# Classification models for heart disease prediction using feature selection and PCA

Anna Karen Gárate-Escamila, Amir Hajjam El Hassani, Emmanuel Andrès

# Classification models for heart disease prediction using feature selection and PCA

Anna Karen GARATE-ESCAMILLA[1*], Amir HAJJAM EL HASSANI[1] and Emmanuel ANDRES [2,3]

[1] Nanomedicine Lab, Univ. Bourgogne Franche-Comte, UTBM, F-90010 Belfort, France
[2] Service de Médecine Interne, Diabète et Maladies métaboliques de la Clinique Médicale B, CHRU de Strasbourg, Strasbourg, France.
[3] Centre de Recherche Pédagogique en Sciences de la Santé, Faculté de Médecine de Strasbourg, Université de Strasbourg (UdS), Strasbourg, France.
*Corresponding author email : anna.garate-escamilla@utbm.fr

**Abstract**
The prediction of cardiac disease helps practitioners make more accurate decisions regarding patients' health. Therefore, the use of machine learning (ML) is a solution to reduce and understand the symptoms related to heart disease. The aim of this work is the proposal of a dimensionality reduction method and finding features of heart disease by applying a feature selection technique. The information used for this analysis was obtained from the UCI Machine Learning Repository called Heart Disease. The dataset contains 74 features and a label that we validated by six ML classifiers. Chi-square and principal component analysis (CHI-PCA) with random forests (RF) had the highest accuracy, with 98.7% for Cleveland, 99.0% for Hungarian, and 99.4% for Cleveland-Hungarian (CH) datasets. From the analysis, ChiSqSelector derived features of anatomical and physiological relevance, such as cholesterol, highest heart rate, chest pain, features related to ST depression, and heart vessels. The experimental results proved that the combination of chi-square with PCA obtains greater performance in most classifiers. The usage of PCA directly from the raw data computed lower results and would require greater dimensionality to improve the results.

**Keywords:** Machine Learning, Heart Disease, Apache Spark, PCA, Feature selection.

## 1 Introduction

The World Health Organization (WHO) [1] lists cardiovascular diseases as the leading cause of death globally with 17.9 million people dying every year. The risk of heart disease increases due to harmful behavior that leads to overweight and obesity, hypertension, hyperglycemia, and high cholesterol [1]. Furthermore, the American Heart Association [2] complements symptoms with weight gain (1-2 kg per day), sleep problems, leg swelling, chronic cough and high heart rate [3]. Diagnosis is a problem for practitioners due the symptoms' nature of being common to other conditions or confused with signs of aging.

The growth in medical data collection presents a new opportunity for physicians to improve patient diagnosis. In recent years, practitioners have increased their usage of computer technologies to improve decision-making support. In the health care industry, machine learning is becoming an important solution to aid the diagnosis of patients. Machine learning is an analytical tool used when a task is large and difficult to program, such as transforming medical record into knowledge, pandemic predictions, and genomic data analysis [4].

Recent studies have used machine learning techniques to diagnose different cardiac problems and make a prediction. Melillo et al. [5] contributed to an automatic classifier for patients with congestive heart failure (CHF) that separates patients with minimal risk from those at high risk. The classification and regression tree (CART) computed a sensitivity and specificity of 93.3% and 63.5%, respectively. Rahhal et al. [6] proposed a deep neural network (DNN) classification of electrocardiogram (ECG) signals to learn the best set of features and improved the performance. Guidi et al. [7] contributed to a clinical decision support system (CDSS) for the analysis of heart failure (HF). They compared the performance of different machine learning classifiers, such as neural network (NN), support vector machine (SVM), a system with fuzzy rules that uses CART, and random forests (RF). The CART model and RF obtained the best performance with an accuracy of 87.6%. Zhang et al. [8] found a NYHA

class for HF from unstructured clinical notes using natural language processing (NLP) and the rule-based method, calculating an accuracy of 93.37%. Parthiban et al. [9] scrutinized an SVM technique to diagnose heart disease in patients with diabetes, obtaining an accuracy of 94.60% and predicting features such as age, blood pressure, and blood sugar.

A major problem of machine learning is the high dimensionality of the dataset [10]. The analysis of many features requires a large amount of memory and leads to an overfitting, so the weighting features decrease redundant data and processing time, thus improving the performance of the algorithm [11-15]. Finding a small set of features characterizes different diseases of health management, genome expression, medical images, and IoT. Dimensionality reduction uses feature extraction to transform and simplify data, while feature selection reduces the dataset by removing useless features [16].

In the literature, the use of feature selection techniques improved the prediction of heart disease. Dun et al. [17] studied the presence of heart disease through deep learning techniques, random forests, logistic regression, and SVM with hyperparameter tuning and feature selection. NN had the best accuracy at 78.3%. Sewak et al. [18] reduced cardiovascular features using the Fisher ranking method, generalized discriminant analysis (GDA), and a binary classifier as extreme learning machine (ELM). They detected coronary heart disease with an accuracy improvement of 100%. Yaghouby et al. [19] classified arrhythmias with heart rate variability (HRV). They achieved 100% accuracy using GDA for feature reduction and multilayer perceptron (MLP) neural network as a classifier. Mohammadzadeh et al. [20] classified 15 features from HRV signal. GDA reduced the features to five and computed 100% precision using SVM.

Principal component analysis (PCA) creates new components that store the most valuable information of the features by capturing a high variance [21]. Recently, several studies have used PCA as a feature extraction technique for classification in health care. Rajagopal et al. [22] compared an automatic classification of cardiac arrhythmia using five different linear and non-linear unsupervised dimensional reduction techniques with the neural network (PNN) classifier. With a minimum of 10 components, fastICA computed an F1 score of 99.83%. Zhang et al. [23] detected breast cancer using an AdaBoost algorithm based on PCA. Negi et al. [24] combined PCA with a feature reduction technique called uncorrelated linear discriminant analysis (ULDA) to obtain the best features that control upper limb motions. Avendaño-Valencia et al. [25] applied PCA to time frequency representations (TFR) to reduce heart sounds and improve performance. Kamencay et al. [26] presented a new method using PCA-KNN called the scale-invariant feature transform (SIFT) descriptor in different medical images, which resulted in an accuracy of 83.6% when training 200 images. Ratnasari et al. [27] reduced X-ray images using a threshold-based ROI and PCA. They obtained the best gray-level threshold of 150.

Earlier studies worked with a heart disease subset of 13 features (Subset-A). The aim of classification was to predict whether a patient had heart disease using, in most cases, the dataset of Cleveland [28]. Some remarkable results were presented: decision tree with an accuracy of 89.1% [29], random forests with an accuracy of 89.2% [30], artificial neural network with an accuracy of 92.7% [30], 89.0% [31], and 89.7% [32], and SVM with an accuracy of 88.0% [32]. GA+NN [33] computed the most notable hybrid model with 94.2% accuracy. PCA+regression and PCA1+NN [34] obtained the best PCA models with an accuracy of 92.0% and 95.2%, respectively.

The classification learning models combined with dimensionality reduction seek to achieve three primary objectives: (i) to learn the best feature representation of the dataset used; (ii) to validate the performance of PCA in conjunction with a feature selection technique; and (iii) to learn the classification model that computes the best performance. Six classifiers compute the 74 features: logistic regression, decision tree, random forest, gradient-boosted tree, multilayer perceptron, and Naïve Bayes.

We propose a model based on chi-square and PCA for the detection of heart disease. Experimental results have shown that PCA delivers a better prediction concerning the high dimensional classification problem by setting the features provided by chi-square, as seen when comparing PCA with raw data. Chi-square ranks the independent features most compatible with the label. We selected k=13 to perform a comparison with the subset of 13 features (Subset-A) used in the literature [28]. For PCA, we tried for the raw data a latent dimensional k greater than a variance of one. This implies k=13 for Cleveland, k=14 for Hungarian, and k=11 for CH (the datasets of Cleveland and Hungarian). Considering the results, the proposed approach improved most of the machine learning techniques.

**Table 1.** Features of the Heart Disease Data set

| Serial Number | Group | Feature Names | Features Descriptions |
|---|---|---|---|
| 1 | | ID | Patient identification number |
| 2 | | CCF | Social security number (replaces this with a dummy value of 0) |
| 3 | Patient record data | AGE | Age in years |
| 4 | Patient record data | SEX | 1=male; 0 =female |
| 5 | Patient record data | PAINLOC | Chest pain location (1=substernal; 0=otherwise) |
| 6 | Patient record data | PAINEXER | 1=provoked by exertion; 0=otherwise |
| 7 | Patient record data | RELREST | 1= relieved after rest; 0=otherwise |
| 8 | Patient record data | PNCADEN | Sum of 5, 6, and 7 |
| 9 | Patient record data | CP | Chest pain type: 1= typical angina; 2=atypical angina; 3=non-angina pain; 4=asymptomatic |
| 10 | Patient record data | TRESTBPS | Systolic blood pressure at rest (in mm Hg on admission to the hospital) |
| 11 | Patient record data | HTN | History of hypertension |
| 12 | Patient record data | CHOL | Serum cholesterol in mg/dl |
| 13 | Patient record data | SMOKE | 1=yes; 0=no (is or is not a smoker) |
| 14 | Patient record data | CIGS | Cigarettes per day |
| 15 | Patient record data | YEARS | Number of years as a smoker |
| 16 | Patient record data | FBS | Fasting blood sugar > 120 mg/dl (1=true; 0= false) |
| 17 | Patient record data | DM | 1=history of diabetes; 0=no such history |
| 18 | Patient record data | FAMHIST | Family history of coronary artery disease (1=yes; 0=no) |
| 19 | Patient record data | RESTECG | Resting electrocardiographic results: 0=normal; 1=having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); 2=showing probable or definite left ventricular hypertrophy by Estes' criteria |
| 20 | Patient record data | EKGMO | Month of exercise ECG reading |
| 21 | Patient record data | EKGDAY | Day of exercise ECG reading |
| 22 | Patient record data | EKGYR | Year of exercise ECG reading |
| 23 | Medication during exercise test | DIG | Digitalis is used during exercise ECG (1=yes; 0=no) |
| 24 | Medication during exercise test | PROP | Beta blocker used during exercise ECG (1=yes; 0=no) |
| 25 | Medication during exercise test | NITR | Nitrates used during exercise ECG (1=yes; 0=no) |
| 26 | Medication during exercise test | PRO | Calcium channel blocker used during exercise ECG (1=yes; 0=no) |
| 27 | Medication during exercise test | DIURETIC | Diuretic used during exercise ECG (1=yes; 0=no) |
| 28 | Exercise test | PROTO | Exercise protocol: 1=Bruce; 2=Kottus; 3=McHenry; 4=Fast Balke; 5=Balke; 6=Noughton; 7=bike 150 kpa min/min; 8=bike 125 kpa min/min; 9=bike 100 kpa min/min; 10=bike 75 kpa min/min; 11=bike 50 kpa min/min; 12=arm ergometer |
| 29 | Exercise electrocardiogram | THALDUR | Duration of exercise test in minutes |
| 30 | Exercise electrocardiogram | THALTIME | Time when ST measure depression was noted |
| 31 | Exercise electrocardiogram | MET | Mets achieved |
| 32 | Exercise electrocardiogram | THALACH | Maximum heart rate achieved |
| 33 | Exercise electrocardiogram | THALREST | Resting heart rate |
| 34 | Exercise electrocardiogram | TPEAKBPS | Peak exercise systolic blood pressure (first of 2 parts) |
| 35 | Exercise electrocardiogram | TPEAKBPD | Peak exercise systolic blood pressure (second of 2 parts) |
| 36 | Exercise electrocardiogram | DUMMY | The same value as trestbps |
| 37 | Exercise electrocardiogram | TRESTBPD | Resting blood pressure |
| 38 | Exercise electrocardiogram | EXANG | Exercise-induced angina (1=yes; 0=no) |

**Table 1.** Features of the Heart Disease Data set (continuation)

| Serial Number | Group | Feature Names | Features Descriptions |
|---|---|---|---|
| 39 | Exercise electrocardiogram | XHYPO | Exercise-induced hypotension (1=yes; 0=no) |
| 40 | Exercise electrocardiogram | OLDPEAK | E'xercise-induced ST depression relative to rest |
| 41 | Exercise electrocardiogram | SLOPE | The slope of the peak exercise ST segment: 1=upsloping; 2=flat; 3=downsloping |
| 42 | Exercise electrocardiogram | RLDV5 | Height at rest |
| 43 | Exercise electrocardiogram | RLDV5E | Height at peak exercise |
| 44 | Cardiac fluoroscopy | CA | Number of major vessels (0-3) colored by fluoroscopy |
| 45 | Cardiac fluoroscopy | RESTCKM | Irrelevant |
| 46 | Cardiac fluoroscopy | EXERCKM | Irrelevant |
| 47 | Cardiac fluoroscopy | RESTEF | Rest radionuclide ejection fraction |
| 48 | Cardiac fluoroscopy | RESTWM | Rest wall motion abnormality: 0=none; 1=mild of moderate; 2= moderate or severe; 3=akinesis or dyskinesis |
| 49 | Cardiac fluoroscopy | EXEREF | Exercise-induce radionuclide ejection fraction |
| 50 | Cardiac fluoroscopy | EXERWM | Exercise-induce wall motion abnormalities |
| 51 | Exercise thallium scintigraphy | THAL | Exercise Thallium heart scan: 3=normal; 6= fixed defect; 7=reversible defect |
| 52 | Exercise thallium scintigraphy | THALSEV | Not used |
| 53 | Exercise thallium scintigraphy | THALPUL | Not used |
| 54 | Exercise thallium scintigraphy | EARLPUL | Not used |
| 55 | Coronary angiograms | CMO | Month of cardiac cath |
| 56 | Coronary angiograms | CDAY | Day of cardiac cath |
| 57 | Coronary angiograms | CYR | Year of cardiac cath |
| 58 | Coronary angiograms | NUM | Diagnosis of heart disease (angiographic disease status):<br>-0= <50% diameter narrowing<br>-1= >50% diameter narrowing<br>(in any major epicardial vessel attributes 59 through 68 are vessels) |
| 59 | Blood Vessels | LMT | Left main truck |
| 60 | Blood Vessels | LADPROX | Proximal left anterior descending artery |
| 61 | Blood Vessels | LADDIST | Distal left anterior descending artery |
| 62 | Blood Vessels | DIAG | Diagonal branches |
| 63 | Blood Vessels | CXMAIN | Circumflex |
| 64 | Blood Vessels | RAMUS | Ramus intermedius |
| 65 | Blood Vessels | OM1 | First obtuse marginal branch |
| 66 | Blood Vessels | OM2 | Second obtuse marginal branch |
| 67 | Blood Vessels | RCAPROX | Proximal right coronary artery |
| 68 | Blood Vessels | RCADIST | Distal right coronary artery |
| 69 | | LVX1 | Not used |
| 70 | | LVX2 | Not used |
| 71 | | LVX3 | Not used |
| 72 | | LVX4 | Not used |
| 73 | | LVF | Not used |
| 74 | | CATHEF | Not used |
| 75 | | JUNK | Not used |

# 2 Materials and Methods

## 2.1 Description of the dataset

The dataset used in the research was the "Heart Disease Dataset" of the UCI Machine Learning Repository [28] as shown in Table 1. It had a label called coronary angiography (NUM) and 74 independent features. NUM specified whether a patient has the presence or absence of heart disease. The presence of heart disease combined the values 1, 2, 3, and 4 from the original datasets. For the examination, patients supplied historical data and were physically examined by practitioners [42]. Three non-invasive tests were part of the protocol: exercise electrocardiogram, exercise thallium scintigraphy, and coronary calcium fluoroscopy. The cardiologist interpreted the coronary angiogram results without knowing the non-invasive results. Previous research [43] has explained some features as well as the complete protocol.

In the literature, a subset of 13 features [28] was used to create an algorithm relevant to clinical situations. The clinical variables considered relevant were AGE, SEX, CP, and TRESTBPS; the routine test data CHOL, FBS, and RESTECG; the exercise electrocardiography test with the features THALACH, EXANG, SLOPE, and OLDPEAK; and the non-invasive test, THAL, and CA. In addition, the label was NUM. For comparison, we call this set of 13 features as "Subset-A".

The datasets used for the analysis were Cleveland, Hungarian, and a combination of both called CH (Cleveland-Hungarian). Table 2 displays the data distribution. Cleveland had a more uniform distribution than Hungarian and CH for both healthy individuals and patients with heart disease.

## 2.2 Proposed approach for the dimensionality reduction and classification

The proposed approach was applied to the three datasets referred in Section 2.1. We pre-processed and cleaned the datasets from Cleveland, Hungarian, and CH, as mentioned in Section 2.3. In addition, some of the features were not considered for the analysis, as stated in Table 3. Further, we performed four type of experiments for analysis. We assessed the raw data first with all six classifiers. In the second experiment, we applied the feature selection technique of chi-square to obtain a unique and reduced set of ranking-based features with the diagnosis of the heart disease (NUM) and validate them with the classifiers. The third test used the reduced datasets obtained by chi-square and then applied PCA. The final experiment was the use of PCA directly from raw data. The validation and analysis module used the performance metrics mentioned in Section 2.6, such as accuracy, precision, recall, F1 score, Matthews correlation coefficient (MCC), and Cohen's Kappa coefficient (κ). The representation of this approach is illustrated in Figure 1.

**Table 2.** Datasets distribution

| Dataset | Total # of instances | Presence HF | Absence HF |
|---|---|---|---|
| Cleveland | 283 | 157 (55%) | 126 (45%) |
| Hungarian | 294 | 188 (64.9%) | 106 (35.1%) |
| CH | 577 | 345 (59.8%) | 232 (40.2%) |

**Table 3.** Features not included

| Category | Features not included in the model |
|---|---|
| Irrelevant | ID (patient identification number), social security number (CCF), PNCADEN (sum of PAINLOC, PAINEXER and RELREST), EKGMO (month of exercise ECG reading), EKGDAY (day of exercise ECG reading), EKGYR (year of exercise ECG reading), CMO (month of cardiac cath), CDAY (day of cardiac cath), CYR (year of cardiac cath) |
| Repeated | DUMMY (same as TRESTBPS) |
| Unexplained | RESTCKM, EXERCKM, THALSEV, THALPUL, EARLOBE, LVX1, LVX,2 LVX3, LVX4, LVF, CATHEF, JUNK, NAME |
| Null data | RESTCKM, EXERCKM, RESTEF (rest raidonuclide ejection fraction), RESTWM (rest wall motion abnormality), EXEREF (exercise radinalid ejection fraction), EXERWM (exercise-induce wall motion abnormalities), THALSEV, THALPUL, EARLOBE |

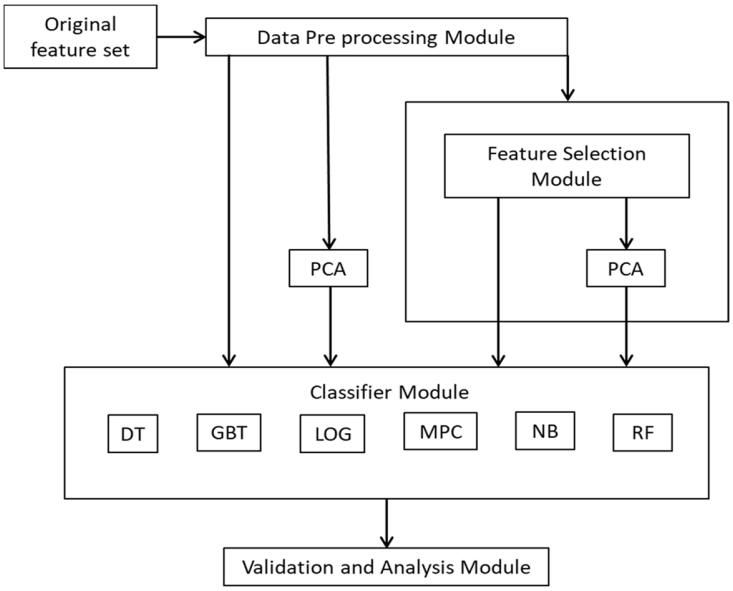


**Figure 1.** Schematic Diagram of Proposed Approach

### 2.3 Preprocessing information and cleansing considerations

The datasets had irrelevant, unexplained, null, or repeated features. Table 3 shows the features not included in the analysis. For this investigation, Cleveland contained 41 features, Hungarian contained 45 features, and CH contained 38 features. The most important considerations for cleansing were to assign a single category for missing values called 'null value' and to create rules that consider data consistency. An example of this is that a patient cannot have cholesterol or age equal to zero. If this occurs, the value will be changed to the 'null value' category.

The complete considerations of the cleaning process were: (1) classes 1, 2, 3, and 4 were converted to the same class (patient with heart disease); (2) null values were replaced by a unique label; (3) zero was unacceptable in continuous results, therefore it was changed as null; (4) if SMOKE was unanswered but CIGS or YEARS were, SMOKE was changed from null to number 1 (patient is a smoker). If CIGS and YEARS had a value of 0, SMOKE was converted to 0; (5) if THALTIME had a value greater than THALDUR, the response was removed; (6) THALACH could not be lower than THALREST; (7) if OLDPEAK had a value of 0, THALTIME was changed to 0; and (8) DUMMY was the same feature as TRESTBPS, so it was eliminated.

### 2.4 Dimensionality reduction

Dimensionality reduction [10] is the process of reducing the number of variables considered. It can be used to extract latent features from raw datasets or to reduce the data while maintaining the structure. This research proposed two different dimensionality reduction methods, for feature selection was selected the chi-square test of independence and for feature extraction, the principal component analysis (PCA).

#### 2.4.1 Chi-square

Chi-square test (CHI) sorts features based on the class and filters out the top features on which depends the class label. ChiSqSelector (CHI) of Apache Spark MLlib is used for feature selection in the model construction. CHI filters the features and sorts them, through repeated iterations, for selection. For this study, we selected the top 13 features using CHI to make a comparison with the literature. Table 4 contemplates the amount of complete data and the correlation between the features and the label. Of the first 13, Cleveland and Hungarian selected 4 vessel features, while CH selected 5; the vessels were LADDIST, RCAPROX, OM1, CXMAIN, and LADPROX. Cleveland used the non-invasive test features, THAL, and CA, while Hungarian selected EXANG. Chest pain values included CP, RELREST, and PAINEXER. The patient records incorporated only CHOL. Exercise

electrocardiogram features indicated THALACH, THALDUR, and ST segment values such as THALTIME, OLDPEAK, and SLOPE. The uncorrelated features involved medications during exercise test, PAINLOC, HTN, SMOKE, FBS, DM, FAMHIST, RESTECG, RAMUS, and OM2. Overall, the common features across the three datasets should be considered as risk factors for heart disease, including CHOL, THALACH, LADDIST, OLDPEAK, THALTIME, RCAPROX, CP, and CXMAIN.

**Table 4.** Features selected by CHI from raw data

| #SF | LF | %DC | Corr | LF | %DC | Corr | LF | %DC | Corr |
|---|---|---|---|---|---|---|---|---|---|
| | Cleveland Dataset | | | Hungarian Dataset | | | CH Dataset | | |
| 1 | CHOL | 100.0 | 0.12 | CHOL | 93.2 | 0.20 | CHOL | 96.5 | 0.17 |
| 2 | THALACH | 100.0 | -0.40 | SLOPE | 35.4 | 0.54 | OLDPEAK | 100.0 | 0.48 |
| 3 | RLDV5E | 100.0 | 0.07 | THALTIME | 35.4 | 0.49 | CP | 100.0 | 0.46 |
| 4 | LADDIST | 100.0 | 0.57 | CXMAIN | 100.0 | 0.59 | CXMAIN | 100.0 | 0.54 |
| 5 | OLDPEAK | 100.0 | 0.42 | LADPROX | 100.0 | 0.56 | THALTIME | 55.0 | 0.38 |
| 6 | THALDUR | 100.0 | -0.25 | EXANG | 99.7 | 0.41 | LADDIST | 100.0 | 0.53 |
| 7 | THALTIME | 75.5 | 0.24 | OLDPEAK | 100.0 | 0.55 | RCAPROX | 100.0 | 0.52 |
| 8 | THAL | 99.3 | 0.44 | THALACH | 99.7 | -0.30 | LADPROX | 100.0 | 0.52 |
| 9 | RCAPROX | 100.0 | 0.51 | CP | 100.0 | 0.51 | EXANG | 99.8 | 0.41 |
| 10 | CP | 100.0 | 0.40 | PAINEXER | 100.0 | 0.54 | THALACH | 99.8 | -0.32 |
| 11 | OM1 | 100.0 | 0.49 | RCAPROX | 100.0 | 0.52 | SLOPE | 67.0 | 0.35 |
| 12 | CA | 99.3 | 0.34 | LADDIST | 100.0 | 0.34 | RLDV5E | 100.0 | 0.11 |
| 13 | CXMAIN | 100.0 | 0.48 | RELREST | 100.0 | 0.45 | OM1 | 100.0 | 0.44 |

#SF= # of selected features; LF=List of Features; %DC=# of data complete; Corr= correlation

### 2.4.2 Principal components

To determine the number of meaningful components to be retained, we select the eigenvalue-one criterion for the analysis. With this, we kept all the components with an eigenvalue greater than 1.00. As individual variables, each component counts for one unit of variance. Therefore, components with an eigenvalue greater than 1.00 stood for a higher variance than their contribution as individual variables. In contrast, components with eigenvalues less than 1.00 contributed less than their individual value and were removed from analysis.

The first 13 components of Cleveland had a variance greater than 1.00 and an accumulated proportion of 0.678. The first two components had a cumulative proportion of 0.246; the amount of variance in component 1 was 5.445 and 4.396 in component 2. The principal components of Hungarian were represented in the first 14 components and had a variance greater than 1.00 and an accumulated proportion of 0.694. The amount of variance in component 1 was 6.340 and in component 2 was 4.451, with a cumulative of 0.240. The first 11 components of CH contained a variance of more than 1.00 and a cumulative proportion of 0.729 of the information. The first two components had an accumulated proportion of 0.399; the amount of variance in component 1 was 14.614 and in component 2 was 4.547. Hence, selecting components with an eigenvalue greater than 1.00 was the best choice, so we selected 13 components for Cleveland, 14 components for Hungarian, and 11 components for CH.

## 2.5 Classifiers proposed

For this research, ML Spark libraries were selected for feature validation. The version of Apache Spark used was 2.2.0 in Java language. MLlib has tools for preprocessing, basic statistics, dimensionality reduction, classification, regression, clustering, and association rules. This work used the CHI for feature selection and PCA for feature extraction. The most important parameter was the “Selection method”, which chose the main features according to CHI as shown in Table 5. The other settings were the default ones.

The classification models use the default value for most of the hyperparameters. The models were: (1) decision tree (DT); (2) gradient-boosted tree (GBT); (3) logistic regression (LOG); (4) multilayer perceptron (MPC); (5)

Naïve Bayes (NB); and (6) random forests (RF). Table 6 describes the parameter settings for each classifier. The GBT and RF trees used several DT parameters as default values, except for Gini impurity. DT hyperparameters are the maximum depth of a tree equal to 5, and the maximum number of bins used when the discretization of continuous features was 32. In addition, LOG had an elasticity of 0.8 and a binomial family parameter. The parameters of MPC were set at a maximum iteration of 100. MPC had two hidden layers, the first with 5 neurons and the second with 4 neurons. The model type selected for NB was multinomial.

The 6 classifiers were run 10 times, the best result was added for this research, and the performance of the label evaluated, in compliance with the percentage of the correct classification. For the experiment, the Heart Disease datasets were divided into two datasets: (1) training dataset with the 70% of the information (80% for training and 20% for validation); and (2) testing dataset with 30% of the information.

**Table 5.** Parameter for feature selection techniques

| Feature selection technique | Basic Hyperarameters |
|---|---|
| ChiSqSelector | Selection method=numTopFeatures;<br>Top features= default set to 50 |

**Table 6.** Parameters tuning for classifier in Apache Spark

| Classifier | Basic Parameters |
|---|---|
| DT | algo="Classification"; numClasses=2; maxDepth=5;<br>minInstancesPerNode="auto"; minInfoGain="auto"; maxBins=32;<br>maxMemoryInMB= 256 MB; subsamlingRate= "auto"; impurity "gini" |
| GBT | Loss="Log Loss"; numIterations="auto"; learningRate="auto";<br>algo="Classification" |
| LOG | numClasses=2; MaxIter=10; RegParam=0.3; ElasticNetParam=0.8;<br>Family="binomial" |
| MPC | Layers="number of features, 2"; BlockSize=128; Seed=1234L;<br>MaxIter=100;<br>two hidden layers, the first with 5 neurons and the second with 4 neurons. |
| NB | Lambda="false"; ModelType="false" |
| RF | numClasses=2; numTrees="auto"; featureSubsetStrategy="false";<br>subsamblingRate="auto"; impurity="gini"; seed="false" |

## 2.6 Evaluation process

The confusion matrix helps practitioners to form a clear idea of whether the results have a high performance. The confusion matrix elements were: (1) true positive (TP), which were patients who had heart disease and were correctly diagnosed; (2) true negative (TN), which were patients who did not have heart disease and were correctly diagnosed; (3) false negative (FN), which were patients who had heart disease and were misdiagnosed; and (4) false positive (FP), which were patients who did not have heart disease and were misdiagnosed. In the medical field, false negatives are the most dangerous predictions.

The different performance metrics were calculated using a confusion matrix. Accuracy (Acc) measured the properly classified instances [1]. The formula for calculating accuracy was given by

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

Precision was the positive predictive value defined by

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall identified the proportion of patients with heart disease given by

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

The F1 score considered a harmonic average between precision in Eq. (2) and recall in Eq. (3) defined by

$$F1\ score = 2\left(\frac{PrecisionxRecall}{Precision+Recall}\right) \tag{4}$$

Matthews correlation coefficient (MCC) was introduced by Brian W. Matthews to predict the performance of protein secondary structure [52]. The results of MCC are in percentage. Therefore, MCC becomes a widely used performance metric in medical research for imbalanced data expressed by

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{5}$$

Cohen's Kappa coefficient (κ) was introduced by J. Cohen [53] in 1960 to correlate the measurement of inter-rater reliability. Kappa measures the percentage of agreement between two raters. The formula to calculate Kappa is represented by

$$\kappa = \frac{p_0 - p_c}{1 - p_c} \tag{6}$$

where $p_0$ is the percent of agreement among raters, as in Eq. (1), and $p_c$ is the chance agreement.


**Figure 2.** Comparison of ML classifiers for Cleveland using accuracy and F1 score

## 3 Results

Significant observations revealed that the use of the selected features of CHI with PCA had the best results with the classifiers across all three datasets in most cases. All performance metrics are in percentage.

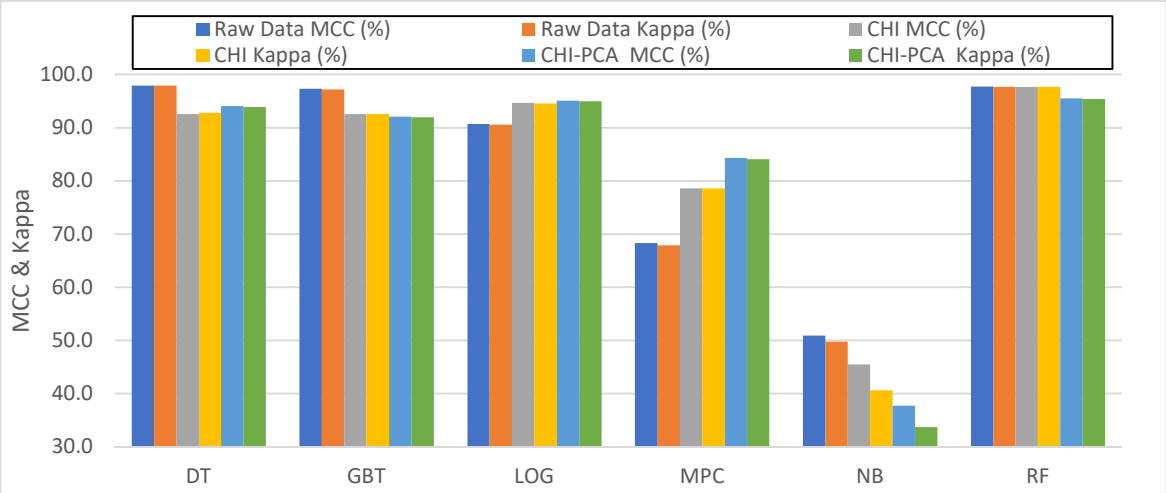


**Figure 3.** Comparison of ML classifiers for Cleveland using MCC and Kappa

### 3.1 Results comparing raw data with CHI-PCA

In this section, we will compare the best results of the raw data with CHI and PCA. For CHI, we selected 13 features, as shown in Table 4. PCA created the principal components using the same CHI features. Overall, Cleveland dataset obtained the best results using CHI-PCA (Figure 2 and Figure 3). Nevertheless, DT and GBT presented better results using raw data. Compared to the raw data, CHI and CHI-PCA improved in the computations of LOG, MPC, NB. However, the performance decreased with DT, and GBT. The greatest improvement was in MPC using the features of CHI-PCA. MPC had an 8.1% accuracy increase, and an F1 score of 9.1%, respectively. RF behavior was the same with raw data and CHI, computing a recall of 100%, an accuracy of 98.9%, an F1 score of 98.8%, an MCC of 97.7%, and a Kappa of 97.7%. CHI-PCA-NB presented the worst value, with an accuracy of 68.4%, an F1 score of 75.7%, an MCC of 37.7%, and a Kappa of 33.7%. GBT presented a pattern when applied with CHI and PCA as shown in Figure 2. The values of MCC and Kappa are consistent among them.

Figure 4 and Figure 5 present the best computations in the Hungarian dataset. The greatest result was CHI-PCA-RF with 99.0% accuracy, 100.0% precision, 96.8% recall, 98.4% F1 score, 97.7% MCC, and 97.6% Kappa. Therefore, CHI-PCA presented the most remarkable performance. GBT, LOG, and RF obtained equivalent results when using CHI-PCA. Even with a lower accuracy, CHI-PCA-GBT computed a perfect recall of 100% and an F1 score of 98.5%, only RF exceeds the result by obtaining a better MCC and Kappa. The computations of DT and NB decreased compared to the raw data. CHI computed the highest values using MPC and DT. Likewise, GBT, LOG, and RF calculated greater results than raw data with CHI features.

Figure 6 and Figure 7 show the performance of the CH dataset. GBT with raw data, CHI-DT, CHI-PCA-LOG, and CHI-PCA-RF computed the highest accuracy of 99.4%. The F1 score was similar in all cases, with a variation of 0.2%. Most models performed better with CHI and CHI-PCA, except for GBT and NB. The greatest improvements between raw data and CHI-PCA were LOG with accuracy increase of 4.5%, F1 score of 4.8%, MCC of 7.5%, and Kappa of 7.5%, respectively. In addition, the values of MCC and Kappa using CHI and CHI-PCA are similar in some cases.

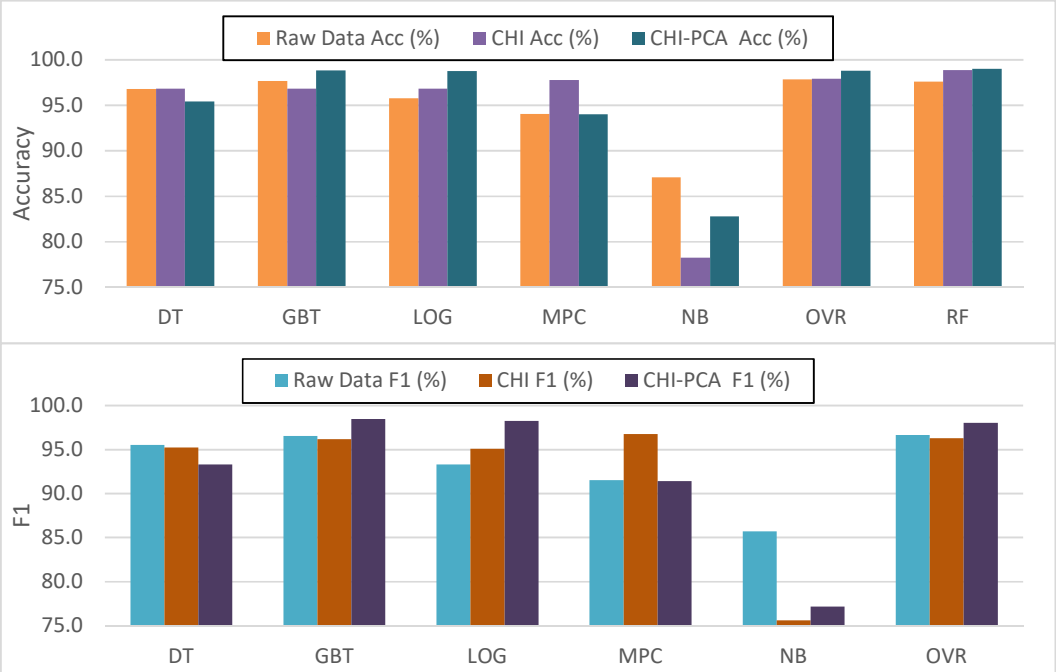


**Figure 4.** Comparison of ML classifiers for Hungarian using accuracy and F1 score

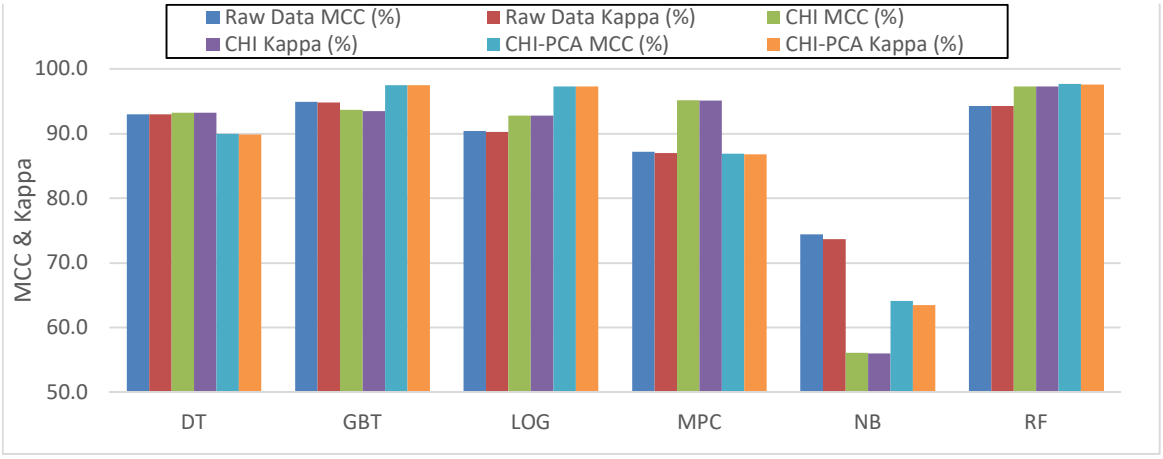


**Figure 5.** Comparison of ML classifiers for Hungarian using MCC and Kappa

### 3.2 Results of the comparison of PCA using raw data and CHI

Table 7 displays a comparison between PCA performance using the features of CHI and the raw data. The use of PCA in raw data had poor results in Cleveland and CH. The performance of the classifiers was reduced to around 30% except for NB and MPC. NB was 3.4% higher in Cleveland and 2.7% higher in Hungarian; MPC computed an accuracy 1.5% greater in Hungarian. Although Hungarian computed lower results in raw data, they were closer to CHI-PCA. Except for MPC, the classifiers were between 4% and 9% lower in accuracy and 6% and 15% inferior in F1 score. As can be seen, PCA retained enough information from the raw data when k was adequate and became less competitive when k was too low or too high. For a large number of features and instances, PCA performance was superior when using CHI features.

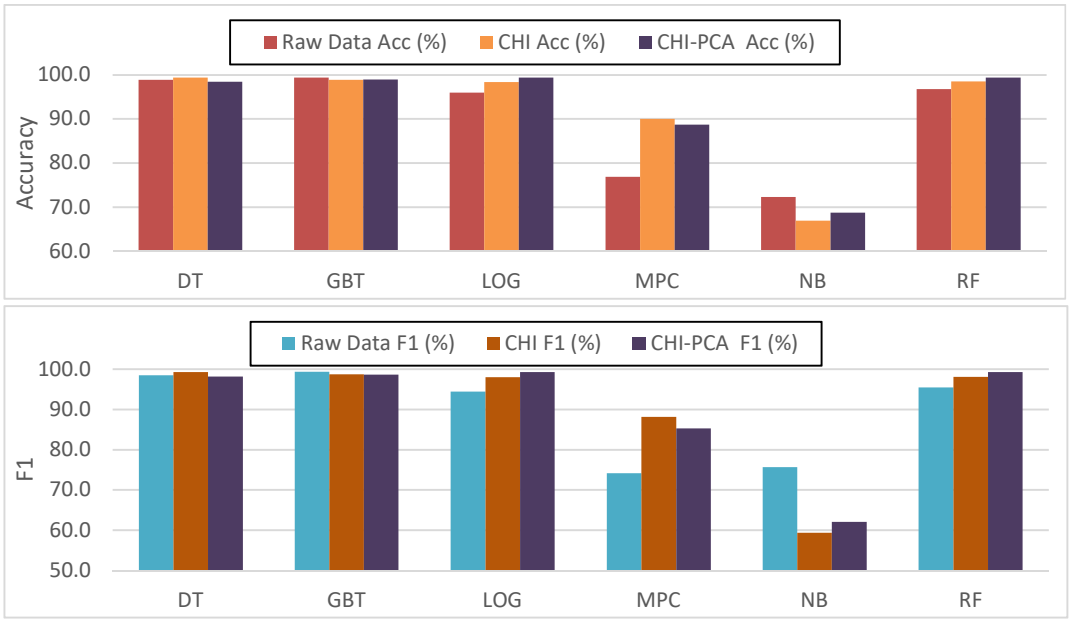


**Figure 6.** Comparison of ML classifiers for CH using accuracy and F1 score

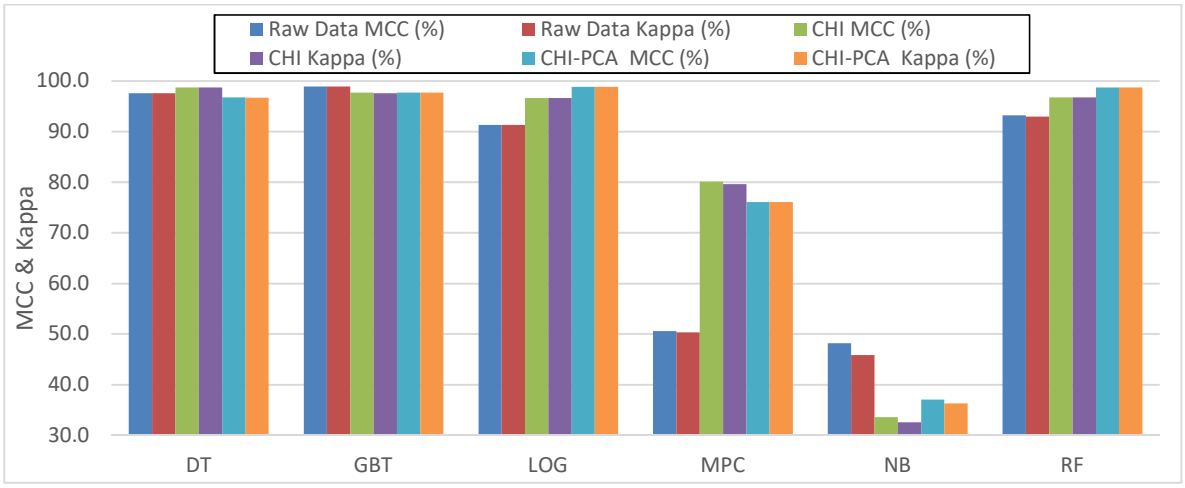


**Figure 7.** Comparison of ML classifiers for CH using MCC and Kappa

### 3.3 Results of the classifiers comparing non-invasive test features

We tested the results obtained by the non-invasive test in Table 8. The features involved were the thallium heart scan (THAL), the number of major vessels colored by fluoroscopy (CA), and whether exercise-induced angina (EXANG). In the case of Cleveland, these features had high data quality. If the results are compared to the Subset-A of Cleveland, some classifiers computed greater values than logistic regression, NB, and SVM. For Hungarian, CA did not exist, and THAL had a weak representation with only 9.5% completed. As a result, a deficient performance is reasonable considering the lack of information. In the case of CH, the representation of THAL was 53.5%. Some classifiers, such as DT, GBT, and NB, achieved competitive results.

**Table 7.** Performance of the raw data and CHI-PCA

| | Performance | DT | GBT | LOG | MPC | NB | RF |
|---|---|---|---|---|---|---|---|
| | | **Cleveland** | | | | | |
| **CHI-PCA** | Accuracy (%) | 97.3 | 96.1 | 97.6 | 92.1 | 68.4 | 98.7 |
| | Precision (%) | 100.0 | 97.1 | 100.0 | 95.2 | 65.0 | 100.0 |
| | Recall (%) | 92.3 | 94.3 | 94.1 | 88.9 | 90.7 | 97.1 |
| | F1 (%) | 96.0 | 95.7 | 97.0 | 92.0 | 75.7 | 98.6 |
| **Raw Data-PCA** | Accuracy (%) | 62.8 | 74.7 | 73.3 | 74.0 | 70.7 | 67.9 |
| | Precision (%) | 57.1 | 59.5 | 60.0 | 74.2 | 67.9 | 77.1 |
| | Recall (%) | 66.7 | 73.3 | 77.8 | 65.7 | 55.9 | 58.7 |
| | F1 (%) | 61.5 | 65.7 | 67.7 | 69.7 | 61.3 | 66.7 |
| | | **Hungarian** | | | | | |
| **CHI-PCA** | Accuracy (%) | 95.5 | 98.8 | 98.8 | 94.0 | 82.8 | 99.0 |
| | Precision (%) | 90.3 | 97.0 | 100.0 | 88.9 | 84.4 | 100.0 |
| | Recall (%) | 96.6 | 100.0 | 96.6 | 94.1 | 71.1 | 96.8 |
| | F1 (%) | 93.3 | 98.5 | 98.2 | 91.4 | 77.1 | 98.4 |
| **Raw Data-PCA** | Accuracy (%) | 88.8 | 89.7 | 94.9 | 95.5 | 78.0 | 93.2 |
| | Precision (%) | 88.0 | 87.0 | 91.4 | 92.6 | 63.6 | 93.1 |
| | Recall (%) | 78.6 | 80.0 | 94.1 | 92.6 | 72.4 | 87.1 |
| | F1 (%) | 83.0 | 83.3 | 92.8 | 92.6 | 67.7 | 90.0 |
| | | **CH** | | | | | |
| **CHI-PCA** | Accuracy (%) | 98.4 | 98.9 | 99.4 | 88.6 | 68.8 | 99.4 |
| | Precision (%) | 100.0 | 97.3 | 100.0 | 87.1 | 70.8 | 100.0 |
| | Recall (%) | 96.3 | 100.0 | 98.6 | 83.6 | 55.3 | 98.6 |
| | F1 (%) | 98.1 | 98.6 | 99.3 | 85.3 | 62.1 | 99.3 |
| **Raw Data-PCA** | Accuracy (%) | 73.7 | 74.3 | 78.6 | 80.5 | 71.5 | 75.1 |
| | Precision (%) | 70.8 | 63.6 | 71.4 | 74.6 | 62.7 | 66.2 |
| | Recall (%) | 52.3 | 60.3 | 70.4 | 72.3 | 88.1 | 70.8 |
| | F1 (%) | 60.2 | 61.9 | 70.9 | 73.4 | 73.2 | 68.5 |

# 4 Discussion

## 4.1 Discussion comparing raw data with CHI-PCA

Promising results were obtained with the use of CHI and PCA. In the first part of Section 3, only DT and GBT lacked improvement in some of the tests. This suggests that tree performance improves when using a large number of features due to supplying more options for the trees. NB dropped the worst results in all the tests. The classifiers LOG had remarkable results when using CHI-PCA. The MPC classifier obtained better results with CHI than CHI-PCA, however the performance was below the rest of the classifiers in most tests due to a network overfit on the training dataset. When we increased the layers or neurons, the performance of the metrics decreased, suggesting that for a small input, such as the datasets in this study, MPC is more stable when using a smaller number of layers. Despite Cleveland, RF made an improvement using CHI-PCA. CHI obtained a remarkable result, and the 13 features selected were prominent for heart disease detection. Significant observations revealed that PCA works best using LOG and CHI using MPC. Overall, LOG, and RF were the classifiers with the best performance and improvement with a smaller number of features.

In most models, the precision value exceeded the recall. Thus, the classifiers computed models that were more sensible to false negatives than false positives. As false positives are examined by practitioners, it is more dangerous to have false negatives. Even so, the value of precision should not be diminished, so it is important to use the F1 score to obtain the optimal balance between precision and recall.

**Table 8.** Performance of non-invasive test values

| Dataset | Performance | DT | GBT | LOG | MPC | NB | RF |
|---|---|---|---|---|---|---|---|
| Cleveland | Acc (%) | 86.6 | 87.7 | 88.0 | 86.8 | 76.1 | 86.2 |
| | Precision (%) | 88.9 | 85.0 | 90.6 | 92.7 | 68.0 | 87.8 |
| | Recall (%) | 82.1 | 85.0 | 82.9 | 80.9 | 85.0 | 81.8 |
| | F1 (%) | 85.3 | 85.0 | 86.6 | 86.4 | 75.6 | 84.7 |
| | MCC (%) | 73.8 | 78.5 | 79.0 | 73.4 | 52.0 | 79.9 |
| | Kappa (%) | 72.7 | 77.9 | 78.7 | 71.4 | 80.8 | 78.8 |
| Hungarian | Acc (%) | 82.9 | 82.3 | 81.9 | 81.5 | 78.7 | 81.3 |
| | Precision (%) | 81.5 | 75.0 | 75.0 | 68.0 | 68.8 | 80.0 |
| | Recall (%) | 71.0 | 72.3 | 67.7 | 73.9 | 71.0 | 68.6 |
| | F1 (%) | 75.9 | 73.8 | 71.2 | 70.8 | 69.8 | 73.8 |
| | MCC (%) | 63.1 | 60.5 | 58.2 | 57.3 | 55.3 | 59.9 |
| | Kappa (%) | 62.8 | 60.5 | 58.1 | 57.1 | 55.2 | 59.4 |
| CH | Acc (%) | 87.3 | 85.6 | 80.2 | 80.5 | 74.9 | 86.4 |
| | Precision (%) | 84.9 | 90.0 | 81.4 | 80.9 | 100.0 | 86.5 |
| | Recall (%) | 78.9 | 76.1 | 66.7 | 63.3 | 40.5 | 81.0 |
| | F1 (%) | 81.8 | 82.0 | 73.3 | 71.0 | 57.7 | 84.0 |
| | MCC (%) | 68.6 | 71.1 | 57.4 | 57.6 | 44.0 | 72.2 |
| | Kappa (%) | 68.6 | 70.4 | 56.8 | 56.7 | 32.5 | 72.1 |

Contrary to accuracy and F1 score, MCC and Kappa show the susceptibility of imbalanced data. In addition, the results were similar between MCC and Kappa in each classifier. LOG computed the best results using CHI-PCA. The Hungarian and CH datasets presented an imbalanced classification problem in which the rate of healthy patients was higher. The raw data in the imbalanced datasets had a greater difference in performance between accuracy and F1 score than the dimensionality reduction results (Table 9). The performance between accuracy and F1 score decreased in the datasets. The difference in Cleveland was not noted in the raw data due to the balance between the two classes. The greatest difference on average was when CHI was used. Even so, each result was lower than the overall average of 0.9%. Hungarian had the biggest difference between accuracy and F1 score due to an imbalance. The CHI and CHI-PCA averages decreased by 0.5% compared to the raw data. CH computed a 1.2% difference in raw data, which was higher than the average. The CHI and CHI-PCA values were 0.5% and 0.7%, respectively. MPC performed better on a balanced dataset such as Cleveland with an average of 0.4%, while Hungarian and CH had the worst performances with 2.0% and 2.7%, respectively. The models with the smallest differences between accuracy and F1 score were GBT, MPC, and RF. NB obtained the poorest results and was excluded from the average results.

**Table 9.** Comparison of the difference in performance between accuracy and F1score

| Dataset | Method | DT | GBT | LOG | MPC | RF | Average |
|---|---|---|---|---|---|---|---|
| Cleveland | Raw data % | 0.3 | 0.0 | 1.2 | 1.1 | 0.1 | 0.54 |
| | CHI % | 0.5 | 2.1 | 1.0 | 0.1 | 0.1 | 0.76 |
| | CHI-PCA % | 1.3 | 0.4 | 0.6 | 0.1 | 0.5 | 0.58 |
| Hungarian | Raw data % | 1.3 | 1.1 | 2.5 | 2.5 | 1.6 | 1.8 |
| | CHI % | 1.6 | 0.6 | 1.7 | 1.0 | 0.8 | 1.14 |
| | CHI-PCA % | 2.2 | 0.3 | 0.6 | 2.6 | 0.6 | 1.26 |
| CH | Raw data % | 0.3 | 0.0 | 1.4 | 2.7 | 1.3 | 1.14 |
| | CHI % | 0.1 | 0.1 | 0.3 | 2.0 | 0.4 | 0.58 |
| | CHI-PCA % | 0.3 | 0.3 | 0.1 | 3.3 | 0.1 | 0.82 |
| Average | | 0.87 | 0.54 | 1.04 | 1.71 | 0.61 | 0.96 |

### 4.2 Discussion of the features selected by chi-square and the PCA results

As in other studies [21, 44-46], the use of PCA after a reduction technique improved the results. The raw dataset produced poorer results in most of the cases (Table 7). The experiment in Section 3.2 compared the performance of raw data with our method. Like the other results, RF improved the computation when using raw data, while MPC

was completely superior using CHI. CHI-PCA outperformed most experiments, especially with LOG as seen in Section 3.1, Section 3.3, and Section 3.4. NB did not present a competitive performance for any of the tests given.

The top 13 features selected by CHI had a great validation for the compilers. The datasets had five vessels. Of these, four of the vessels were part of the left main coronary artery (LAD), considered the most important because it supplies more than half of the blood to the heart. The vessels were the proximal left anterior descending artery (LADPROX), the distal left anterior descending artery (LADDIST), the first obtuse marginal branch (OM1), and the circumflex (CXMAIN). Remaining was the proximal right coronary artery (RCAPROX), which is part of the right coronary artery (RCA). For the non-invasive, the selector considered THAL and CA with a high ranking. Features related to risk factors were not highly ranked by the selector, except for cholesterol. Physicians obtained other features that are part of the exercise test and correlated with heart disease and ST segment values. Taken together, the information obtained by the different tests aided in the diagnosis of heart disease and must be considered for model prediction.

The risk factors that performed best in this study and that are cited by the WHO and the American Heart Association were high blood cholesterol, chest discomfort, inadequate physical activity (seen in the exercise electrocardiogram features). Other features, such as history of hypertension, smoking and the fasting blood sugar, were not complete and were difficult to compare with WHO and American Heart Association standards.

### 4.3 Discussion of the invasive and non-invasive test

The vessels' models were enough to achieve a great result. The invasive test's limitation was its exclusive use in patients with a previous heart attack, severe chest pain, abnormal electrocardiogram or stress test. The non-invasive features performed poorly, so they must be completed with more information. In the literature, other studies worked with the non-invasive testing. This study [51] compared psychological and physiological factors to predict angina on an exercise treadmill test (ETT), concluding that these factors are important in the prediction of exercise angina. Another study [59] concluded that there are sex differences in the experience of chest pain (pain features) and the prediction of exercise-induced angina, while [60] included some painful and non-painful sensations in the relationship with exercise-induced ischemia in women but not in men. [61] concluded that patients are more prone to have long-term survival from preoperative thallium scanning and coronary revascularization before major vascular surgery. Further studies should be conducted with the non-invasive variables.

### 4.4 Discussion of the results with the literature

We compared our results with earlier studies in Table 10 using Cleveland dataset. Our approach achieved greater results with raw data and the use of CHI with PCA. The accuracy, precision, and recall are used for the comparison. Hung et al. [48] classified the data using linear support vector machine (SVM), Naïve Bayes, and logistic regression. The best result was SVM obtaining 89.98% accuracy for raw data. Compared to our results, the values are lower than our classifiers, except for NB and MPC.

**Table 10.** Comparison with other studies using the Cleveland dataset

| Author | Method | Accuracy | Precision | Recall | Features |
|---|---|---|---|---|---|
| Our study | ChiSqSelector+PCA and RF | 98.7% | 100.0% | 97.1% | 13 |
| Shamosollahi, et al., 2019 [47] | C&RT | 92.6% | 92.6% | 90.4% | 20 |
| Shamosollahi, et al., 2019 [47] | ANN | 90.4% | 97.1% | 80.8% | 20 |
| Mounika Naiudu, et al., 2012 [49] | K-mean based MAFIA with ID3 | 85.0% | 80.0% | 85.0% | NA |
| Miao, et al., 2016 [50] | Adaptative Boosting | 80.14% | 81.5% | 71.0% | 29 |

**Table 11.** Comparison with other studies using CH dataset

| Author | Method | Accuracy | Precision | Recall | Features |
|---|---|---|---|---|---|
| Our study | Gradient-boosted Tree (GBT) | 99.4% | 100.0% | 98.7% | Raw data |
| Our study | ChiSqSelector+PCA and Logistic Regression | 99.4% | 97.6% | 100.0% | 13 |
| Hung, et al., 2018 [48] | Linear SVM | 89.9% | NA | 87.0% | Raw data |
| Hung, et al., 2018 [48] | Infinite Latent Feature Selection (ILFS) | 90.7% | NA | 91.0% | 39 |

Each of the classifiers that used CHI-PCA outperformed the literature. Using a dimensionality reduction technique, CHI with PCA and RF classifier computed the best result using 13 features. Shamosollahi, et al. [47] used clustering to determine the k number. After, they performed decision tree and artificial neural network (a hidden layer with 3 nodes). The best result was the C&RT decision tree with an accuracy of 92.6%, and neural networks computing 90.4% accuracy. Mounika Naidu et al. [49] proposed the use of K-mean based on Maximal Frequent Itemset Algorithm (MAFIA) with ID3. The data was clustered using K-means algorithm with k value as 2, then MAFIA used the relevant cluster of 13 features and the ID3. The result of the experimentation was 85.0% accuracy. Miai, et al. [50] used an adaptative Boosting algorithm with an accuracy of 80.14% on Cleveland dataset. In addition, the authors computed on the Hungarian dataset an accuracy of 89.12%, which is below our models, except NB. The results of Cleveland were superior in all the metrics in the literature.

Table 11 shows the comparative performance of CH with the literature. Hung et al. [48] performed the feature selection techniques of Infinite Latent Feature Selection (ILFS), Sort features according to pairwise correlations (CFS), Feature Selection and Kernel Learning for Local Learning-Based Clustering (LLCFS), and PCA. ILFS performed the best computation with 90.65% accuracy and was the classifier that used the least number of features. CFS, LLCFS, and PCA computed 89.93% accuracy, but they need at least 55 features to achieve that result. They used the datasets of Cleveland, Hungarian, and Switzerland. Our method outperformed the literature with a smaller number of features, such as in the case of Cleveland and Hungarian datasets.

**Table 12.** Best models accuracy with other studies using the Subset-A of Cleveland

| Author | Method | Accuracy | Tool | Features | Dataset |
|---|---|---|---|---|---|
| Khanna, et al., 2015 [54] | Logistic Regression | 84.80% | - | 13 | 50% training, 50% testing |
| Khan, et al., 2016 [55] | Decision Tree C4.5 | 89.10% | WEKA | 13 | 70% training, 30% testing |
| Khan, et al., 2016 [55] | Random Forest | 89.25% | WEKA | 13 | 70% training, 30% testing |
| Khanna, et al., 2015 [54] | SVM (linear) | 87.60% | - | 13 | 50% training, 50% testing |
| Kodati, et al., 2018 [56] | Naive Bayes | 83.70% | WEKA | 13 | - |
| Uyar, et. al., 2017 [57] | GA based on RFNN | 96.63% | - | 13 | 85% training, 15% testing |
| Santhanam, et al., 2013 [58] | PCA1 +FFNN | 95.20% | - | 8 ±1 | - |
| Alotaibi, 2019 [62] | SVM | 92.30% | - | 13 | 10-fold cross-validation |
| Latha & Jeeva, 2019 [63] | Majority vote with NB, BN, RF and MP | 85.48% | - | 9 | - |
| Gupta et al., 2019 [64] | MIFH (Factor analysis of mixed data + RF) | 93.44% | - | 28 | - |

Table 12 contains the best models from the literature considering different methods using the features of Subset-A. Some studies had high accuracy [57, 58, 62, 64], but we must consider that most of them only used accuracy, and important metrics such as precision, recall, and F1 score were not considered for evaluation. Furthermore, the dataset division was not mentioned in the articles [56, 58, 63, 64], or the testing dataset was small, as in [57], with only 45 test instances, which can lead to an error in the results. The non-hybrid models computed promising results for the trees [55], but SVM [62] was the only one with more than 90% accuracy. There is a gap in the SVM results between [54] and [62] of 4.7% with the only difference in the dataset, this must be address and verify in the future. According to the hybrid models comparison, when the dimensionality reduction was used, a better prediction of heart disease was obtained. Similar to our model, some of the hybrid models in the literature outperformed the others with the use of RF. Our CHI-PCA logistic regression model had the second greatest improvement when compared with raw data, this can be observed in [54] where the performance was the worst among the non-hybrid models.

Based on these results, our model outperformed those in the literature. It is important when practitioners can only work with three or four times less than the given number of features and achieve competitive results compared to full features. Our method helps to reduce unnecessary patients' attributes and reduce the amount of data.

## 5 Conclusion

In this paper, we proposed the use of a chi-square (CHI) with PCA to improve the prediction of machine learning models. The goal for the classifier was to predict whether a patient has heart disease. Use of complete features is not feasible when the system resources need to be considered. In this study, we successfully applied dimensionality reduction techniques to improve the raw data results. For the 74 features given, we selected three groups of features and achieved the best performance. It was found that among the classifiers, CHI-PCA with RF had the maximum performance, with 98.7% accuracy for Cleveland, 99.0% accuracy for Hungarian, and 99.4% accuracy for CH. Our aim is to find the best dimensionality reduction method for the prediction of heart disease in terms of performance, for this reason, CHI-PCA was the most consistent and preferable method.

From the analysis, chi-square derived features of anatomical and physiological relevance, such as cholesterol, maximum heart rate, chest pain, features related to ST depression, and heart vessels. Our method can be employed in many real-life applications or in other medical diagnoses to analyze great amounts of data and identify the risk factors involved in different diseases. Our main limitation is the difficulty to extend these findings on heart disease due to the small sample size. For future developments, we plan to apply our method to a larger dataset and perform the analysis of some other disease with different feature selection techniques.

## Acknowledgments

This study was funded by the Consejo Nacional de Ciencia y Tecnología (CONACYT- Mexico; grant number: 568729) and by the Institute of Innovation and Technology Transfer of Nuevo Leon, Mexico.

## References

1. Cardiovascular diseases (CVDs). (2019, July 16). Retrieved from http://www.who.int/cardiovascular_diseases/en/.
2. American Heart Association. Classes of Heart Failure. (2018, August 11). Retrieved from https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/classes-of-heart-failure.
3. Heart Failure. (2018, June 19). Retrieved from http://www.heart.org/HEARTORG/Conditions/HeartFailure/Heart-Failure_UCM_002019_SubHomePage.jsp.
4. Shalev-Shwartz, S., Ben-David, S. (2016). *Understanding machine learning: From theory to algorithms*. New York: Cambridge University Press.
5. Melillo, P., Luca, N. D., Bracale, M., & Pecchia, L. (2013). Classification Tree for Risk Assessment in Patients Suffering From Congestive Heart Failure via Long-Term Heart Rate Variability. *IEEE Journal of Biomedical and Health Informatics*, 17(3), 727-733. doi:10.1109/jbhi.2013.2244902.

6. Rahhal, M. A., Bazi, Y., Alhichri, H., Alajlan, N., Melgani, F., & Yager, R. (2016). Deep learning approach for active classification of electrocardiogram signals. *Information Sciences*, 345, 340-354. doi:10.1016/j.ins.2016.01.082.
7. Guidi, G., Pettenati, M. C., Melillo, P., & Iadanza, E. (2014). A Machine Learning System to Improve Heart Failure Patient Assistance. *IEEE Journal of Biomedical and Health Informatics*, 18(6), 1750-1756. doi:10.1109/jbhi.2014.2337752.
8. Zhang, R., Ma, S., Shanahan, L., Munroe, J., Horn, S., & Speedie, S. (2017). Automatic methods to extract New York heart association classification from clinical notes. *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. doi:10.1109/bibm.2017.8217848.
9. Parthiban, G., & Srivatsa, S. K. (2012). Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients. *International Journal of Applied Information Systems*, 3(7), 25-30. doi:10.5120/ijais12-450593.
10. Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55(10)*, 78. doi:10.1145/2347736.2347755.
11. Wettschereck, D., & Dietterich, T. G. (1995). *Machine Learning*, 19(1), 5-27. doi:10.1023/a:1022603022740.
12. Wettschereck, D., Aha, D.W., & Mohri, T. (1997). *Artificial Intelligence Review*. 11: 273. https://doi.org/10.1023/A:1006593614256.
13. Yang, M., & Nataliani, Y. (2018). A Feature-Reduction Fuzzy Clustering Algorithm Based on Feature-Weighted Entropy. *IEEE Transactions on Fuzzy Systems,* 26(2), 817-835. doi:10.1109/tfuzz.2017.2692203.
14. Chen, R., Sun, N., Chen, X, Yang, M., & Wu, Q. (2018). Supervised Feature Selection With a Stratified Feature Weighting Method. *IEEE Access*, vol. 6, pp. 15087-15098. doi: 10.1109/ACCESS.2018.2815606.
15. Imani, M., & Ghassemian, H. (2015). Feature Extraction Using Weighted Training Samples. *IEEE Geoscience and Remote Sensing Letters*, 12(7), 1387-1391. doi:10.1109/lgrs.2015.2402167.
16. Liu, H., & Motoda, H. (1998). Feature Extraction, Construction and Selection: A Data Mining Perspective. *Springer Science-Busniess Media*, LLC. New York. doi: 10.1007/978-1-4615-5725-8
17. Dun B., Wang E., Majumder S. (2016). Heart Disease Diagnosis on Medical Data Using Ensemble Learning.
18. Singh, R. S., Saini, B. S., & Sunkaria, R. K. (2018). Detection of coronary artery disease by reduced features and extreme learning machine. *Clujul Medical*, 91(2), 166. doi:10.15386/cjmed-882.
19. Yaghouby, F., Ayatollah, A., & Soleimani, R. (2009). Classification of Cardiac Abnormalities Using Reduced Features of Heart Rate Variability Signal. *World Applied Sciences Journal,* 6 (11), 1547-1554.
20. Asl, B. M., Setarehdan, S. K., & Mohebbi, M. (2008). Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal. *Artificial Intelligence in Medicine*, 44(1), 51-64. doi:10.1016/j.artmed.2008.04.007.
21. Guyon, I., Gunn, S., & Nikravesh, M. (2008). Feature Extraction: Foundations and Applications. Springer Science + Business Media. Netherlands.
22. Rajagopal, R., & Ranganathan, V. (2017). Evaluation of effect of unsupervised dimensionality reduction techniques on automated arrhythmia classification. *Biomedical Signal Processing and Control*, 34, 1-8. doi:10.1016/j.bspc.2016.12.017.
23. Zhang, D., Zou, L., Zhou, X., & He, F. (2018). Integrating Feature Selection and Feature Extraction Methods With Deep Learning to Predict Clinical Outcome of Breast Cancer. *IEEE Access*, 6, 28936-28944. doi:10.1109/access.2018.2837654.
24. Negi, S., Kumar, Y., & Mishra, V. M. (2016). Feature extraction and classification for EMG signals using linear discriminant analysis. *2016 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA)*. doi:10.1109/icaccaf.2016.7748960.
25. Avendano-Valencia, D., Martinez-Tabares, F., Acosta-Medina, D., Godino-Llorente, I., & Castellanos-Dominguez, G. (2009). TFR-based feature extraction using PCA approaches for discrimination of heart murmurs. *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. doi:10.1109/iembs.2009.5333772.
26. Kamencay, P., Hudec, R., Benco, M., & Zachariasova, M. (2013). Feature extraction for object recognition using PCA-KNN with application to medical image analysis. *2013 36th International Conference on Telecommunications and Signal Processing (TSP)*. doi:10.1109/tsp.2013.6614055.
27. R, R. N., Susanto, A., Soesanti, I., & Maesadji. (2013). Thoracic X-ray features extraction using thresholding-based ROI template and PCA-based features selection for lung TB classification purposes. *2013 3rd International Conference on Instrumentation, Communications, Information Technology and Biomedical Engineering (ICICI-BME)*. doi:10.1109/icici-bme.2013.6698466.

28. UCI Heart Disease Data set. (2018, September 26). Retrieved from http://archive.ics.uci.edu/ml/datasets/heart+disease, last accessed 2018/06/20.
29. Sen, S. K. (2017). Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms. *International Journal Of Engineering And Computer Science*. doi:10.18535/ijecs/v6i6.14.
30. Khan, S.S. (2016). Prediction of Angiographic Disease Status using Rule Based Data Mining Techniques. *Biological Forum – An International Journal*, 8(2), pp 103-107.
31. Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. *Expert Systems with Applications*, 36(4), 7675-7680. doi:10.1016/j.eswa.2008.09.013.
32. Srinivas, K., Rao, G. R., & Govardhan, A. (2010). Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. *2010 5th International Conference on Computer Science & Education.* doi:10.1109/iccse.2010.5593711.
33. Amma, N. G. (2012). Cardiovascular disease prediction system using genetic algorithm and neural network. 2012 *International Conference on Computing, Communication and Applications*. doi:10.1109/iccca.2012.6179185.
34. Santhanam, T., & Ephzibah, E. P. (2013). Heart Disease Classification Using PCA and Feed Forward Neural Networks. *Mining Intelligence and Knowledge Exploration Lecture Notes in Computer Science*, 90-99. doi:10.1007/978-3-319-03844-5_10.
35. Hastie, T., Tibshirani, R. & Friedman, J. (2017). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
36. Marsland, S. (2015). *Machine Learning: An algorithmic perspective*. Boca Raton, FL: CRC Press.
37. Breiman, L. (2001). Random Forest. *Machine Learning*, Kluwer Academic Publishers 45, 5-32.
38. Friedman, J. H. (1999). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics. 29. 1189-1232. 10.2307/2699986.
39. Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2), 251-257. doi:10.1016/0893-6080(91)90009.
40. Khalaf, A. F., Owis, M. I., & Yassine, I. A. (2015). A novel technique for cardiac arrhythmia classification using spectral correlation and support vector machines. *Expert Systems with Applications*, 42(21), 8361-8368. doi:10.1016/j.eswa.2015.06.046.
41. Shlens, J. (2014). A Tutorial on Principal Component Analysis. *International Journal of Remote Sensing*. 51(2).
42. Detrano, R., Salcedo, E. E., Hobbs, R. E., & Yiannikas, J. (1986). Cardiac cinefluoroscopy as an inexpensive aid in the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 57(13), 1041-1046. doi:10.1016/0002-9149(86)90671-5.
43. Detrano, R., Jánosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., et al. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5), 304-310. https://doi.org/10.1016/0002-9149(89)90524-9.
44. Wang, T., Liu, F., Zhou, W., Zhu, X., & Guan, S. (2016). Neural incremental attribute learning based on principal component analysis. *2016 IEEE International Conference on Big Data Analysis (ICBDA)*. doi:10.1109/icbda.2016.7509830.
45. Deventer, H. V., Cho, M. A., Mutanga, O., Naidoo, L., & Dudeni-Tlhone, N. (2015). Reducing Leaf-Level Hyperspectral Data to 22 Components of Biochemical and Biophysical Bands Optimizes Tree Species Discrimination. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6), 3161-3171. doi:10.1109/jstars.2015.2424594.
46. Anindita, N., Nugroho, H. A., & Adji, T. B. (2017). A Combination of multiple imputation and principal component analysis to handle missing value with arbitrary pattern. *2017 7th International Annual Engineering Seminar (InAES*). doi:10.1109/inaes.2017.8068537.
47. Shamosollahi, M, Badiee, A., & Ghazanfari M. (2019). Using Combined Descriptive and Predictive Methods of Data Mining for Coronary Artery Disease Prediction: a Case Study Approach. *Journal of Artificial Intelligence & Data Mining (JAIDM*). 7(1), 47-58.
48. Le, H. M., Tran, T. D., & Tran, L. V. (2018). Automatic Heart Disease Prediction ing Feature Selection And Data Mining Technique. *Journal of Computer Science and Cybernetics*, 34(1), 33-48. doi:10.15625/1813-9663/34/1/12665.
49. Naidu, M., & Rajendra C. (2012). Detection of health care using datamining concepts through web. *International Journal of Advanced Research in Computer Engineering & Technology*, 1(4).
50. H., K., H., J., & J., G. (2016). Diagnosing Coronary Heart Disease using Ensemble Machine Learning. *International Journal of Advanced Computer Science and Applications*, 7(10). doi:10.14569/ijacsa.2016.071004.

51. Bekkouche, N. S., Wawrzyniak, A. J., Whittaker, K. S., Ketterer, M. W., & Krantz, D. S. (2013). Psychological and physiological predictors of angina during exercise-induced ischemia in patients with coronary artery disease. *Psychosomatic medicine*, 75(4), 413–421. doi:10.1097/PSY.0b013e31828c4cb4
52. Matthews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et biophysica acta*. 1975; 405(2):442–451. https://doi.org/10.1016/0005-2795(75)90109-9.
53. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46. doi:10.1177/001316446002000104.
54. Khanna, D., et al. 2015. Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease. *International Journal of Machine Learning and Computing*, vol. 5, no. 5, pp. 414–419.
55. Khan, S.S. 2016. Prediction of Angiographic Disease Status using Rule Based Data Mining Techniques. *Biological Forum – An International Journal*, 8(2), pp 103-107.
56. Kodati, S. "Analysis of Heart Disease using in Data Mining Tools Orange and Weka". *Global Journal of Computer Science and Technology*, vol 18-1, 2018.
57. Uyar, Kaan, and Ahmet Ilhan. Diagnosis of Heart Disease Using Genetic Algorithm Based Trained Recurrent Fuzzy Neural Networks. *Procedia Computer Science*, vol. 120, 2017, pp. 588–593.
58. Santhanam, T., and E. P. Ephzibah. "Heart Disease Classification Using PCA and Feed For-ward Neural Networks." *Mining Intelligence and Knowledge Exploration Lecture Notes in Computer Science*, 2013, pp. 90–99.
59. D'Antono, B., & Dupuis, G., Fleet, R., Marchand, A., and Burelle, D. (2003). Sex differences in chest pain and prediction of exercise-induced ischemia. *The Canadian journal of cardiology*. 19. 515-22.
60. D'Antono, B., Dupuis, G., Fortin, C., Arsenault, A., & Burelle, D. (2006). Detection of exercise-induced myocardial ischemia from symptomatology experienced during testing in men and women. *The Canadian journal of cardiology*, 22(5), 411–417. doi:10.1016/s0828-282x(06)70927-8
61. Giora Landesberg, Yacov Berlatzky, Moshe Bocher, Ron Alcalai, Haim Anner, Tatyana Ganon-Rozental, Myron H. Luria, Inna Akopnik, Charles Weissman, Morris Mosseri, A clinical survival score predicts the likelihood to benefit from preoperative thallium scanning and coronary revascularization before major vascular surgery, *European Heart Journal*, Volume 28, Issue 5, March 2007, Pages 533–539, https://doi.org/10.1093/eurheartj/ehl390
62. Alotaibi, F. S. (2019). Implementation of Machine Learning Model to Predict Heart Failure Disease. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 10(6). http://dx.doi.org/10.14569/IJACSA.2019.0100637
63. Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203. https://doi.org/10.1016/j.imu.2019.100203
64. Gupta, A., Kumar, R., Singh Arora, H., & Raman, B. (2020). MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis. *IEEE Access*, *8*, 14659–14674. https://doi.org/10.1109/ACCESS.2019.2962755