Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2776-2786 | ISSN:2251-6727

The Effect of Feature Subset Optimization on the Precision of Cardiovascular Disease Predictive Models

[1]Teja Sri Dharma Reddy Vanukuri, [2]Sohail Shaik, [3]Bala Akash Mutthavarapu, [4]K. B. V. Brahma Rao, [5]Sai Teja Naidu Vadranam, [6]S.Hrushi Kesava Raju

[1]C.S.Student, Department of Computer Science and Engineering, Koneru Lakshmaiah Education, Foundation(KLEF), Vaddeswaram, Guntur Andhra Pradesh

[2]C.S.Student, Department of Computer Science and Engineering, Koneru Lakshmaiah Education, Foundation(KLEF), Vaddeswaram, Guntur Andhra Pradesh

[3]C.S.Student, Department of Computer Science and Engineering, Koneru Lakshmaiah Education, Foundation(KLEF), Vaddeswaram, Guntur Andhra Pradesh

[4]Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education, Foundation(KLEF), Vaddeswaram,Guntur, Andhra Pradesh

[5]C.S.Student, Department of Computer Science and Engineering, Koneru Lakshmaiah Education, Foundation(KLEF), Vaddeswaram, Guntur,Andhra Pradesh

[6]Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education, Foundation(KLEF), Vaddeswaram, Guntur, Andhra Pradesh

(Received: 07 January 2024 Revised: 12 February 2024 Accepted: 06 March 2024)

KEYWORDS

Heart disease; Mortality; Diagnosis; Non-invasive medical techniques; Artificial intelligence; Machine learning; Cardiovascular disease; Risk factors; Datasets; Correlation; Feature selection; Optimal reduced feature subset; Classification models; Experimental analysis; Accuracy.

ABSTRACT:

In the last 10 years, heart disease has arisen as a significant worldwide health problem that substantially influences mortality rates. To avoid patients from any harm, prompt and accurate testing is vital. In order to recognize cardiovascular illness, non-invasive medical technologies—especially those that make use of artificial intelligence and machine learning—have become more crucial. However, heart illness prediction is still hard, particularly in light of the intricacy of vast volumes of medical data. With the purpose of boosting the accuracy of heart disease diagnosis, this study tries to find relevant risk indicators in high-dimensional datasets. Two distinct heart disease files with varied medical features were employed in order to accomplish this. The link and correlation between these features and heart disease was the principal focus of the investigation. After that, a filter-based feature selection strategy was incorporated to these datasets, providing a reduced feature group that may be utilized to identify heart disease. Many machine learning classification models that incorporated both reduced and complete feature groups were constructed for trial analysis. Learned models were assessed using precision, the Receiver Operating Characteristics (ROC) curve, and the F1-Score. The classification results indicated that important attributes had a substantial effect on accuracy, indicating that the models performed considerably better even with small amounts of data. Compared to models trained on the whole feature set, the reduced feature set considerably enhances classification accuracy while requiring less training time. This paper shows the relevance of feature selection in enhancing the performance of machine learning models for heart disease prediction.

I. Introduction

A World Health Organization (WHO) estimate suggests that heart illness is become increasingly common worldwide, with 17.90 million heart disease-related deaths projected in 2016—more than 30% of all fatalities globally [1]. Remarkably, 55% of cardiac patients pass away in the first three years, even if the annual cost of treatment amounts for just 4% of overall medical expenses [2]. Acknowledging this new trend, early heart disease identification and treatment are crucial to controlling the condition and decreasing healthcare expenses.

The background of medical research has radically transformed as a result of recent technology discoveries [3, 4]. It is an incredible accomplishment and a promise of a bright future that machine learning (ML) has emerged as a significant tool in the treatment of cardiovascular disease [5]. Machine learning (ML) models based on statistical analysis of input data may be employed to effectively diagnose cardiac sickness from vast volumes of electronic medical data supplied by low-cost smart devices [6, 7].

To reduce overfitting, ML models must be trained on a range of data sets [8]. However, integrating unequal data quality results in the curse of dimensionality and necessitates ongoing rewriting [9, 10]. Medical datasets typically comprise recurrent and related characteristics that enhance complexity without supplying the prediction function crucial information, which might cause the prediction function to be off [11]. Essential feature selection becomes important in medical diagnosis to reduce model complexity, increase prediction accuracy, and shorten training time in order to handle these issues [12, 13].

Due to these developments, feature selection algorithms are presently commonly applied in research on heart disease and stroke [14, 15, 16]. This work delivers a substantial contribution by extensively assessing medical symptoms in two separate datasets connected with heart disease. It focuses on the linkages and interdependencies between the various components of cardiac disease. It makes use of a filter-based feature selection technique to identify which medical signals are most effective for anticipating cardiac difficulties. Appropriate models are generated using a number of machine learning (ML) classification approaches, including Random Forest (RF), Decision Tree (DT), Naive Bayes (NB), Multi-Layer Perceptron (MLP), and Logistic Regression (LR).

Both full and reduced feature groups are applied to examine the classification models in order to learn how feature selection effects performance. The source code used in this study is publicly accessible on GitHub in compliance with the standards for reproducible research [1,].

II. LITERATURE SURVEY

While cardiac disease detection has improved using machine learning (ML), ML prediction models confront a fundamental difficulty owing to the high information complexity. In order to overcome this challenge, feature selection looks to be a crucial strategy. It focuses on determining the most critical components that strongly effect the course of disease. This strategy is useful for minimizing the probability of medical complications as well as enhancing forecast accuracy.

Zhang et al. employed filter-based approaches including Fisher score, information gain, and standard deviation to construct a new feature selection strategy called Weighting-and-Ranking-Based Hybrid Feature Selection (WRHFS) in an important research effort [17]. Nine of the 28 essential input parameters for the prediction of heart strokes were successfully discovered with this strategy. A separate research [18] applied the Latent Feature Selection (ILFS) approach to uncover critical risk variables for good heart disease prediction from a vast feature space. Despite using just half of the qualities from a group of fifty, our probabilistic latent graph-based feature selection technique fared better than the competitors.

Using the glow-worm swarm optimization approach, a separate research [19] presented a feature selection procedure for discovering quality characteristics in electronic health records (EHR). Notably, six traits—including high blood pressure and alkaline phosphatase—were revealed to be essential for identifying stroke. The authors of [20] concentrated on generating the most significant EHR signals to predict the early-stage risk of heart disease-related mortality. The recursive feature elimination (RFE) and minimum redundancy maximum relevance (mRmR) approaches generated features with an average accuracy of 80%.

Using the Cardiovascular Health Study (CHS) dataset, Singh et al. [21] constructed an effective stroke prediction strategy that obtained 97.7% accuracy by upgrading a feature set using principle component analysis (PCA) and the Decision Tree (DT) method. In order to uncover key variables for heart disease, [22] applied a wrapper-based Genetic Algorithm (GA), which produced an 88.34% accuracy with a reduced feature set.

To diagnose cardiac illness, ML and deep learning algorithms were integrated in a novel way [23]. With an astounding accuracy of 98.56%, the feature group was created using the least absolute shrinkage and selection operator (LASSO) penalty technique. A comprehensive machine learning (ML)-based heart disease detection system [5] displays higher classification performance in terms of processing time and accuracy by applying feature selection techniques.

Using a restricted set of attributes, this study [24] provides a two-stage feature group recovery strategy to identify heart disease. The XGBoost prediction coupled with a winding approach yielded the best projected outcomes. Correct feature group selection is critical, as proved by a study of several classifiers in [25], where the sequential minimum optimization (SMO) classifier had the highest accuracy of 86.468%.

Nonetheless, a persistent concern in current investigation is the absence of a systematic technique to assist identify input features for forecast models. By integrating two datasets of patients with heart illness, extensively investigating the elements associated to heart disease, and assessing the utility and accuracy of each parameter in the forecast job, this work answers a gap in the literature. A complete analysis of feature selection techniques' influence on prediction performance delivers essential information on which classifiers are authorized to be used for the specified purpose.

III. METHODOLOGY

This study paper stresses the critical function of feature selection in boosting the accuracy of heart disease categorization. The prospective strategy for heart disease prediction is illustrated in Figure 1.

A. Datasets

In this experiment, we utilized the Framingham dataset and the Cardiovascular Disease (CVD) dataset individually. The objective was to design a machine learning (ML)-based system for heart disease monitoring and to look at how various variables impact how heart disease develops. Critical medical indicators including "age," "hypertension," "glucose levels," "blood pressure," and "cholesterol," all of which are associated to the development of heart disease, are included in these data, which are gathered from a number of sources. There were two significant difficulties that impeded the dataset selection. We started by researching the variety of medical processes, specifically the multiple clinical approaches utilized to recognize heart attacks. Since varied sources provide varying volumes of data and feature collections, the second reason the datasets were selected was based on data availability. The approach presented datasets that supplied a large quantity of data volume while having some degree of feature closeness. This technique delivers a detailed assessment of clinical routes and medical components, establishing the foundation for a thorough evaluation of heart disease risk.

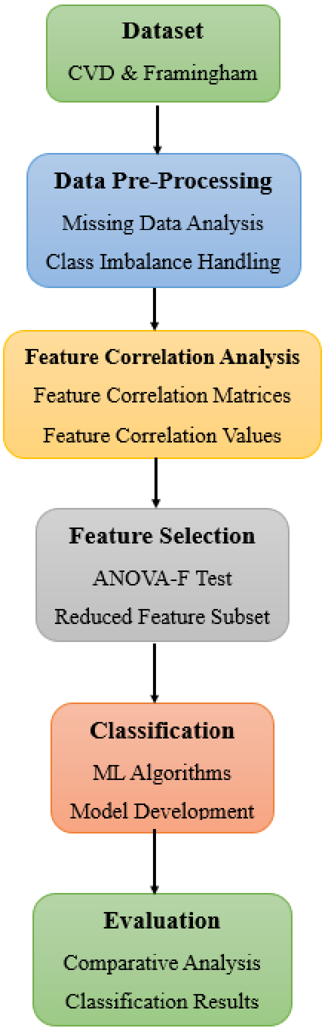
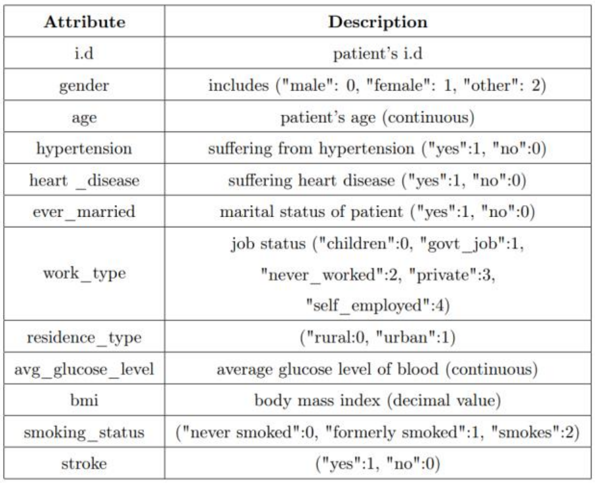


Fig 1. Diagram of the Suggested Methodology Showing Each Step in the Prediction of Heart Disease

B. Cardio Vascular Disease (CVD)

An open dataset source is now providing the CVD dataset, which McKinsey & Company received during their healthcare hackathon, to the larger public. 29,072 patient records with 12 data attributes are available in this collection. These eleven features are broad clinical indicators that are explored as input variables. The objective attribute that indicates whether or not a patient has had a stroke is the 12th feature, which is designated "stroke." For a thorough comprehension, Table 1 gives a detailed explanation of all the data components included in the CVD dataset.

Table I: Features Overview of CVD Dataset

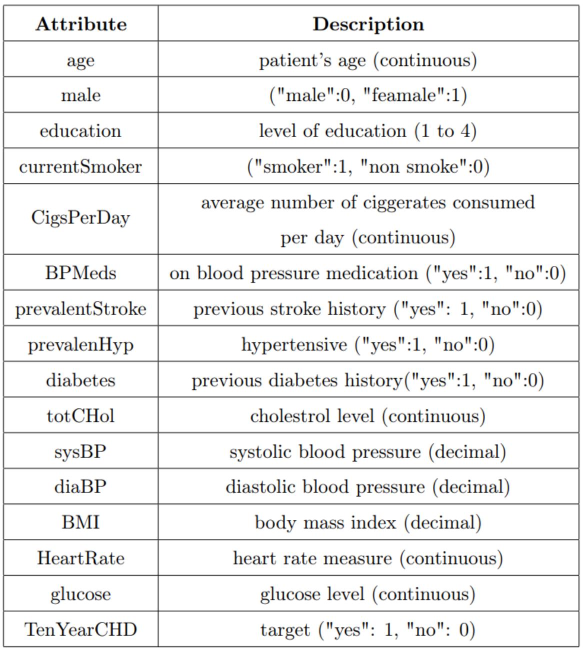


Attribute	Description
i.d	patient's i.d
gender	includes ("male": 0, "female": 1, "other": 2)
age	patient's age (continuous)
hypertension	suffering from hypertension ("yes":1, "no":0)
heart _disease	suffering heart disease ("yes":1, "no":0)
ever_married	marital status of patient ("yes":1, "no":0)
work_type	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)

C. *Cardiovascular*

The Framingham dataset, which may be supplied on the Kaggle website, is based on recent medical studies done among residents of Framingham, Massachusetts. This dataset tries to investigate a patient's 10-year risk of having coronary heart disease (CHD). It will largely be employed in classifying responsibilities. There are fifteen components in the compilation, each of which represents a different risk factor, and 4,240 patient data points. The Framingham dataset's 14 input components were applied to look into the 10-year risk of CHD. A full description of each data item is provided in Table 2.

Table 2: Features Overview of Framingham Dataset



Attribute	Description
age	patient's age (continuous)
male	("male":0, "feamale":1)
education	level of education (1 to 4)
currentSmoker	("smoker":1, "non smoke":0)
CigsPerDay	average number of ciggerates 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	cholestrol 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)

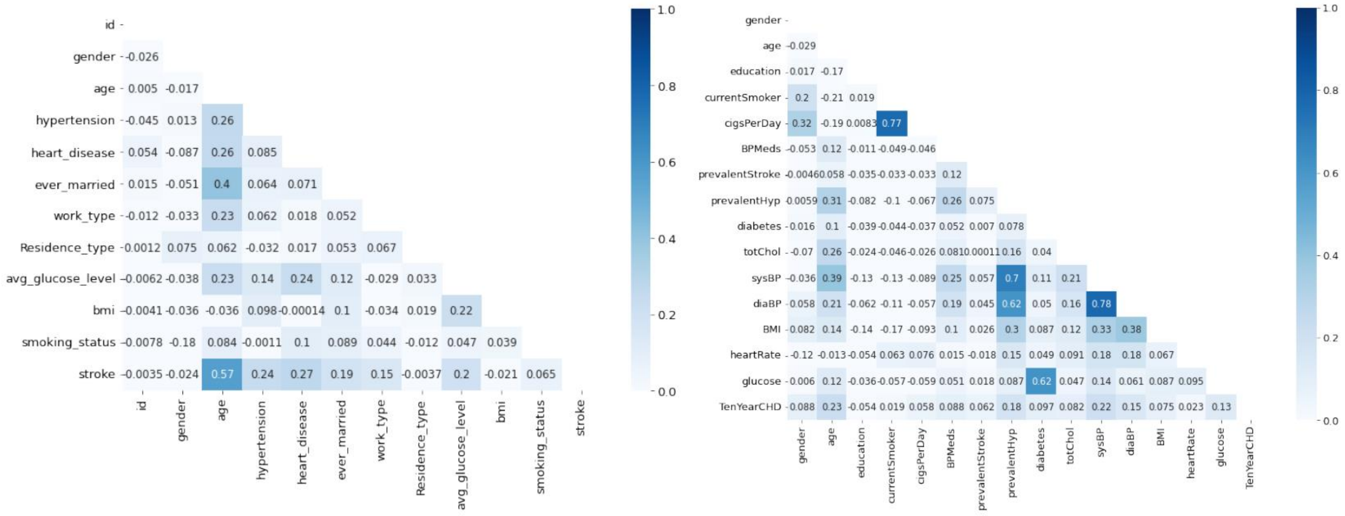
D. *Data Preparations*

Since data pretreatment enhances the accuracy and speed of machine learning algorithms and speeds up data analysis, it is a critical stage in the process [26]. To address challenges like missing values and class mismatch in the final datasets, a variety of preparatory measures were done. For instance, the Framingham dataset had 645 null values with 4,240 patient records, but the CVD dataset included 43,400 patient records with 14,754 missing or null values. In medical records, null values—which imply unknown or uncollected data—are regularly encountered due to practitioners' limited or inadequate observation. Data replacement algorithms are effective at filling up gaps in data, but their value in healthcare applications—especially for sickness detection—is dubious [27]. All entries with null values were completely eliminated from both datasets in this analysis to prevent any accuracy problems.

Additionally, in both datasets, there was a skewed class distribution. Out of 29,072 patients in the CVD dataset, only 548 experienced stroke symptoms; however, out of 3,101 occurrences in the Framingham dataset, 557 had a high risk of CHD. During the training of ML models, classification issues may develop due to the underlying unpredictability of datasets [30]. In order to overcome this, a "Random Downsampling" technique was devised, resulting to the split of the population into two classes: the "minority," which included persons with cardiac issues, and the "majority," which included those without symptoms. A balanced sample of 1,096 observations was constructed for the CVD dataset by keeping 548 minority observations and randomly choosing 28,524 majority examples. Applying a similar method to the Framingham dataset provided 1,114 balanced observations and 557 minority observations. A more effective analysis of feature value and sickness classification is attainable with this method.

E. *Feature Correlation Examination*

A good method for figuring out the fundamental linkages between various data variables within a dataset is feature correlation. Finding the interdependence of data components and recognizing how each component impacts the final variable may be done with the implementation of this analytical approach [31]. Correlation coefficients between distinct data quality and the goal illness were generated inside the feature matrix M, given as M = [v1, v2,..., vq], where vectors with q characteristics are represented by v1, v2,..., vq. The length of each vector, denoted by the symbol p, shows a whole medical course of therapy on a specific day. Figure 2 illustrates the estimated association values for each sample.



(a) CVD Dataset

(b) Framingham Dataset

Fig 2. Correlation Analysis Results for Medical Features and Heart Disease Prediction

Figure 2 demonstrates that the chosen variable "stroke" was positively related with four variables in the CVD dataset: "age," "hypertension," "heart disease," and "avg glucose lvl." The correlation values for these qualities were 0.57, 0.24, 0.27, and 0.2, respectively. Analogously, the Framingham dataset's features 'age,"sysBP,' 'prevalentHyp,' 'diabBP,' and 'glucose' exhibited positive association values of 0.23, 0.22, 0.18, and 0.15, respectively, which agree with the predicted output feature 'TenYearCHD.'

In all datasets, there is little or no relationship between non-medical metrics like "gender," "bmi," "heart rate," and others that are associated to lifestyle aspects like smoking habits, education, social status, and living standards and the output feature. Generally, in both datasets, established medical variables such as "age," "hypertension," and "glucose" are strongly connected with the result, suggesting their importance as risk factors.

Medical examination demonstrates that substantial anomalies in the heart and blood vessels are brought on by aging. Age-related changes, especially a slower heart rate during physical exercise compared to younger persons, may minimize the risk of cardiac disease, according to the National Heart, Lung, and Blood Institute [32]. It has long been recognized that high blood pressure increased the risk of stroke, ischemic heart disease, and renal failure [33]. Elevated blood pressure levels, which constrict arteries and restrict oxygen and blood flow to the heart, are the feature of hypertension and may certainly exacerbate heart disease. Due to blood vessel constriction brought on by high blood glucose levels, which damages the heart and blood vessels and exacerbates heart disease, diabetics are more prone to experience early-stage heart disease [34]. Over time, this operation could provoke a heart attack.

F. Variable Selection

Finding the medical parameters that increase the prognosis accuracy of heart disease is the key emphasis of this research. Choosing a subset of the most relevant characteristics from a bigger pool of original data that have the largest effect on the output is known as feature selection, and it is a crucial component of machine learning. Enhanced predictive performance, shorter prediction model processing times, quicker data collection, and higher data quality are all advantages of feature selection.

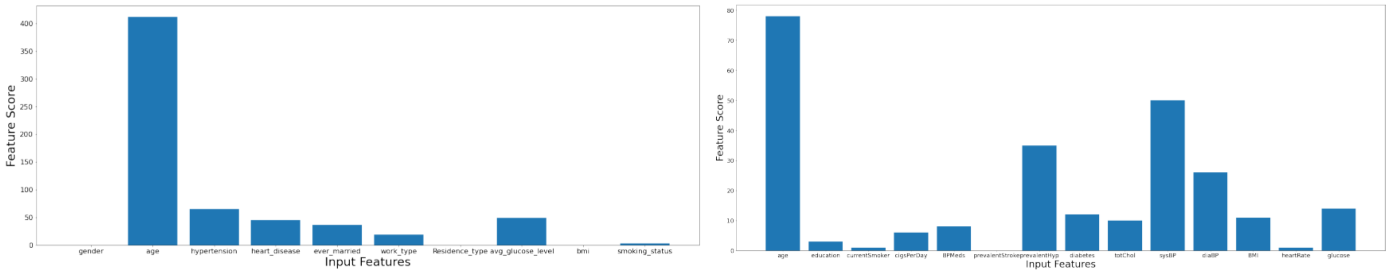
In this work, the ANOVA-F test—a filter-based feature selection method—was applied to reveal the most critical attributes from two datasets. In filter-based feature selection algorithms, important correlations or links between input and target attributes are determined by statistical methodologies. One quantitative statistical strategy for evaluating whether the means of numerous data sets are from the same distribution is ANOVA. A single statistical test called the ANOVA-F test evaluates each attribute to the target feature to discover statistically significant associations.

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

$$variance\_within\_groups = \frac{\sum_{i=1}^{S}\sum_{p=1}^{j_i}(K_{ip} - \bar{K}_i)^2}{(N-S))}$$

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

The f_classif() function from the scikit-learn package was used to develop the ANOVA-F test in Python. This approach, which was built in collaboration with the SelectKBest class, prioritizes features according to user ratings. The ANOVA-F test is run in this situation for assessment. The overall sample size (N), the number of groups (S), the number of observations in each group (ji), the group sample means (K i), the general mean of the data (K), and the pth observation in the ith group (Kip) are all included in the ANOVA-F equation.



(a) CVD Dataset

(b) Framingham Dataset

Fig 3. Significance Scores for Each Feature in Both Datasets.

The feature significance scores for both datasets generated from the ANOVA-F test are provided in Figures 3(a) and (b). The statistical analysis reveals that the variables "age," "hypertension," "heart disease," and "avg_glucose_lvl" are the most important in predicting "stroke." On the other hand, factors with little to no predictive value for "stroke" include "female," "bmi," "residence_type," and "smoking_status." Figure 3(b) demonstrates which factors received the highest scores for the 'TenYearCHD' outcome: 'age,' 'prevalentHyp,' 'diabetes,''sysBP,' 'diaBP,' and 'glucose'.

These findings confirm the connection data in section 3.3 and suggest that blood pressure, age, fructose, and hypertension all have a large effect on the risk of heart disease. The features revealed by the ANOVA-F test are also connected to probable risk factors for heart disease, according to the American Heart Association [37].

IV. Assessment Metrics

We applied three widely-known performance assessment metrics—Accuracy, F1-score, and ROC—to examine the effectiveness of our ML classification models [38]. Four important components make up the Confusion Matrix, a crucial tool for assessing classification models: The four potential outcomes are as follows: (1) True Positive (TP), which indicates that a patient has heart disease; (2) True Negative (TN), which indicates that a patient does not have heart disease; (3) False Negative (FN), which indicates that a patient is incorrectly classified as not having heart disease; and (4) False Positive (FP), which indicates that a patient is incorrectly classified as having heart disease. In the medical sector, False Negatives (FN) are very damaging misclassifications. The following formula is used to determine a collection of size n's accuracy:

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

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

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

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

To measure a model's efficacy in classification, the Receiver Operating Characteristic (ROC) graphs contrast the "true positive rate" and "false positive rate" in the ML model output.

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

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

V. RESULT & DISCUSSIONS

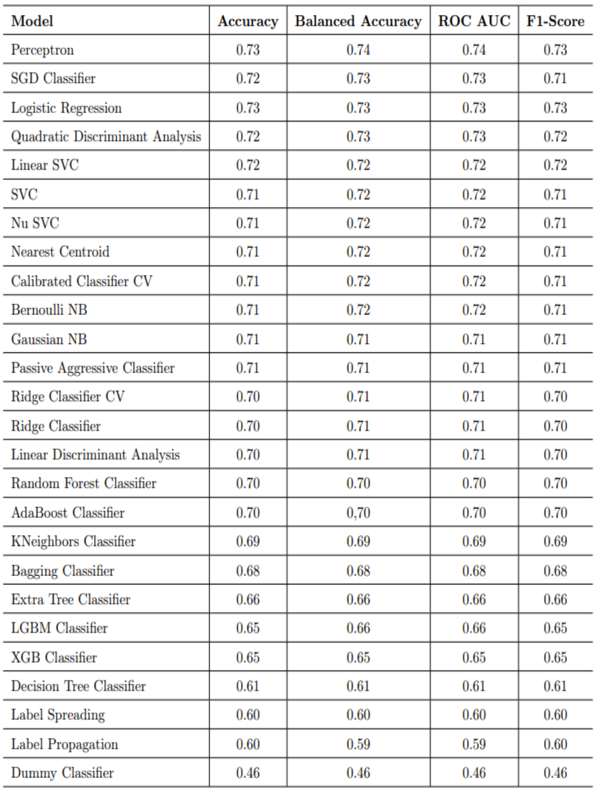
We will study the selected categorization models from numerous viewpoints in this chapter. In order to assess whether models perform better for one dataset or the other, we first examined each model's performance individually for the two datasets using all of the available variables. After that, we looked at how the models performed with the provided set of features to examine how the feature selection technique influences the accuracy of the classifiers. The performance of the classifier was assessed using the Accuracy, F1-score, and ROC assessment criteria.

A. Classification outcomes employing the entire feature set

This chapter evaluated every machine learning model on both datasets that predicted the binary illness result using all relevant parameters. The whole dataset—80% for training and 20% for testing—was applied to train the forecast models. 10.98 iterations per second (it/s) were required for training prediction models on the CVD dataset and 24.20 iterations per second (it/s) on the Framingham dataset, respectively, in terms of computation time. Tables 3 and 4 exhibit the binary classification findings for both datasets' diagnosis of heart disease.

Examining Table 3's classification findings, the MLP for the CVD dataset had the maximum accuracy at 0.73, followed by a ROC of 0.74 and an F1-Score of 0.73. When the whole feature set was used, other models including LR, SVC, and RF also displayed high prediction accuracy. The enhanced accuracy of MLP may perhaps be attributable to its abilities for recognizing patterns within complex medical information. Moreover, without any preceding domain knowledge, our neural network model performs fairly well at generalizing data. The dummy classifier, on the other hand, had the lowest accuracy of 0.46 for diagnosing myocardial stroke. This was reportedly because it based too much on simple notions, which may not be useful for solving real-world circumstances.

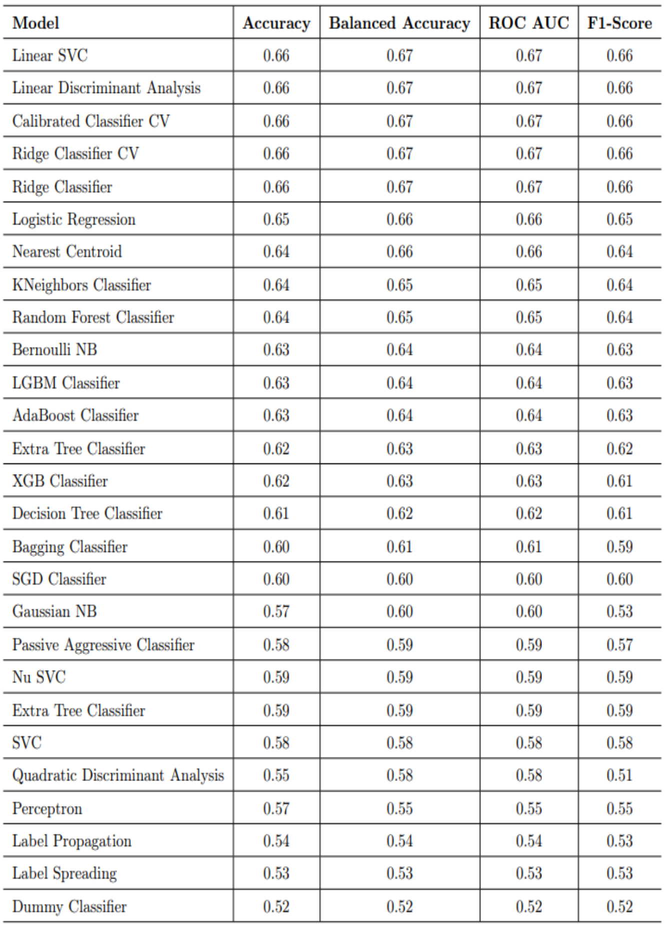
Table 3: Outcome of Different ML Models for CVD Dataset Using Complete Feature Set



Model	Accuracy	Balanced Accuracy	ROC AUC	F1-Score
Perceptron	0.73	0.74	0.74	0.73
SGD Classifier	0.72	0.73	0.73	0.71
Logistic Regression	0.73	0.73	0.73	0.73
Quadratic Discriminant Analysis	0.72	0.73	0.73	0.72
Linear SVC	0.72	0.72	0.72	0.72
SVC	0.71	0.72	0.72	0.71
Nu SVC	0.71	0.72	0.72	0.71
Nearest Centroid	0.71	0.72	0.72	0.71
Calibrated Classifier CV	0.71	0.72	0.72	0.71
Bernoulli NB	0.71	0.72	0.72	0.71
Gaussian NB	0.71	0.71	0.71	0.71
Passive Aggressive Classifier	0.71	0.71	0.71	0.71
Ridge Classifier CV	0.70	0.71	0.71	0.70
Ridge Classifier	0.70	0.71	0.71	0.70
Linear Discriminant Analysis	0.70	0.71	0.71	0.70
Random Forest Classifier	0.70	0.70	0.70	0.70
AdaBoost Classifier	0.70	0,70	0.70	0.70
KNeighbors Classifier	0.69	0.69	0.69	0.69
Bagging Classifier	0.68	0.68	0.68	0.68
Extra Tree Classifier	0.66	0.66	0.66	0.66
LGBM Classifier	0.65	0.66	0.66	0.65
XGB Classifier	0.65	0.65	0.65	0.65
Decision Tree Classifier	0.61	0.61	0.61	0.61
Label Spreading	0.60	0.60	0.60	0.60
Label Propagation	0.60	0.59	0.59	0.60
Dummy Classifier	0.46	0.46	0.46	0.46

Using the same approaches, Table 4 illustrates the findings of the classification process for the Framingham dataset. The best accuracy score was 0.66, with a ROC of 0.67 and an F1-Score of 0.66; the accuracy scores were not as amazing. Similar findings were achieved using numerous methodologies including LR, the ridge classifier, and Linear Discriminant Analysis (LDA). The vast range of values across the data variables might be the explanation for the comparably weak findings. We retained the original feature values to avoid creating substantial obstacles in medical research, even though feature scaling is known to normalize data within a specified range and maybe increase model performance [39, 40].

Table 4: Outcome of Various ML Models for Framingham Dataset Using Full Feature Set



Model	Accuracy	Balanced Accuracy	ROC AUC	F1-Score
Linear SVC	0.66	0.67	0.67	0.66
Linear Discriminant Analysis	0.66	0.67	0.67	0.66
Calibrated Classifier CV	0.66	0.67	0.67	0.66
Ridge Classifier CV	0.66	0.67	0.67	0.66
Ridge Classifier	0.66	0.67	0.67	0.66
Logistic Regression	0.65	0.66	0.66	0.65
Nearest Centroid	0.64	0.66	0.66	0.64
KNeighbors Classifier	0.64	0.65	0.65	0.64
Random Forest Classifier	0.64	0.65	0.65	0.64
Bernoulli NB	0.63	0.64	0.64	0.63
LGBM Classifier	0.63	0.64	0.64	0.63
AdaBoost Classifier	0.63	0.64	0.64	0.63
Extra Tree Classifier	0.62	0.63	0.63	0.62
XGB Classifier	0.62	0.63	0.63	0.61
Decision Tree Classifier	0.61	0.62	0.62	0.61
Bagging Classifier	0.60	0.61	0.61	0.59
SGD Classifier	0.60	0.60	0.60	0.60
Gaussian NB	0.57	0.60	0.60	0.53
Passive Aggressive Classifier	0.58	0.59	0.59	0.57
Nu SVC	0.59	0.59	0.59	0.59
Extra Tree Classifier	0.59	0.59	0.59	0.59
SVC	0.58	0.58	0.58	0.58
Quadratic Discriminant Analysis	0.55	0.58	0.58	0.51
Perceptron	0.57	0.55	0.55	0.55
Label Propagation	0.54	0.54	0.54	0.53
Label Spreading	0.53	0.53	0.53	0.53
Dummy Classifier	0.52	0.52	0.52	0.52

B. Outcome of Classification with Limited Features

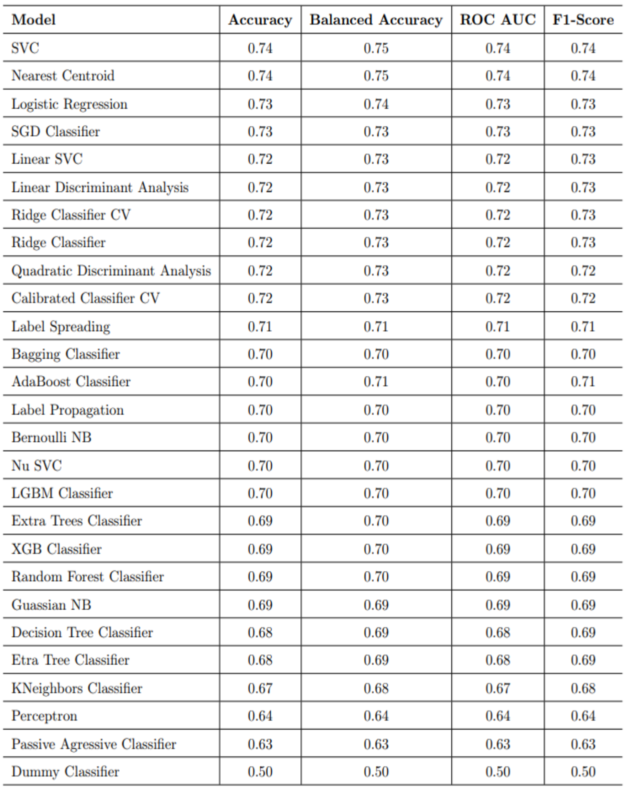
To find novel biomarkers and examine the effect of feature selection technique on classification accuracy, we opted to extract the most essential features from the complete feature space based on individual feature scores. As illustrated in Figures 3(a) and (b), the ANOVA-F test was

done to identify which variables had the most influence on the outcome for each sample.

Based on feature evaluations, four variables were selected from a cohort of eleven for the CVD dataset: age, heart disease, hypertension, and average milliliter glucose level. For the Framingham dataset, the ANOVAF test discovered five factors out of a total of fifteen: age, prevalentHyp, sysBp, diaBp, and glucose. Next, using simply the qualities supplied as inputs, the effectiveness of each classification model was tested. Table 5 demonstrates the category.

Outputs of every model employing the little feature group of the CVD dataset. Even with fewer characteristics, the research demonstrated that ML models performed better than models that looked at the whole feature set. With only four input characteristics, the SVC model attained the maximum accuracy (0.74 F1-Score and 0.74 ROC). Table 6 demonstrates that when the whole feature set was applied, the greatest accuracy was 0.71, exceeding all accuracy values reported for the Framingham dataset. Moreover, models trained on the smaller feature set executed data processing operations more quickly: the Framingham dataset required 15.52 iterations per second (it/s), while the CVD models needed 3.86 iterations per second (it/s).

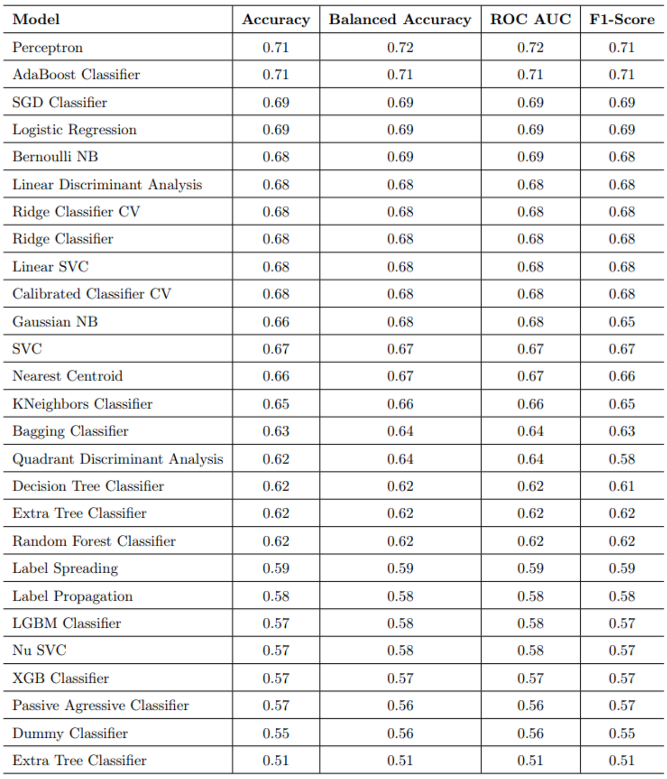
Table 5: Outcome of Various ML Models for CVD Dataset Using Reduced Feature Set



Model	Accuracy	Balanced Accuracy	ROC AUC	F1-Score
SVC	0.74	0.75	0.74	0.74
Nearest Centroid	0.74	0.75	0.74	0.74
Logistic Regression	0.73	0.74	0.73	0.73
SGD Classifier	0.73	0.73	0.73	0.73
Linear SVC	0.72	0.73	0.72	0.73
Linear Discriminant Analysis	0.72	0.73	0.72	0.73
Ridge Classifier CV	0.72	0.73	0.72	0.73
Ridge Classifier	0.72	0.73	0.72	0.73
Quadratic Discriminant Analysis	0.72	0.73	0.72	0.72
Calibrated Classifier CV	0.72	0.73	0.72	0.72
Label Spreading	0.71	0.71	0.71	0.71
Bagging Classifier	0.70	0.70	0.70	0.70
AdaBoost Classifier	0.70	0.71	0.70	0.71
Label Propagation	0.70	0.70	0.70	0.70
Bernoulli NB	0.70	0.70	0.70	0.70
Nu SVC	0.70	0.70	0.70	0.70
LGBM Classifier	0.70	0.70	0.70	0.70
Extra Trees Classifier	0.69	0.70	0.69	0.69
XGB Classifier	0.69	0.70	0.69	0.69
Random Forest Classifier	0.69	0.70	0.69	0.69
Guassian NB	0.69	0.69	0.69	0.69
Decision Tree Classifier	0.68	0.69	0.68	0.69
Etra Tree Classifier	0.68	0.69	0.68	0.69
KNeighbors Classifier	0.67	0.68	0.67	0.68
Perceptron	0.64	0.64	0.64	0.64
Passive Agressive Classifier	0.63	0.63	0.63	0.63
Dummy Classifier	0.50	0.50	0.50	0.50

Our findings were further strengthened by a comparison with previously published theories [41, 42], in which the identical datasets were investigated using the whole feature set and the accuracy results were either lower or equivalent to what we obtained using the reduced feature set. Overall, the experiment's findings suggest that supplying just the most relevant information greatly boosts the performance of machine learning models. Additionally, while using the smaller feature set to train classification models, fewer computing iterations per second (it/s) were discovered. These experimental findings show the significance of feature selection techniques, proving that these approaches increase the performance of multidimensional machine learning models in addition to lowering feature space.

Table 6: Outcome of Various ML Models for Framingham Dataset Using Reduced Feature Set



Model	Accuracy	Balanced Accuracy	ROC AUC	F1-Score
Perceptron	0.71	0.72	0.72	0.71
AdaBoost Classifier	0.71	0.71	0.71	0.71
SGD Classifier	0.69	0.69	0.69	0.69
Logistic Regression	0.69	0.69	0.69	0.69
Bernoulli NB	0.68	0.69	0.69	0.68
Linear Discriminant Analysis	0.68	0.68	0.68	0.68
Ridge Classifier CV	0.68	0.68	0.68	0.68
Ridge Classifier	0.68	0.68	0.68	0.68
Linear SVC	0.68	0.68	0.68	0.68
Calibrated Classifier CV	0.68	0.68	0.68	0.68
Gaussian NB	0.66	0.68	0.68	0.65
SVC	0.67	0.67	0.67	0.67
Nearest Centroid	0.66	0.67	0.67	0.66
KNeighbors Classifier	0.65	0.66	0.66	0.65
Bagging Classifier	0.63	0.64	0.64	0.63
Quadrant Discriminant Analysis	0.62	0.64	0.64	0.58
Decision Tree Classifier	0.62	0.62	0.62	0.61
Extra Tree Classifier	0.62	0.62	0.62	0.62
Random Forest Classifier	0.62	0.62	0.62	0.62
Label Spreading	0.59	0.59	0.59	0.59
Label Propagation	0.58	0.58	0.58	0.58
LGBM Classifier	0.57	0.58	0.58	0.57
Nu SVC	0.57	0.58	0.58	0.57
XGB Classifier	0.57	0.57	0.57	0.57
Passive Agressive Classifier	0.57	0.56	0.56	0.57
Dummy Classifier	0.55	0.56	0.56	0.55
Extra Tree Classifier	0.51	0.51	0.51	0.51

VI. CONCLUSION

Heart disease is an important worldwide cause of death that is increasing worse, hence improved prediction approaches are essential for early treatment. The objective of this study is to foresee heart illness and describe important components. The major purpose is to investigate at how feature selection strategies impact how successfully machine learning (ML) models work, notably with the utilization of online and Framingham heart disease datasets. Data pre-processing, which includes updates, deletions, and balances, is the initial phase in the inspection process. Then, in order to reveal key features for effective heart disease prediction, the ANOVA-F test is carried out applying a filter-based feature selection approach. Age, blood pressure, glucose levels, hypertension, past experience with heart disease, and other variables are significant risk factors for heart disease in both datasets, above and beyond normal conditions, according to individual feature scores from the ANOVA-F test.

In addition, classification tests are done using both entire and reduced feature sets in order to explore how chosen characteristics impact prediction accuracy across multiple machine learning models. The Framingham heart disease dataset (0.66) and the complete feature set for CVD (0.73) produced the most accurate findings. For both datasets, accuracy rises to 0.71 and 0.75 when the reduced feature set is used. The results indicate that ML models outperformed fully featured models even with a decreased feature set. This outstanding study highlights how feature selection strategies with fewer features and better processing speeds may properly identify heart disease. Consequently, feature selection reduces processing barriers and boosts prediction model performance by focusing on the most significant heart disease-related data.

The business ultimately hopes to boost forecast accuracy by analyzing different machine learning and deep learning models. To give more relevant feature groups for medical study, fresh datasets will be applied in comparison studies, and inventive feature selection processes will be investigated.

References

[1] Mansoul, Abdelhak, and Baghdad Atmani. "Representative Case-Based Retrieval to Support Case-Based Reasoning for Prediction." International Journal of Strategic Decision Sciences (IJSDS) 12, no. 3 (2021): 14-35.

[2] Meneguin, Silmara, and Debora Santana. "QUALIDADE DE VIDA RELACIONADA À SAÚDE EM GESTANTES CARDIOPATAS: REVISÃO DE LITERATURA." Revista Saúde-UNG-Ser 10, no. 3/4 (2016): 84-93.

[3] Abougreen, Arij Naser. "Big Medical Data Analytics Under Internet of Things." In Efficient Data Handling for Massive Internet of Medical Things: Healthcare Data Analytics, pp. 25-44. Cham: Springer International Publishing, 2021.

[4] Hartono, Ambran, Lizky Azka Dewi, Elvan Yuniarti, and Salsabila Tahta Hirani Putri. "Machine Learning Classification for Detecting Heart Disease with K-NN Algorithm, Decision Tree and Random Forest." EKSAKTA: Berkala Ilmiah Bidang MIPA 23, no. 04 (2023): 513-522.

[5] Zhang, Haoyue, Zheng Ouyang, and Wenpeng Zhang. "Advances in mass spectrometry for clinical analysis: Data acquisition, interpretation and information integration." TrAC Trends in Analytical Chemistry (2023): 117380.

[6] Raghavendra, S., Vasudev Parvati, R. Manjula, Ashok Kumar Nanda, Ruby Singh, D. Lakshmi, and S. Velmurugan. "DLMNN Based Heart Disease Prediction with PD-SS Optimization Algorithm." Intelligent Automation and Soft Computing 35, no. 2 (2023): 1353-1368.

[7] Maulidah, Nurlaelatul. "LOGISTIC REGRESSION DENGAN PRINCIPAL COMPONENT ANALYSIS (PCA) UNTUK MEMPREDIKSI PENYAKIT CARDIOVASCULAR."

[8] Hu, Li, Anli Yan, Hongyang Yan, Jin Li, Teng Huang, Yingying Zhang, Changyu Dong, and Chunsheng Yang. "Defenses to Membership Inference Attacks: A Survey." ACM Computing Surveys (2023).

[9] Bhuvanya, R., and M. Kavitha. "Offline and Online Performance Evaluation Metrics of Recommender System: A Bird's Eye View." Machine Learning Paradigm for Internet of Things Applications (2022): 113-146.

[10] Abhinaya, D. "Prediction of groundnut yield using principal component analysis of weather parameters." (2021).

[11] Sinha, Dibakar, and Ashish Sharma. "Various Feature Selection Techniques Used for Predicting and Diagnosing Heart Disease." In Smart Healthcare for Sustainable Urban Development, pp. 214-234. IGI Global, 2022.

[12] Clemente, Fabiana, Gonçalo Martins Ribeiro, Alexandre Quemy, Miriam Seoane Santos, Ricardo Cardoso Pereira, and Alex Barros. "ydata-profiling: Accelerating data-centric AI with high-quality data." Neurocomputing 554 (2023): 126585.

[13] Deng, Qifang, and Toktam Mahmoodi. "Kinesthetic Data Reduction Techniques in Bilateral Teleoperation Systems." PhD diss., King’s College London, 2023.

[14] Luo, Kuei-Hau, Chih-Hsien Wu, Chen-Cheng Yang, Tzu-Hua Chen, Hung-Pin Tu, Cheng-Hong Yang,

and Hung-Yi Chuang. "Exploring the association of metal mixture in blood to the kidney function and tumor necrosis factor alpha using machine learning methods." Ecotoxicology and Environmental Safety 265 (2023): 115528.

[15] Ramanna, Sheela, Negin Ashrafi, Evan Loster, Karen Debroni, and Shelley Turner. "Rough-set based learning: Assessing patterns and predictability of anxiety, depression, and sleep scores associated with the use of cannabinoid-based medicine during COVID-19." Frontiers in Artificial Intelligence 6 (2023): 981953.

[16] Shamrock Nwosu, Chidozie, Soumyabrata Dev, Peru Bhardwaj, Bharadwaj Veeravalli, and Deepu John. "Predicting Stroke from Electronic Health Records." arXiv e-prints (2019): arXiv-1904.

[17] Toppo, Shashikala, Mr Apurv Verma, and Vijayant Verma. "INVESTIGATING THE IMPLICATIONS OF FEATURE SELECTION ON THE ACCURACY OF HEART DISEASE PREDICTIONS."

[18] THIỆU, I. GIỚI. "DATA MINING IN HEALTHCARE SYSTEM ON PATIENTS CLINICAL SYMPTOMS DATASET."

[19] Dhivya, P., P. Rajesh Kanna, K. Deepa, and S. Santhiya. "Square Static–Deep Hyper Optimization and Genetic Meta-Learning Approach for Disease Classification." IETE Journal of Research (2023): 1-10.

[20] Çifci, A., M. İlkuçar, and İ. Kırbaş. "What makes survival of heart failure patients? Prediction by the iterative learning approach and detailed factor analysis with the SHAP algorithm." In Explainable Artificial Intelligence for Biomedical Applications, pp. 101-121. River Publishers, 2023.

[21] Lee, Jonathan, Kartono Apriliyandi, Erwin Gunawan, Shilvia Meidhi Honova, Brigita Vanessa Salim, Maria Susan Anggreainy, and Ivan Halim Parmonangan. "Comparative Analysis of Neural Network Method on Stroke Prediction."

[22] Gopu, Krishna Lava Kumar, and Suthendran Kannan. "Deep Graph Neural Network with Fish-Inspired Task Allocation Algorithm for Heart Disease Diagnosis." Journal of Current Science and Technology 13, no. 2 (2023): 392-411.

[23] Almonteros, Jayrhom R., Junrie B. Matias, and Joanna Victoria S. Pitao. "Forecasting Students' Success to Graduate Using Predictive Analytics." International Journal of Computing and Digital Systems 14, no. 1 (2023): 1-xx.

[24] Baseer, Abdul, Zulfiqar Ali, Maryam Ilyas, and Mahrukh Yousaf. "A new Monte Carlo Feature Selection (MCFS) algorithm-based weighting scheme for multi-model ensemble of precipitation." Theoretical and Applied Climatology (2023): 1-12.

[25] Toppo, Shashikala, Mr Apurv Verma, and Vijayant Verma. "INVESTIGATING THE IMPLICATIONS OF FEATURE SELECTION ON THE ACCURACY OF HEART DISEASE PREDICTIONS."

[26] Tang, Yifan, M. Rahmani Dehaghani, and G. Gary Wang. "Comparison of Transfer Learning based Additive Manufacturing Models via A Case Study." arXiv preprint arXiv:2305.11181 (2023).

[27] Orlu, Glory Urekwere, Rusli Bin Abdullah, Zeinab Zaremohzzabieh, Yusmadi Yah Jusoh, Shahla Asadi, Yousef AM Qasem, Rozi Nor Haizan Nor, and Wan Mohd Haffiz bin Mohd Nasir. "A systematic review of literature on sustaining decision-making in healthcare organizations amid imperfect information in the Big Data era." Sustainability 15, no. 21 (2023): 15476.

[28] Zhang, Fan, Melissa Petersen, Leigh Johnson, James Hall, Raymond F. Palmer, Sid E. O'Bryant, and Health and Aging Brain Study (HABS–HD) Study Team. "A Machine Learning-Based Multiple Imputation Method for the Health and Aging Brain Study–Health Disparities." In Informatics, vol. 10, no. 4, p. 77. MDPI, 2023 and Decision Making 22 (1) (2022) 1–14.

[29] Washington, Ariel. "The development of a culturally-informed cervical cancer screening and prevention mhealth intervention for African American women." (2020).

[30] Parente, Mimmo, Luca Rizzuti, and Mario Trerotola. "A profitable trading algorithm for cryptocurrencies using a Neural Network model." Expert Systems with Applications 238 (2024): 121806.

[31] ROHMA, DIANA. "IMPLEMENTASI ALGORITMA NAÏVE BAYES UNTUK SISTEM REKOMENDASI TOPIK PENELITIAN PADA

JURUSAN ILMU KOMPUTER UNIVERSITAS LAMPUNG BERBASIS KLASIFIKASI." (2021).

[32] Ahmad o‘g‘li, Abdullo Akramov, and Isaqulov Ulug'bek Alisher o'g. "FALLO TETRADASI: TA’RIFI, KELIB CHIQISH SABABLARI, MORFOLOGIYASI, DIAGNOZTIKASI, DAVOLASH VA OQIBATLARINI CHUQUR O’RGANISH." Yangi O'zbekistonda Tabiiy va Ijtimoiy-gumanitar fanlar respublika ilmiy amaliy konferensiyasi 1, no. 4 (2023): 5-8.

[33] Aune, Dagfinn, Wentao Huang, Jing Nie, and Yafeng Wang. "Hypertension and all-cause and cause-specific mortality: An outcome-wide association study of 67 causes of death in the National Health Interview Survey." (2021).

[34] Huxley, Rachel, Federica Barzi, and Mark Woodward. "Excess risk of fatal coronary heart disease."

[35] Van Zyl, Corne, Xianming Ye, and Raj Naidoo. "Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP." Applied Energy 353 (2024): 122079.

[36] Martins, Mariana, Ana Rafaela Oliveira, Solange Martins, José Pedro Vieira, Pedro Perdigão, Ana Rita Fernandes, Luís Pereira de Almeida et al. "A Novel Genetic Variant in MBD5 Associated with Severe Epilepsy and Intellectual Disability: Potential Implications on Neural Primary Cilia." International Journal of Molecular Sciences 24, no. 16 (2023): 12603.

[37] Liang, Zhen, Yingyue Lou, Yulei Hao, Hui Li, Jiachun Feng, and Songyan Liu. "The Relationship of Astrocytes and Microglia with Different Stages of Ischemic Stroke." Current Neuropharmacology 21, no. 12 (2023): 2465-2480.

[38] Niea, Yuhao, Xiatong Lib, Quentin Palettac, Max Aragone, Andea Scotta, and Adam Brandta. "Open-Source Ground-based Sky Image Datasets for Very Short-term Solar Forecasting, Cloud Analysis and Modeling: A Comprehensive Survey."

[39] da Silva, Patrícia Mayara Moura, Edgar Ramos Vieira, Edgard Morya, and Fabrícia Azevêdo. "5.3 Artigo 03: Clusters of patients with diabetes type 2: an explora-tory unsupervised machine learning approach based on gait para-meters." Aprendizagem de máquina aplicada à execução da marcha em diabéticos tipo 2 (2023): 79.

[40] Dai, Shuang. "A Distributed and Real-time Machine Learning Framework for Smart Meter Big Data." PhD diss., University of Essex, 2023.