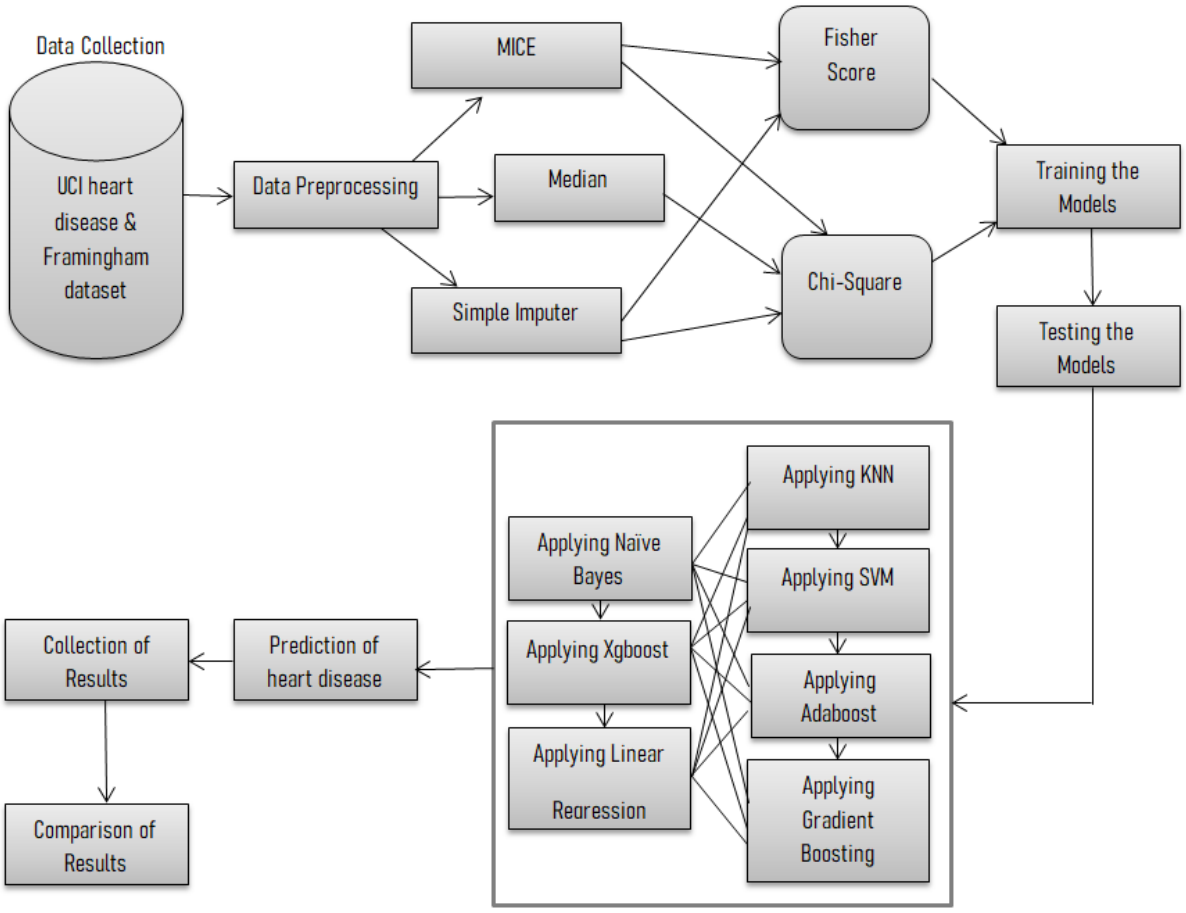

Figure 3.1 Flowchart of our methodology

3.9 SUMMARY

We covered the research methodology in this chapter.

CHAPTER IV

RESULTS AND DISCUSSION

Experimental Setup and Analysis

We need powerful computing power for our heart disease prediction research, which uses a machine learning method. Numerous libraries from the Scikit Learn package were used with our data set. As we've already discussed, our research projects frequently employ many methods, and the initial data set was sizable and had many rows. Prior to starting our research, we must initialize our data.

4.1 DATASET

The machine learning approach we are using in our research on heart disease prediction calls for powerful computing power. The Scikit Learn package had a number of packages that were employed with our data set. We initially started by uploading our merged dataset as a CSV file to Google Collaboratory. Fig.4.1 shows the snippet of the original CSV file that we have collected.

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	...	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	...	321	0	0	182	0	0.0	2	0	2	1
1	1	39	4.0	0	0.0	0.0	0	0	0	195.0	...	199	0	1	179	0	0.0	2	0	2	1
2	1	39	4.0	0	0.0	0.0	0	0	0	195.0	...	199	0	1	179	0	0.0	2	0	2	1
3	1	39	4.0	0	0.0	0.0	0	0	0	195.0	...	220	0	1	152	0	0.0	1	0	2	1
4	1	39	4.0	0	0.0	0.0	0	0	0	195.0	...	220	0	1	152	0	0.0	1	0	2	1
5	1	39	4.0	0	0.0	0.0	0	0	0	195.0	...	220	0	1	152	0	0.0	1	0	2	1
6	1	39	4.0	0	0.0	0.0	0	0	0	195.0	...	321	0	0	182	0	0.0	2	0	2	1
7	1	39	4.0	0	0.0	0.0	0	0	0	195.0	...	199	0	1	179	0	0.0	2	0	2	1
8	1	39	4.0	0	0.0	0.0	0	0	0	195.0	...	220	0	1	152	0	0.0	1	0	2	1
9	1	39	4.0	0	0.0	0.0	0	0	0	195.0	...	219	0	1	140	0	1.2	1	0	3	0
10	1	39	4.0	0	0.0	0.0	0	0	0	195.0	...	321	0	0	182	0	0.0	2	0	2	1
11	1	39	4.0	0	0.0	0.0	0	0	0	195.0	...	219	0	1	140	0	1.2	1	0	3	0
12	1	39	4.0	0	0.0	0.0	0	0	0	195.0	...	219	0	1	140	0	1.2	1	0	3	0
13	1	39	4.0	0	0.0	0.0	0	0	0	195.0	...	219	0	1	140	0	1.2	1	0	3	0
14	0	39	2.0	1	9.0	0.0	0	0	0	226.0	...	321	0	0	182	0	0.0	2	0	2	1
15	0	39	2.0	1	9.0	0.0	0	0	0	226.0	...	199	0	1	179	0	0.0	2	0	2	1
16	0	39	2.0	1	9.0	0.0	0	0	0	226.0	...	199	0	1	179	0	0.0	2	0	2	1
17	0	39	2.0	1	9.0	0.0	0	0	0	226.0	...	220	0	1	152	0	0.0	1	0	2	1
18	0	39	2.0	1	9.0	0.0	0	0	0	226.0	...	220	0	1	152	0	0.0	1	0	2	1
19	0	39	2.0	1	9.0	0.0	0	0	0	226.0	...	220	0	1	152	0	0.0	1	0	2	1

20 rows × 29 columns

Figure 4. 1 Our merged dataset

4.2 DATA PREPROCESSING

We first checked the missing values in our dataset in the data preprocessing stage. We dealt with duplicate values, and outlier detection and examined the target variable's distribution.

```
male                0
age                 0
education         856
currentSmoker       0
cigsPerDay        264
BPMeds            518
prevalentStroke     0
prevalentHyp        0
diabetes            0
totChol           451
sysBP               0
diaBP               0
BMI               133
heartRate          10
glucose          3315
TenYearCHD          0
sex                 0
cp                  0
trestbps            0
chol                0
fbs                 0
restecg             0
thalach             0
exang               0
oldpeak             0
slope               0
ca                  0
thal                0
target              0
dtype: int64
```

Figure 4. 2 Checking null values on our dataset

We attempted to fill in these missing values using the mice imputation technique, the median, and the simple imputer.

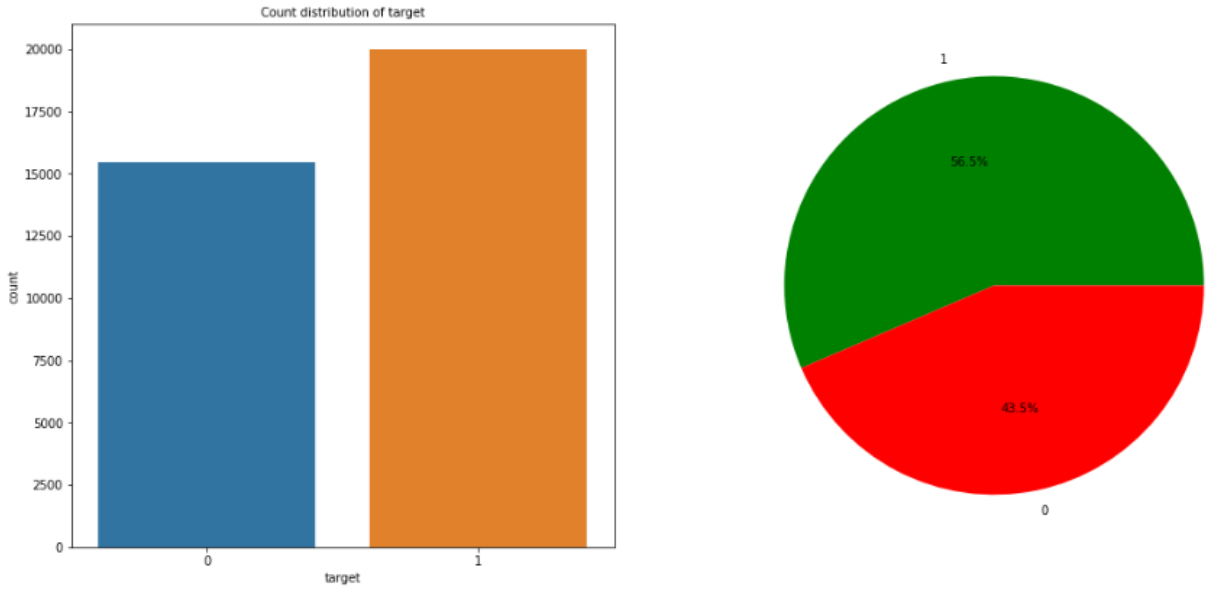

Figure 4. 3 Distribution of Target

It is evenly distributed. In other words, the proportion of negative cases does not exceed the proportion of positive cases. so there would be no class imbalance problems as a result.

4.3 FEATURE SELECTION

The attributes we considered when assessing our model are listed in this section. These are the Fisher score and the Chi-square.

Table 4. 1 Feature Scores in Chi-Square

Selected features	**1st Approach**	**2nd Approach**	**5th Approach**
Specs	**Scores**	**Scores**	**Scores**
'oldpeak'	79342.009209	80223.424515	82238.827344
'thalach’	63733.962335	64256.063172	65038.219223
'chol'	10021.956380	10077.352205	10312.938997
'age'	9138.126939	9247.136988	9169.273190
'ca'	7919.605565	7999.288933	8165.924007
'sysBP'	7351.638283	7526.322810	7556.729500
'cp'	7290.259264	7411.865062	7382.532552
'BMI'	5682.293051	5566.729205	6087.528471
'exang'	4237.292495	4270.877533	4351.795517
'trestbps'	3108.750340	3108.707828	3641.720100
'totChol'	1973.613718	1945.132670	2222.026472
'cigsPerDay'	1275.643377	1387.724323	1309.692083
'slope'	1019.172696	1028.828005	1079.831594
'thal'	705.616089	712.603738	721.085961
'diaBP'	668.672906	668.776842	666.868488

We have included the results using the Chi-square approach along with the chosen attributes. Chi-square was used in our first, second, and third approaches.

```
target              0.000000e+00
oldpeak             0.000000e+00
trestbps            0.000000e+00
age                 0.000000e+00
BMI                 0.000000e+00
sysBP               0.000000e+00
totChol             0.000000e+00
exang               0.000000e+00
cp                  0.000000e+00
thalach             0.000000e+00
ca                  0.000000e+00
cigsPerDay         1.538577e-222
slope              9.415282e-180
diaBP              2.013805e-125
thal               6.115064e-125
restecg             3.611733e-48
prevalentHyp        5.697899e-21
education           4.247250e-14
glucose             9.188459e-13
currentSmoker       9.497780e-11
BPMeds              1.710507e-08
diabetes            5.972721e-06
heartRate           6.280550e-03
prevalentStroke     1.472506e-01
fbs                 4.357430e-01
male                5.118887e-01
dtype: float64
```

Figure 4. 4 f-scores using Fisher score

Here we can see f-scores after applying the feature selection technique fisher score. We selected the top 15 features for our 3rd and 4th approaches by setting scores equal to zero to give us the maximum likelihood estimate of the Parameter as the Fisher score is a gradient of the log-likelihood function.

We generated our final dataset, which will be used for training and testing the models, using the features chosen by the chi-square and fisher score.

4.4 COMPARISON BETWEEN DIFFERENT MODELS BASED ON EVALUATION METRICS

Several matrices, including the Confusion Matrix, Classification Accuracy, Area under Curve (AUC), and Cross-validation score, are taken into account while assessing the performance of any algorithm.

4.4.1 CONFUSION MATRIX

Confusion matrices show counts between expected and actual values. The confusion matrix not only explains errors the classifier is making, but also the types of errors that are being made. Confusion metrics are illustrated below for the five proposed approaches where the number of classifiers used are seven namely Adaboost, Gradient boosting, KNN, Logistic Regression, Naive Bayes, SVM, and Xgboost. Therefore, the correct predictions and the types of errors being made by each classifier are shown in the figures for the corresponding five approaches.

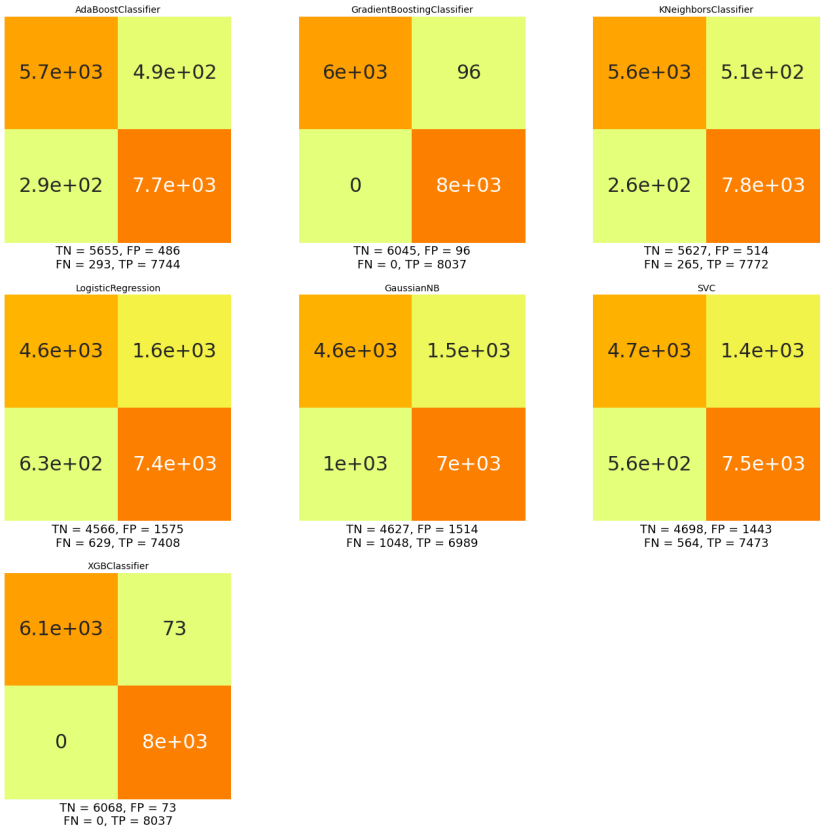

Figure 4. 5 Confusion matrix of 1st approach

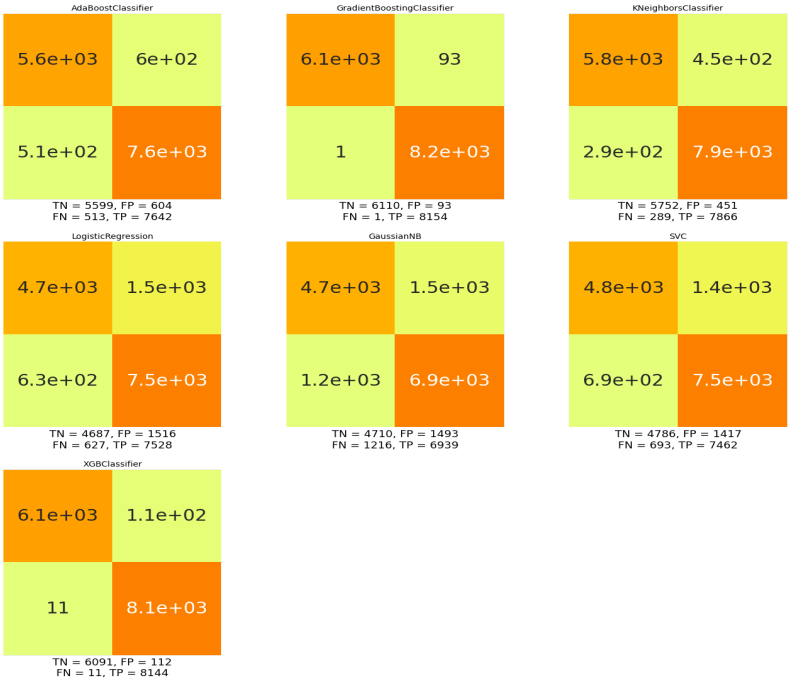

Figure 4. 6 Confusion matrix of 2nd approach

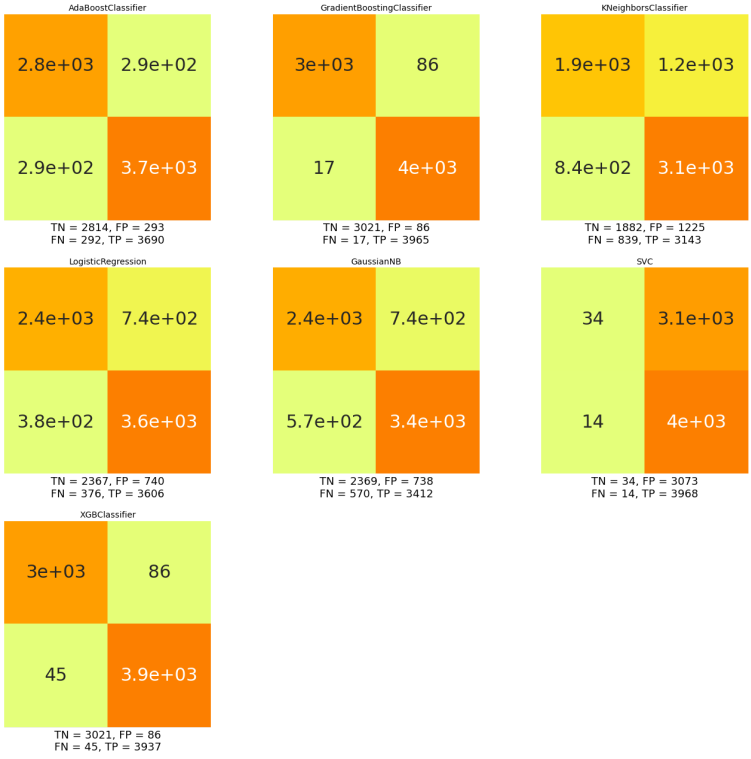

Figure 4. 7 Confusion matrix of 3rd approach

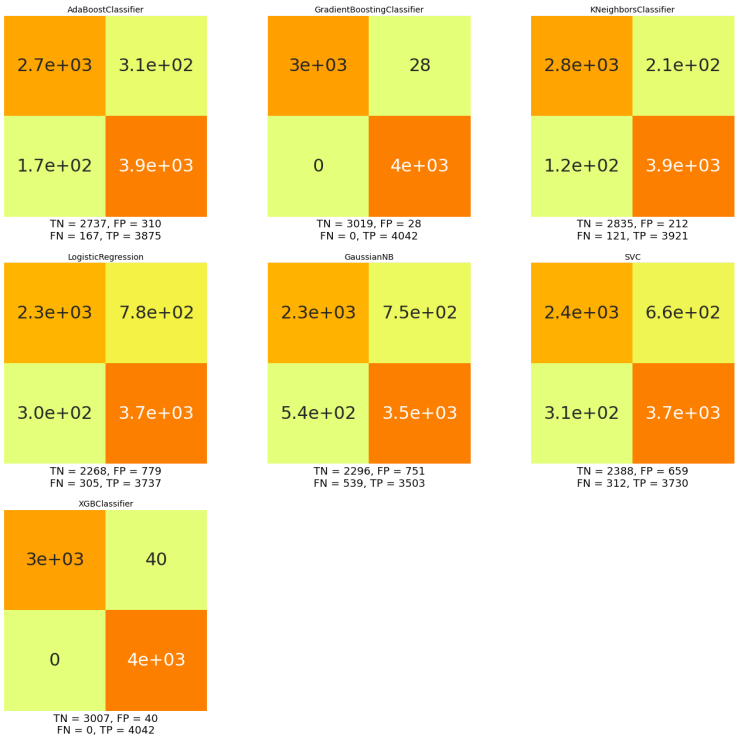

Figure 4. 8 Confusion matrix of 4th approach

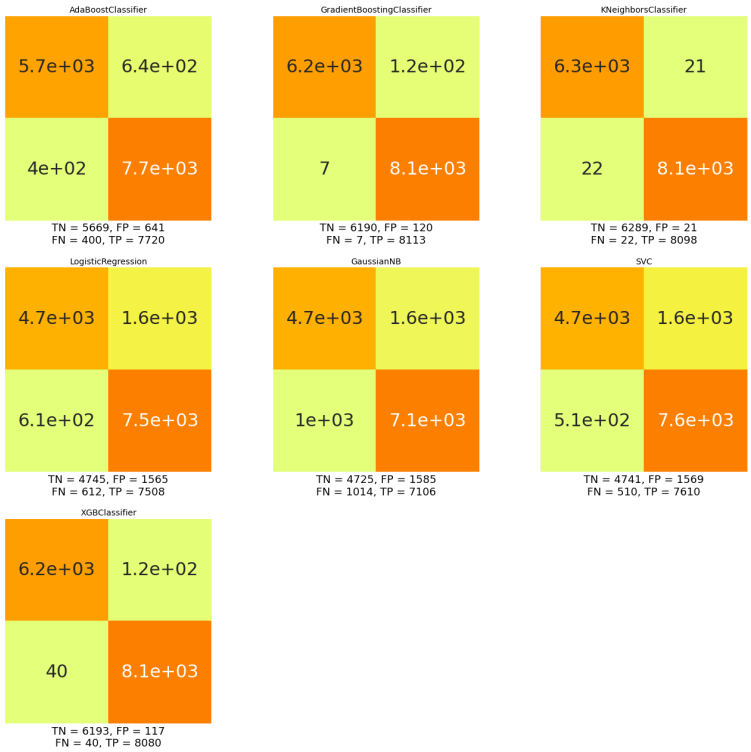

Figure 4. 9 Confusion matrix of 5th approach

From the above five approaches, for the **first approach** Gradient boosting and Xgboost give the most correct predictions that correctly classified TP and TN values are higher whereas wrongly classified FP and FN values are minimal. Similarly for the **second, third, fourth, and fifth approaches** but again for the fifth approach, K-Nearest Neighbor also gives the most correct predictions and least wrong predictions.

4.4.2 AUC Score and ROC Curve

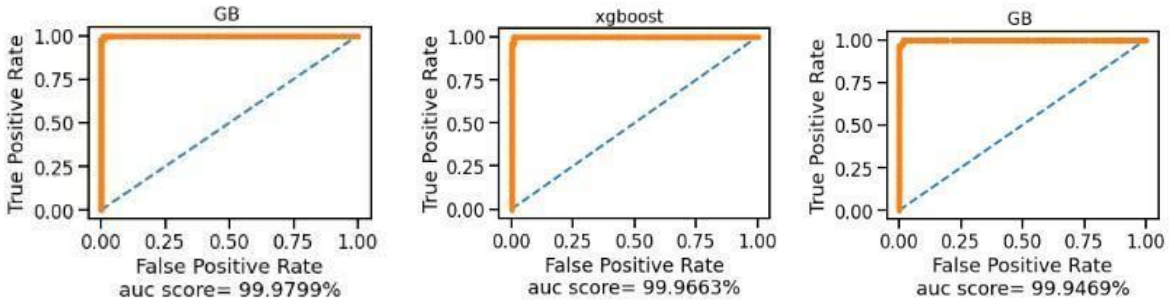

Figure 4. 10 Highest AUC scores of 1st, 2nd & 3rd approach

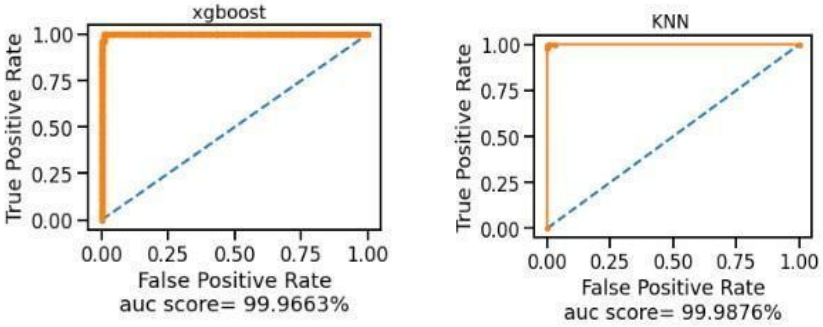

Figure 4. 11 Highest AUC scores of 4th & 5th approach

The capacity of a classifier to distinguish between classes is evaluated by the Area Under the Curve (AUC), which behaves as a summary of the ROC curve. With better AUC, the model does a better job of distinguishing between the desirable and undesirable groups. We have displayed each approach's highest AUC score. Here, we can observe that in the first and third approaches, GB has the highest AUC score. XGB, however, has the highest AUC score in the second and fourth approaches. Finally, the fifth approach has the highest AUC score for KNN.

4.4.3 Cross Validation

Table 4. 2 Cross-Validation scores

Approaches	ADAB	GB	KNN	LR	NB	SVM	XGB
1st Approach	0.935593	0.993545	0.938983	0.845786	0.813374	0.849872	0.990852
2nd Approach	0.930126	0.994263	0.939296	0.843090	0.814313	0.856820	0.993276
3rd Approach	0.908944	0.985224	0.709621	0.845747	0.818134	0.565947	0.979757
4th Approach	0.931266	0.994498	0.949640	0.846170	0.815630	0.857244	0.993511
5th Approach	0.927506	0.988818	0.996211	0.851684	0.824701	0.851499	0.990205

When a model is created on a portion of input data and then assessed on a small subsection of raw data that has never been utilized earlier, this procedure is referred to as cross-validation. For the 1st approach, we got the highest score on GB, 99.35%, and the lowest on NB, 81.33%. In the 2nd and 3rd approaches, we got the highest score on GB, 99.42% and 98.52% respectively. NB and SVC gave the lowest scores on the 2nd and 3rd approaches respectively. Lastly, in the 4th and 5th approaches, we got the highest score on GB and KNN.

4.5 PREDICTION ACCURACY

Table 4. 3 Accuracy comparison table of models

Approaches	ADAB	GB	KNN	LR	NB	SVM	XGB
1st Approach	92.22%	99.35%	94.85%	85.07%	81.13%	85.30%	99.14%
2nd Approach	94.51%	99.32%	94.51%	84.45%	81.93%	85.84%	99.49%
3rd Approach	91.75%	98.55%	70.88%	84.26%	81.55%	56.45%	98.15%
4th Approach	93.27%	99.61%	95.30%	84.71%	81.80%	86.30%	99.44%
5th Approach	97.79%	99.12%	99.70%	84.91%	81.99%	85.59%	98.91%

The above table provides comprehensive information on the accuracy results obtained using different methods. In the 1st approach, we got a higher accuracy on GradientBoostingClassifier, 99.35%. 99.49% of the XGB Classifier was the highest accuracy in the 2nd approach. In the 3rd and 4th approaches, we got higher accuracy on GradientBoostingClassifier which were 98.55% and 99.61%. Lastly, we got higher accuracy on KNN Classifier which was 99.70%.

It can be seen that the gradient boosting technique achieved remarkably high accuracy.

4.6 COMPARISON WITH PREVIOUS WORK

Figure 4. 12 Comparison with EXISTING WORKS

Research	Accuracy	Classifier
Bharti et al [2]	84.8%	KNN
Md. Nayeem et al [15]	87.36%	
Our 5th Approach	99.70%	

Research	Accuracy	Classifier
Bharti et al [2]	88.89%	GB
Ch. Anwar ul Hassanet et al [13]	95.83%	
Our 4th Approach	99.61%	

Research	Accuracy	Classifier
Bharti et al [2]	83.2%	SVM
Ch. Anwar ul Hassanet et al [13]	84.97%	
Our 4th Approach	86.30%	

Research	Accuracy	Classifier
Bharti et al [2]	71.4%	XGB
Md. Nayeem et al [15]	92.37%	
Ch. Anwar ul Hassanet et al [13]	88.25%	
Louridi et al [6]	92.37	
Our 2nd Approach	99.49%	

Research	Accuracy	Classifier
Louridi et al [6]	91.66%	ADAB
Our 5th Approach	97.79%	

These tables compare the results of our five approaches to the earlier research. We used seven machine learning classifiers, with five of them providing improved accuracy. These include KNN, GB, SVM, XGB, NB, LR, and ADAB. Our fifth approach to the KNN classifier had the highest prediction accuracy, at 99.70%, compared to 84.8% and 87.36% for Bharti et al [2] and

Md Nayeem et al [15]. Bharti et al [2] and Ch. Anwar ul Hassan et al[13] respectively achieved 88.89% and 95.83% in the GB classifier. On the other hand, the fourth approach, out of the five, had the highest accuracy, at 99.61%. The accuracy of our fourth approach was at its highest, 86.30%, whereas that of the other two current approaches was slightly lower, at 83.2% and 84.97%. Our fifth strategy for ADAB had the best accuracy, at 97.79%. Finally, among Bharti et al. [2], Md Nayeem et al. [15], Ch. Anwar ul Hassan et al. [13], and Louridi et al. [6], our second technique outperforms them all in XGB.

4.7 DISCUSSION

Increase in the prediction accuracy is the key task of our approach. We conducted the experiment to identify the algorithm that has the greatest impact on the increase in prediction accuracy. We worked with five different supervised models using unique missing value imputation techniques with feature selection methods, Fisher score & Chi-square. Fisher score & Chi-square. The machine learning algorithms and ensemble learning methods are applied to selected features. Furthermore, we compared the performance of different machine learning algorithms and ensemble learning methods. Models are assessed using a variety of techniques, including accuracy, confusion matrix, recall, precision, F-measure, cross validation, and ROC analysis. Finally, our five models achieved higher prediction accuracy than other previous works. Overall, the machine learning algorithms showed good performance on average of 80% and above. On the other hand ensemble learning methods showed very good performance on average of 90% and above.

4.8 SUMMARY

Beginning with an explanation of how the training and testing datasets were handled and what kind of process was utilized for such experimental analysis, the contents of this chapter have been organized in a way that covers the process from start to finish. We talked about the dataset, dataset preparation, as well as the experimental outcomes of our suggested approach.

CHAPTER V

CONCLUSION AND FUTURE WORKS

5.1 CONCLUSION

Using seven ML classification modeling strategies, cardiovascular disease prediction models have been created. The following techniques were used to build the proposed models: gradient boosting, adaboost, Xgboost, Naive bayes, SVM, KNN, and logistic regression. The three ensemble classifiers such as Adaboost, Xgboost, and Gradient boosting gave high levels of accuracy like 97.79%, 99.49%, and 99.61% respectively. Moreover, KNN also gave a high level of accuracy which is 99.70%. More training data increases the possibility that the model will accurately determine whether or not a person has a cardiovascular abnormality. We can predict the patient quickly and more accurately by applying these computer-aided procedures, and the cost can be greatly decreased. Since machine learning algorithms are more robust and can make forecasts better than humans, they enable us to deal with a variety of medical databases, which is advantageous for both patients and medical practitioners. As a result, this project assists us in forecasting the patients who will be diagnosed with heart diseases by cleaning the dataset and applying ML classifiers, and to get an accuracy of an average of 80% and above on our five proposed approaches better than the previous works.

5.2 CONTRIBUTION OF THE THESIS

Utilizing widely used and rarely used feature selection approaches, to increase the prediction accuracy is the research's main contribution.
We used different missing value imputation methods namely MICE, Simple imputer, and median to improve the classifiers' accuracy.
In this study, we used five different approaches and compared the prediction accuracy of each approach.

5.3 LIMITATION

While the research work’s main focus is on the most precise heart disease diagnosis, other models can be compared in order for the timely identification of heart disease accurately.

5.4 FUTURE WORK

Going forward, to make the model more adaptable to various feature selection algorithms and more resilient to datasets with sizable amounts of missing data, we aim to further generalize it. Another potential strategy is to use Deep Learning algorithms. This study's primary objective was to advance previous work by creating the model in a creative and original approach, as well as to make the model practical and simple to apply in real-world scenarios.

REFERENCES

[1] https://www.ncbi.nlm.nih.gov/books/NBK45688/ 22 January 2023

[2] Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S. and Singh, P., 2021. Prediction of heart disease using a combination of machine learning and deep learning. *Computational intelligence and neuroscience*, *2021*.

[3] https://www.cdc.gov/heartdisease/facts.htm October 14, 2022

[4]https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) 11 June 2021

[5] Folsom A R, Prineas R J, Kaye S A & Soler J T (1989). Body fat distribution and self-reported prevalence of hypertension, heart attack, and other heart disease in older women. International journal of epidemiology, 18(2), 361-7.

[6] Louridi, N., Douzi, S. and El Ouahidi, B., 2021. Machine learning-based identification of patients with a cardiovascular defect. *Journal of Big Data*, *8*(1), pp.1-15.

[7] Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M.F. and Ullah, N., 2022. A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms. *Mobile Information Systems*, *2022*.

[8] Jindal, H., Agrawal, S., Khera, R., Jain, R. and Nagrath, P., 2021. Heart disease prediction using machine learning algorithms. In *IOP conference series: materials science and engineering*

[9] Gubbala, S.B., 2022. Heart Disease Prediction Using Machine Learning Techniques.

[10]S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques", Proc. IEEE/ACS Int. Conf. Computer. Syst. Appl., pp. 108-115, Mar. 2008.

[11] Shrivastava, R., Vigneshwaran, P. and Soni, A.K., 2022. Heart Disease Prediction Analysis Through Ensemble Learning Model. In *Proceedings of International Conference on Communication and Artificial Intelligence* (pp. 449-458). Springer, Singapore.

[12] Jabbar, M.A., Deekshatulu, B.L. and Chandra, P., 2016. Intelligent heart disease prediction system using random forest and evolutionary approach. *Journal of network and innovative computing*, *4*(2016), pp.175-184.

[13] Hassan, C.A.U., Iqbal, J., Irfan, R., Hussain, S., Algarni, A.D., Bukhari, S.S.H., Alturki, N. and Ullah, S.S., 2022. Effectively Predicting the Presence of Coronary Heart Disease Using

Machine Learning Classifiers. *Sensors*, *22*(19), p.7227.

[14] Karthick, K., Aruna, S.K., Samikannu, R., Kuppusamy, R., Teekaraman, Y. and Thelkar, A.R., 2022. Implementation of a Heart Disease Risk Prediction Model Using Machine Learning. *Computational and Mathematical Methods in Medicine*, *2022*.

[15] Nayeem, M.J.N., Rana, S. and Islam, M.R., 2022. Prediction of Heart Disease Using Machine Learning Algorithms. *European Journal of Artificial Intelligence and Machine Learning*, *1*(3), pp.22-26.

[16] Chen, J.I.Z. and Hengjinda, P., 2021. Early prediction of coronary artery disease (CAD) by machine learning method-a comparative study. *Journal of Artificial Intelligence*, *3*(01), pp.17-33.

[17] Yahaya, L., Oye, N.D. and Garba, E.J., 2020. A comprehensive review on heart disease prediction using data mining and machine learning techniques. *American Journal of Artificial Intelligence*, *4*(1), pp.20-29.

[18] Nissa, N., Jamwal, S. and Mohammad, S., 2021. Heart Disease Prediction using Machine Learning Techniques. *Wesleyan Journal of Research*, *13*(67).

[19] Salhi, D.E., Tari, A. and Kechadi, M., 2020, December. Using machine learning for heart disease prediction. In the International *Conference on Computing Systems and Applications* (pp. 70-81). Springer, Cham.

[20] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F.J.M., Ignatious, E., Shultana, S., Beeravolu, A.R. and De Boer, F., 2021. Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. *IEEE Access*, *9*, pp.19304-19326.

[21] Shetgaonkar, P. and Aswale, S., Heart Disease Prediction using Data Mining Techniques.

[22] Pal, M., Parija, S., Panda, G., Dhama, K. and Mohapatra, R.K., 2022. Risk prediction of cardiovascular disease using machine learning classifiers. *Open Medicine*, *17*(1), pp.1100-1113.

[23] Saqlain, S.M., Sher, M., Shah, F.A., Khan, I., Ashraf, M.U., Awais, M. and Ghani, A., 2019. Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. *Knowledge and Information Systems*, *58*(1), pp.139-167.

[24] Yazdani, A., Varathan, K.D., Chiam, Y.K., Malik, A.W. and Wan Ahmad, W.A., 2021. A novel approach for heart disease prediction using strength scores with significant predictors. *BMC medical informatics and decision making*, *21*(1), pp.1-16.

[25] Alsaffar, M., Alshammari, A., Alshammari, G., Aljaloud, S., Almurayziq, T.S., Abdoun, F.M. and Abebaw, S., 2021. Machine learning for ischemic heart disease diagnosis aided by evolutionary computing. *Applied Bionics and Biomechanics*, *2021*.

[26] Dhande, B., Bamble, K., Chavan, S. and Maktum, T., 2022. Diabetes & Heart Disease Prediction Using Machine Learning. In *ITM Web of Conferences* (Vol. 44, p. 03057). EDP Sciences. (Vol. 1022, No. 1, p. 012072). IOP Publishing.

[27] Anbuselvan, P., 2020. Heart disease prediction using machine learning techniques. *Int. J. Eng. Res. Technol*, *9*, pp.515-518.

[28] Nikhar, S. and Karandikar, A.M., 2016. Prediction of heart disease using machine learning algorithms. *International Journal of Advanced Engineering, Management and Science*, *2*(6), p.239484.

[29] Vijayashree, J. and SrimanNarayanaIyengar, N.C., 2016. Heart disease prediction system using data mining and hybrid intelligent techniques: A review. *International Journal of Bio-Science and Bio-Technology*, *8*(4), pp.139-148.