ISSN 2063-5346

INVESTIGATING THE IMPLICATIONS OF FEATURE SELECTION ON THE ACCURACY OF HEART DISEASE PREDICTIONS

Shashikala Toppo
M.Tech. Scholar
School of Engineering and I.T., MATS University, Raipur

Mr. Apurv Verma
Assistant Professor, School of Engineering and I.T., MATS University, Raipur

Dr. Vijayant Verma
Associate Professor
School of Engineering and I.T., MATS University, Raipur

Article History: Received: 03.04.2023 **Revised:** 25.04.2023 **Accepted:** 15.05.2023

Abstract

Coronary heart disease has emerged as one of the most debilitating conditions that can have a substantial influence on the quality of human life. During the past decade, it has become one of the main causes of death among individuals all over the world, making it one of the leading causes of mortality overall. It is important to arrive at a correct diagnosis of heart disease in a timely manner in order to protect patients from suffering additional complications. The use of non-invasive medical methods, such as those based on artificial intelligence and other forms of machine learning, has become increasingly common in recent years in the area of medicine. In particular, machine learning makes use of a number of algorithms and methods that are not only popular but also quite helpful in effectively identifying cardiac disease in a shorter length of time. However, determining whether or not someone has cardiac disease is a difficult undertaking. As the amount of medical datasets continues to grow, it has become increasingly difficult for medical practitioners to grasp the intricate feature interactions and to accurately forecast disease. In light of this, the purpose of this study is to extract from a dataset with a high number of dimensions the risk variables that contribute the most to an accurate classification of heart disease while simultaneously reducing the number of associated comorbidities. We have employed two different datasets on heart illness, each with its own unique set of medical characteristics, so that we may conduct a more comprehensive study. In the first step of our investigation, we examined the connection and interdependence of many medical characteristics in relation to heart disease. The second step that we took was to use a filter-based feature selection approach on both datasets. This allowed us to choose the characteristics that were most important (an optimum reduced feature subset) when it came to diagnosing heart disease. In the end, many different classification models for machine learning were tested utilising a complete and a reduced features subset as inputs for the experimental analysis. Accuracy, the Receiver Operating Characteristics (ROC) curve, and the F1-Score were the metrics that were utilised in the analysis of the trained classifiers. The classification outcomes of the models demonstrated that there is a significant influence that relevant characteristics have on the accuracy of the categorization. The performance of the classification models increased dramatically with a decreased training period in comparison to models that were trained on complete feature sets, even if the amount of features included in the models was reduced.

Keyword: *Coronary heart disease, Mortality, Patients, Diagnosis, Artificial Intelligence, Machine Learning, Accuracy, Receiver Operating Characteristics (ROC) Curve, F1-Score, Classification Models.*

1. Introduction

The prevalence of heart disease is increasing precipitously in every region of the world. Heart disease was responsible for the deaths of nearly 17.90 million individuals in 2016, according to a study report that was issued by the World Health Organisation (WHO) [1]. This much number is responsible for around 30 percent of all fatalities throughout the world. Nearly 55 percent of heart patients pass away within the first three years of their diagnosis, and the expenditures of treating heart disease account for around 4 percent of the yearly spending on healthcare. [2]. Keeping in mind the rising numbers, it is extremely important to diagnose and treat this dangerous condition in a precise and timely manner. Doing so is highly important for both the prevention of sickness and the efficient use of medical resources.

The area of medical sciences has witnessed a significant progress throughout the course of time [3, 4], which may be attributed to the current technological breakthroughs. Particularly, the use of machine learning, also known as ML, has seen widespread usage in the field of cardiovascular medicine, which has led to the creation of a potential sector [5]. The fundamental building blocks of machine learning are known as models. These models start with some kind of input data, such as text or photos, and then apply various statistical tests and mathematical optimisations in order to provide the desired prediction outcomes (such as disease, no disease, or neutral). [6] The basic framework of ML is built on models. ML models may be trained on tonnes of raw electronic medical data acquired from low-cost wearable devices, which enables more efficient detection of heart disease with less resources and greater accuracy [7].

To prevent themselves from becoming overly accurate, machine learning models need to be trained using a substantial amount of data [8]. The inclusion of a high number of data characteristics, on the other hand, is not essential because of issues connected to the "curse of dimensionality" [9, 10]. The majority of medical datasets cover aspects that are both related and redundant to one another. Unnecessary characteristics do not provide any relevant information to the prediction job, and they also cause noise in the description of the goal (output class), which leads to mistakes in the prediction process [11]. In addition, these characteristics make machine learning models more complicated, which in turn makes the system function more slowly owing to the higher amount of time spent training it. Only those features that are closely connected with the objective should be selected/identified from datasets and supplied as inputs to ML models [12]. This will help overcome the "curse of dimensionality," which occurs when there are too many features. The selection of relevant features can be helpful in performance improvement by reducing the complexity of the model and enhancing prediction accuracy, both of which are extremely essential in the field of medical diagnostics [13].

In the field of cardiovascular disorders and strokes, feature selection strategies are currently seeing widespread use [14], [15], and [16]. This can be attributed to the several advantages discussed earlier.

The following is a list of the contributions that this research has made:

- In order to provide a more comprehensive examination of medical characteristics, the study makes use of two datasets of individuals suffering from heart disease that come from separate sources.
- To carry out the correlation and dependency analysis between the various aspects of the datasets in terms of heart disease.
- The use of a filter-based method to the process of selecting medical characteristics, with the goal of locating those that are most helpful in the diagnosis and prognosis of heart disease.
- A variety of machine learning classification models, including Logistic Regression (LR), Decision Tree (DT), Naive Bayes (NB), Random Forest (RF), and Multi Layer Perceptron (MLP), among others, are applied to the datasets in order to determine which models are most suited to solve the problem.

In order to investigate the effect that feature selection has on the overall performance of the classification models, these models were evaluated using both the complete set of features and a subset that had less features.

2. Related work

This literature review examines a collection of research papers focused on diverse topics within the field of data analysis and machine learning applied to healthcare. The selected references cover a range of subjects, including missing data imputation techniques, feature selection methods, prediction of cardiovascular diseases, and the impact of class imbalance on model performance. The review provides a comprehensive analysis of each paper, discussing their methodologies, findings, and contributions to the respective domains. By exploring these studies, we gain valuable insights into the challenges and advancements in data analysis techniques and their application in healthcare decision-making. The synthesis of these papers enhances our understanding of the state-of-the-art methodologies and provides valuable knowledge for researchers and practitioners in the fields of medical informatics, machine learning, and healthcare analytics.

Buda et al. (2018) conducted a systematic study on the class imbalance problem in convolutional neural networks (CNNs). The research focuses on the challenges posed by imbalanced datasets in CNN-based classification tasks. The authors investigate the impact of class imbalance on model performance and propose strategies to mitigate its effects. The study provides valuable insights into the handling of imbalanced datasets in CNNs and offers recommendations to improve classification accuracy in such scenarios.

Gopika (2018) presented a correlation-based feature selection algorithm for machine learning. The study addresses the problem of selecting relevant features from high-dimensional datasets. The proposed algorithm utilizes correlation measures to identify the most informative features for machine learning models. The author demonstrates the effectiveness of the algorithm in reducing dimensionality and improving the performance of machine learning classifiers. The study provides a valuable contribution to feature selection techniques in machine learning applications.

Haq et al. (2018) proposed a hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. The research focuses on developing an integrated approach that combines multiple machine learning techniques to enhance the accuracy of heart disease prediction models. The authors compare and evaluate the performance of different machine learning algorithms in predicting heart disease. The study demonstrates the potential of the proposed framework in improving the accuracy of heart disease prediction and providing decision support in healthcare settings.

Gavhane et al. (2018) investigated the prediction of heart disease using machine learning techniques. The study explores the application of machine learning algorithms for early detection and prediction of heart disease. The authors evaluate the performance of different classifiers and feature selection methods in heart disease prediction. The study highlights the potential of machine learning models in assisting medical professionals in diagnosing and managing heart disease.

Thara et al. (2019) focused on auto-detection of epileptic seizure events using a deep neural network with different feature scaling techniques. The study explores the use of deep neural networks to detect epileptic seizure events. The authors investigate different feature scaling techniques and evaluate their impact on the performance of the deep neural network model. The study demonstrates the effectiveness of deep learning approaches in automating the detection of epileptic seizure events.

Beunza et al. (2019) compared machine learning algorithms for clinical event prediction, specifically the risk of coronary heart disease. The study evaluates and compares the performance of various machine learning algorithms in predicting clinical events related to coronary heart disease. The authors assess the accuracy and reliability of the algorithms and provide insights into their suitability for clinical event prediction tasks. The study contributes to the selection and application of appropriate machine learning algorithms in the field of cardiovascular disease prediction.

Saranya and Asha (2019) presented a survey on big data analytics in healthcare. The study provides an overview of the applications and techniques of big data analytics in healthcare. The authors discuss various aspects, including data collection, processing, analysis, and privacy concerns in healthcare big data analytics. The survey highlights the potential benefits and challenges associated with utilizing big data in healthcare, offering insights into current trends and future directions.

Remeseiro and Bolon-Canedo (2019) presented a comprehensive review of feature selection methods in medical applications. The study focuses on the application of feature selection techniques in the field of medicine. The authors review various feature selection methods and their effectiveness in medical data analysis. The study provides insights into the strengths, limitations, and suitability of different feature selection techniques for medical applications.

Nwosu et al. (2019) addressed the prediction of stroke from electronic health records (EHRs). The study explores the use of EHR data for predicting stroke. The authors develop a predictive model using machine learning techniques and evaluate its performance in stroke prediction. The study demonstrates the potential of utilizing EHRs for early detection and prediction of stroke.

Zhang et al. (2019) focused on stroke risk detection and proposes an improved hybrid feature selection method. The study addresses the challenge of identifying relevant features for stroke risk detection. The authors propose an enhanced feature selection method and apply it to stroke risk detection models. The study shows that the improved feature selection approach can enhance the accuracy and efficiency of stroke risk detection.

Mishra et al. (2019) discussed the application of statistical tests, including Student's t-test, analysis of variance (ANOVA), and analysis of covariance (ANCOVA). The study highlights the utility of these statistical tests in medical research and data analysis. The authors provide an overview of the concepts, assumptions, and interpretations of these statistical tests, emphasizing their relevance in clinical and research settings.

Gokulnath and Shantharajah (2019) proposed an optimized feature selection method based on a genetic approach and support vector machine (SVM) for heart disease prediction. The study focuses on feature selection to improve the accuracy of heart disease prediction models. The authors employ a genetic algorithm to select the most informative features, and SVM is used as the classification model. The study demonstrates the effectiveness of the proposed approach in selecting relevant features for accurate heart disease prediction.

Stavseth et al. (2019) examined the impact of missing data on conclusions drawn from categorical questionnaire data. The study compares six different imputation methods for handling missing data and evaluates their effectiveness. The authors highlight the importance of appropriate handling of missing data and discuss the implications of different imputation techniques on research findings. The study provides insights into the potential biases introduced by various imputation methods in the analysis of categorical questionnaire data.

Bommert et al. (2020) addressed the benchmark for filter methods for feature selection in high-dimensional classification data. The study evaluates different filter methods for feature selection in high-dimensional datasets. The authors compare the performance of various filter methods and provide insights into their effectiveness. The benchmarking results can guide researchers in selecting appropriate feature selection methods for high-dimensional classification tasks.

Nalluri et al. (2020) focused on chronic heart disease prediction using data mining techniques. The study applies data mining techniques to develop a predictive model for chronic heart disease. The authors explore various data mining algorithms and evaluate their performance in predicting chronic heart disease. The study highlights the potential of data mining techniques in predicting and managing chronic heart disease.

Kumar et al. (2020) presented an analysis and prediction of cardiovascular disease using machine learning classifiers. The study employs machine learning classifiers to analyze and predict cardiovascular disease. The authors compare the performance of different machine learning algorithms and evaluate their accuracy in predicting cardiovascular disease. The study emphasizes the role of machine learning in cardiovascular disease prediction.

Aremu et al. (2020) proposed a machine learning approach to circumvent the curse of dimensionality in discontinuous time series machine data. The study addresses the challenges posed by high-dimensional, discontinuous time series machine data. The authors apply a machine learning approach to handle this type of data, enabling effective analysis and prediction. The study provides insights into mitigating the curse of dimensionality in the context of time series machine data.

Pathan et al. (2020) focused on identifying stroke indicators using rough sets. The study applies rough set theory to identify indicators of stroke. The authors propose a methodology that utilizes rough sets to analyze and extract relevant features related to stroke. The study highlights the potential of rough set theory in identifying important indicators for stroke diagnosis and prediction.

Pavithra and Jayalakshmi (2020) presented a review of feature selection techniques for predicting diseases. The study provides an overview of various feature selection techniques used in disease prediction models. The authors discuss the strengths and limitations of different techniques and highlight their applications in disease prediction. The review serves as a comprehensive guide for researchers in selecting appropriate feature selection techniques for disease prediction tasks.

Jain et al. (2021) validated clustering frameworks for electric load demand profiles. The study compares the performance of several clustering algorithms for electric load demand profiles and identifies the most effective algorithms for this task. The study concludes that clustering algorithms can be a useful tool for analyzing electric load demand profiles.

Zhang et al. (2021) proposed a heart disease prediction model based on the embedded feature selection method and deep neural network. The study uses a large dataset of patients with heart disease to develop and test a machine learning model that can predict the risk of heart disease with high accuracy. The study concludes that the proposed model can be a useful tool for predicting the risk of heart disease.

Das et al. (2021) focused on estimating ground-level nitrogen dioxide concentration from satellite data. The study explores the use of satellite data to estimate air pollution levels. The authors propose a methodology and model for estimating ground-level nitrogen dioxide concentration based on satellite data. The study concludes that the proposed approach can provide valuable insights into air pollution monitoring and management.

Reddy et al. (2021) presented a study on heart disease risk prediction using machine learning classifiers with attribute evaluators. The study aims to develop a model for predicting the risk of heart disease using machine learning techniques. The authors compare different machine learning classifiers and attribute evaluators to identify the most effective combination. The study concludes that the proposed approach can improve the accuracy of heart disease risk prediction.

Wang et al. (2021) investigated the influence of dimensionality reduction on heart disease prediction. The study examines the impact of dimensionality reduction techniques on the performance of heart disease prediction models. The authors evaluate different dimensionality reduction methods and assess their effect on the accuracy of heart disease prediction. The study concludes that dimensionality reduction can have a significant impact on the performance of heart disease prediction models.

Al Mehedi Hasan et al. (2021) focused on identifying prognostic features for predicting heart failure using a machine learning algorithm. The study aims to identify the most important features that contribute to the prediction of heart failure. The authors employ a machine learning algorithm to analyze a dataset and identify the prognostic features. The study concludes that the identified features can be valuable for predicting heart failure.

Hasan and Bao (2021) compared different feature selection algorithms for cardiovascular disease prediction. The study explores the effectiveness of various feature selection algorithms in predicting cardiovascular disease. The authors evaluate and compare the performance of different algorithms to identify the most suitable approach. The study concludes that certain feature selection algorithms can significantly improve the accuracy of cardiovascular disease prediction models.

Sachan et al. (2021) presented a study on evidential reasoning for preprocessing uncertain categorical data for trustworthy decisions, with applications in healthcare and finance. The study focuses on handling uncertain categorical data and proposes an evidential reasoning approach to preprocess the data. The authors apply this approach to healthcare and finance domains, demonstrating its effectiveness in improving decision-making. The study highlights the importance of addressing uncertainty in categorical data for reliable decision-making.

Wang et al. (2022) explored the application of machine learning missing data imputation techniques in clinical decision-making, using the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example. The study concludes that machine learning can be a useful tool in missing data imputation and can improve the accuracy of clinical decision-making.

Dev et al. (2022) presented a predictive analytics approach for stroke prediction using machine learning and neural networks. The study uses a large dataset of stroke patients to develop and test a machine learning model that can predict stroke risk with high accuracy. The study concludes that machine learning can be an effective tool for predicting stroke risk.

Sivapalan et al. (2022) proposes a lightweight neural network called ANNet for ECG anomaly detection in IoT edge sensors. The study focuses on developing a machine learning model that can be implemented on low-power IoT edge devices for real-time ECG anomaly detection. The study concludes that ANNet can achieve high accuracy in ECG anomaly detection with low computational cost.

Despite their relevance, one major drawback of existing works on heart disease prediction is the lack of systematic guidance when selecting the input features for the development of prediction models which is an important aspect in terms of predictive performance. Previous research proposals chose features mostly in an impromptu manner without incorporating latest medical research findings. Mostly the focus is on the prediction models and their final prediction performance. However, a very less attention is paid on the correlation between different medical features and their individual importance in the prediction of heart disease. A few works present analysis of medical features but for the purpose of heart disease detection only. This research aims at addressing the ineffective feature selection in previous studies on heart disease prediction. Two heart disease patient datasets collected from different sources were utilized in this research to cover a broader study of features related to heart disease and to identify various medical procedures. To further analyze the role of each parameter in the prediction task, we obtain the interdependence and importance of the collected set of medical features. A detailed analysis of ML models trained on both full and selected feature set is provided to analyze the impact of feature selection techniques on the prediction performance as well as the identification of suitable classifiers for the specified problem.

3. Proposed methodology

In this study work, the relevance of feature selection in the precise categorization of heart disease is brought to light. Figure 1 illustrates the process that would be followed by the suggested approach for predicting heart disease.

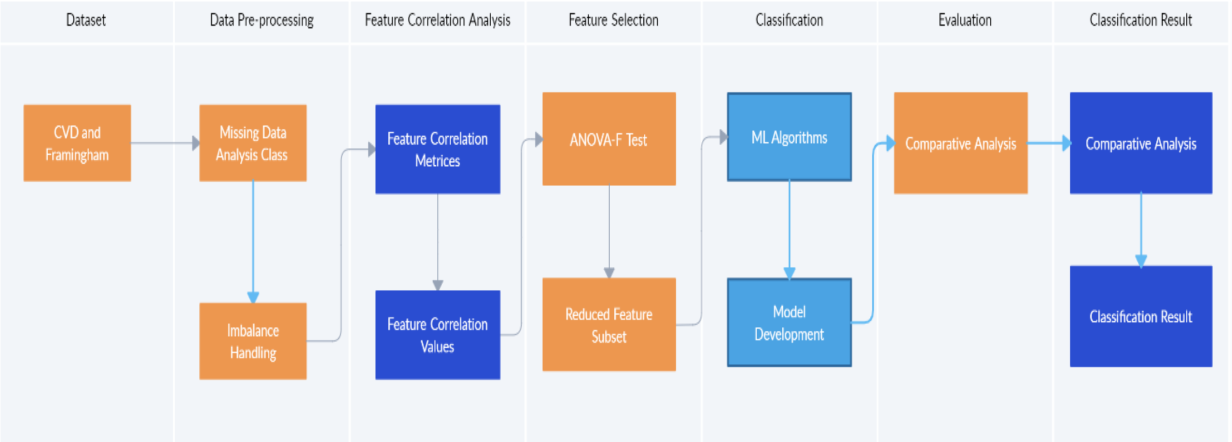

Figure 1: A flowchart of the suggested technique that outlines each stage in the process of predicting cardiovascular disease.

3.1. Datasets

This work made use of two datasets known as cardiovascular disease (CVD) and Framingham to investigate the influence of various variables on the incidence of heart disease and to construct a machine learning-based method for detecting heart disease. The research makes use of two different datasets to conduct a more comprehensive investigation of the many clinical pathways and medical characteristics that are used in the diagnosis of heart attacks and strokes. The datasets were obtained from a variety of different locations. The datasets included a number of important medical characteristics, including "age," "hypertension," "glucose levels," "blood pressure," and "cholesterol," amongst others, all of which are strongly associated with the development of illness and provide a great deal of flexibility for the investigation of heart disease. The datasets that were used were selected using two different criteria. The variability of the medical processes was the first criteria, and the purpose of this research was to investigate the various medical procedures and the significance of each aspect in the context of heart disease. Second, the availability of the data was taken into consideration while selecting the datasets. The quantity of data and the characteristics included within the datasets obtained from a variety of sources varies greatly. As a result, we have selected datasets that provide a sufficient amount of data and have a certain degree of similarity in terms of the properties they include.

3.1.1. CVD

The cardiovascular disease dataset was gathered as part of McKinsey & Company's healthcare hackathon, and it is under the company's management. The dataset is available from a free dataset repository, and it contains 29072 patient observations and 12 data characteristics. Eleven of them are typical clinical symptoms, and they are regarded as input characteristics. The twelfth feature, which is referred to as "stroke," is the target feature, and it indicates whether or not a patient has had a stroke. Table 1 contains an exhaustive explanation of all the data characteristics that make up the CVD dataset.

3.1.2. Framingham

The Framingham dataset was created during an ongoing cardiovascular study involving the residents of Framingham, Massachusetts, and is available on the Kaggle website.4 The dataset is most commonly used in classification tasks to determine whether or not a patient has a chance of developing coronary heart disease (CHD) in the next ten years. The dataset includes 4,240 patient records and 15 attributes, each of which represents a different risk factor. 14 different input factors were analysed in order to determine the decisional feature, which was the 10-year risk of coronary heart disease. The description of the data characteristics included in the Framingham dataset may be found in Table 2.

Table 1: A description of the characteristics included inside the CVD dataset.

Attribute	Description
i.d	patient’s i.d
gender	includes (“male”: 0, “female”: 1, “other”: 2)
age	patient’s age (continuous)
hypertension	suffering from hypertension (“yes”:1, “no”:0)
heart _disease	suffering heart disease (“yes”:1, “no”:0)

ever_married	marital status of patient (“yes”:1, “no”:0)
work_type	job status (“children”:0, “govt_job”:1, “never_worked”:2, “private”:3, “self_employed”:4)
residence_type	(“rural:0, ”urban”:1)
avg_glucose_level	average glucose level of blood (continuous)
bmi	body mass index (decimal value)
smoking_status	(“never smoked”:0, “formerly smoked”:1, “smokes”:2)
stroke	(“yes”:1, “no”:0)

Table 2: Detailed explanation of the characteristics of the Framingham dataset.

Attribute	Description
age	patient’s age (continuous)
male	(“male”:0, “female”:1)
education	level of education (1 to 4)
currentSmoker	(“smoker”:1, “non smoke”:0)
CigsPerDay	average number of cigarettes consumed per day (continuous)
BPMeds	on blood pressure medication (“yes”:1, “no”:0)
prevalentStroke	previous stroke history (“yes”: 1, “no”:0)
prevalenHyp	hypertensive (“yes”:1, “no”:0)
diabetes	previous diabetes history(“yes”:1, “no”:0)
totCHol	cholesterol level (continuous)
sysBP	systolic blood pressure (decimal)
diaBP	diastolic blood pressure (decimal)
BMI	body mass index (decimal)
HeartRate	heart rate measure (continuous)
glucose	glucose level (continuous)
TenYearCHD	target (“yes”: 1, “no”: 0)

3.2. Pre-processing

Data pre-processing is one of the essential steps in the machine learning life cycle since it simplifies data analysis, which in turn improves the algorithms' precision and efficiency [26]. Because the acquired dataset had issues with missing values and class imbalances, we ran several pre-processing processes to clean it up. In the case of the CVD dataset, there were a total of 43400 patient entries in the dataset, out of which 14754 had values that were either null or missing. On the other hand, the Framingham dataset had 4240 patient records, of which 645 values were blank. A value of null does not always suggest that the item does not exist; rather, it indicates that the value's location is uncertain. In medical datasets, the null or missing value is often caused by a lack of collection or the practitioner may not consider the observation since the medical test is believed to be of poor yield for the patient. In other words, the null or missing value is generally due to one of these two reasons. Methods of data imputation are beneficial when it comes to the management of missing data; nevertheless, their use in the medical area is restricted, and their particular effectiveness for illness diagnosis is unclear [27]. Because the usual techniques of data imputation are not adequate to represent the missing data complexity in health care applications [28], [29], researchers often do not consider the observations with missing values and purposefully discard the incomplete instances. However, the selection of the appropriate data imputation techniques can likely only be made with the assistance of an in-depth understanding of the condition in question. In accordance with the study that was presented, we removed all of the observations from both datasets that had a null value in order to prevent any accuracy biases.

In addition, by looking at the distribution of the classes, we can see that both datasets are very imbalanced in their natural state. Only 548 out of 29,072 patients in the CVD dataset were found to have diseases related to stroke, whereas 28,524 patients did not have any occurrences of stroke. Out of the 3101 patient records in the Framingham dataset, only 557 demonstrated a risk for coronary heart disease (CHD). During the process of training machine learning models [30], classification mistakes might occur as a result of the imbalanced nature of the datasets. Because of this, we decided to apply a method known as "Random Down-Sampling" in order to reduce the negative consequences that were brought on by the imbalanced data. We created two classes, one called the "majority" class, and one called the "minority" class. Patients who had cardiac illness were classified as belonging to a minority group, whilst patients who had no symptoms were classified as belonging to a majority group. In the instance of the CVD dataset, 548 observations were counted as part of the minority class, while the remaining 28,524 were taken into account as belonging to the majority class. We produced a balanced dataset consisting of 1096 observations by choosing 548 random instances from a total of 28,524 majority cases and all 548 observations from the minority class. The same approach was carried out on the Framingham dataset, which resulted in the generation of 557 random observations from 3101 majority instances, for a grand total of 1114 observations in the form of a balanced dataset. Two datasets of equal size were generated in this way for the purpose of conducting an effective investigation into the significance of the characteristics and the categorization of the illness.

3.3. Analysis of the association of features

A strategy that assists in gaining a grasp of the underlying connections between the many data characteristics that are included within a dataset is known as feature correlation. The correlation of features may be helpful in a variety of contexts, including the determination of the interdependencies between the data characteristics and the manner in which each feature influences the output feature. We were able to determine the correlation values between the data features by computing the correlation coefficients of the feature matrix M, which had the dimensions p and q and was represented as follows: M = [v1, v2,... , vq], where $v1$, $v2$, ... , and vq are the vectors that have a total of q characteristics each. The value p denotes the length of the vector, whereby each vector represents a finished medical treatment at a certain point in time. Figure displays the calculated correlation values that were found between various medical characteristics and the illness that was being studied for each dataset

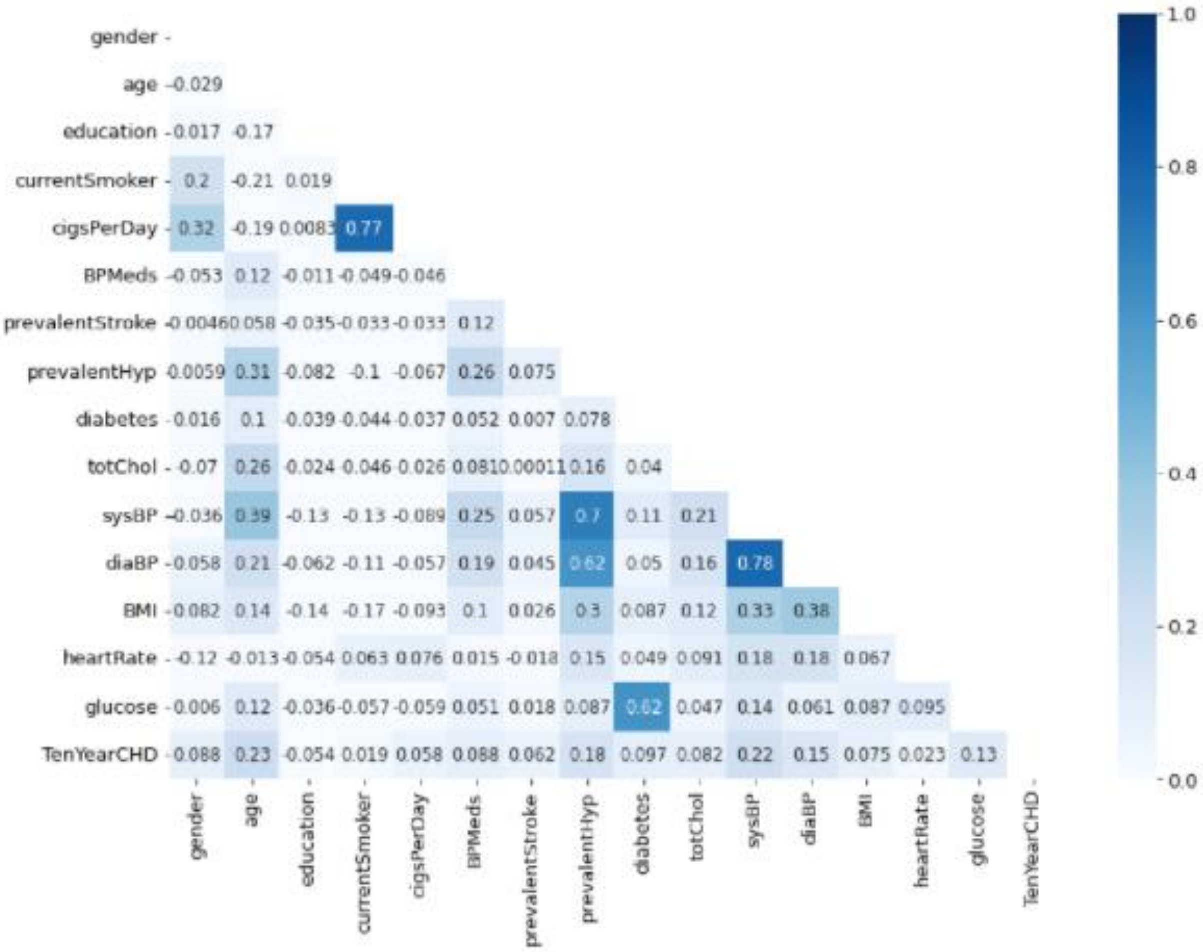

Figure 2: The values of each medical feature's connection with the target cardiac disease in both sets of data.

The evidence shown in Figure 2 out of the 11 characteristics that make up the CVD dataset have a positive connection with the decision feature, which is referred to as "stroke." When connected with'stroke,' the characteristics 'age,' 'hypertension,' and 'heart_disease,' and 'avg_glucose_lvl' had values of 0.57, 0.24, 0.27, and 0.2, respectively, indicating a substantial association between the two. Similarly, for framingham dataset, features ‘age’, ‘sysBP’, ‘prevalentHyp’, ‘diabBP’ and ‘glucose’ showed positive values of 0.23, 0.22, 0.18, 0.15 and reflect the motif of the desired output feature ‘TenYearCHD’. factors such as "gender," "bmi," and "heart rate," in addition to other non-medical factors such as smoking habits, education, social position, and living standards, exhibited a very low correlation with the output in both datasets, indicating that these features had either no influence on the output or a very low effect on the output. In general, the common medical characteristics that are included in both datasets, such as "age," "hypertension," and "glucose," are strongly connected with the result and may be regarded to be the significant risk factors.

According to the results of several medical studies, significant alterations may be seen occurring in both the heart and the blood arteries as a natural consequence of ageing. For instance, the pace at which your heart beats during any kind of physical exercise is not as quick as it might be when you were younger. According to a reliable source cited by the National Heart, Lung, and Blood Institute, the natural ageing process may increase a person's likelihood of developing cardiovascular disease. It is well known that hypertension is a risk factor for having a stroke, having ischemic heart disease, or having renal impairment. The reason of blood pressure that is higher than the usual range is hypertension. The increased blood pressure levels make the arteries less elastic, which lowers the flow of oxygenated blood and blood away from the heart, which increases the risk of developing a heart disease. Patients who have diabetes have a higher risk of developing coronary artery disease at an earlier stage. Diabetes generates high levels of glucose in the blood, which leads to a greater contraction of the blood vessels that regulate your heart and blood vessels, which ultimately leads to heart disease. This procedure, if left unchecked, might eventually result in a heart attack.

3.4. Choice of available features

The primary objective of this study is to identify the clinical characteristics that, when combined, have the potential to enhance the accuracy of heart disease prediction. The process of picking a subset of the most relevant characteristics from among a larger collection of original information, which ultimately have the greatest impact on the result, is referred to as feature selection. The benefits of feature selection include, but are not limited to: an increase in predictive performance; a reduction in the amount of computing time required by prediction models; an enhancement of data quality; and an efficient data gathering procedure.

In this study, we have used a filter-based feature selection approach known as the ANOVA-F test in order to determine the characteristics from both datasets that are the most significant. Filter-based feature selection strategies make use of statistical methods such as similarity, dependency, information, and distance to highlight the significant dependencies or correlations that exist between the input data and the target features. The Analysis of Variance (ANOVA) is a set of parametric statistical models and the estimate processes that go along with them. Its purpose is to establish whether or not the means of two or more sets of data are from the same distribution. The F-test, also known as the F-statistic, is a group of statistical tests that calculates the ratio of variance values, such as the variance of two independent samples, using certain statistical procedures. The F-test is also known as the F-statistic. An ANOVA f-test is a sort of F-statistic that refers to the Analysis of Variance (ANOVA) approach. It is a kind of univariate statistical test that compares each characteristic to the target feature in order to determine whether or not there is a link between the two that is statistically significant. The majority of the time, ANOVA is used in classification problems involving numerical input characteristics and categorical target features.

Using the f_classif() function that is made available by the scikit-learn module, the ANOVA-F test may be carried out in the python programming language. By using the SelectKBest class, the f_classif() method is called upon to determine which characteristics are the most significant (those that have the highest values). The technique known as selectKBest is made accessible by the scikit-learn library. It accepts a scoring function as an argument and then ranks the features according to the scores. In this case, the scoring function is f_classif(), often known as the ANOVA-F test, and we have constructed the SelectKBest class in order to determine which dataset characteristics are the most significant. The following equation may be used to get the ANOVA-F values:

$$variance\_between\_groups = \frac{\sum_{i=1}^{j} j_i(\bar{K}_i - \bar{K})^2}{(S-1)}$$

$$variance\_within\_groups = \frac{\sum_{i=1}^{s}\sum_{p=1}^{j_i} j_i(\overline{K}_{ip} - \overline{K}_\iota)^2}{(N-S)}$$

$$F\_value = \frac{\text{variance\_between\_groups}}{\text{variance\_within\_groups}}$$

where N is the total number of people in the sample, S is the total number of groups, ji is the total number of observations in the jth group, ki is the sample mean for the ith group, k is the overall mean of the data, and kip is the pth observation in the ith group out of S groups.

The results of the ANOVA-F test, which were used to calculate the feature significance scores, are shown in Figure. 3a and 3b, respectively, for both sets of data. In accordance with the data shown in Figure 3(a), the most essential characteristics for predicting'stroke' are ones that exhibit adequate scores when connected with the result. These features are 'age,' 'hypertension,' 'heart_disease,' and 'avg_glucose_lvl,' respectively. However, features ‘gender’, ‘bmi’, ’residence_type’ and ’smoking_status’ showed less or 0 significance for the feature ‘stroke’. When looking at 3 (b), we can see that the features 'age,' 'prevalentHyp,' 'diabetes,''sysBP,' 'diaBP,' and 'glucose' acquire the highest scores when compared to the other characteristics of the dataset that are associated to 'TenYearCHD,' which is a term that refers to the ten-year risk of developing CHD. When we look at the importance values of the features for each dataset, we can see that they are very similar to the correlation results that are listed in Section 3.3. This means that in the majority of cases, the features that are associated with age, hypertension, glucose, and blood pressure have a significant influence in the prediction of heart disease. According to the American Heart Association, the characteristics that were found using the ANOVA-F test are also mentioned as possible risk factors for heart disease.

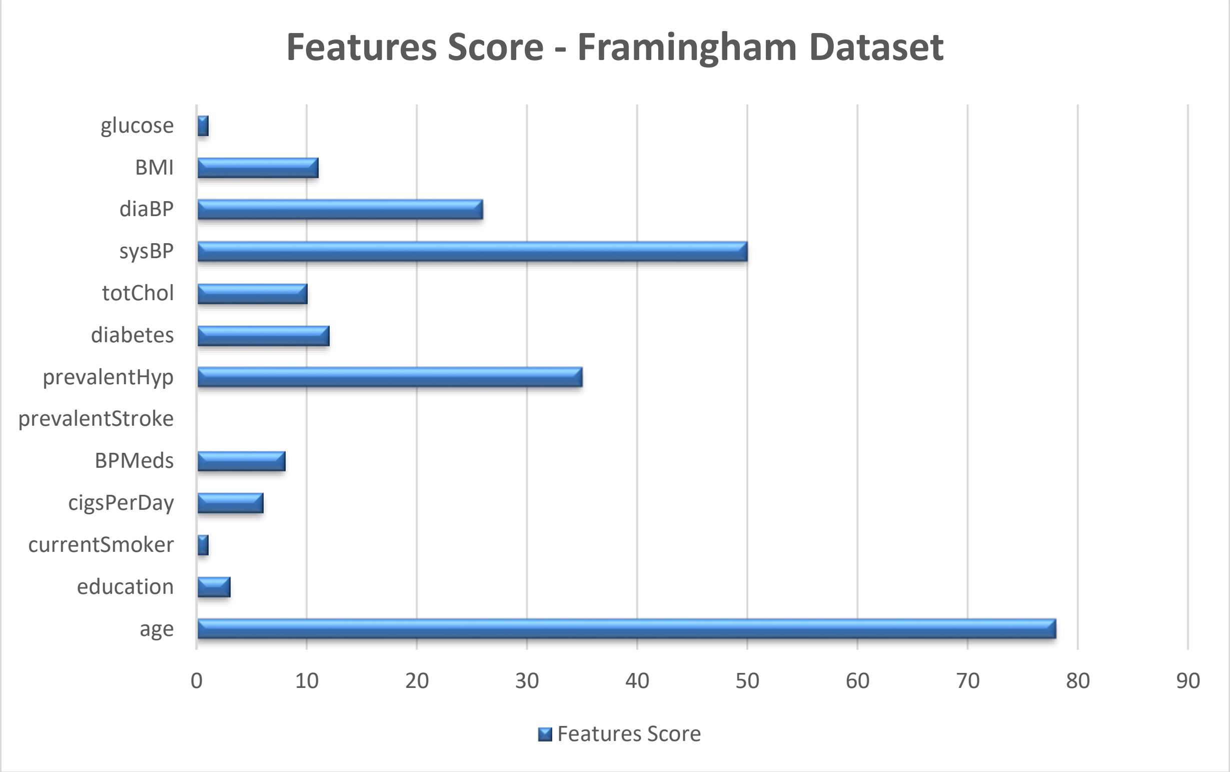

Figure 3: Score of features assigned to each characteristic for consideration in Framingham Dataset.

4. Evaluation matrices

In order to assess the effectiveness of machine learning classification models, we made use of three of the most used performance evaluation metrices, namely accuracy, F1-score, and ROC. A confusion matrix is a table that provides practitioners of machine learning with the ability to explain the performance of a classification model. The four categories that make up the confusion matrix are used to determine the performance matrices of a classifier. These categories are as follows: (1) True Positive (TP) test result that correctly classifies the presence of heart disease in patient; (2) True Negative (TN) test result that correctly classifies the absence of heart disease in patient; (3) False Negative (FN) test result that incorrectly classifies that a particular patient does not have heart disease; and (4) False Positive (FP) test result which incorrect Within the context of the medical industry, FN are regarded as the most dangerous forecasts. The accuracy of a given dataset of size n may be assessed using the formula:

$$Accuracy = (TP + TN)/(TP + FP + FN + TN)$$

The F1-Score is calculated by finding the harmonic mean of the Precision and Recall scores.

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

$$F1 - Score = 2(Precision \times Recall)/(Precision \times Recall)$$

The Receiver Optimistic Curves, often known as ROC, are used to investigate the classification abilities of a model. It assesses the "true positive rate" and the "false positive rate" in the output of a machine learning model.

$$TPR = TP / (TP + FN)$$

$$FPR = FP / (FP + TN)$$

5. The results as well as the conversations

In this part, we will evaluate the performance of the chosen categorization models from a variety of vantage points before moving on. To begin, we tested each model's performance independently on both datasets using all of the characteristics available in order to determine which models perform particularly well on each dataset. Second, we analysed the influence of the feature selection approach on the accuracy of the classifiers by evaluating the performance of the models using the chosen set of features and then evaluating how well the models performed. The Accuracy, F1-score, and ROC assessment matrices were used in order to examine the classifiers' overall performance.

5.1. The results of classification utilizing all available features

In this part of the article, all of the ML models were evaluated using the whole feature set in order to make a prediction about the binary illness outcome. All of the prediction models were trained using the whole dataset, with 80% of the data used for training and 20% used for testing. When training prediction models, the CVD dataset required 10.98 iterations per second (it/s), whereas the Framingham dataset required 24.20 iterations per second (it/s). This indicates a significant difference in the amount of total computing time required. The results of the ML model's binary classification analyses for predicting heart disease are shown in Figure 4 and 5, respectively, for both datasets.

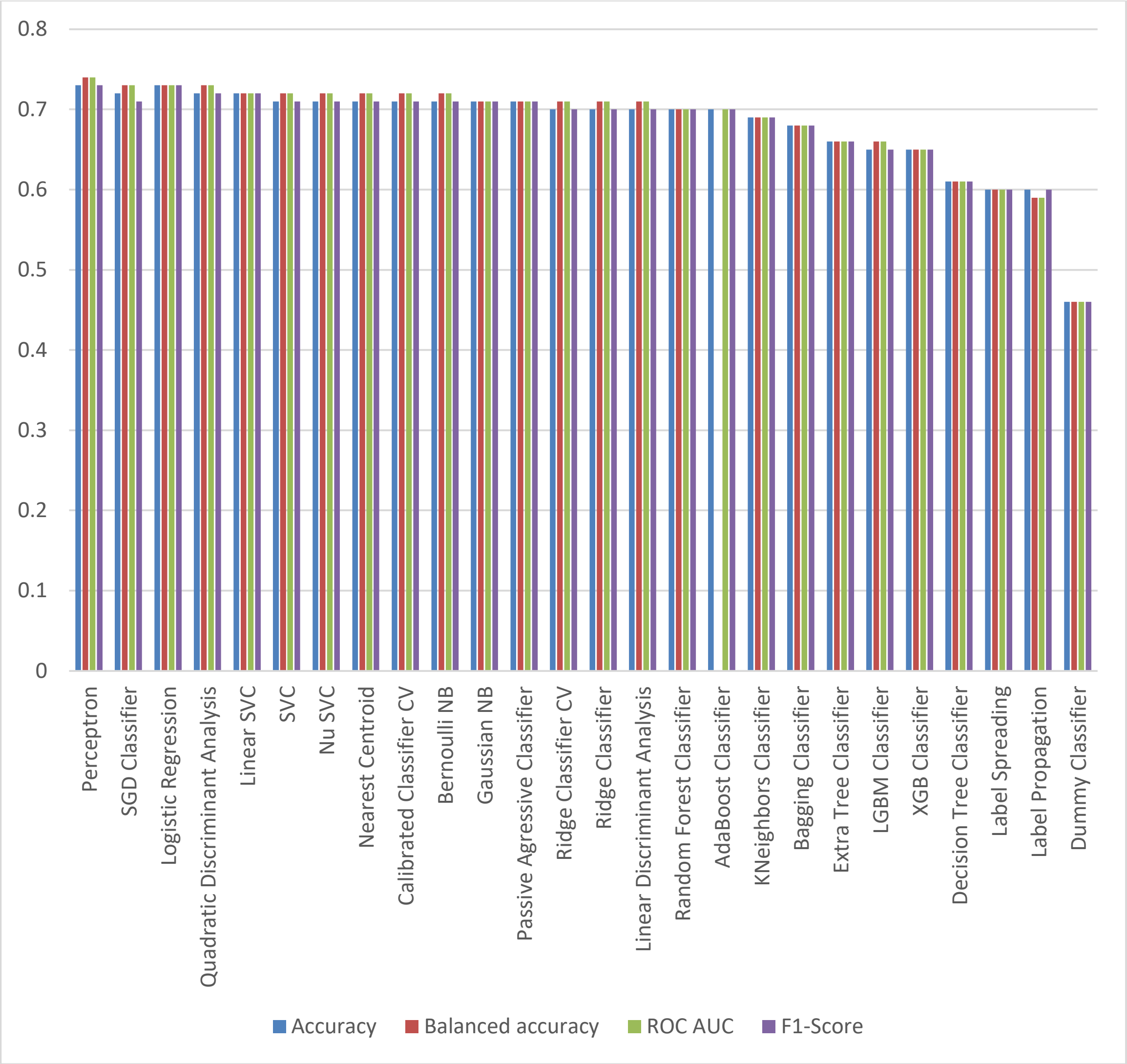

Figure 4: Classification results obtained from a variety of ML models using the whole feature set for the CVD dataset.

When looking at the classification results that are provided in Figure 4, the MLP algorithm for the CVD dataset got the best accuracy, which was 0.73, and it also had a ROC score of 0.74 and an F1-score of 0.73. Other classifiers, including as LR, SVC, and RF, functioned well and offered good prediction accuracy using the whole feature set. MLP was one of these classifiers. Because MLP is so effective at extracting patterns from convoluted medical datasets, it has been able to attain a higher level of accuracy than competing methods. In addition, this network model is effective at generalising data even when it lacks previous domain expertise. When it came to predicting heart attacks, the dummy classifier had the poorest results possible, with just a 0.46 accuracy rate. One of the possible explanations for the disappointing classification results is that the dummy classifier bases its predictions on too simplistic principles, which are of little use when addressing issues that arise in the actual world.

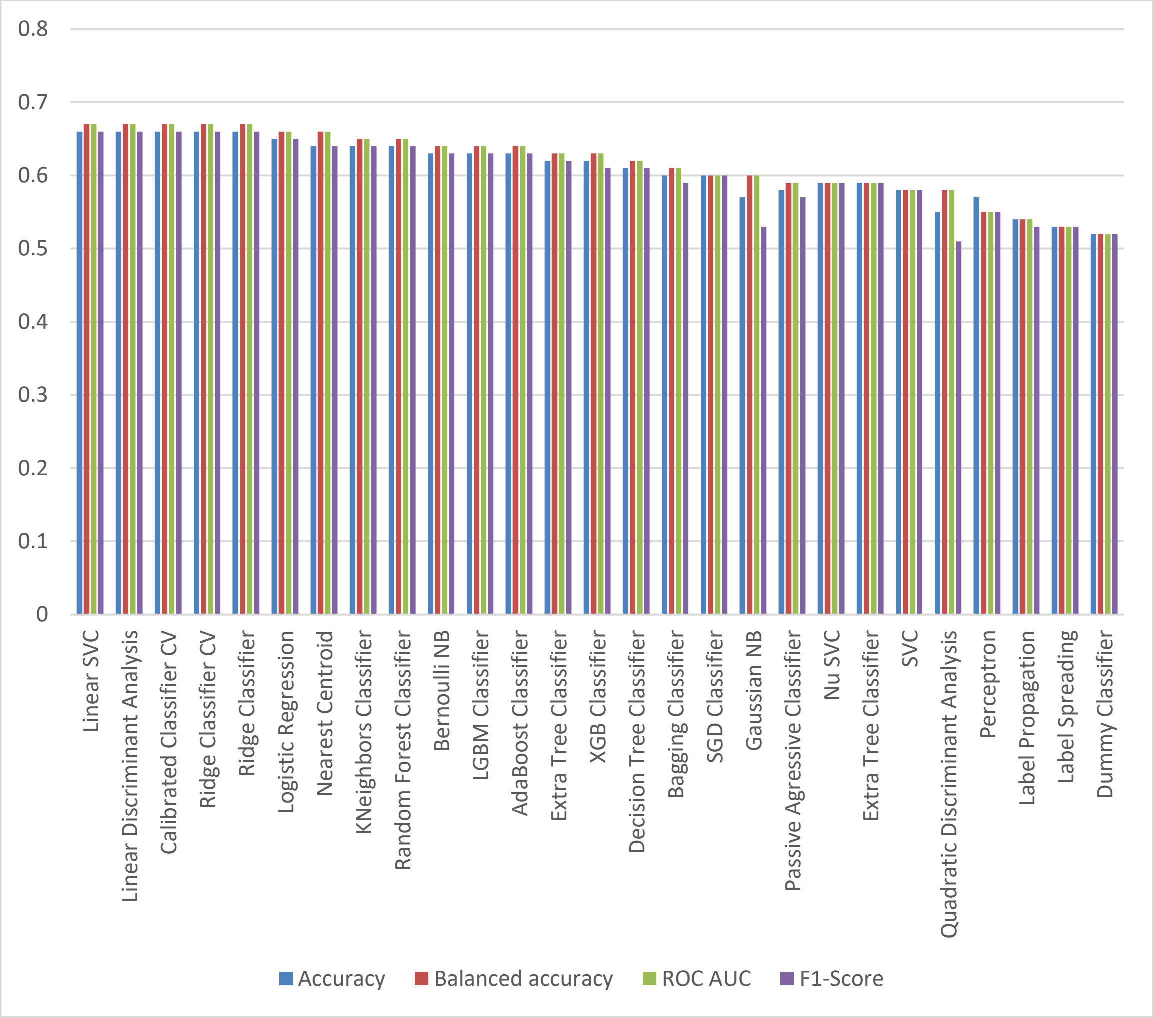

Figure 5: Results of multiple machine learning models' classifications of the Framingham dataset utilizing its whole feature set.

Figure 5 displays the categorization results obtained using the same methods for the Framingham dataset. The maximum accuracy that could be reached was 0.66, while the ROC was 0.67 and the F1-score was 0.66. These values do not indicate particularly excellent accuracy. Other methods such as linear discriminant analysis (LDA), linear regression (LR), and ridge classifier performed in a manner that was comparable. It's possible that the wide range of values between the data attributes is to blame for the unsatisfactory findings. Scaling of features helps to normalise data within a certain range, which may enhance the outcomes of models in general. However, any approach for manipulating the data that is used in medical research has the potential to generate major biases; as a result, we have not altered any of the feature values.

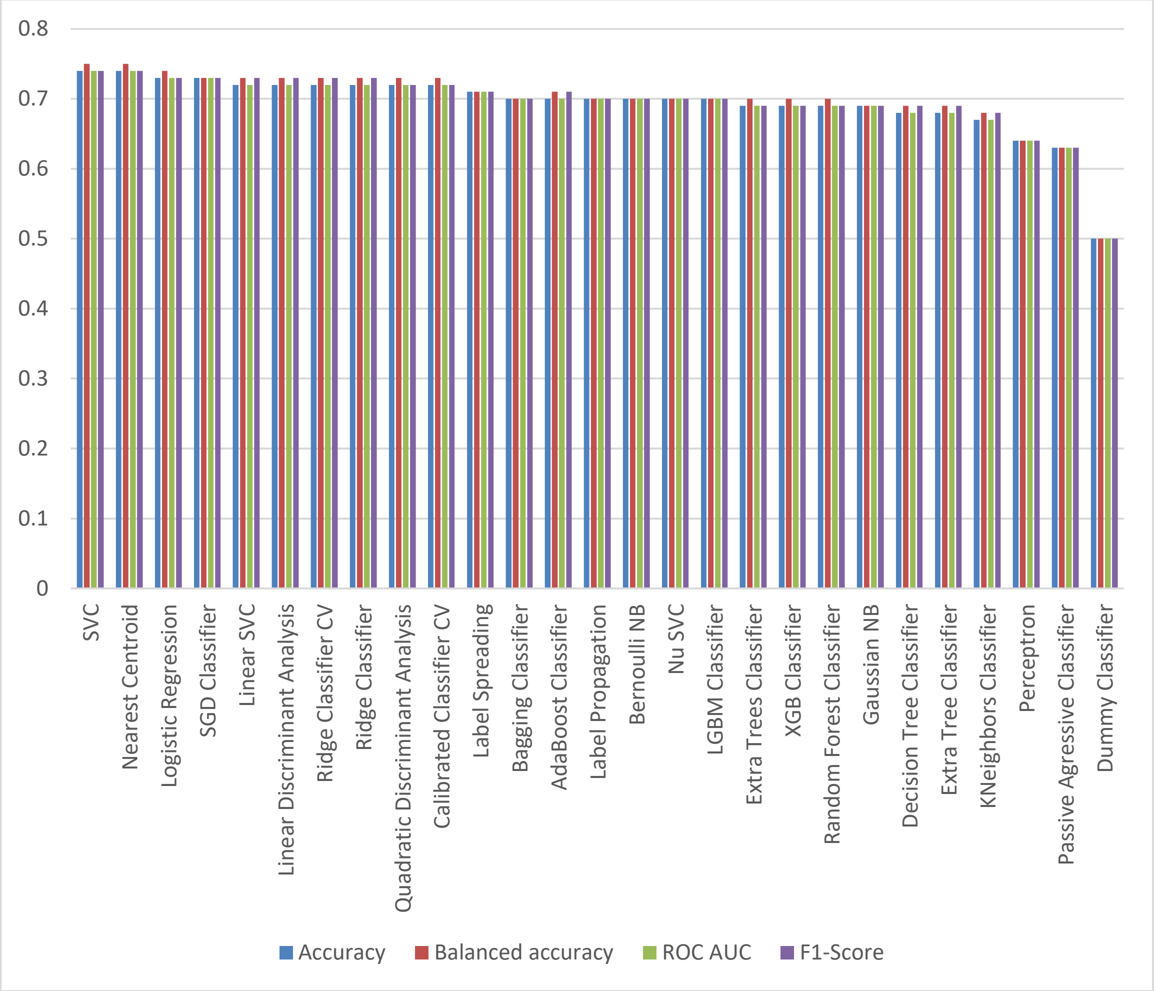

Figure 6: Results of classification performed by a variety of ML models on the CVD dataset using a reduced feature set.

5.2. The results of classification with a more limited feature set

Because we wanted to determine the possible biomarkers and investigate the effect that the method of feature selection had on the accuracy of classification, we chose the most prevalent features from the whole feature space using the individual feature scores as our guide. The ANOVA-F test was used to determine, as shown in Figure, which aspects of each dataset had the most influence on the final result. 3(a) and (b). Out of the 11 characteristics that were considered for the cardiovascular disease dataset, the following four were chosen: age, hypertension, heart disease, and average glucose level. Age, prevalentHyp, sysBp, diaBp, and glucose were the only characteristics from the Framingham dataset that were selected after taking the feature weights derived from the ANOVA-F test into consideration. The dataset had a total of 15 features. We tested the performance of each classification model by feeding it just the characteristics that were chosen to be included. The classification performance of each model is shown in Figure 6, which uses the CVD dataset's reduced feature subset as its data source. According to the findings of the study, the performance of ML models was superior to that of models that made use of the whole feature set even after the number of features had been restricted. The SVC model with just 4 input characteristics got the best accuracy, which was 0.74. This model also had an F1-score of 0.74 and a ROC of 0.74. Taking into consideration the findings in Figure 7, the maximum accuracy that could be reached is 0.71, which is greater than any of the other accuracy values that were obtained utilising the whole feature set for the Framingham dataset.

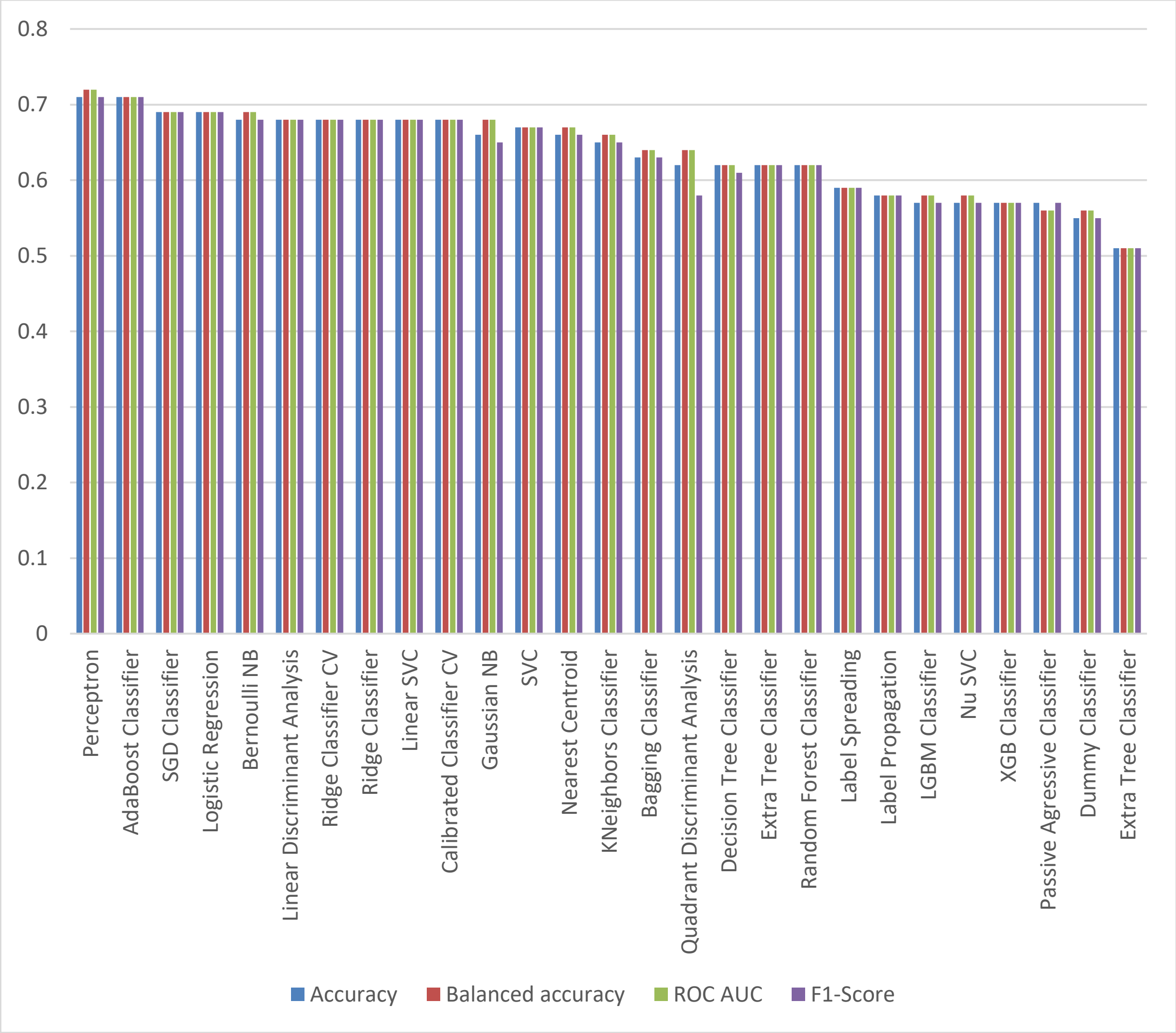

Figure 7: The classification results of multiple machine learning models applied to the Framingham dataset using the decreased feature set.

In addition, the models that were trained with a smaller feature set required much less computational time, with just 3.86 iterations per second (it/s) when using CVD and 15.52 iterations per second (it/s) while using the framingham dataset. Our findings have also been confirmed by comparing our work with previous published proposals, in which the same datasets were employed with full feature sets, and the resultant accuracy results were either less than or equivalent to the results that we achieved using reduced feature set. This comparison helped us determine that our findings are accurate. In general, the findings of the experiments demonstrated that the performance of the ML models improved significantly when they were restricted to employing just the relevant characteristics. In addition, while training classification models with the smaller feature set, a lower number of computational iterations per second (it/s) was seen in action. These experimental findings clarify the notions regarding the influence that feature selection strategies have, namely that it not only decreases the size of the feature space, but it also enhances the performance of machine learning models in a variety of different ways.

6. Conclusions and suggestions for further research

Heart disease is the most deadly illness that is fast on the rise and has become one of the leading causes of mortality all over the globe. If appropriate treatment approaches are used at the early phases of this condition, it may be possible to considerably lessen the harm that is produced by the illness. The prediction of cardiac disease and the identification of its most relevant characteristics are the subjects of this study. The primary objective of this research project is to investigate the effect that different feature selection strategies have on the overall performance of ML models. This research was carried out using the CVD dataset as well as the Framingham heart disease dataset, both of which may be found online. In the beginning of our investigation, we carried out a stage known as data preprocessing. our step consisted of transforming the data, cleaning the data, and balancing the data. In the second step of the process, we employed a filter-based feature selection approach called the ANOVA-F test to determine which datasets included the most relevant information for making accurate predictions on heart disease. The individual feature scores were used in conjunction with the ANOVA-F test in order to determine which characteristics were the most related to the results of both datasets. Using both datasets, we discovered that age, hypertension, hyperglycemia, prior heart disease, and blood pressure were shown to represent the most significant risk factors for heart disease, other than the conventional variables. This was the case even though the traditional factors were considered to be the most important risk factors. In addition, the classification tests were carried out using the whole feature sets in addition to the reduced feature sets in order to investigate the impact that certain characteristics had on the level of accuracy achieved by the different machine learning prediction models. When using all of the available features, the greatest level of accuracy that could be attained was 0.73 for cardiovascular disease and 0.66 for the Framingham heart disease dataset. Following the use of the simplified feature set, the accuracy improved to 0.75 for one dataset and 0.71 for the other. According to the findings of the study, the performance of ML models was superior to that of models that used a complete feature set even after the amount of features used in the models was reduced. The findings of the experiments indicate that by using a strategy for selecting characteristics, we are able to effectively identify heart disease despite the limited number of variables that are being used and in a shorter amount of time. We are able to draw the conclusion that by utilising the feature selection method, only the most significant aspects that are associated with heart disease are picked. This helps to minimise the amount of computational complexity and enhance the accuracy of prediction models. In the work that we want to do in the future, we are going to strive to improve the accuracy of our predictions by using a wide variety of machine learning and deep learning models in order to come up with the most effective model that is practical for the detection of heart disease. As part of our ongoing effort, we will evaluate the quality of our analysis using supplementary data sets. In addition, we will attempt to employ more than one approach for selecting features in order to generate more practicable feature subsets that are more directly associated with medical research.

References

[1] Wang, H., Tang, J., Wu, M., Wang, X., & Zhang, T. (2022). Application of machine learning missing data imputation techniques in clinical decision making: Taking the

discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example. BMC Medical Informatics and Decision Making, 22(1), 13. https://doi.org/10.1186/s12911-022-01752-6

[2] Dev, S., Wang, H., Nwosu, C. S., Jain, N., Veeravalli, B., & John, D. (2022). A predictive analytics approach for stroke prediction using machine learning and neural networks. Healthcare Analytics, 2, article 100032. https://doi.org/10.1016/j.health.2022.100032

[3] Sivapalan, G., Nundy, K. K., Dev, S., Cardiff, B., & John, D. (2022). ANNet: A lightweight neural network for ECG anomaly detection in IoT edge sensors. IEEE Transactions on Biomedical Circuits and Systems, 16(1), 24–35. https://doi.org/10.1109/TBCAS.2021.3137646

[4] Jain, M., AlSkaif, T., & Dev, S. (2021). Validating clustering frameworks for electric load demand profiles. IEEE Transactions on Industrial Informatics, 17(12), 8057–8065. https://doi.org/10.1109/TII.2021.3061470

[5] Zhang, D., Chen, Y., Chen, Y., Ye, S., Cai, W., Jiang, J., Xu, Y., Zheng, G., & Chen, M. (2021). Heart disease prediction based on the embedded feature selection method and deep neural network. Journal of Healthcare Engineering, 2021, 6260022. https://doi.org/10.1155/2021/6260022

[6] Das, B. P., Pathan, M. S., Lee, Y. H., & Dev, S. (2021). Estimating ground-level nitrogen dioxide concentration from satellite data. In Proceedings of the Photonics and Electromagnetics Research Symposium (PIERS) (pp. 1176–1182). IEEE Publications. https://doi.org/10.1109/PIERS53385.2021.9694752

[7] Reddy, K. V. V., Elamvazuthi, I., Aziz, A. A., Paramasivam, S., Chua, H. N., & Pranavanand, S. (2021). Heart disease risk prediction using machine learning classifiers with attribute evaluators. Applied Sciences, 11(18), 8352. https://doi.org/10.3390/app11188352

[8] Wang, G., Lauri, F., & Hajjam El Hassani, A. H. (2021). A study of dimensionality reduction's influence on heart disease prediction 12th International Conference on Information, Intelligence, Systems y Applications, IISA (pp. 1–6). IEEE Publications. https://doi.org/10.1109/IISA52424.2021.9555550

[9] Al Mehedi Hasan, M., Shin, J., Das, U., & Yakin Srizon, A. (2021). Identifying prognostic features for predicting heart failure by using machine learning algorithm 11th International Conference on Biomedical Engineering and Technology (pp. 40–46). https://doi.org/10.1145/3460238.3460245

[10] Hasan, N., & Bao, Y. (2021). Comparing different feature selection algorithms for cardiovascular disease prediction. Health and Technology, 11(1), 49–62. https://doi.org/10.1007/s12553-020-00499-2

[11] Sachan, S., Almaghrabi, F., Yang, J.-B., & Xu, D.-L. (2021). Evidential reasoning for preprocessing uncertain categorical data for trustworthy decisions: An application on healthcare and finance. Expert Systems with Applications, 185, article 115597.

https://doi.org/10.1016/j.eswa.2021.115597

[12] Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics and Data Analysis, 143, article 106839. https://doi.org/10.1016/j.csda.2019.106839

[13] Nalluri, S., Saraswathi, R. V., Ramasubbareddy, S., Govinda, K., & Swetha, E. (2020). Chronic heart disease prediction using data mining techniques Data Engineering and Communication. Technology. Springer, 903–912.

[14] Kumar, N. K., Sindhu, G. S., Prashanthi, D. K., & Sulthana, A. S. (2020). Analysis and prediction of cardio vascular disease using machine learning classifiers 6th International Conference on Advanced Computing and Communication Systems, ICACCS (pp. 15–21). IEEE Publications. https://doi.org/10.1109/ICACCS48705.2020.9074183

[15] Aremu, O. O., Hyland-Wood, D., & McAree, P. R. (2020). A machine learning approach to circumventing the curse of dimensionality in discontinuous time series machine data. Reliability Engineering and System Safety, 195, article 106706. https://doi.org/10.1016/j.ress.2019.106706

[16] Pathan, M. S., Jianbiao, Z., John, D., Nag, A., & Dev, S. (2020). Identifying stroke indicators using rough sets. IEEE Access, 8, 210318–210327. https://doi.org/10.1109/ACCESS.2020.3039439

[17] Pavithra, V., & Jayalakshmi, V. (2020). Review of feature selection techniques for predicting diseases. In Proceedings of the 5th International Conference on Communication and Electronics Systems, ICCES (pp. 1213–1217). IEEE Publications.

[18] Thara, D., PremaSudha, B., & Xiong, F. (2019). Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. Pattern Recognition Letters, 128, 544–550. https://doi.org/10.1016/j.patrec.2019.10.029

[19] Beunza, J. J., Puertas, E., García-Ovejero, E., Villalba, G., Condes, E., Koleva, G., Hurtado, C., & Landecho, M. F. (2019). Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). Journal of Biomedical Informatics, 97, article 103257. https://doi.org/10.1016/j.jbi.2019.103257

[20] Saranya, P., & Asha, P. (2019). Survey on big data analytics in health care International Conference on Smart Systems and Inventive Technology, ICSSIT (pp. 46–51). IEEE Publications. https://doi.org/10.1109/ICSSIT46314.2019.8987882

[21] Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. Computers in Biology and Medicine, 112, article 103375. https://doi.org/10.1016/j.compbiomed.2019.103375

[22] Nwosu, C. S., Dev, S., Bhardwaj, P., Veeravalli, B., & John, D. (2019). Predicting stroke from electronic health records. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference. Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE

Publications, 2019, 5704–5707. https://doi.org/10.1109/EMBC.2019.8857234

[23] Zhang, Y., Zhou, Y., Zhang, D., & Song, W. (2019). A stroke risk detection: Improving hybrid feature selection method. Journal of Medical Internet Research, 21(4), article e12437. https://doi.org/10.2196/1243

[24] Mishra, P., Singh, U., Pandey, C. M., Mishra, P., & Pandey, G. (2019). Application of Student's t-test, analysis of variance, and covariance. Annals of Cardiac Anaesthesia, 22(4), 407–411. https://doi.org/10.4103/aca.ACA_94_19

[25] Gokulnath, C. B., & Shantharajah, S. P. (2019). An optimized feature selection based on genetic approach and support vector machine for heart disease. Cluster Computing, 22(S6), 14777–14787. https://doi.org/10.1007/s10586-018-2416-4

[26] Stavseth, M. R., Clausen, T., & Røislien, J. (2019). How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. SAGE Open Medicine, 7, article 2050312118822912. https://doi.org/10.1177/2050312118822912

[27] Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

[28] Gopika, N. ME. (2018). A.M.K. Correlation based feature selection algorithm for machine learning 3rd International Conference on Communication and Electronics Systems, ICCES (pp. 692–695). IEEE Publications.

[29] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems, 2018, 1–21. https://doi.org/10.1155/2018/3860146

[30] Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, K. (2018). Prediction of heart disease using machine learning International Conference on Electronics, Communication and Aerospace Technology, ICECA (pp. 1275–1278). IEEE Publications. https://doi.org/10.1109/ICECA.2018.8474922.