Tilburg University

A COMPARATIVE STUDY OF MACHINE LEARNING TECHNIQUES IN PREDICTING HEART DISEASE FOR THE USA RESIDENTS

JESUTOMIWO OGUNLERE

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

NUMBER

U155987

COMMITTEE

dr. Görkem Saygili
dr. Richard Starmans

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

Word count : 8,608

DATE

June 30, 2022

A COMPARATIVE STUDY OF MACHINE LEARNING TECHNIQUES IN PREDICTING HEART DISEASE FOR THE USA RESIDENTS

JESUTOMIWO OGUNLERE

CONTENTS

Abstract

The World Health Organization (WHO) asserts that heart disease is a disorder of the heart and blood vessels and the leading global cause of death. According to the Centers for Disease Control and Prevention (CDC), 659,000 people die yearly from heart disease in the USA. Death is recorded every 36 seconds due to heart disease in the USA. If heart diseases are discovered early, one in every three deaths related to heart disease is preventable. However, the traditional techniques deployed to detect heart diseases, such as medical diagnosis, scans, and sensors, are expensive, time-consuming, and prone to misdiagnosis. Although data mining has emerged as an alternative technique to diagnose heart disease, its predictive performance could be improved. Accurate and early prediction of heart disease would lead to early detection, preventive management, and fewer fatalities in the USA This study explores the information gain, XGBoost feature importance method, and the integration of machine learning models to enhance predictive performance in heart disease classification for USA residents. Random Forest, XGBoost, and Multilayer Perceptron were integrated with hard and soft voting techniques on a USA dataset released by CDC. The data was massively imbalanced. However, the data imbalance was addressed by oversampling, under-sampling, and SMOTE sampling. The main finding indicated that general health conditions, sex, and age are significant factors for heart disease among US residents. The soft voting classifier, with random oversampling technique, predicts better than other classifiers with 74.57% accuracy, 78.29% recall, 22.53% precision, 35.00% f1 score, and 76.25% AUCROC.

Keywords: Feature selection, Hard voting, Heart disease, Information Gain, Machine learning, Multilayer Perceptron, Oversampling, Random Forest, SMOTE, Soft voting, Undersampling, XGBoost.

data source, code & ethics statement

Work on this thesis did not involve data collection from human participants or animals. I fully acknowledge that I have no legal claim to this data. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. Moreover, the author is the owner of the code used for this study.

1 INTRODUCTION

The World Health Organization (WHO, 2021) described cardiovascular diseases (CVDs) as disorders of the heart and blood vessels. Heart diseases are the foremost cause of death worldwide, with CVDs amassing 32% of international deaths in 2019. Therefore, it is essential to prematurely detect cardiovascular ailments as this could help monitor the disease as early as possible with medications and counseling (WHO, 2021). In support of the above study, the Centers for Disease Control and Prevention (CDC) asserted heart disease to be a significant cause of death for most tribes and races in the United States of America. Furthermore, the research informed that heart disease in the United States has a high fatality rate of one out of every four (25%) deaths. The fatality amasses 659,000 deaths yearly, despite annual spending of over 363 Billion dollars to eradicate the disease (CDC, 2021).

Medical diagnoses have been the dominant approach to examining heart disease. However, it is time-consuming and arduous to discover the abnormalities in the sensors and scans (Ali et al., 2020). Moreover, though it is pertinent to timely predict heart disease, Although it is pertinent to timely predict heart disease, it is a very complex process that necessitates the aid machine and advanced technology (Islam, Rafa, & Kibria, 2020). In support of Islam et al. (2020), Erdoğan and Güney (2020) affirmed that medical diagnosis of heart disease is a complicated task because the diagnosis can be mistaken for other ailments that show similar symptoms like nausea, chest pains, and shortness of breath.

It is occasionally impossible to save a life when much damage has been done to the vital organs. These damages occur when patients could not process any significant symptom of the ailment early, or the disease was not detected (Khan & Mondal, 2020). According to CDC (2019), various factors such as age, family history, health condition, and general lifestyle may increase the risk for heart disease; they referred to these as the risk factors. About 47% of all Americans have at least one of three key factors for heart disease mainly; high blood pressure (HBP), smoking, and high cholesterol (CDC, 2021), these were further supported in studies (Alalawi & Manal, 2021), (WHO, 2021). Other factors such as alcohol consumption, lack of physical exercise, and unhealthy diet have additionally been affirmed to be risk factors for heart disease diagnosis (CDC, 2021; Srinivas & Katarya, 2022; WHO, 2021). In line with this information, the data science approach solves many issues on early diagnosis in medicine (Bashir, Khan, Khan, Anjum, & Bashir, 2019). Machine learning techniques can be used to identify heart disease symptoms and predict future heart disease occurrences. Therefore, machine learning would be beneficial in reducing cases and casualties of the disease.

An increasing number of people are getting infected, but there remain minimal ways to effectively detect heart disease from basic information (Nazri, Das, & Promi, 2021). Islam et al. (2020) purported that an abundance of heart disease data is being generated in healthcare sectors. However, these data have not been well utilized to determine the concealed knowledge for efficient decision-making. These data can be used to obtain valuable insights by implementing machine learning algorithms because the data can be too complex for human understanding. Still, machine learning algorithms are suitable for finding the patterns and hidden details in the datasets that could help diagnosis and decision-making. Researchers have attempted different techniques to detect heart diseases early. Still, there is no silver bullet to this problem yet, leaving much room for exploration in finding a solution to detecting heart diseases.

1.1 *Scientific and Societal Relevance*

Machine learning is an innovative technology with a high potential to substantially impact the health sector, particularly in the early detection of heart diseases (Ed-Daoudy & Maalmi, 2019). Machine learning is used to predict heart disease in this project, specifically coronary heart disease (CHD) and myocardial infarction (MI) or heart attack. The CDC (2021) propounded CHD is the most common heart disease in the USA, and at least one person gets affected by Mi every 40 seconds in the US.This study proposes a technique for detecting coronary heart disease and myocardial infarction in the US and research the most recent data from CDC. The proposed method will help create a new baseline for predicting heart diseases for USA residents. Other than for the United States residents, it would benefit the health sector and other heart disease analyses. This study would extend the findings of existing studies on heart diseases. This analysis will be helpful for health practitioners, health awareness organizations, government health policies, and all humans.

Researchers have explored several machine learning methods for predicting heart disease. Nevertheless, predictive performance requires improvement (Bashir et al., 2019). A poor diagnosis can result in giving the patient the wrong treatment, which can have severe consequences. The proper treatment can be administered by improving the quality of heart disease diagnosis, leading to recovery or better health conditions (Zunaidi et al., 2018). In other to enhance the performance of machine learning models in predicting heart disease, it has been suggested to combine models with a voting technique for better performance (Khan & Mondal, 2020; Raza, 2019).

1.2 *Research Questions*

This study intends to examine if assembling algorithms will lead to a higher predictive power than single algorithms on US data. The voting classifier will aggregate the results from the combined models to make a final classification. For instance, this can be seen from a democratic point of view, such as taking the majority outcome class from the classifiers' predictions to decide the final class of the instances. This brings the research question of this thesis:

Would the hybrid model consisting of Random Forest, XGBoost, and Multilayer Perceptron, combined with the soft voting technique, predict heart disease better than each of the individual classifiers for the United States residents?

The hybrid model is a fusion of distinct models that are expected to conjointly predict better than any of the individual constituents (Ali et al., 2020). It is expected that the predictive performance of the individual models will be enhanced when combined with a voting technique to predict heart disease for the residents of the USA. Subsequently, this leads to more accurate diagnoses of heart disease and a higher rate of early detection.

Before heart disease is detected in the United States, it is essential to know the significant factors that will help identify heart disease among United States residents. This leads to the first sub-research question:

RQ1 ***What are the key factors of heart diseases in the United States?***

The outcome of this research will help identify the early symptoms for diagnosing heart diseases in the USA. In addition, the result will help to determine the vital features for predicting heart disease in the main research question and ensuing sub-question. These symptoms identified would potentially lead to heart disease, and knowing them will help reduce or prevent heart diseases in the USA.

After finding the factors of heart disease in US data, it is essential to know the voting technique that will achieve the best result on United States data. A hard vote takes the majority predicted class as the final output. In contrast, a soft voting technique takes the average probability values of predictions to decide the final class. This leads to the second sub-research question:

RQ2 ***Does the hard voting technique predict heart disease better than the soft voting for United States residents?***

Identifying the better voting technique between soft and hard voting techniques will help in solving the main research question, and the result will be the proposed voting technique for future works in predicting heart disease for US residents.

2 LITERATURE REVIEW

The advancement of artificial intelligence has led to machine learning being explored as a feasible data-driven approach to detecting heart disease. This section will highlight the findings, proposals, and analyses of what has been done to detect the presence of heart disease. In the context of this study, Coronary heart disease and heart attack or Myocardial Infarction (MI) are studied and referred to as heart disease.

2.1 *Detecting heart disease*

Machine learning classification is associated with many attributes; the significance level varies for the attributes in a classification task (Omuya, Okeyo, & Kimwele, 2021). Moreover, all data variables are not created equally, nor do all variables have the same contribution to the dependent variable(Jadhav, He, & Jenkins, 2018). In furtherance, because medical data often include various features, it is crucial to identify the important factors that will aid the diagnoses of heart disease (Amin, Chiam, & Varathan, 2019).

A study that put forward an approach for discovering the significant factors when predicting heart attacks was by Ramesh, Madhavi, Reddy, Somasekar, and Tan (2021). To accurately predict heart attack, the authors emphasized that it is essential to develop mechanisms to identify the valuable variables that lead to a heart attack. They proposed the information gain-based (IGFS) method to detect the vital variables. Sex, the maximum heart rate achieved by patients, chest pain (angina), and fasting blood sugar were the factors their study found to be associated with a heart attack. Support Vector Machine and Random Forest performed best with an equal accuracy of 88% when IG was implemented in their study. From this study, it can be denoted that IG could be used to discern the influential factors of heart disease.

Furthermore, the same technique was explored in another survey executed by Ali et al. (2020). The information gain (IG) technique removed irrelevant attributes in their study. They predicted heart disease with ensembled deep learning algorithms consisting of feed-forward Neural Networks with backpropagation. The information gain (IG) technique distinguished the relevant attributes in their study. The researchers predicted heart disease with ensembled deep learning algorithms consisting of feed-forward Neural Networks with backpropagation. The authors predicted with essential factors such as blood sugar, cholesterol, blood pressure, oxygen saturation, clinical examination, physical activities, heart rate, respiration rate, smoking, diabetes, and medical history.

Aside from applying statistical methods such as the aforementioned reviewed works, researchers are examining the importance of features on an algorithmic level. For example, Alam, Rahman, and Rahman (2019) did a study on ten datasets, including three heart disease datasets. The authors used the Random Forest algorithm on each dataset to rank features according to their importance. In a similar study, Zafari, Langlois, Zulkernine, Kosowan, and Singer (2022) predicted chronic obstructive pulmonary disease (COPD) with Multilayer Neural Network and XGBoost models. The XGB performed best in their research. After prediction, the scholars used the XGB to examine the vital role of features in model performance. The study found characteristics such as health conditions and age as highly important.

The reviewed studies did not explore different methods to verify feature importance. The authors did not assess each factor's attributed rank or importance score through both statistical and algorithmic evaluation methods. In addition, the data used to implement the technique in the reviewed studies differ from the data used in this study. For instance, Ali et al. (2020) researched using non-structured data from sensors and scans. Also, Ramesh et al. (2021) and Alam et al. (2019) carried out their research on the aged data, with the most recent data in the studies published over 21 years (2001) and the highest instance in all is about 300 records. However, for this study, a recent structured textual data gathered in 2020 with a substantial volume of about 300,000 records would be used to predict heart disease.

Additionally, these studies revealed that information gain (IG) could successfully extract influential features when predicting heart disease. Equally, feature importance could be derived from algorithmic performance. Contrarily, limited studies have given attention to the essential attributes in predicting heart disease (Amin et al., 2019). The IG method and algorithmic feature importance could be applied to USA data. Therefore, this presents an area of research to determine the main factors of heart disease in the USA using the IG and model feature importance methods. In addition, the valuable factors of heart disease in the existing studies will serve as a baseline for comparing the influential factors of this study.

2.2 *Heart disease prediction*

Researchers and professionals have been exploring various techniques to predict heart disease in patients. Previous studies have found heart disease diagnosis based on machine learning to be a flexible and cheap approach, and it helps to enhance data-driven decisions (Ahsan & Siddique, 2022).

Chang, Bhavani, Xu, and Hossain (2022) built a health monitoring application system that uses machine learning to predict the chances of humans developing heart disease. Random Forest, the highest of their framework, had an accuracy of 82.19% . This study further inferred that machine learning has the potential to detect heart diseases. A similar outcome was observed when Rajdhan, Agarwal, Sai, Ravi, and Ghuli (2020) designed a machine learning model that compared four algorithms on heart disease classification. Random Forest had the best performance with an accuracy of 90.16% . It can be deduced from these studies that Random Forest can predict heart disease with a high level of performance.

A contrasting finding has been observed in other studies. Another algorithm that has performed well in previous research is the XGBoost. Rezaei, Woodward, Ramírez, and Munroe (2021) predicted heart arrhythmias with XGBoost attaining 87.22% f1 score, 88.55% sensitivity, and 85.95% specificity. XGBoost was compared to Random Forest and Extra Tree (ET) classifiers (Budholiya, Shrivastava, & Sharma, 2020). After evaluating the models with five different metrics, including accuracy, f1 score, and AUC (Area Under Curve), it was observed that the fine-tuned XGBoost surpassed Random Forest. Based on the above studies, it can be argued that XGBoost can successfully predict heart disease.

A new research area has been ventured into when predicting heart disease. With the development and successes made by deep learning algorithms in areas like image analysis and medical imaging, researchers are exploring the domain of deep learning to predict heart diseases. An artificial Neural Network is a recent machine learning model that has been successful in heart disease studies (Dominguez-Morales, Jimenez-Fernandez, Dominguez-Morales, & Jimenez-Moreno, 2017; Kumar & Kumar, 2021; Sarmah, 2020).

Waqar et al. (2021) compared Artificial Neural Network to Random Forest, Naïve Bayes, Logistic Regression, Support Vector Machine, and KNN to predict heart attack on a SMOTE data. The study results indicated that Artificial Neural Networks outperformed other classifiers with an accuracy of 100. In another research on coronary heart disease (CHD) prediction, Ahmadi, Weckman, and Masel (2018) found Multilayer Perceptron as the most suitable algorithm after comparing it with C5.0 (Decision Tree). The findings of these studies help to note the potential of Artificial Neural Network in predicting heart disease.

In conclusion, an insight from the studies in this section is that when the prediction of heart disease with machine learning models is examined, their performances might vary across different studies. The same algorithms may not perform consistently in a different study or on separate data. A study that highlights this phenomenon is Alalawi and Manal (2021). The authors designed a survey to predict heart disease on two datasets using the same machine learning and deep learning models. The models were evaluated based on recall, precision, f-score, and accuracy. With 94% accuracy, Random Forest outperformed the XGBoost, Neural Network, and other classifiers on the first dataset. In contrast, Gradient Boosting performed best on the second dataset with 73% accuracy.

Still, there has not been an established algorithm that constantly supersedes other algorithms in predicting heart disease. Moreover, studies have not identified the algorithm that best predicts heart disease for USA residents. Therefore, this work aims to discover the best algorithm to predict heart disease on USA data.

2.3 *Alternate approach to predicting heart disease*

Several works have been carried out to predict heart disease, and different levels of performance metrics have been attained (Ahsan & Siddique, 2022). Many researchers continuously work towards better accuracy when predicting heart disease; recent studies have started integrating algorithms to improve model performance (Ansari, Alankar, & Kaur, 2020).

These assertions were supported by Tama, Im, and Lee (2020) when the scholars aimed to predict coronary heart disease With a hard voting technique, the ensemble of Random Forest, XGBoost, and Gradient Boosting Machine outperformed the individual classifiers in the experiment. Another study that strengthened the previous postulations was Raza (2019), the author predicted heart disease by combining Logistic Regression, Multilayer Perceptron, and Naïve Bayes algorithms with hard voting. The ensemble method in their study achieved an accuracy of 88.88%, which was a higher accuracy when the results were compared to using lone classifiers. These studies imply that hard combining models with hard voting has successfully predicted heart disease better than single models in some research.

Besides hard voting, soft voting is another voting system that has been explored to classify heart disease in people. The soft voting model was more successful than the individual algorithms with an accuracy of 90.93% when gradient boosting, Random Forest, and Extra Trees algorithms were compared to the ensemble (Sherazi, Bae, & Lee, 2021). This study conveys the prospect of soft voting to predict heart disease better than solitary algorithms.

Given the above studies, it can be argued that an ensemble of algorithms combined with a voting technique has the potential to diagnose heart diseases. Thus, this could be extended to heart disease prediction on US data. However, the limitation of the above studies was the choice of the voting system, which was mainly predetermined. Further, the authors did not state the motivation behind the chosen voting strategy. Nor do the studies compare different voting methods to decide which would lead to a better predictive performance in their studies. Equally important, there have not been many studies that predicted heart disease with soft voting. Therefore, this study will compare the two proposed voting strategies to determine the better voting strategy for predicting heart disease for the United States inhabitants.

3 METHOD

This section is structured to explain, justify and give the motivation behind each method adapted in this study. As well, this section gives insight into the experimental procedure and evaluation metrics.

3.1 *Feature importance*

The Information Gain (IG) is used to identify the influential attributes of heart disease in this research. IG is used to evaluate and score features based on the information the features provide to differentiate between the heart disease and no heart disease classes in the dataset. IG has a good reputation for selecting the vital features for heart disease classification (Amin et al., 2019; Nandhini & Tamilselvi, 2020; Ramesh et al., 2021). IG is more beneficial than other feature ranking methods because it quantifies and accounts for feature importance by returning the scores and ranking of all features. Other methods return either ranking or subsets of features. Moreover, it is independent of any model used for prediction (Jadhav et al., 2018). The formula for calculating information gain is presented below:

$$IG(H, f) = E(H) - E(H|f) \tag{1}$$

Equation (1) gives the mutual information between the heart disease variable (H) and the feature being examined (f). $E(H)$ is the entropy of the heart disease category, and $E(H|F)$ is the entropy of the heart disease category given a feature in the heart disease dataset. If there is no dependency between them, the information gain will be 0. If the score is greater than 0, the feature can provide some information about the heart disease category; thus, there is dependency (Jamei et al., 2022) IG values vary from 0 to 0.5 in most applications (Bhat & Dutta, 2021).

3.2 *XGBoost (XGB)*

XGB is a tree-boosting algorithm that transforms features to build a tree as a week learner. It is an iterative process of sequentially adding and updating the weak learners (trees) to form a more robust model as an ensemble. The misclassification made during a prediction quantifies the loss function of the next prediction. The newly added tree helps to increase the predictive potential of the model by learning and reducing the loss function of its predecessor.

XGBoost has a high reputation in studies compared to other algorithms due to its ability to exceedingly boost the performance of the model in terms of result, computational speed, and model efficiency; this is achieved by parallel and distributed processing to reduce computing time (Srinivas & Katarya, 2022). Nevertheless, XGB is highly sensitive to hyper-parameters Budholiya et al. (2020); Pan, Zheng, Guo, and Luo (2022). The XGBoost is expected to perform well in predicting heart disease for the US inhabitants because it uses a highly accurate estimation of the learning objective. More so, the data used for this experiment is vastly imbalanced, which increases the possibility of overfitting, howbeit XGB enhances performance and generalizability by using a regularization term to prevent overfitting.

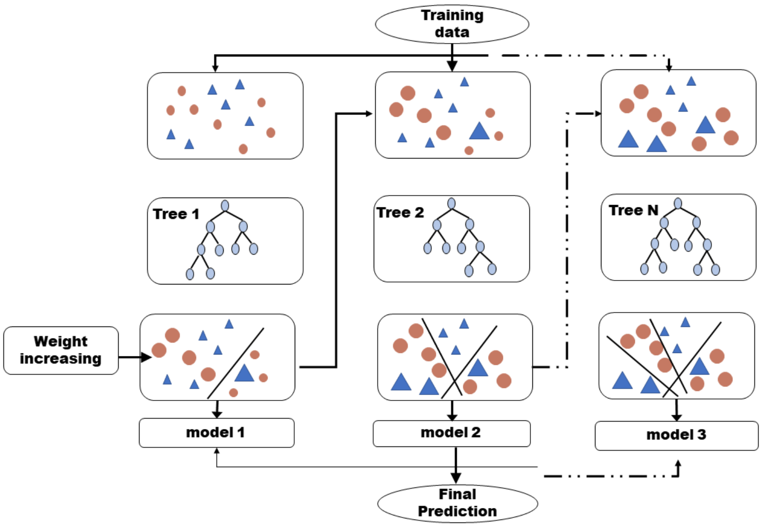

Figure 1: The working mechanism of xgboost

In Figure 1, the weak learners (trees) obtain the features (shapes) from the training dataset and tries to classify the instances of the features into the same classes for the same shapes. The misclassification of the first weak learner (tree 1) is used to optimize the performance of tree two. Tree two decides its decision boundary and tries to minimize the loss of tree one, and the same process applies to tree three. Higher weights are assigned to the samples that contributed more, and the final prediction is a weighted sum of the predictions.

3.3 *Random Forest (RF)*

Random Forest (RF) algorithm uses random samples from the dataset to form trees. Then, RF combines the trees with the bagging technique to form an ensemble (forest). It is an iterative process whereby each tree in the forest classifies the data instance in the test data. The final decision is based on the most predicted class label by the forest.

RF is considered versatile and an algorithm that can be easily used (Chang et al., 2022). The prediction becomes more accurate as the trees increase (Ashish, Kumar, & Yeligeti, 2021). Contrarily, the excessive number of trees makes it prone to overfitting, i.e., learning the patterns too well, making the model fail to generalize properly on new data (Latha & Jeeva, 2019). For this study, the Rf will be tuned to ensure the number of trees does not overfit, and hyperparameter tuning has been propounded to overcome the overfitting of RF (Latha & Jeeva, 2019).

Although RF and XGB train the trees simultaneously and make decisions based on aggregate predictions of trees, RF trains the trees with different random subsets of features from the training data. In contrast, XGB trains all trees on the same set of features from the training dataset. Additionally, RF computes results after all trees have been built (Li, Lin, Lei, & Wei, 2022). Conversely, XGB calculates and uses the results to train the next tree during the training stage, i.e., boosting (Srinivas & Katarya, 2022). For this study, Random Forest is expected to perform well in predicting heart disease for US residents because it does not require many hyperparameters (Demir & Sahin, 2022), and it does not necessarily need its hyperparameters to be tuned before getting a commendable result (Chang et al., 2022). The working procedure of RF is demonstrated in Figure 2.

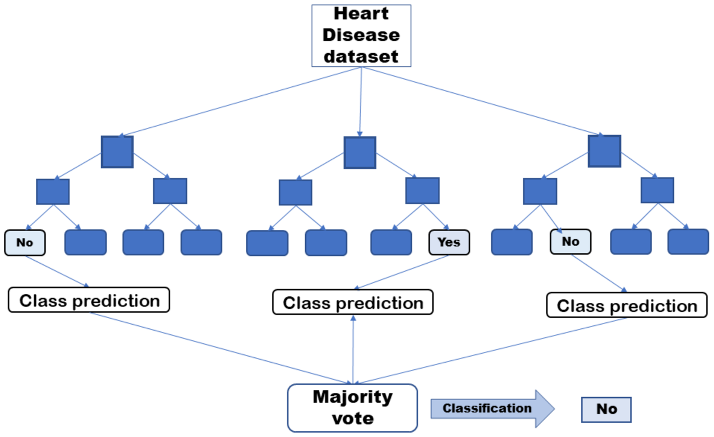

Figure 2: The visualization of a Random Forest model with three trees

3.4 *Multilayer Perceptron (MLP)*

MLP is a feed-forward artificial neural network classifier. It is a set of mathematical algorithms built to emulate the human brain, consisting of three or more fully connected layers that perform specific functions. The layers are connected to the neurons (Perceptrons) by links to form a neural network (Jeyaranjani, Rajkumar, & Kumar, 2021).

MLP predicts by performing calculations based on the input values passed through a network of neurons in the input layer. The input layer provides its output as an input to the hidden layer(s), and the output layer takes input from the hidden layer(s) and produces the prediction. The networks use weights and a bias term to adjust and optimize the calculations; this helps to minimize the prediction error. MLP computes the outcome of each neuron with an activation function. MLP uses a backpropagation mechanism with the derivative of the difference between the predicted and outcome to learn from the training data. MLP is a single, highly complex model compared to RF and XGB, which are ensembles of models (a committee of weak models as trees) (Ching, Zou, Wu, So, & Chen, 2022). MLP is highly sensitive to hyperparameters, loss function, and activation function, coupled with the black box problem of lack of transparency in decision-making. (Baral, Alsadoon, Prasad, Al Aloussi, & Alsadoon, 2021). The MLP structure can be seen in Figure 3.

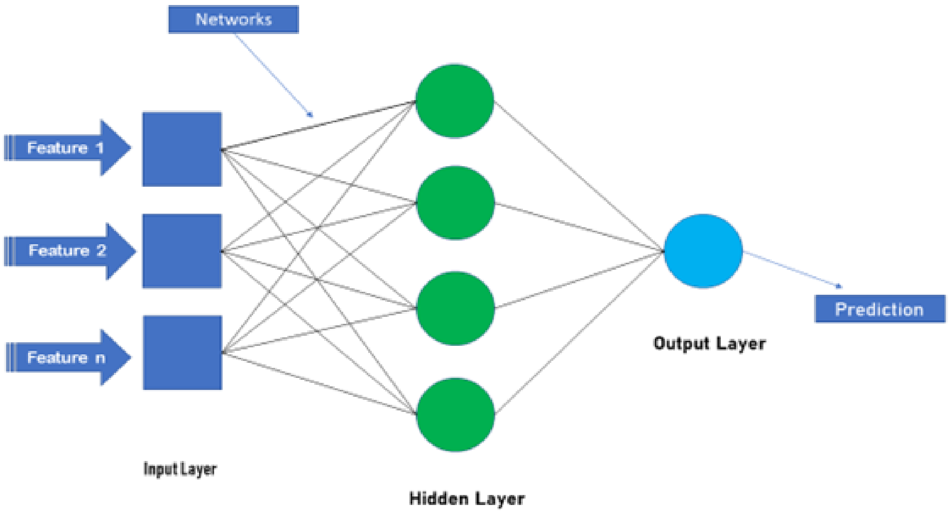

Figure 3: A Multilayer Perceptron with one hidden layer

MLP is used in this study because it can uncover implicit patterns, allowing the detection of patients' conditions that are uneasy to discover (Rojas, Olivera, & Vidal, 2022). Additionally, MLP can approximate functions of any type ranging from complex predictions to categorical predictions(Alalawi & Manal, 2021).

3.5 *Voting classifier*

A Voting classifier is an ensemble meta-classifier and a multi-expert approach that concurrently trains and combines the predictions of multiple classifiers. Eventually, the voting classifier combines the forecasts and takes a final decision based on the majority outcome (Naji et al., 2021). A voting classifier can either use the hard or soft voting technique to take the final decision (Kumari, Kumar, & Mittal, 2021). Hard voting takes the majority predicted outcome as the final decision, and the soft voting classifier decides the final class based on the class with the highest probability values from the predictions.

3.6 *Proposed model*

The proposed model is an ensemble soft voting meta-classifier that integrates Random Forest, XGBoost, and Multilayer Perceptron machine learning models to classify heart disease. All the individual models will contribute to the decision-making of the hybrid meta-classifier, and the final decision (classification) is taken by soft voting. This Vote technique is equivalently associated with Bayes' minimum error rule for classification

(Kumari et al., 2021). That is, combining classifiers will help to reduce the risk of random classifications. Thus, the classification error will be minimized. Additionally, a vote-based ensemble model is proposed for this study as it could act as a multi-expert procedure in reducing the erroneous diagnosis of heart disease (Chandra, Verma, Singh, Jain, & Netam, 2021). The working mechanism of the proposed model is presented in Figure 4.

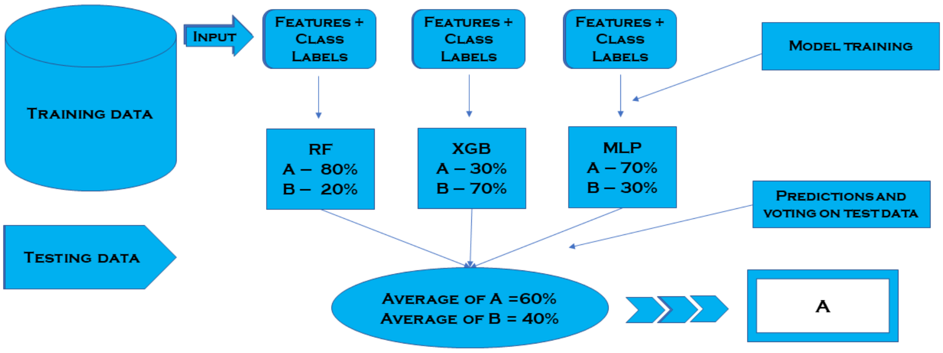

Figure 4: Illustration of ensemble soft voting classifier for heart disease prediction

Figure 4 describes the training data consisting of the selected features and the class labels (heart disease, no heart disease) used to train a hybrid model comprising Random Forest, XGBoost, and Multilayer Perceptron. After the models have learned the latent patterns in the training data, they make predictions on unseen data (test data). The class with the highest probability from the models' aggregate predictions will be the final predicted class.

3.7 *Experimental Setup*

The experimental workflow of this experiment consists of ten stages that can be sectioned into four phases. The exploratory and cleaning phase is the initial phase of exploring and cleaning the data. The pre-processing phase involves feature selection with IG, data transformation, normalization, and data partitioning. The model building, training, hyperparameter tuning and predictions is done in the modeling stage, and the concluding stage is the results and analysis stage. The steps in the experimental pipeline are depicted in Figure 5.

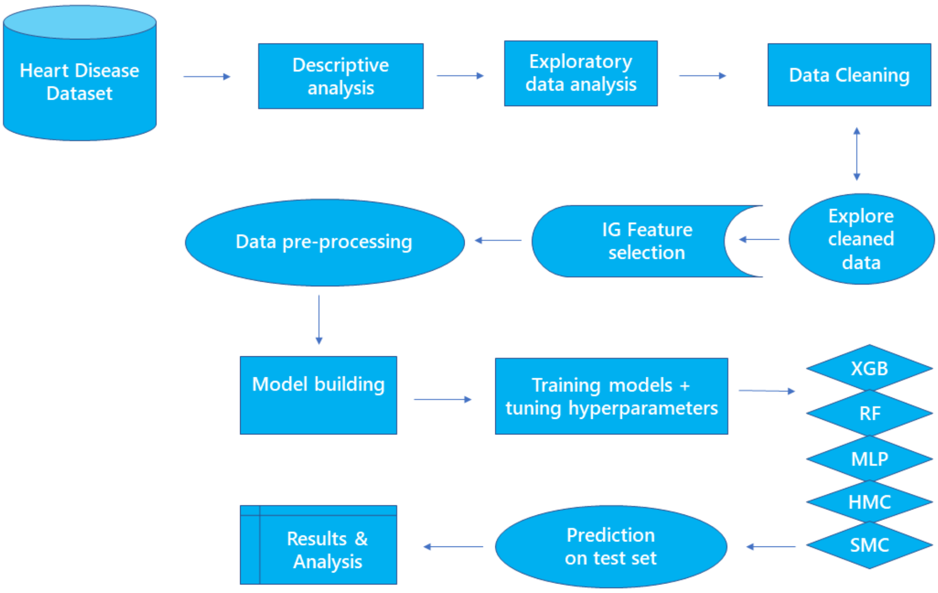

Figure 5: Experimental pipeline of the study

3.8 *Dataset*

The dataset used to predict heart disease is the 2020 Behavioural Risk Factor Surveillance (BRFSS) data (CDC, 2021). The BRFSS is an alliance project by every state in the USA and its territories. BFRSS is conducted by the Centers for Disease Control and Prevention (CDC). The aim is to gather details on health-related risk attributes, abysmal health conditions, access to medical services, and information on the foremost factors that cause diseases and deaths among the United States residents above the age of 18. The data was collected in 2020 through Computer-Assisted Telephone Interview (CATI) systems and published in 2021 by CDC. The data consist of 401,958 records and 300 attributes.

The information and meaning of all records and attributes can be found in the LLCP codebook (CDC, 2021). (Pytlak, 2022) selected a subset of the data comprising potentially related attributes to heart disease and published it on Kaggle, an online community for data scientists and machine learning. There are 18 attributes and 319,795 records in the heart disease dataset. The details of the fields in the dataset are in Appendix A (3).

The BFRSS (CDC) dataset is used for this work because it provides the necessary attributes needed to fulfill the aim of this study. Such as the heart disease category, potential risk factors of heart disease, and the recency of the dataset. Also, the dataset details are specific to the intended population of this study.

The dataset is highly imbalanced. It consists dominantly of the residents without the disease 91.44% (292,422) and only 8.56% (27,373) have heart disease. The dataset was resampled to overcome this caveat. The resampling procedure is discussed in the experimental procedure sub-section. The proportion of heart disease in the dataset can be seen in Figure 6.

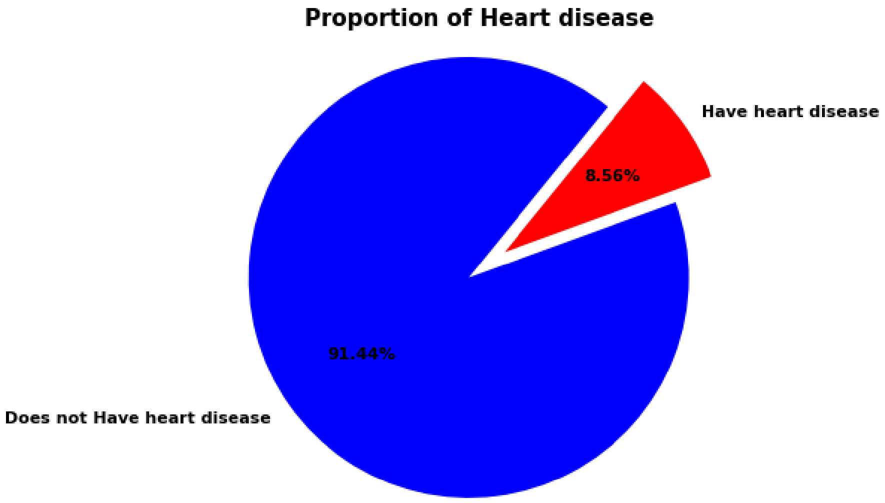

Figure 6: The proportion of heart disease instances in the BFRSS dataset

The likelihood of having heart disease gets higher as US residents grow older, this is demonstrated in Figure 7.

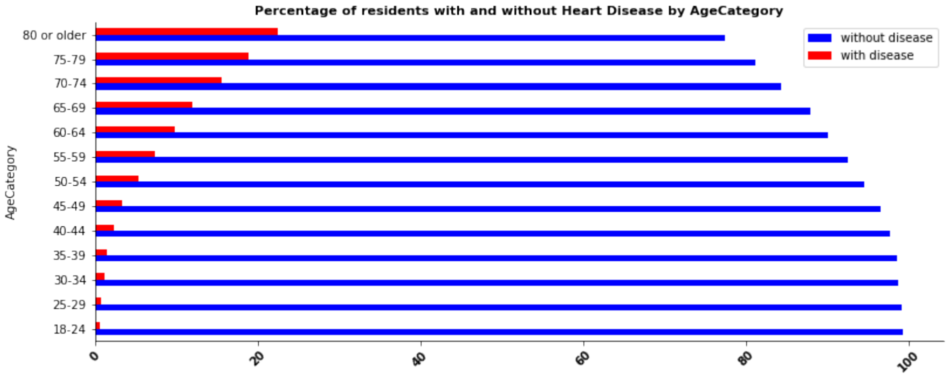

Figure 7: The exploratory analysis of the relationship between heart disease category and age category.

Although there are more females (167,805) than males (151,990) in the dataset, the males develop heart disease than the females, as seen in Figure 8.

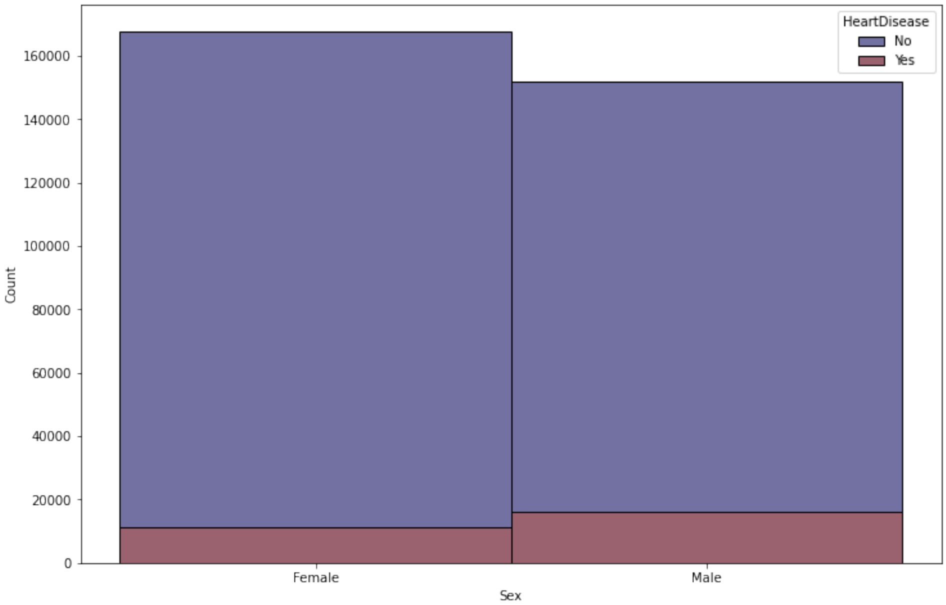

Figure 8: Heart disease by gender status

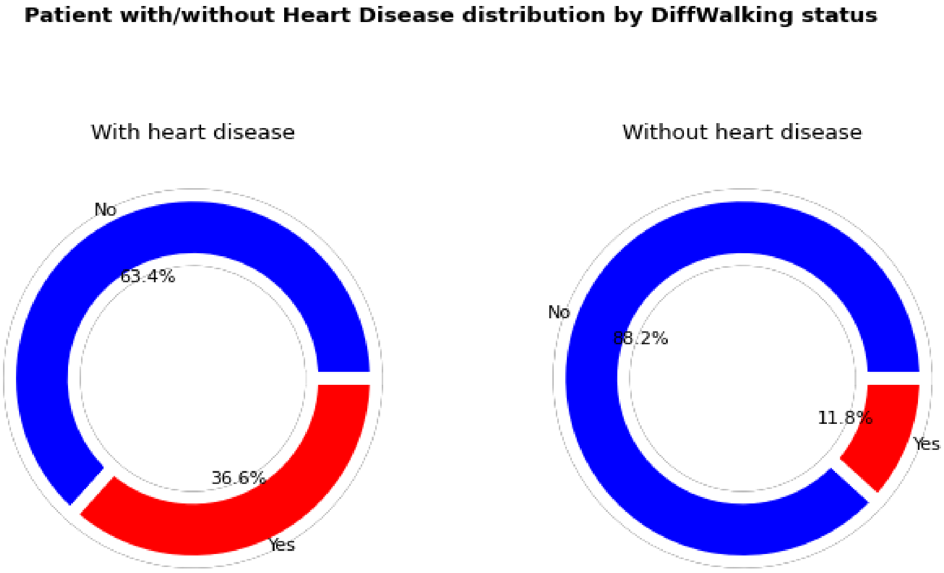

Figure 9: Visualization of the relationship between heart disease category and difficulty walking or difficulty climbing the stairs category

Note: The residents who find it difficult to walk or climb the stairs are in the yes category while those without difficulty are in the no category as represented by Figure 9.

3.9 *Data cleaning and processing on the dataset*

The data were explored, cleaned, and processed using Jupyter Notebook with Python (Van Rossum & Drake Jr, 1995) version 3.7.13 on the Google Colaboratory environment. There was no missing value in the dataset. There were outliers in the numerical categories, BMI, Physical health, mental health, and sleep time. The outliers were not eliminated because eradicating them could make the models lose some important and distinguished information that will help achieve the objective of this work. For instance, people that recorded that their physical health was not good for the whole 30 days in a month are distant from the mean of 3.4 days in the dataset. These patient may be sick, and taking them out can affect the result of the analysis negatively. Some patients recorded an average of fewer than 4 hours and up to 24 hours of sleep. They may have severe health conditions, chronic illness, or sleeping disorders, which may be vital to the aim of this study. Figure 10 describes the average sleeping hours of the residents within 24 hours.
newpage

3.10 *Experimental procedure*

Firstly, the categorical features and the target column were encoded from text to numeric format. The heart disease column was encoded with LabelEncoder(), and other categorical columns were encoded with OrdinalEncoder(). The purpose of encoding the variables from text to numeric format is to enable the machine learning models to process the data for predictions (Budholiya et al., 2020). The encoding functions are in the preprocessing class in the scikit-learn library (Pedregosa et al., 2011)

Afterward, the information gain method was implemented on the dataset by building two functions. The first function applies the mutual_info_classif(), mutual_info_classif() takes two arguments; the first argument takes any given variable in the dataset, and the second argument takes the heart disease column. After, the second function iteratively applies the earlier function to all variables in the dataset. The variable names were appended to an empty list, and IG values were appended to another empty list. Finally, the appended lists of variable names and IG scores were converted into a data frame object using the pd.Dataframe function from the pandas libraryMacKay and Mac Kay (2003).

The dataset with was sectioned into 80% train and 20% test samples using the train_test_split() function from scikit-learn (Pedregosa et al., 2011), with a random state of 42. This was done to create a dataset for generalization beyond the training data. The dataset was partitioned before resampling as a robust measure to avoid data leakage from the test dataset to the training set. Next, the features contain values of different ranges, and the StandardScaler() function was used to normalize the train and test sets. Normalization gives all parameters unbiased influence for the models to perform better (Ruchay et al., 2022). After that, the training dataset was further partitioned into 80% train and 20% validation datasets to create validation data for hyperparameter tuning.

3.11 *Data resampling*

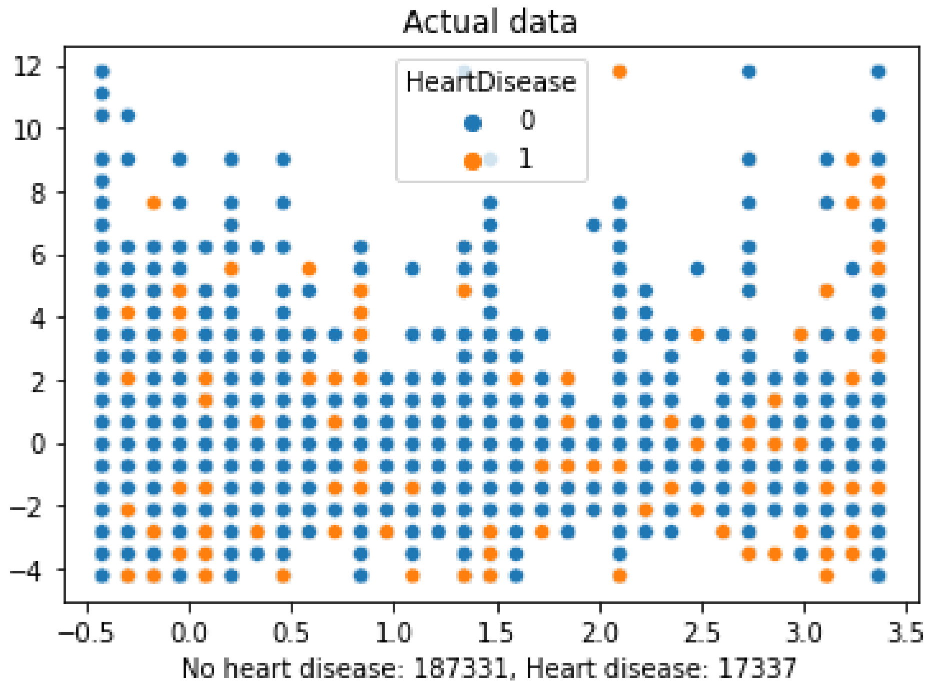

Figure 10: The dataset before resampling

Random oversampling (ROS), synthetic minority oversampling (SMOTE), and random undersampling (RUS) techniques were applied to the training set to overcome the challenge of class imbalance. The random oversampling method randomly duplicates the data samples of the minority class with replacement, increasing the sample size of the minority class to that of the majority class in the training dataset (Zhu, Zhou, & Zhang, 2021).

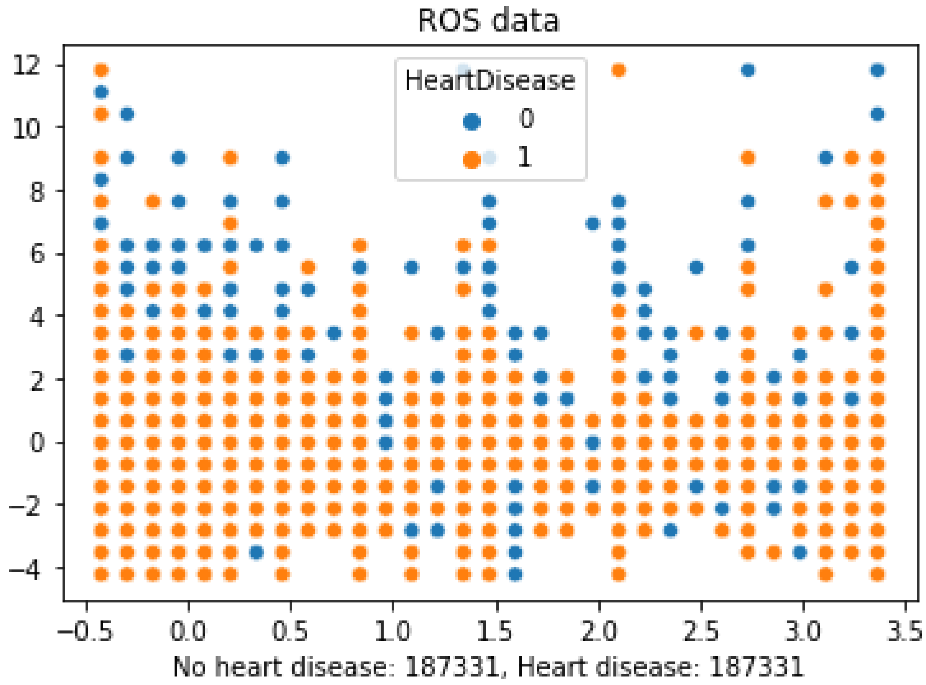

Figure 11: The dataset after random oversampling

SMOTE generates new samples of the minority class by linear interpolation between close instances of the minority class on the feature space. It uses a K Nearest Neighbour (KNN) technique. A sample of the minority class is chosen at random, the k nearest datapoints (neighbors) are found, a neighbor is chosen at random, and a synthetic sample is generated by selecting a random point between the two data points on the feature space (Kaisar & Chowdhury, 2022).

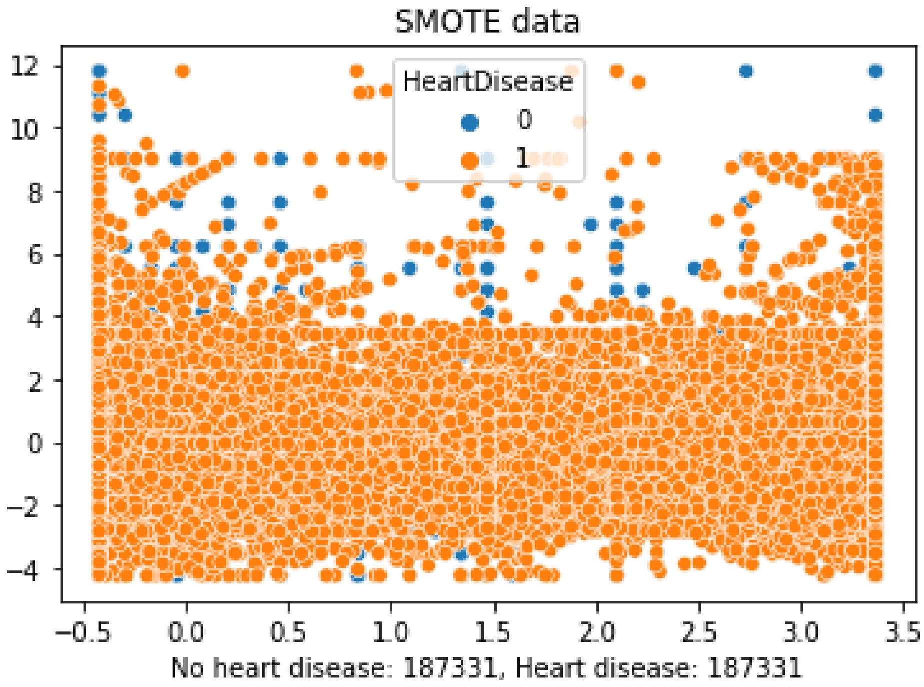

Figure 12: The dataset after SMOTE oversampling

The random undersampling method randomly takes out instances from the majority class to downsize the majority class to the dimension of the minority class (B. Liu & Tsoumakas, 2020). These are the most widely used techniques for resampling (Guo, Zhuang, Sun, & Qin, 2020; Sağlam & Cengiz, 2022).

The dataset was resampled with a random state of 42 each using RandomOversampler().fit_resample(). SMOTE().fit_resample, and RandomUndersampler.fit_resample() functions from the imblearn library in scikit-learn (Pedregosa et al., 2011).

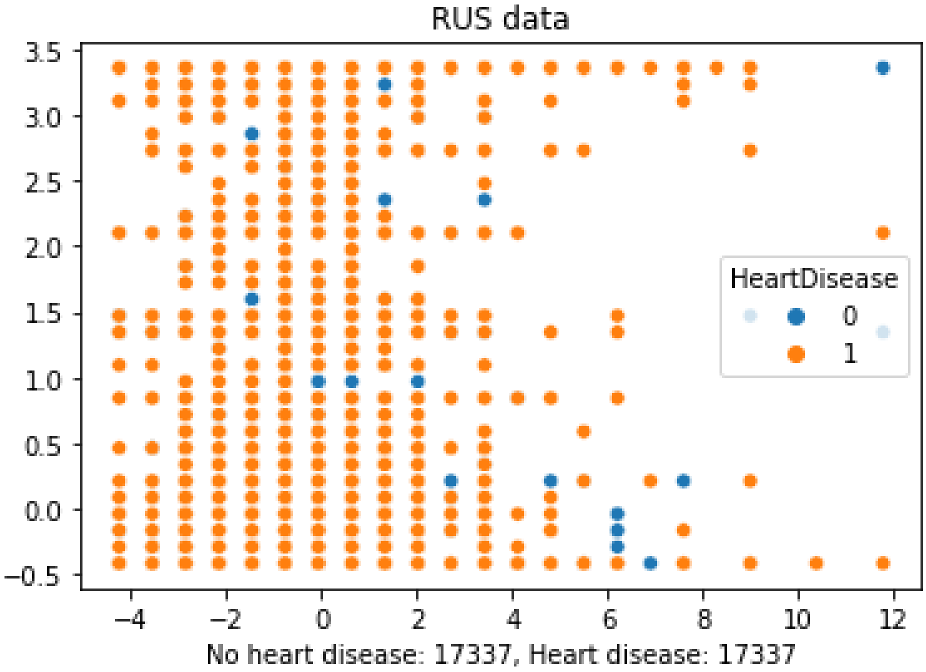

Figure 13: The dataset after SMOTE oversampling

3.12 *Hyperparameter tunning*

The hyperparameters were tuned with a ten cross-fold validation for each individual model. The hyperparameter tuning was done on the validation dataset, which was not resampled, with GridSearchCV() library from scikit-learn (Pedregosa et al., 2011). Except for XGBoost, hyperopt bayesian optimization tunning was used to obtain the optimal hyperparameters for XGB. This tunning approach yielded the best performance for the XGB model after the result was compared to grid search tuning. The full details of the tested hyperparameters and adopted values are described in Appendix B(4). The default parameters were used for the constituents of the voting classifiers, as it yielded better performance than tuned hyperparameters. In addition, the weight of the voting classifiers were manually tuned with a gradual increase of 0.1 for each constituent. The adopted parameters for the Voting classifiers are presented in Appendix C (5).

The five models, namely Rf, XGB, MLP, hard voting classifier, and soft voting classifier, were distinctly trained on all the features and on all the training data samples: imbalanced data, ROS, SMOTE, and RUS. After, each model made four predictions on the test set, which amasses to twenty predictions. A robustness check was carried out to ensure the dataset generalizes above the training data. The result is presented in the following section.

3.13 *Feature importance*

The XGB model was used to get the feature importance because XGB performed best among all the lone classifiers in this work. Initially, the features and values were extracted from the XGB model by using XGBClassifier().get_booster().get_score(importance_type='weight') function from XGBClassifier(), and it was stored into an assigned variable named feature_importance. After that, the feature names were appended into an empty list by using list(feature_importance.keys()) function. Likewise, the values were appended into another empty list by using the list(feature_importance.values()) function. The extracted feature names and importance scores were saved into a dataframe object before plotting a bar graph of the features and respective importances.

3.14 *Evaluation method*

Although accuracy is the most commonly used metric, it exhibits biases toward the majority class (Gu et al., 2022). Given the imbalanced nature of the BFRSS dataset, the results will be accessed mainly with AUCROC, F-score, recall, precision, and the number of misclassifications. The recall is used because it accesses the number of positive cases that is correctly classified (Douzas, Bacao, & Last, 2018), the recall score ranges from 0% to 100%. AUCROC (Area under the Receiver Characteristics) has been adopted since it compares the performances of models (Khan & Mondal, 2020; Zalikha, El-Othmani, & Shah, 2022; Zea-Vera et al., 2021). A model with an AUCROC score above 0.5 is considered a better performer than a model that predicts randomly (W. Liu, Chen, & Hu, 2022). The F-score evaluates the balance between the correctly classified instances in the majority and minority classes (Sağlam & Cengiz, 2022). The precision is adopted because it assesses the rate of correct positive class predictions (Sambasivam & Opiyo, 2021). These metrics were implemented with recall_score(), roc_auc_score(), precision_score(), and f1_score() functions in the Scikit-learn library (Pedregosa et al., 2011). . The accuracy was implemented using the accuracy_score() function from the Scikit-learn library(Pedregosa et al., 2011).

True positive (TP) = These are the correct diagnosis of heart disease. True Negative (TN) = These samples are rightly diagnosed as not having heart disease False Positive = This is the category of patients that are wrongly diagnosed of having heart disease False Negative (FN) = These residents have heart disease but have been misdiagnosed as healthy residents.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + TrueNegative + FalsePositive + FalseNegative}$$

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$F1 - measure = 2\frac{Precision * Recall}{Precision + Recall}$$

$$AUCROC = \frac{1 + TruePositiveRate - FalsePositiveRate}{2}$$

$$Misclassification = FalseNegative + FalsePositive$$

The data level performance was done by using the performance on the imbalanced data as a baseline to evaluate the performance on the resampled data. The between-model comparison was made by comparing the best performances of the individual models. The individual models′ results were used as a baseline to evaluate the ensemble models′ performances.

4 RESULTS

This section presents the results for the research question and the sub-research questions. The HMC represents the Hard voting Meta Classifier, the SMC represents the soft voting Meta Classifier, miss is the amount of misclassified instances, and the %miss represents the percentage of misclassified records.

4.1 *Key factors of heart disease in the USA*

The first sub-research question is about the determinants of heart disease in the USA. The result is presented in Figure 14 and Figure 15, and the values are presented in Appendix D (6) and Appendix E (7). All features have an IG value greater than zero. This implies they could all provide some information to predict heart disease. The two methods highly ranked the general health conditions, sex, and age factors. These features produced the most information to detect heart disease in both techniques, as they ranked among the top 30% factors for both methods. In contrast, skin cancer, alcohol intake, mental health, and asthma were the common least substantial factors in predicting heart disease for both systems. Although, factors such as diabetes, stroke, race, and physical health did not commonly emerge as highly influential factors. However, the features were relevant in detecting heart disease in both techniques. Also, the physical activity factor did not perform any role in predicting heart disease by XGB. Nevertheless, IG considered physical activity as an influential factor for heart disease.

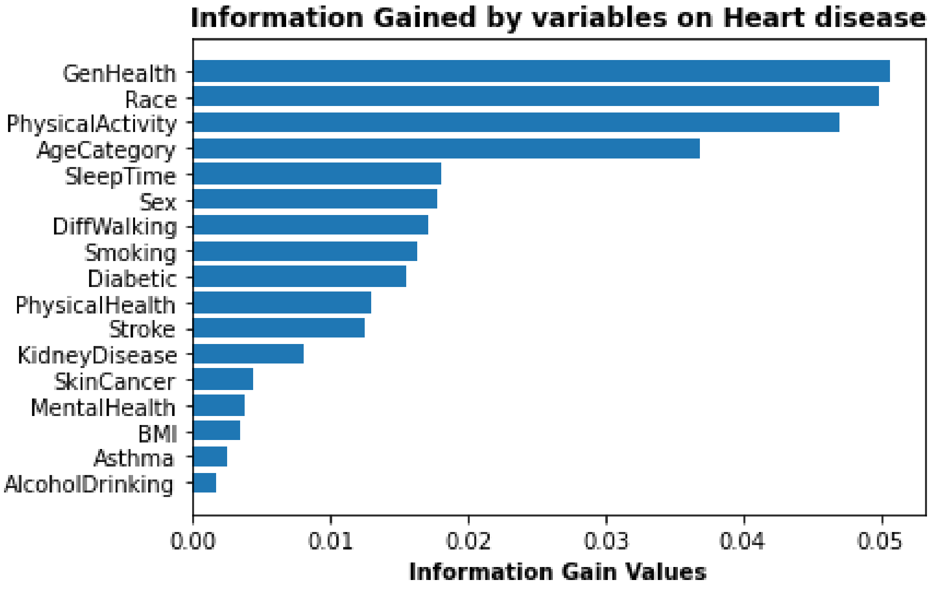

Figure 14: The information gain value of the features in the dataset

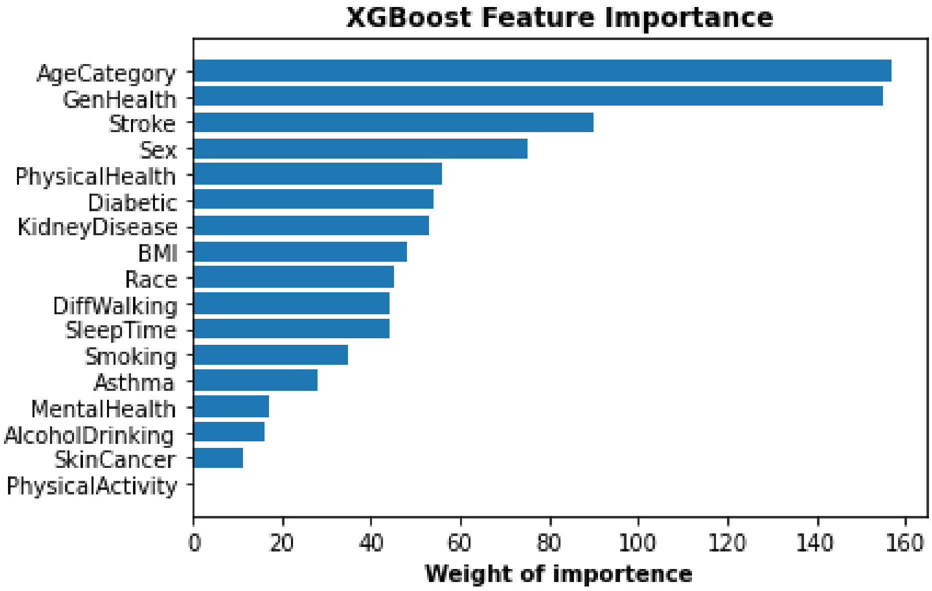

Figure 15: The XGB feature importance chart

4.2 *The outcome of researching the second sub-research question.*

The second sub-research question posed an inquiry on the best voting mechanism for predicting heart disease for United States inhabitants. The result of the experiment is illustrated in Table 1. The hybrid models performed better on the resampled data than on the actual data. Both meta-classifiers predicted best on the random oversampled data. Nevertheless, they best predicted the heart disease class on the SMOTE data. Although HMC outperformed SMC on recall and AUCROC, SMC was more precise, had a higher F1 score, and had fewer misclassified instances. Therefore, the soft voting performed better than the hard voting technique in this study.

Model	Data	Accuracy	Recall	Precision	F1	AUROC	Miss	%Miss
HMC	Actual	91.41%	7.30%	57.06%	12.94%	53.39%	5,491	8.59%
	ROS	73.58%	79.67%	22.04%	34.53%	76.33%	16,896	26.42%
	SMOTE	11.28%	99.48%	8.93%	16.39%	51.16%	56,745	88.72%
	RUS	73.50%	79.72%	21.99%	34.47%	76.31%	16,952	26.50%
SMC	Actual	91.38%	6.56%	55.95%	11.75%	53.03%	5,514	8.62%
	ROS	74.57%	78.29%	22.53%	35.00%	76.21%	16,266	25.43%
	SMOTE	44.07%	95.01%	13.02%	22.90%	67.10%	35,775	55.93%
	RUS	72.84%	80.33%	21.64%	34.09%	76.23%	17,369	27.16%

Note: Number of test instances = 63,959. Positive heart disease instances = 5,592.

Table 1: Result of the second sub-research question

4.3 *The outcome of researching the main research question*

The main research question enquires if the combination of models with soft voting has higher predictive performance than the single models. Table 2 presents the outcome of the baseline (individual classifiers) and the metrics obtained from the ensembled models. Generally, the models performed better on the resampled data than on the actual data, and the precision rates were low compared to other metrics. The models predicted the class with heart disease best on the SMOTE data, especially XGB, with an almost perfect recall rate of 99.89%. However, when all other metrics are considered, the overall performance was better on the random oversampled data (ROS) except for the MLP.

The MLP appeared to be less sensitive to the data resampling methods, with not many variations between the performances on the resampled data compared to other models. The MLP results on the resampled data were almost inseparable. Hence, the fewer misclassified instances were used to decide the better model, which is the performance of MLP on SMOTE data. The XGB model with random oversampling had the best performance among the independent models. However, when the individual models are compared to the meta-classifiers, the XGB on ROS data performed better than the HMC on ROS data. This is because the XGB obtained higher metrics than HMC in all evaluated metrics. Considering SMC, XGB had better recall and AUCROC rates than the SMC on ROS data. Howbeit, SMC recorded a better precision rate, exceeding f1 score, and a fewer misclassification rate than XGB.

Model	Data	Accuracy	Recall	Precision	F1	AUROC	Miss	%Miss
RF	Actual	91.36%	3.61%	60.12%	6.82%	51.69%	5,524	8.64%
	ROS	73.26%	79.15%	21.74.%	34.11%	75.92%	17,103	26.74%
	SMOTE	47.20%	91.63%	13.33%	23.28%	67.29%	33,769	52.80%
	RUS	72.37%	80.76%	21.39%	33.82%	76.16%	17,671	27.62%
XGB	Actual	91.43%	7.69%	57.56%	13.57%	53.57%	5,479	8.57%
	ROS	73.75%	79.72%	22.16%	34.68%	76.45%	16,790	26.25%
	SMOTE	9.86%	99.89%	8.83%	16.23%	50.56%	57,651	90.14%
	RUS	73.60%	79.69%	22.05%	34.54%	76.35%	16,888	26.40%
MLP	Actual	91.35%	5.78%	54.84%	10.45%	52.66%	5,535	8.65%
	ROS	73.50%	79.67%	21.98%	34.46%	76.29%	16,947	26.50%
	SMOTE	74.00%	78.27%	22.12%	34.50%	75.94%	16,623	25.99%
	RUS	73.54%	79.22%	21.94%	34.36%	76.11%	16,925	26.46%
HMC	Actual	91.41%	7.30%	57.06%	12.94%	53.39%	5,491	8.59%
	ROS	73.58%	79.67%	22.04%	34.53%	76.33%	16,896	26.42%
	SMOTE	11.28%	99.48%	8.93%	16.39%	51.16%	56,745	88.72%
	RUS	73.50%	79.72%	21.99%	34.47%	76.31%	16,952	26.50%
SMC	Actual	91.38%	6.56%	55.95%	11.75%	53.03%	5,514	8.62%
	ROS	74.57%	78.29%	22.53%	35.00%	76.21%	16,266	25.43%
	SMOTE	44.07%	95.01%	13.02%	22.90%	67.10%	35,775	55.93%
	RUS	72.84%	80.33%	21.64%	34.09%	76.23%	17,369	27.16%

Note: Number of test instances = 63,959. Positive heart disease instances = 5,592.

Table 2: Result of the main research question

4.4 *Robustness check on models' performances.*

A robustness check was carried out on the performance of the models to ascertain no substantial difference between the training and test errors. The test results of the robustness check are presented in Appendix F (8). A general observation is that the models overfit the smote data because there are considerable differences between the train and test accuracies on the smote data. For example, XGB performance on the smote data exhibits a high level of overfitting with a difference of 81.43 between the train and test accuracies. Similarly, HMC, to a large extent, did not generalize performance on SMOTE data. However, an exception to the general overfitting on SMOTE data is MLP. MLP captured the general trend in SMOTE data and generalized it to SMOTE test data.

4.5 *Error analysis and generalization above the training set*

An error analysis was conducted to ensure the models do not perform by random guessing. The visualization of the error analysis can be found in Appendix G (21). There is an overall trend of random predictions by the models on the imbalanced data; this can be further verified from the confusion matrix plots that the models assigned most cases to the majority class. Also, the models had a similar discriminative threshold on the ROS and RUS data; this signifies they performed similarly on both data. In addition, the model performed best on the ROS and RUS data. The AUCROC plots of XGB and HMC on SMOTE data indicate that models did not generalize from train data. Both models predicted almost all instances by randomly assigning the cases to the positive class.

5 DISCUSSION

The current study investigated a research question and two sub-research questions to help know the prominent factors of heart disease in the USA. Additionally, to uncover the best algorithm and technique for predicting heart disease in the USA. Experiments were conducted to answer these questions. The results are presented in the preceding section and are further discussed in this section

5.1 *What are the key factors of heart disease in the United States*

The most influential factors of heart disease in this study are general health conditions, age, and sex. The findings corresponded to the discovery of Ali et al. (2020); CDC (2019); Chandra et al. (2021); Ramesh et al. (2021); Srinivas and Katarya (2022). In addition, Gao, Chen, Sun, and Deng (2019) confirmed sex as a fundamental factor of heart disease. The authors proclaimed that men are more at risk of having coronary heart disease and heart attack than women. However, females have a worse prognosis and higher fatality rate from heart disease than men. Also, older people have been proclaimed to be at higher risk of heart disease in the USA due to a decline in sex hormones: testosterone and estrogen. Further, the study claimed that an increase in age comes with declining cardiovascular activities, which subsequently increases the risk of heart disease (Rodgers et al., 2019). Iantorno (2020) affirmed that the general health condition had been a principal factor of heart disease. In addition, the scholar described general health conditions to be compounded by many other changeable lifestyle factors such as smoking, alcohol, healthy eating and exercising.

Alcohol intake was not a significant feature for predicting heart disease in this study. This finding contradicts the assertions of existing studies (Ahsan & Siddique, 2022; Latha & Jeeva, 2019) . An explanation for this discovery could be derived from Iantorno (2020). The author explained that factors such as alcohol do not directly affect heart disease. However, they are causal mechanisms that lead to factors of heart disease. For example, the authors perceived that drinking alcohol could lead to gaining extra weight (higher BMI) or raising blood pressure, which could later lead to heart disease. However, feature importance mechanisms cannot vastly account for these indirect relationships, as the mechanisms only examine the direct connections between variables (Dokeroglu, Deniz, & Kiziloz, 2022). Thus, the IG and XGB consider alcohol intake not influential in predicting heart disease due to the unsubstantial direct-dependent relationship between alcohol and heart disease.

5.2 *Does the hard voting technique predict heart disease better than the soft voting for United States residents?*

In the study, soft voting predicted heart disease better than hard voting for United States residents. Making decisions based on the probability values of predictions rather than the predicted class label proved to be the more efficient technique. It could be assumed that soft voting has a lower tendency to propel unsure predictions towards a class, as it decides based on the confidence of the voters (probability-based decision-making). Contrarily, hard voting could propel unsure predictions toward a class as it decides based on the binary outputs of voters and does not consider the confidence of voters. The outcome of this experiment contradicts the findings of Raza (2019) that a hard voting strategy should be the foremost approach to predicting heart disease.

A notable area of influence on the result is the assigned weights of the estimators and the effect of weight tuning. Tuning the weight assigned to the XGB constituent from 1 to 7.7 increased the voting influence of XGB. Consequently, a higher recall rate was observed in both models. The outcome of this experiment contradicts the findings of Raza (2019) that a hard voting strategy should be the foremost approach to predicting heart disease.

5.3 *Would the hybrid model consisting of Random Forest, XGBoost, and Multilayer Perceptron, combined with the soft voting technique, predict heart disease better than each of the individual classifiers for the United States residents?*

The hybrid model consisting XGBoost, Artificial Neural Network, and Random Forest, combined with soft technique, predicted heart disease better than each of the individual classifiers for the United States residents. This result is consistent with Sherazi et al. (2021), which support that soft voting predicts heart disease better than the individual voting technique. A probable cause for this outcome is that each algorithm has its specific mastery. Therefore, combining more than an algorithm helps integrate the diverse expertise of the constituents to improve the process. Fan, Lung, Ajila, et al. (2018) validated this claim by emphasizing that an ensemble generally outperforms single models, as they have different capabilities to recognize trends in data. Thus, the combination of distinct expertise and divergent data learning skills forms a powerful amalgamation that supersedes the abilities of individual models.

Suppose future studies aim to predict heart disease. This study proposes SMC with SMOTE sampling when the priority is majorly on high recall, and other metrics are less important. Among the models that did not predict randomly (AUCROC), SMC on Smote had the best recall rate. However, if the algorithm's overall performance is considered, and the high rate of misdiagnosing healthy people as high-risk patients is unwanted. In that case, the SMC with random oversampling is proposed.

5.4 *Error analysis and robustness checks.*

The models did not accurately predict the minority class on the imbalanced data. The likely reason is that the models trained on data with few instances of the minority class. This could make the models not capture the limited trends that categorize instances to the minority class. Gu, Tian, Li, and Jiang (2022) confirmed this suspicion that, at default, the models assume a slight difference between the sample sizes of the classes. Thus, the models tend to learn instances of the majority class mainly. The performance was lesser on SMOTE data than on other data. Most misclassifications and widest margins between the train and test accuracies were on SMOTE data. Chen, Zhang, Huang, Wu, and Luo (2022) proclaimed that SMOTE interpolates unrepresentative samples and noise.Zhang, Yu, Zhou, Huan, and Yang (2022) bolstered the argument by claiming that SMOTE does not examine the internal structure of training data and distribution, which results in non-robust and unstable predictions. However MLP did not overfit on SMOTE data. This could be because MLP has a high potential to approximate functions that are uneasy for other models to execute, and also its high ability to discern implicit patterns (Alalawi & Manal, 2021; Rojas et al., 2022).

6 research limitations, further recommendations

A limitation of this research was that a substantial portion of the data instances belonged to the group that did not have heart disease. Also, the dataset might not accurately represent the intended target as it accounts for 27,373 (0.0015%) out of the 18.2 million real positive cases of CHD in the US (CDC, 2021). This challenge made analyzing heart disease more complex in this study. For example, the influence of the factors in predicting heart disease may differ when more positive cases are observed. Also, training machine learning algorithms with imbalanced data led to imprecise classification results, which may mislead diagnoses or treatment. Suppose health organizations could endeavor to provide inclusive datasets with higher positive cases. Consequently, it will lead to more certainty that the factors discovered in analyses are common and generalizes over a larger population of the studied class. Equally, it will eradicate the bias of machine learning algorithms towards a particular class when predicting heart disease, leading to more precise outputs.

Another challenge in this study was no established threshold or method to decide if the information gain value of a feature is adequate for the variable to be a valuable predictor. Therefore, future studies can explore ways to determine a threshold that will aid in knowing if a variable's IG value is enough to be considered beneficial when predicting heart disease.

Likewise, a constraint experienced was the required time to train and tune the models, specifically the RF, MLP, and hybrid models. Suppose the computational time to train and predict a large volume of data is considered, bearing that it is needed to save lives or prevent the severity of cases as soon as possible. In that case, future studies can research a more efficient and faster real-time prediction for the inhabitants of the USA and other heart disease analyses.

Future can study heart disease using a larger unbiased dataset on a global level or in other countries.

7 THEORETICAL AND PRACTICAL IMPLICATIONS

This project contributed to the existing studies by extending the work on heart disease. This was achieved by determining the crucial factors of heart disease in the US and proposing a technique for reducing heart diseases in the USA. The results from this analysis would provide more insights and information to health practitioners and society in general. It further proves that artificial intelligence could develop societal health.

This study provides bases for further exploration, baselines, and methodology for scholars and researchers. A practical impact could be that this research provides clinical researchers, institutions, and health organizations with a blueprint for implementing research on causal analysis, clinical diagnosis, and predictive study on diseases.

In addition, this study could provide a basis to selectively target people for computerized tomography (CT) scans and coronary angiograms. Hence, saving the time and finances of other patients not at risk of heart disease. Also, it could lead to a lesser need for CT scans for suspicious patients; this will help circumvent needless side effects resulting from the tests and scans.

This study could help aid the detection of heart disease in regions where they do not have enough clinics or few medical practitioners for instance, refugee camps.

The feature importance method could improve the interpretability and explainability of model performance. Furthermore, the proposed technique of this study could be used for medical diagnosis and to aid decision-making on medical diagnosis.

8 CONCLUSION

The aimed to research the important factors that lead to heart disease in the USA and found age, sex and general health conditions to be very influential in detecting heart disease in the US. Also, Soft and hard voting techniques were evaluated to predict heart disease in this research. This soft voting classifier predicted heart disease better than hard voting and other models. This study therefore take the first steps in predicting heart disease with voting meta classifiers for the residents of USA and dicerning factors of heart disease by a double assessment of statistical method and on algorithm level. This study proposed a soft voting hybrid model for the prediction and diagnosis of heart disease. Also, this work explored ways to prevent more cases of the disease. The IG and XGB feature importance methods were able to select the important factors of heart disease in the USA. This study proposed a soft voting hybrid model for the prediction and diagnosis of heart disease. Hopefully, this work familiarises people, care professionals, and researchers with applying supervised machine learning to problems and gets to develop this study's work.

9 ACKNOWLEDGMENTS

I would like to express my immense appreciation to my family. They supported me and encouraged me throughout the thesis. Also, special gratitude goes to dr. Görkem Saygili. He provided me with guidance and support during this thesis period. Finally, I appreciate my friends and colleagues Itoro Unang and Jesse Berger for their help and advice.

REFERENCES

Ahmadi, E., Weckman, G. R., & Masel, D. T. (2018). Decision making model to predict presence of coronary artery disease using neural network and c5. 0 decision tree. *Journal of Ambient Intelligence and Humanized Computing*, *9*(4), 999–1011.

Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, 102289.

Alalawi, H. H., & Manal, S. A. (2021). Detection of cardiovascular disease using machine learning classification models. *International Journal of Engineering Research & Technology (IJERT) ISSN*, 2278–0181.

Alam, M. Z., Rahman, M. S., & Rahman, M. S. (2019). A random forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked*, *15*, 100180.

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

Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*, *36*, 82-93. Retrieved from https://www.sciencedirect.com/science/article/pii/S0736585318308876 doi: https://doi.org/10.1016/j.tele.2018.11.007

Ansari, M. F., Alankar, B., & Kaur, H. (2020). A prediction of heart disease using machine learning algorithms. In *International conference on image processing and capsule networks* (pp. 497–504).

Ashish, L., Kumar, S., & Yeligeti, S. (2021). Ischemic heart disease detection using support vector machine and extreme gradient boosting method. *Materials Today: Proceedings*.

Baral, S., Alsadoon, A., Prasad, P., Al Aloussi, S., & Alsadoon, O. H. (2021). A novel solution of using deep learning for early prediction cardiac arrest in sepsis patient: enhanced bidirectional long short-term memory (lstm). *Multimedia Tools and Applications*, *80*(21), 32639–32664.

Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019). Improving heart disease prediction using feature selection approaches. In *2019 16th international bhurban conference on applied sciences and technology (ibcast)* (pp. 619–623).

Bhat, P., & Dutta, K. (2021). A multi-tiered feature selection model for android malware detection based on feature discrimination and information gain. *Journal of King Saud University-Computer and Information Sciences*.

Budholiya, K., Shrivastava, S. K., & Sharma, V. (2020). An optimized xgboost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University-Computer and Information Sciences*.

CDC. (2019). *Know your risk for heart disease.* Retrieved from https://www.cdc.gov/heartdisease/risk_factors.htm

CDC. (2021). *Behavioral risk factor surveillance system.* Retrieved from https://www.cdc.gov/brfss/annual_data/annual_2020.html

CDC. (2021). *Llcp 2020 codebook report.* https://www.cdc.gov/brfss/annual$_{d}ata/2020/pdf/codebook20_{l}lcp - v2 - 508.pdf$.

Chandra, T. B., Verma, K., Singh, B. K., Jain, D., & Netam, S. S. (2021). Coronavirus disease (covid-19) detection in chest x-ray images using majority voting based classifier ensemble. *Expert systems with applications*, *165*, 113909.

Chang, V., Bhavani, V. R., Xu, A. Q., & Hossain, M. (2022). An artificial intelligence model for heart disease detection using machine learning algorithms. *Healthcare Analytics*, *2*, 100016.

Chen, Q., Zhang, Z.-L., Huang, W.-P., Wu, J., & Luo, X.-G. (2022). Pf-smote: A novel parameter-free smote for imbalanced datasets. *Neurocomputing*.

Ching, P. M. L., Zou, X., Wu, D., So, R. H. Y., & Chen, G. (2022). Development of a wide-range soft sensor for predicting wastewater bod5 using an extreme gradient boosting (xgboost) machine. *Environmental Research*, *210*, 112953.

Demir, S., & Sahin, E. K. (2022). Comparison of tree-based machine learning algorithms for predicting liquefaction potential using canonical correlation forest, rotation forest, and random forest based on cpt data. *Soil Dynamics and Earthquake Engineering*, *154*, 107130.

Dokeroglu, T., Deniz, A., & Kiziloz, H. E. (2022). A comprehensive survey on recent metaheuristics for feature selection. *Neurocomputing*.

Dominguez-Morales, J. P., Jimenez-Fernandez, A. F., Dominguez-Morales, M. J., & Jimenez-Moreno, G. (2017). Deep neural networks for the recognition and classification of heart murmurs using neuromorphic auditory sensors. *IEEE transactions on biomedical circuits and systems*, *12*(1), 24–34.

Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. *Information Sciences*, *465*, 1–20.

Ed-Daoudy, A., & Maalmi, K. (2019). Performance evaluation of machine learning based big data processing framework for prediction of heart disease. In *2019 international conference on intelligent systems and advanced computing sciences (isacs)* (pp. 1–5).

Erdoğan, A., & Güney, S. (2020). Heart disease prediction by using machine learning algorithms. In *2020 28th signal processing and communications applications conference (siu)* (pp. 1–4).

Fan, X., Lung, C.-H., Ajila, S. A., et al. (2018). Using hybrid and diversity-based adaptive ensemble method for binary classification. *International Journal of Intelligence Science*, *8*(03), 43.

Gao, Z., Chen, Z., Sun, A., & Deng, X. (2019). Gender differences in cardiovascular disease. *Medicine in Novel Technology and Devices*, *4*, 100025.

Gu, Q., Tian, J., Li, X., & Jiang, S. (2022). A novel random forest integrated model for imbalanced data classification problem. *Knowledge-Based Systems*, 109050.

Guo, L., Zhuang, Z., Sun, Y., & Qin, W. (2020). Data-driven rated power prediction of diesel engines using improved multi-class imbalanced learning method. *Procedia Manufacturing*, *51*, 324–329.

Iantorno, M. (2020). *Heart disease - risk factors.* https://medlineplus.gov/ency/patientinstructions/000106.htm.

Islam, M. T., Rafa, S. R., & Kibria, M. G. (2020). Early prediction of heart disease using pca and hybrid genetic algorithm with k-means. In *2020 23rd international conference on computer and information technology (iccit)* (pp. 1–6).

Jadhav, S., He, H., & Jenkins, K. (2018). Information gain directed genetic algorithm wrapper feature selection for credit rating. *Applied Soft Computing*, *69*, 541–553.

Jamei, M., Karbasi, M., Alawi, O. A., Kamar, H. M., Khedher, K. M., Abba, S., & Yaseen, Z. M. (2022). Earth skin temperature long-term prediction using novel extended kalman filter integrated with artificial intelligence models and information gain feature selection. *Sustainable Computing: Informatics and Systems*, *35*, 100721.

Jeyaranjani, J., Rajkumar, T. D., & Kumar, T. A. (2021). Coronary heart disease diagnosis using the efficient ann model. *Materials Today: Proceedings*.

Kaisar, S., & Chowdhury, A. (2022). Integrating oversampling and ensemble-based machine learning techniques for an imbalanced dataset in dyslexia screening tests. *ICT Express*.

Khan, M. I. H., & Mondal, M. R. H. (2020). Effectiveness of data driven diagnosis of heart disease. In *2020 11th international conference on electrical and computer engineering (icece)* (pp. 419–422).

Kumar, N., & Kumar, D. (2021). Machine learning based heart disease diagnosis using non-invasive methods: a review. In *Journal of physics: Conference series* (Vol. 1950, p. 012081).

Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, *2*, 40–46.

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203.

Li, H., Lin, J., Lei, X., & Wei, T. (2022). Compressive strength prediction of basalt fiber reinforced concrete via random forest algorithm. *Materials Today Communications*, *30*, 103117.

Liu, B., & Tsoumakas, G. (2020). Dealing with class imbalance in classifier chains via random undersampling. *Knowledge-Based Systems*, *192*, 105292.

Liu, W., Chen, Z., & Hu, Y. (2022). Xgboost algorithm-based prediction of safety assessment for pipelines. *International Journal of Pressure Vessels and Piping*, *197*, 104655.

MacKay, D. J., & Mac Kay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Naji, M. A., El Filali, S., Bouhlal, M., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Breast cancer prediction and diagnosis through a new approach based on majority voting ensemble classifier. *Procedia Computer Science*, *191*, 481–486.

Nandhini, C. U., & Tamilselvi, P. (2020). Hybrid framework of id3 with multivariate attribute selection for heart disease analysis. *Materials Today: Proceedings*, *33*, 3918–3921.

Nazri, R. A., Das, S., & Promi, R. T. H. (2021). Heart disease prediction using synthetic minority oversampling technique and soft voting. In *2021 international conference on automation, control and mechatronics for industry 4.0 (acmi)* (pp. 1–6).

Omuya, E. O., Okeyo, G. O., & Kimwele, M. W. (2021). Feature selection for classification using principal component analysis and information gain. *Expert Systems with Applications*, *174*, 114765.

Pan, S., Zheng, Z., Guo, Z., & Luo, H. (2022). An optimized xgboost method for predicting reservoir porosity using petrophysical logs. *Journal of Petroleum Science and Engineering*, *208*, 109520.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Pytlak, K. (2022). *Personal key indicators of heart disease.* Retrieved from https://www.kaggle.com/datasets/kamilpytlak/

personal-key-indicators-of-heart-disease

Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. (2020). Heart disease prediction using machine learning. *International Journal of Research and Technology*, *9*(04), 659–662.

Ramesh, G., Madhavi, K., Reddy, P. D. K., Somasekar, J., & Tan, J. (2021). Improving the accuracy of heart attack risk prediction based on information gain feature selection technique. *Materials Today: Proceedings*.

Raza, K. (2019). Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. In *U-healthcare monitoring systems* (pp. 179–196). Elsevier.

Rezaei, M. J., Woodward, J. R., Ramírez, J., & Munroe, P. (2021). A novel two-stage heart arrhythmia ensemble classifier. *Computers*, *10*(5), 60.

Rodgers, J. L., Jones, J., Bolleddu, S. I., Vanthenapalli, S., Rodgers, L. E., Shah, K., . . . Panguluri, S. K. (2019). Cardiovascular risks associated with gender and aging. *Journal of cardiovascular development and disease*, *6*(2), 19.

Rojas, M. G., Olivera, A. C., & Vidal, P. J. (2022). Optimising multilayer perceptron weights and biases through a cellular genetic algorithm for medical data classification. *Array*, *14*, 100173.

Ruchay, A., Kober, V., Dorofeev, K., Kolpakov, V., Dzhulamanov, K., Kalschikov, V., & Guo, H. (2022). Comparative analysis of machine learning algorithms for predicting live weight of hereford cows. *Computers and Electronics in Agriculture*, *195*, 106837.

Sağlam, F., & Cengiz, M. A. (2022). A novel smote-based resampling technique trough noise detection and the boosting procedure. *Expert Systems with Applications*, *200*, 117023.

Sambasivam, G., & Opiyo, G. D. (2021). A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks. *Egyptian Informatics Journal*, *22*(1), 27–34.

Sarmah, S. S. (2020). An efficient iot-based patient monitoring and heart disease prediction system using deep learning modified neural network. *Ieee access*, *8*, 135784–135797.

Sherazi, S. W. A., Bae, J.-W., & Lee, J. Y. (2021). A soft voting ensemble classifier for early prediction and diagnosis of occurrences of major adverse cardiovascular events for stemi and nstemi during 2-year follow-up in patients with acute coronary syndrome. *PLoS One*, *16*(6), e0249338.

Srinivas, P., & Katarya, R. (2022). hyoptxg: Optuna hyper-parameter optimization framework for predicting cardiovascular disease using xgboost. *Biomedical Signal Processing and Control*, *73*, 103456.

Tama, B. A., Im, S., & Lee, S. (2020). Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble. *BioMed Research International*, *2020*.

Van Rossum, G., & Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Waqar, M., Dawood, H., Dawood, H., Majeed, N., Banjar, A., & Alharbey, R. (2021). An efficient smote-based deep learning model for heart attack prediction. *Scientific Programming*, *2021*.

WHO. (2021). *Cardiovascular diseases ̆(cvds).* Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Zafari, H., Langlois, S., Zulkernine, F., Kosowan, L., & Singer, A. (2022). Ai in predicting copd in the canadian population. *Biosystems*, *211*, 104585.

Zalikha, A. K., El-Othmani, M. M., & Shah, R. P. (2022). Predictive capacity of four machine learning models for in-hospital postoperative outcomes following total knee arthroplasty. *Journal of Orthopaedics*, *31*, 22–28.

Zea-Vera, R., Ryan, C. T., Havelka, J., Corr, S. J., Nguyen, T. C., Chatterjee, S., . . . Ghanta, R. K. (2021). Machine learning to predict outcomes and cost by phase of care after coronary artery bypass grafting. *The Annals of Thoracic Surgery*.

Zhang, A., Yu, H., Zhou, S., Huan, Z., & Yang, X. (2022). Instance weighted smote by indirectly exploring the data distribution. *Knowledge-Based Systems*, *249*, 108919.

Zhu, L., Zhou, X., & Zhang, C. (2021). Rapid identification of high-quality marine shale gas reservoirs based on the oversampling method and random forest algorithm. *Artificial Intelligence in Geosciences*, *2*, 76–81.

Zunaidi, W. H. A. W., Saedudin, R. R., Shah, Z. A., Kasim, S., Seah, C. S., & Abdurohman, M. (2018). Performances analysis of heart disease dataset using different data mining classifications. *International Journal on Advanced Science, Engineering and Information Technology*, *8*(6), 2677–2682.

10 APPENDICES

APPENDIX A

Feature	Meaning	Range
HeartDisease	Having coronary heart disease, myocardial infarction.	Yes, No
BMI	BMI	12.02 – 94.8
Smoking	Resident that smoked up to 100 cigarettes (5 packs) in their lifetime	Yes, No
AlchoholDrinking	Men taking above 14 and women taking above 7 alcoholic drinks weekly	Yes, No
Stroke	If the respondent ever had a stroke	Yes, No
PhysicalHealth	How many days in past 30 days has respondents been physically ill or injured	0 - 30
MentalHealth	How many days in the past 30 days has respondents' mental health not been ideal	0 - 30
DiffWalking	Respondent has difficulty in walking or climbing stairs	Yes, No
Sex	The gender of the respondent	Male, Female
AgeCategory	The age category of respondent	14 categorical levels
Race	Ethnicity 6	categories
Diabetic	If the respondent has ever been informed to have diabetes	4 levels
PhysicalActivity	Residents that engage in physical activity asides their regular job	Yes, No
GenHealth	How the adults rate their health in general	Yes, No
SleepTime	How many hours of sleep the resident gets within 24 hours	1 - 24
Asthma	If the adult ever had asthma	Yes, No
KidneyDisease	If respondent has ever been informed of having kidney disease	Yes, No
SkinCancer	If the respondent ever had skin cancer	Yes, No

Table 3: The feature importance in the dataset.

APPENDIX B

Model	Model parameter	Tested values	Adopted value
Rf	max_depth	5, 10, 15, 20	10
	criterion	entropy, gini	gini
	max_features	auto, sqrt	sqrt
	min_samples_leaf	4, 6, 8	4
	min_samples_split	5, 7,10	10
	random_state		42
XGB	n_estimators	50,100,150,200	150
	max_depth	3, 18, 1 (3-18, by 1 step)	4
	min_child_weight	0, 10, 1 (0-10, by 1 step)	4
	reg_alpha	40,180,1 (40-180, by 1 step)	49.0
	reg_lambda	0, 1	0.6020967245261853
	gamma	1, 9	4.175519458278 37
	colsample_bytree	0.5, 1	0.7149849459607005
	random_state		42
MLP	activation	tanh, relu	relu
	hidden_layer_sizes	10, 30, 10	20
	solver	sgd, adam	sgd
	alpha	0.0001, 0.05	0.0001
	learning_rate	constant,adaptive	adaptive
	random_state		42

Table 4: List of tested hyper parameters

APPENDIX C

Model	Hyper parameter	constituents	Weight	Random_state
HMC	Default values	RF	1	42
	Default values	XGB	7.7	42
	Default values	MLP	1	42
SMC	Default values	RF	1	42
	Default values	XGB	7.7	42
	Default values	MLP	1	42

Table 5: List of hybrid models parameters

APPENDIX D

Feature	Information gain
GenHealth	0.0507
Race	0.0499
PhysicalActivity	0.0470
AgeCategory	0.0368
SleepTime	0.0181
Sex	0.0178
DiffWalking	0.0172
Smoking	0.0163
Diabetic	0.0155
PhysicalHealth	0.0130
Stroke	0.0124
KidneyDisease	0.0081
SkinCancer	0.0044
MentalHealth	0.0037
BMI	0.0035
Asthma	0.0026
AlcoholDrinking	0.0017

Table 6: The information gain value of the features in the dataset.

APPENDIX E

Feature	Weight
AgeCategory	157
GenHealth	155
Stroke	90
Sex	75
PhysicalHealth	56
Diabetic	54
KidneyDisease	53
BMI	48
Race	45
DiffWalking	44
SleepTime	44
Smoking	35
Asthma	28
MentalHealth	17
AlcoholDrinking	16
SkinCancer	11
PhysicalActivity	0

Table 7: The XGB feature importance in the dataset.

APPENDIX F

Model	Train accuracy	Test accuracy	Effect size
Random Forest Actual	91.95%	91.36%	0.59
Random Forest ROS	73.99%	73.26%	0.73
Random Forest SMOTE	84.55%	47.20%	37.35
Random Forest RUS	78.03 %	72.37%	5.66
XGB Actual	91.74%	91.43%	0.31
XGB ROS	77.19%	73.75%	3.44
XGB SMOTE	91.29%	9.86%	81.43
XGB RUS	76.62%	73.60%	3.02
MLP Actual	91.66%	91.35%	0.31
MLP ROS	76.62%	73.50%	3.12
MLP SMOTE	77.75%	74.00%	3.75
MLP RUS	76.50%	73.54%	2.96
HMC Actual	91.75%	91.41%	0.34
HMC ROS	76.93%	73.58%	3.35
HMC SMOTE	88.54%	11.28%	77.26
HMC RUS	76.81%	73.50%	3.31
SMC Actual	92.13%	91.38%	0.75
SMC ROS	81.20%	74.57%	6.63
SMC SMOTE	90.92%	44.07%	46.85
SMC RUS	79.92%	72.84%	7.08

Table 8: Model's robustness check

appendix g

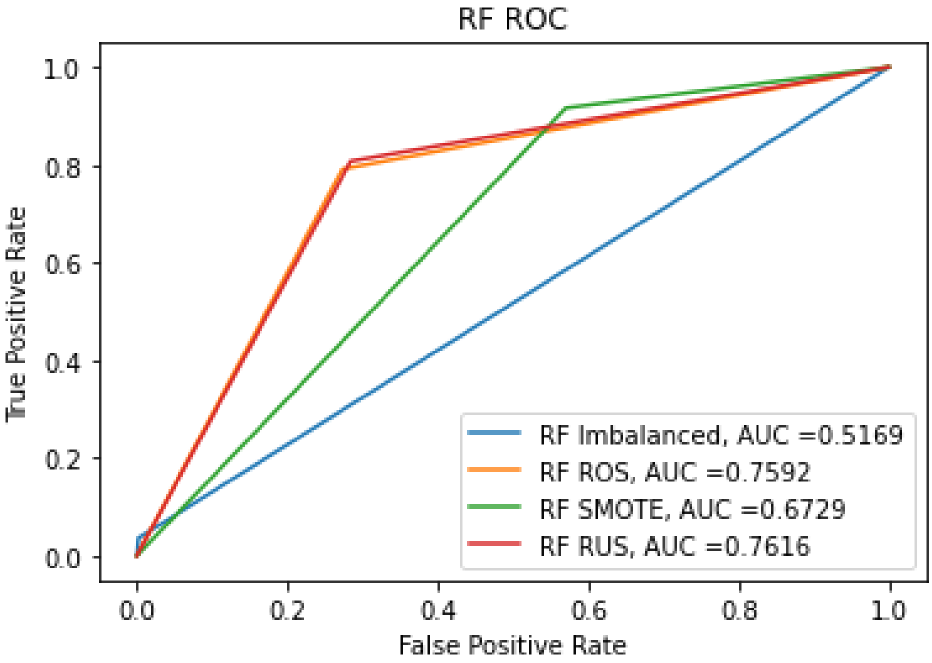

Figure 16: The ROCAUC of RF

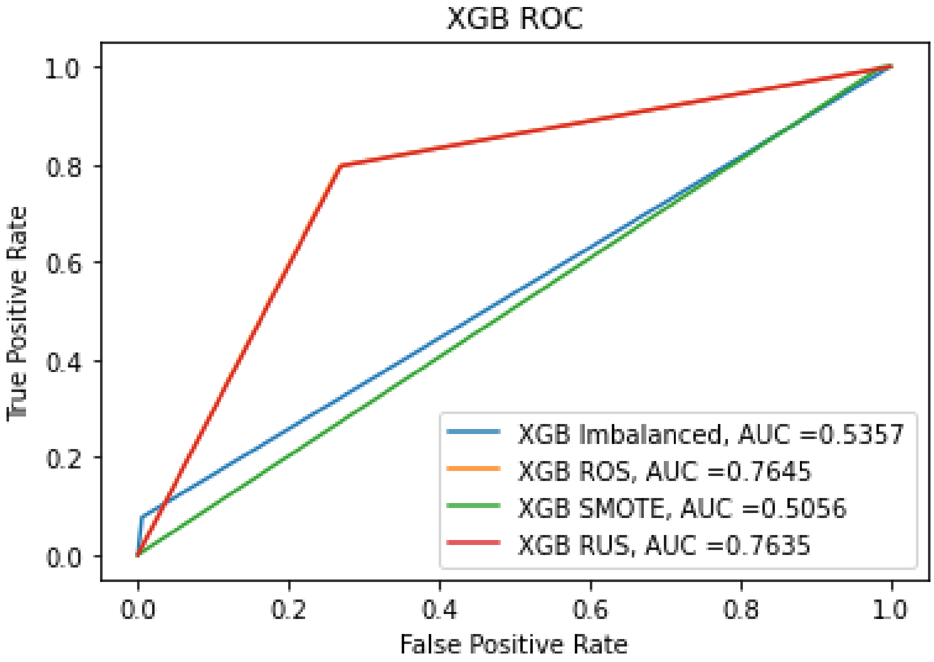

Figure 17: The ROCAUC of XGB

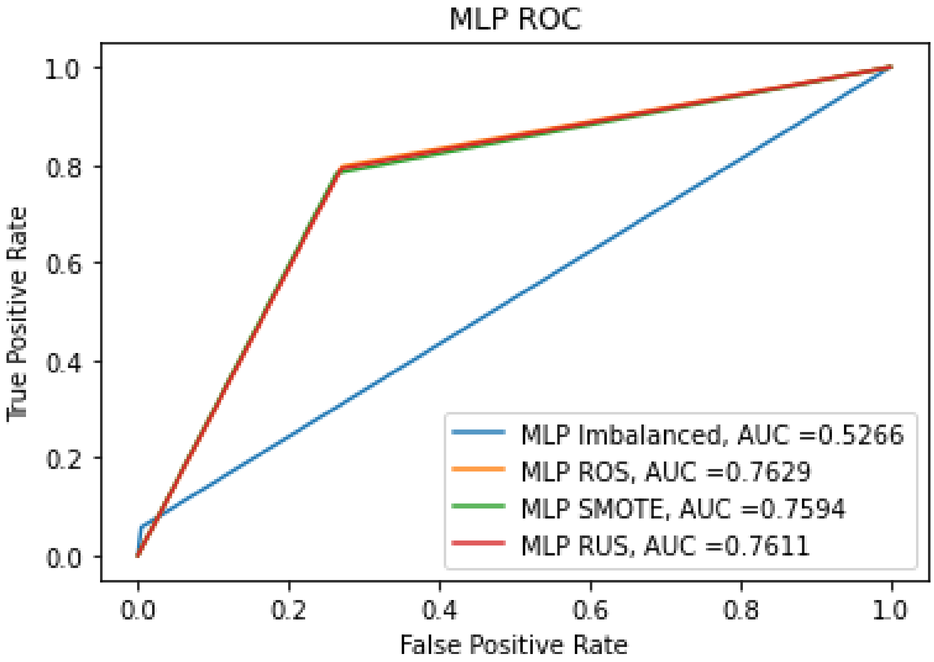

Figure 18: The ROCAUC MLP

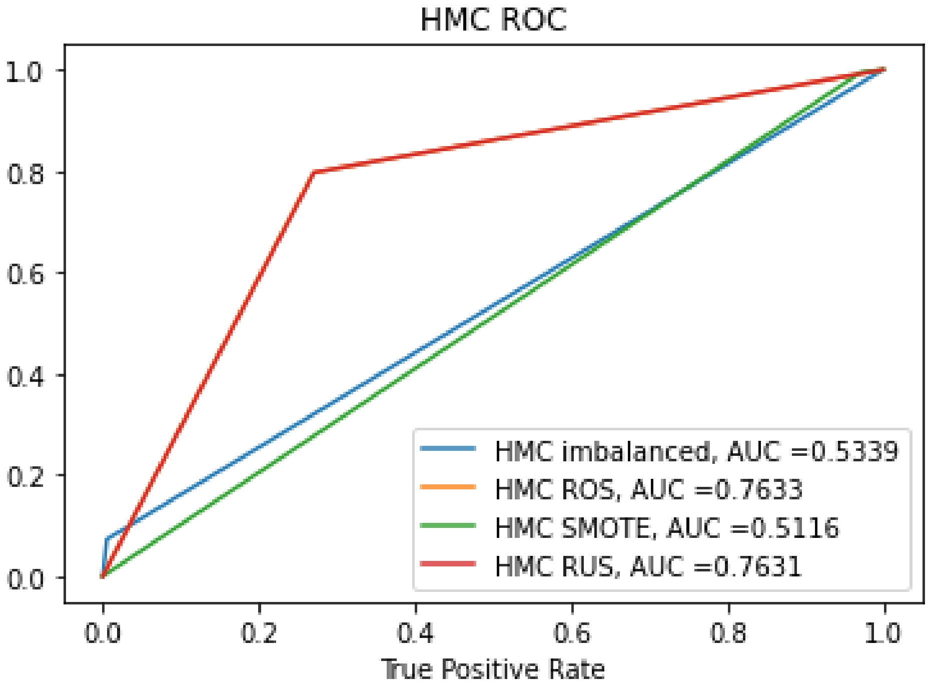

Figure 19: The ROCAUC HMC

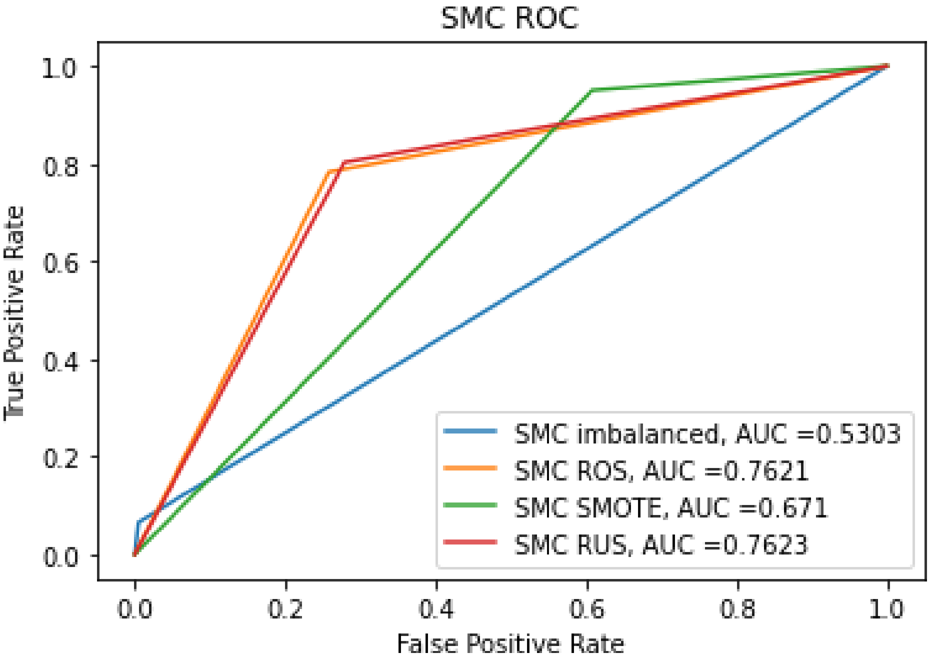

Figure 20: The ROCAUC SMC

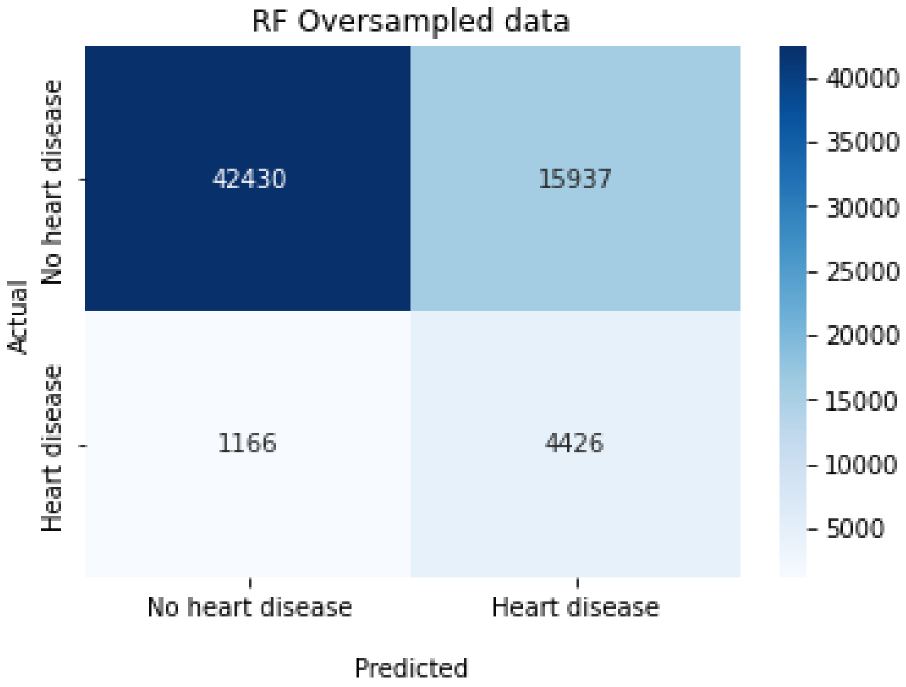

Figure 21: The confusion matrix of RF ROS data

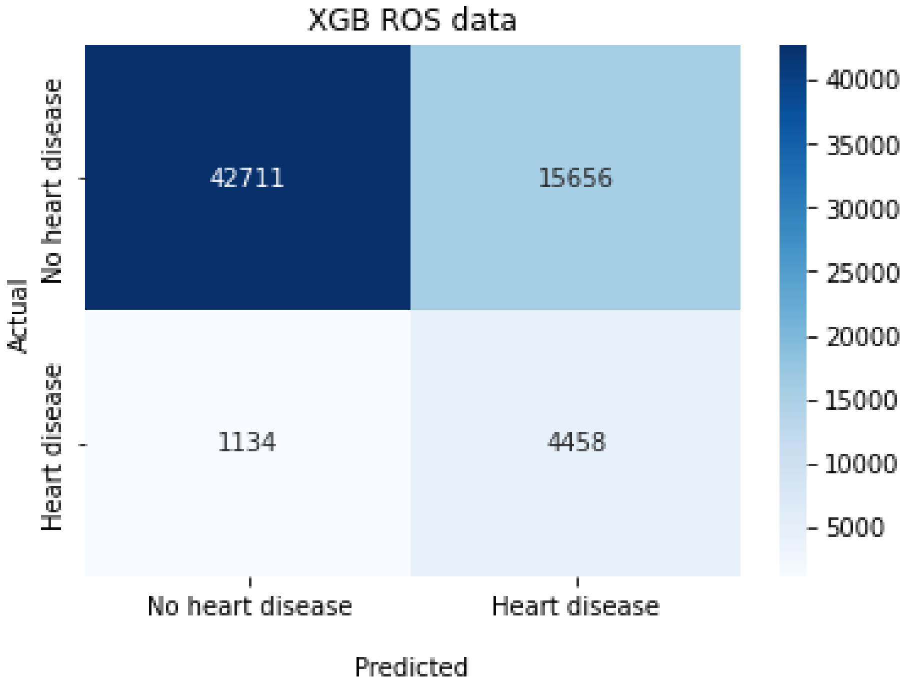

Figure 22: The confusion matrix of XGB ROS

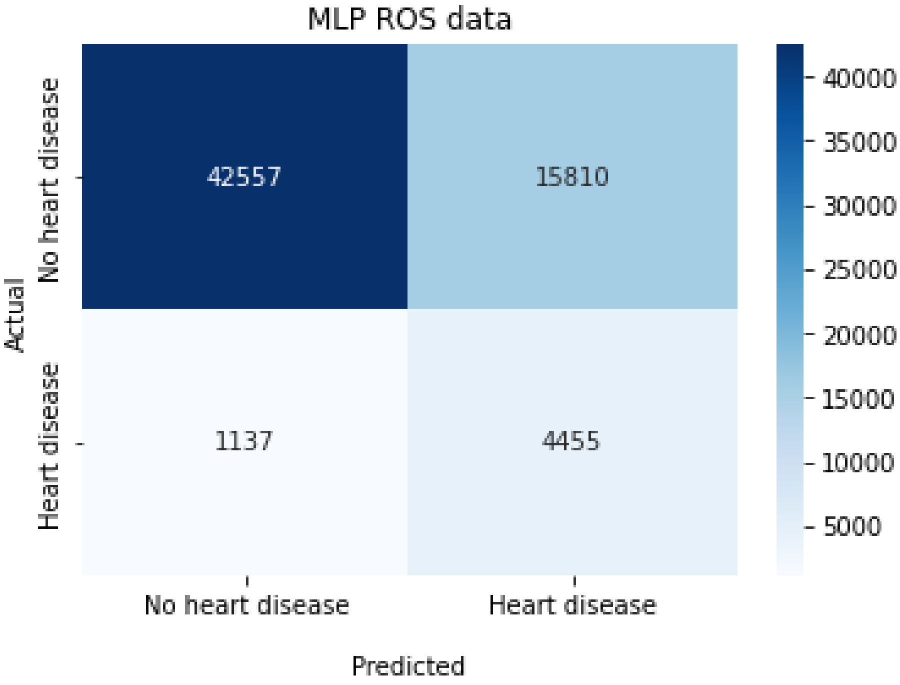

Figure 23: The confusion matrix of MLP ROS

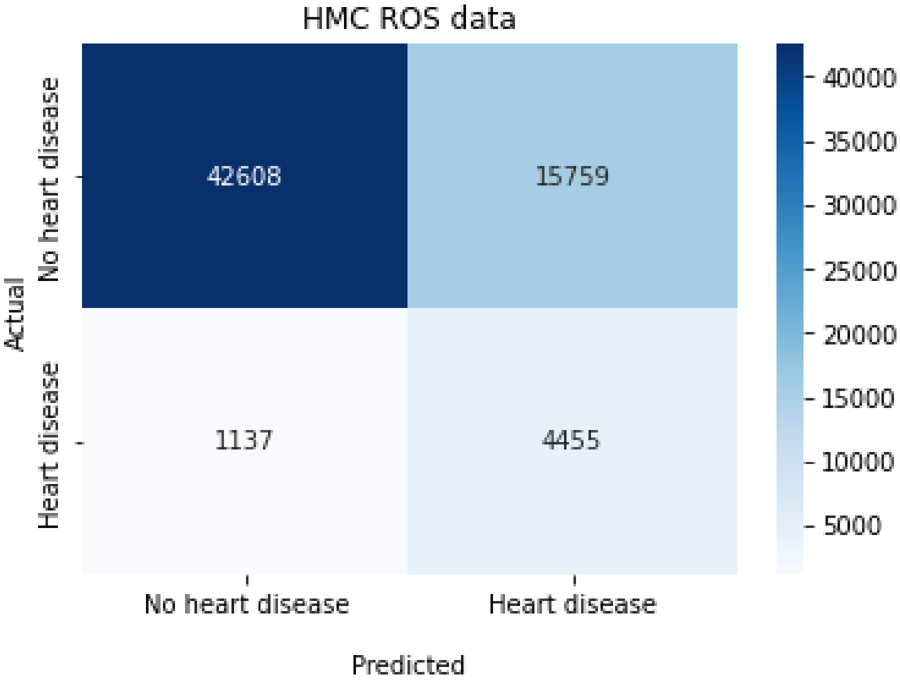

Figure 24: The The confusion matrix of HMC ROS

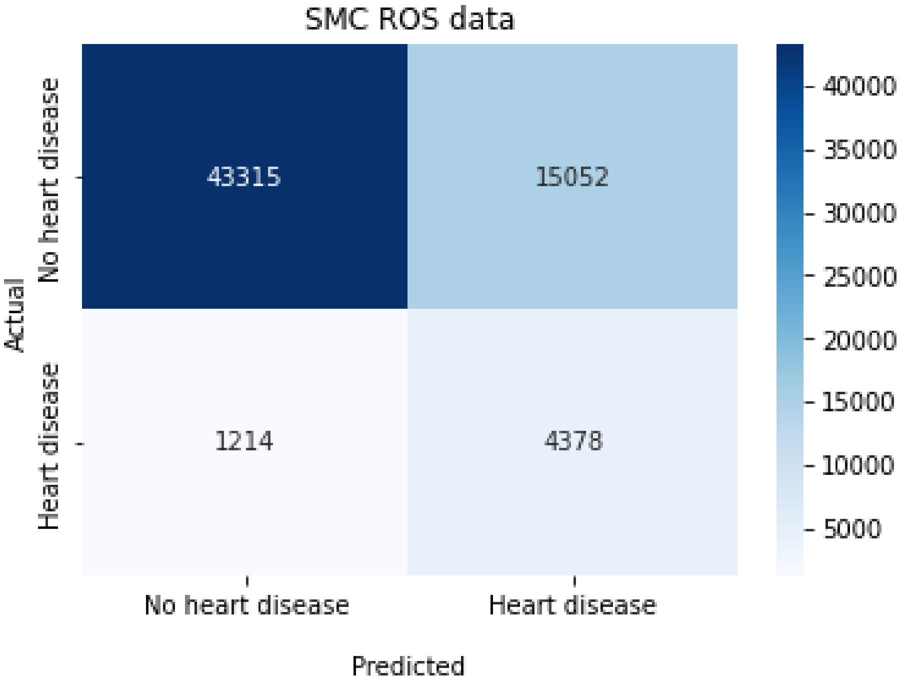

Figure 25: The confusion matrix of SMC ROS

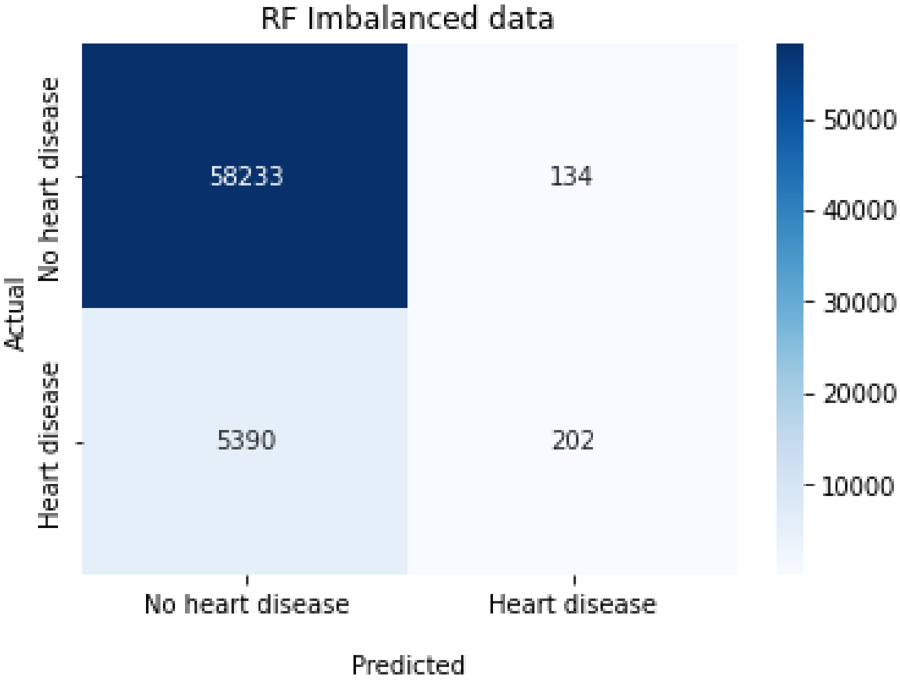

Figure 26: The confusion matrix of RF imbalanced data

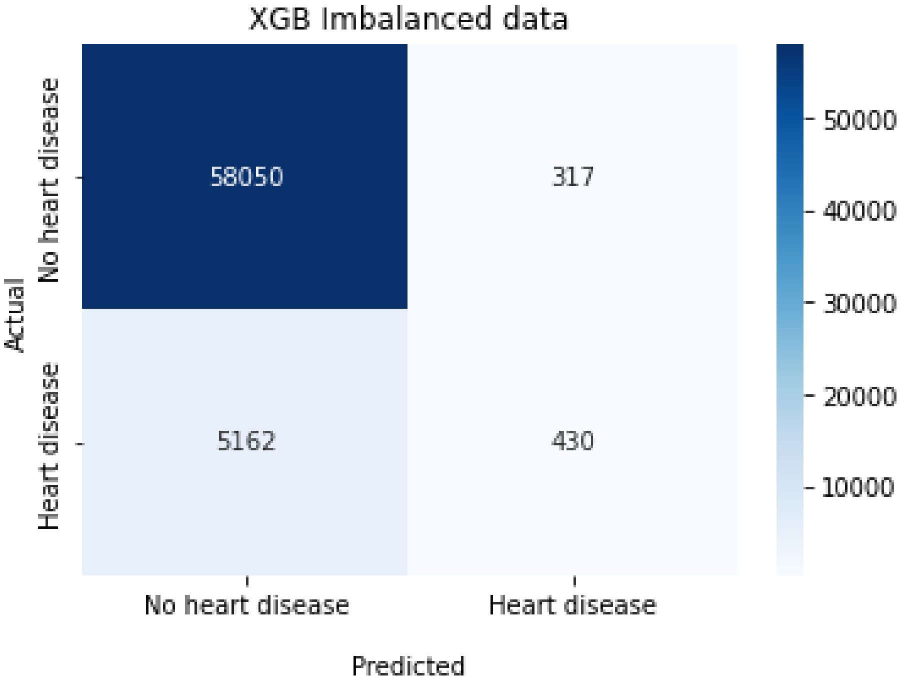

Figure 27: The confusion matrix of XGB imbalanced data

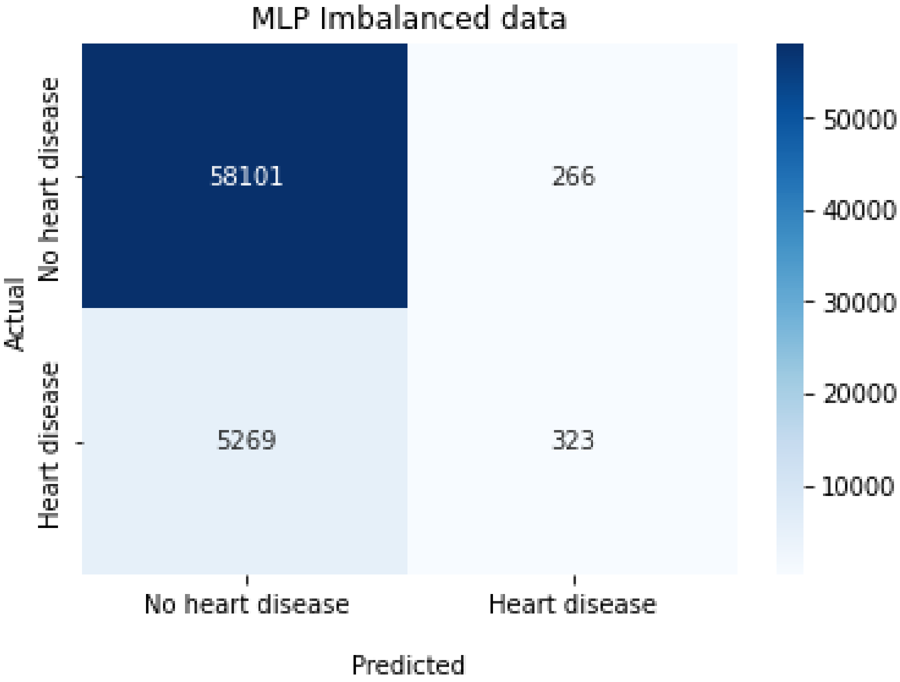

Figure 28: The confusion matrix of MLP imbalanced data

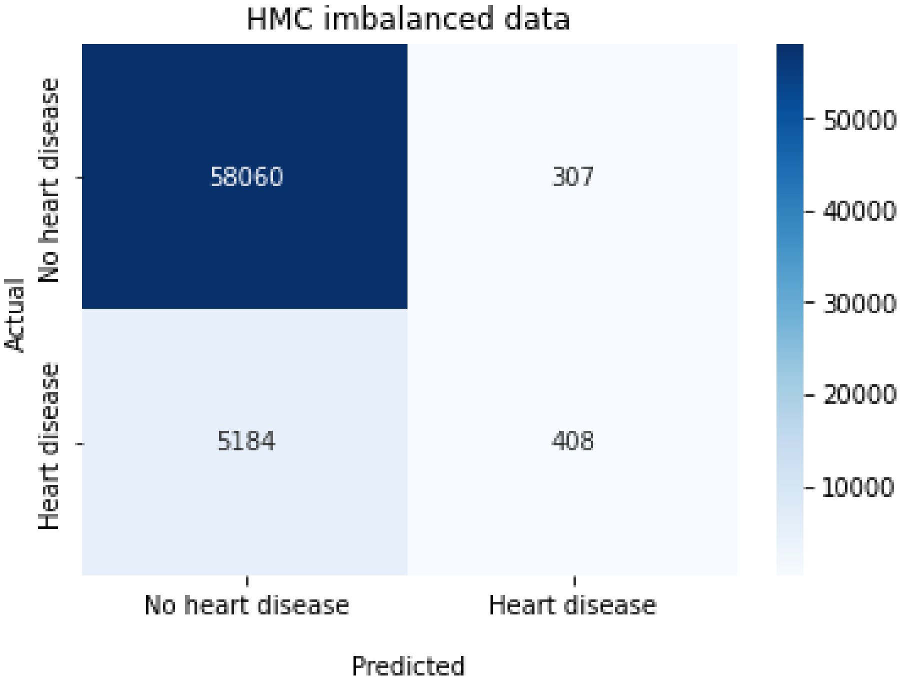

Figure 29: The confusion matrix of HMC imbalanced data

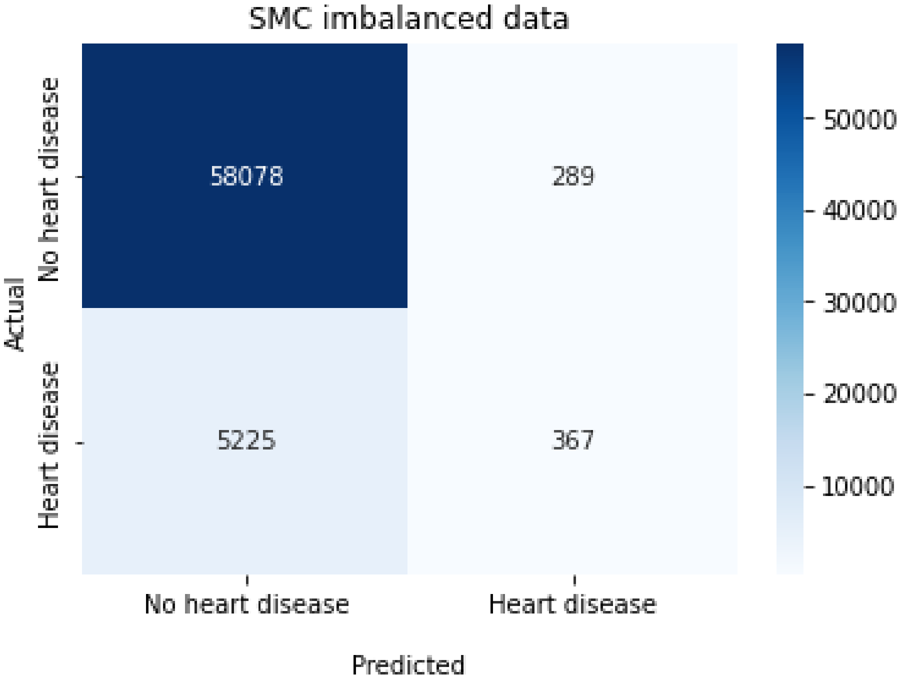

Figure 30: The confusion matrix of SMC imbalanced data

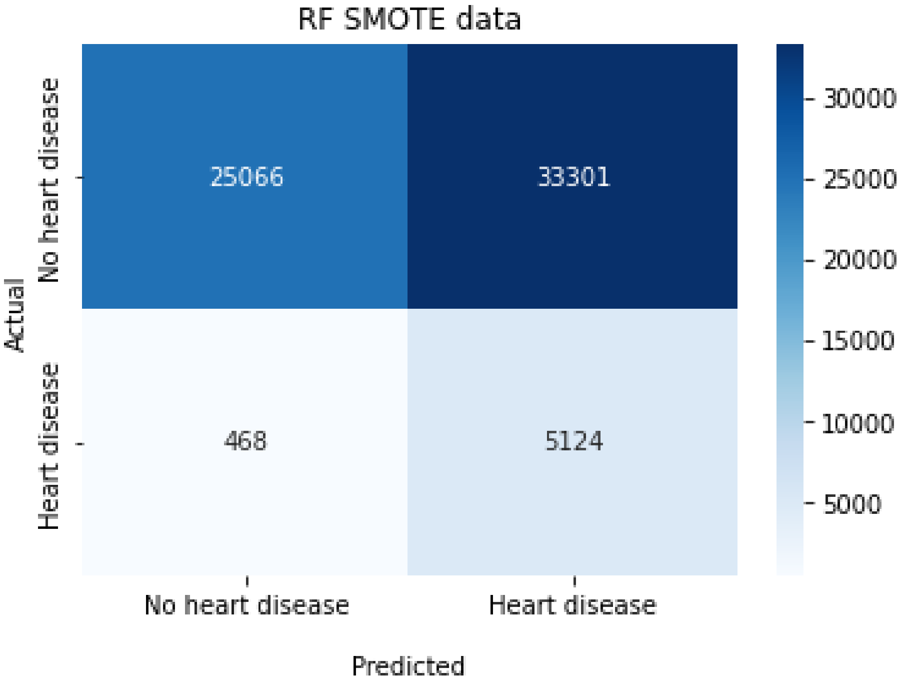

Figure 31: The confusion matrix of RF SMOTE data

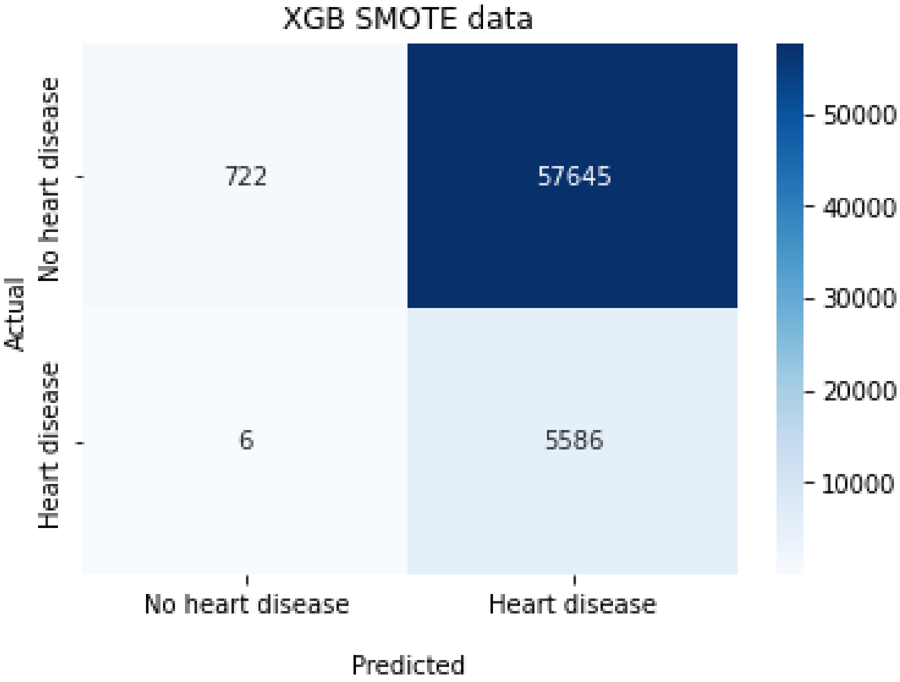

Figure 32: The confusion matrix of XGB SMOTE data

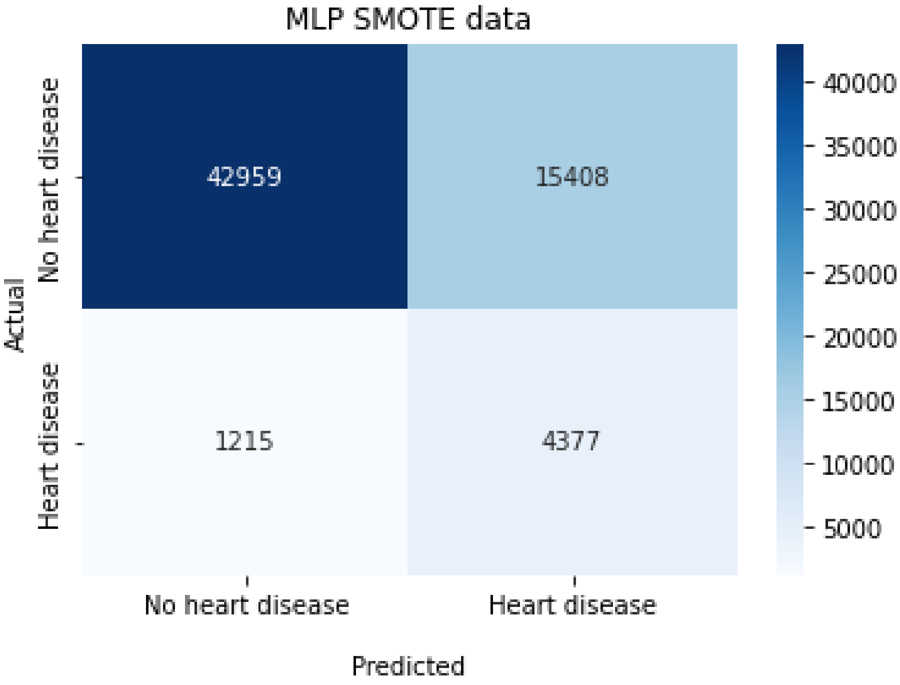

Figure 33: The confusion matrix of MLP SMOTE data

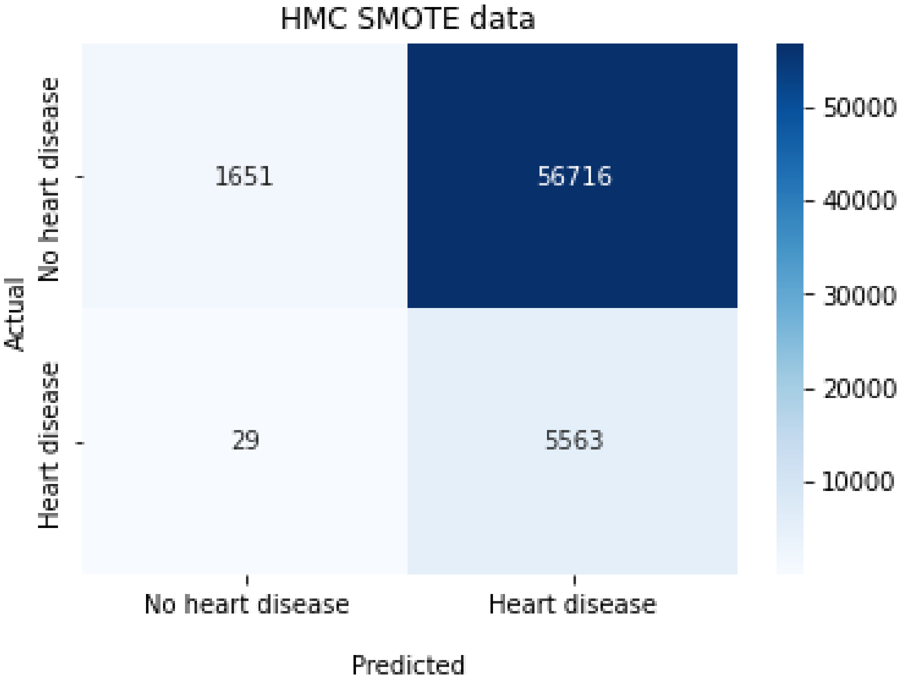

Figure 34: The confusion matrix of HMC SMOTE data

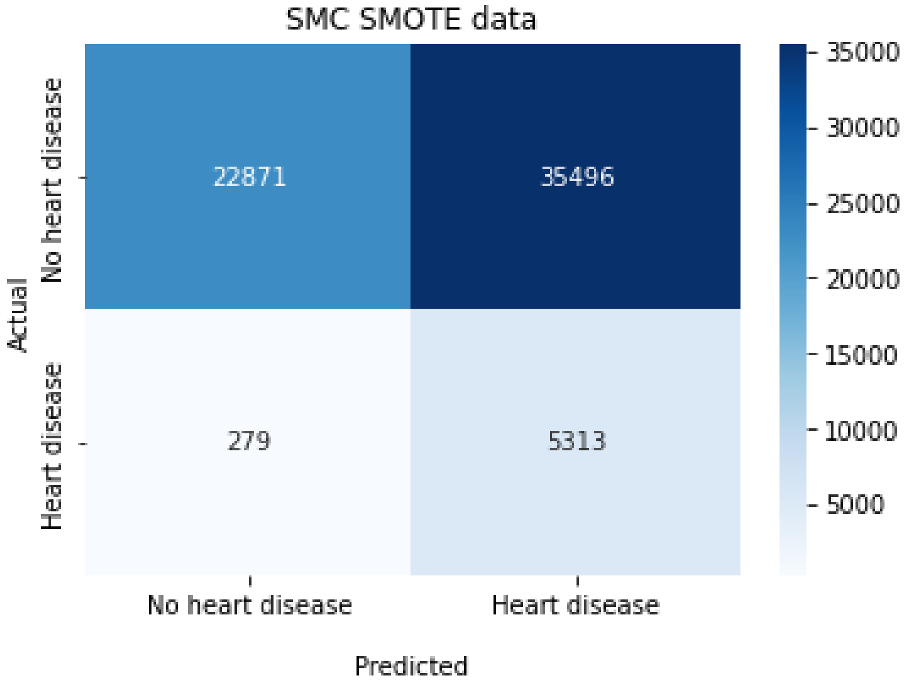

Figure 35: The confusion matrix of SMC SMOTE data

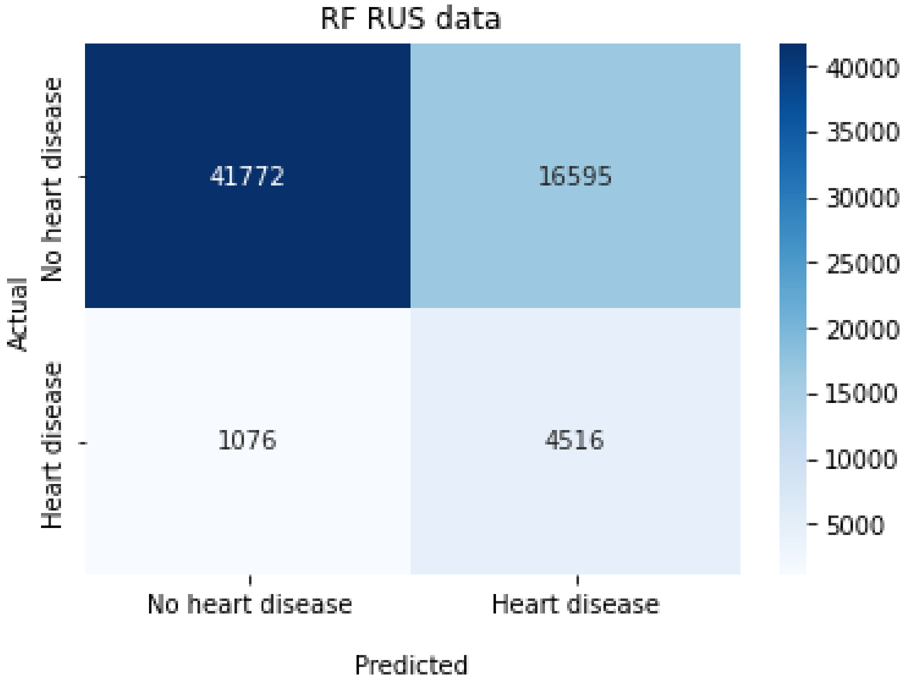

Figure 36: The confusion matrix of RF RUS data

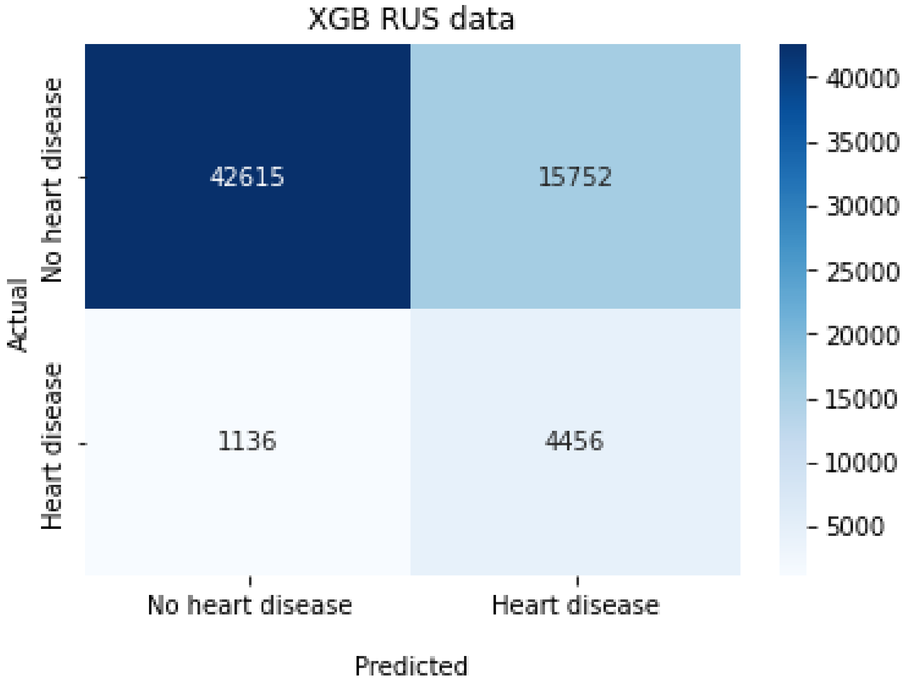

Figure 37: The confusion matrix of XGB RUS data

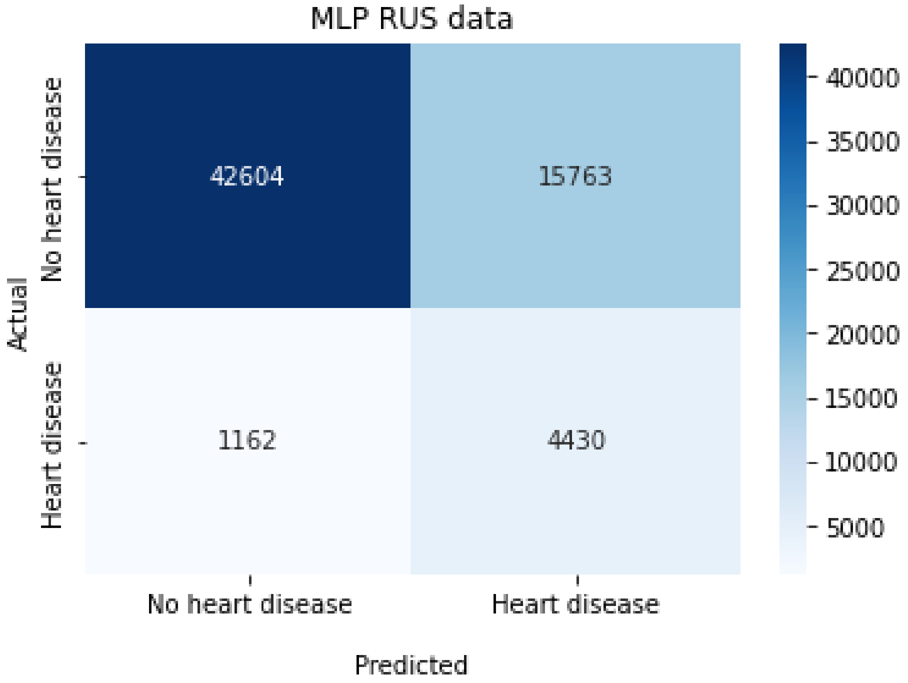

Figure 38: The confusion matrix of MLP RUS data

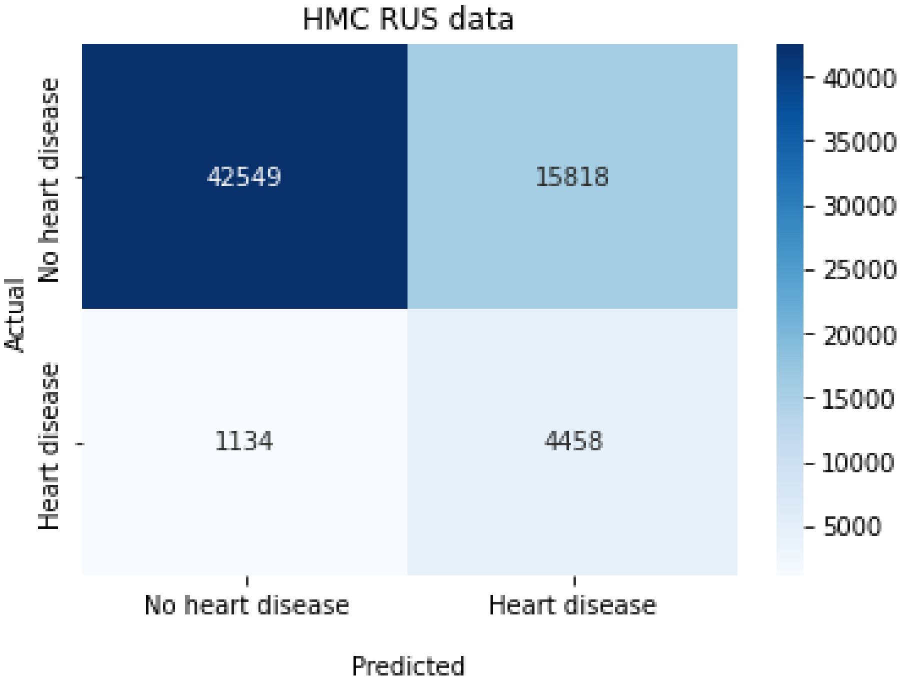

Figure 39: The confusion matrix of HMC RUS data

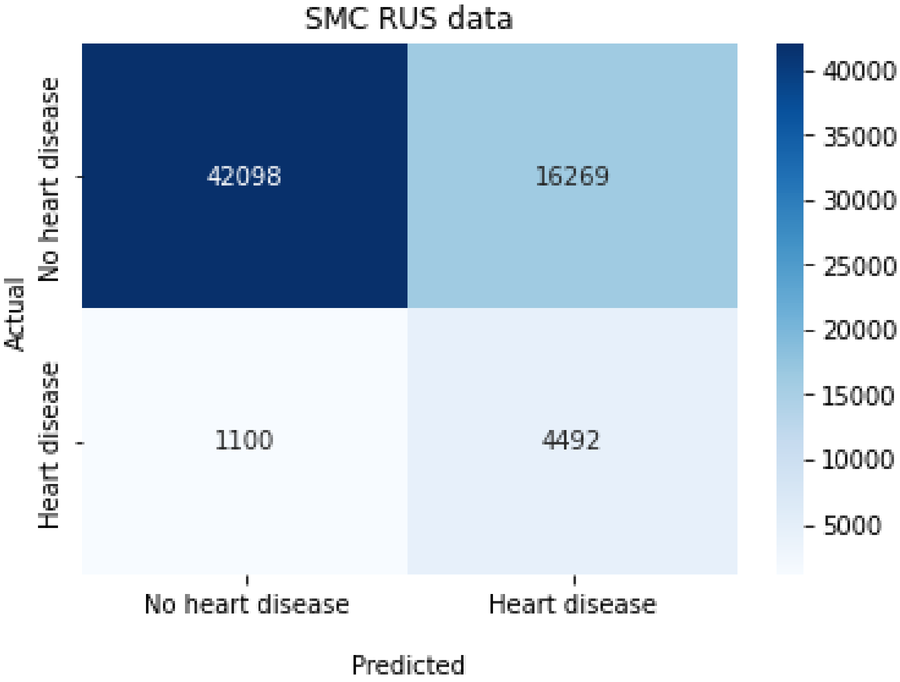

Figure 40: The confusion matrix of SMC RUS data