Identify Best Learning Method for Heart Diseases Prediction Under impact of Different Datasets Characteristics

Zahraa Ch. Oleiwi
Faculty of Computer Science and Mathematics, University of Kufa, Najaf, Iraq
zahraac.albadri@student.uokufa.edu.iq
https://orcid.org/0000-0001-9969-7394

Ebtesam N. AlShemmary
IT Research and Development Center, University of Kufa, Najaf, Iraq
dr.alshemmary@uokufa.edu.iq
https://orcid.org/0000-0001-7500-9702

Salam Al-augby
Faculty of Computer Science and Mathematics, University of Kufa, Najaf, Iraq
salam.alaugby@uokufa.edu.iq
https://orcid.org/0000-0001-8247-9497

DOI: http://dx.doi.org/10.31642/JoKMC/2018/100104

Received Aug. 16, 2022. Accepted for publication Sept. 28, 2022

Abstract— This paper introduces an experimental study of the heart disease datasets characteristics impact on the performance of classification algorithms in the aim of identifying the best algorithm for each dataset under its characteristics. The performance of five machine learning algorithms (logistic regression (LR), K-Nearest Neighbor (KNN), Decision tree (DT), Random Forest (RF), and support vector machine (SVM)), single layer neural network (ANN), and deep neural network (DNN), has been evaluated using five heart disease datasets under four data complexity measurement: number of samples (dataset size), number of features (dimension of dataset), Data sparsity measures, and correlation of features. All datasets have been processed and normalized then the mutual information-based feature selection method was used to solve the overfitting problem. The results show that in general, the machine learning especially the Random Forest algorithm achieves high classification accuracy than deep learning network. In other hand, the high sparsity and less mutual information of dataset has large impact on degradation of the performance of classification algorithms than other characteristics of data.

Keywords— Cardiovascular diseases; Deep learning; Data sparsity; Machine learning; Random Forest

I. INTRODUCTION

Coronary artery disease, arrhythmias, and other congenital heart defects are all examples of heart disease. Cardiovascular disease is a condition that causes blood vessels to become clogged, resulting in heart attack/angina/stroke. Prediction of cardiovascular disease is an important concern in clinical data analysis because heart disease has become one of the most common causes of death. People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established disease) need early detection and management where in a machine learning model can be of great help[1].

The general steps of prediction techniques are data acquisition and preprocessing, features extraction and selection, classification and prediction methods[1].

What kind of data do cardiologists use for heart diseases diagnosing, and how can we analyze this data to illustrate its complexity and characteristics. Is there any effect of these factors on the performance of the classification techniques and how that requires an optimization step for this data as a pre-process before using it to train the classifier models. Where the proper features extraction and determining significant data are the main challenges in classification problem.

So, the essential solution to this problem is to analysis the datasets and measure its complexity then using feature selection method as one way to improve the quality and information importance of datasets in aims of increasing convergence speed and classification accuracy especially since this matter is related to human life

The datasets used for heart diseases either structured datasets or raw Electrocardiograph (ECG) signals[2]. An ECG signal is a non-stationary, quasi-periodic, multicomponent, and low amplitude of several mV biomedical signals which are characterized by noise susceptibility, and variability among individuals. The ECG is a graphical recording of the human heart muscles electrical activity. The electrocardiogram is used to investigate some types of abnormal heart function including arrhythmias and conduction disturbances, as well as heart morphology (e.g., the orientation of the heart in the chest cavity,

hypertrophy, and evolving myocardial ischemia or infarction). It is also useful for assessing performance of pacemakers[3].
In structured datasets, high blood pressure, high blood cholesterol, smoking, age, environment, occupation, family history, genetics, lifestyle habits, other medical conditions, race or ethnicity, sex, chronic inflammation, and diabetes, The buildup of plaques inside larger coronary arteries, molecular changes associated with aging, chest pain type, maximum heart rate achieved, and the slope of the peak exercise ST segment are most common features to use in diagnosing coronary heart disease[2].
From raw ECG signals different features can be extracted such as Time domain features (temporal/morphological features e.g., heart rate, ST Segment, slope, main wave amplitude, QT, PP, and RR interval ratio), frequency domain features (e.g., transformation coefficients and average power (AP)), time-frequency domain features (e.g., instantaneous frequency IF and PSD), and chaotic and nonlinear features of heart rate variability (HRV) signals[3][4].
An artificial intelligent technique is determined according to type, complexity and characteristic of datasets used.
In this study multiple models of machine and deep learning are used for different datasets to analysis the impact of datasets complexity on the classification accuracy. The aim of this work best methods have been determined as recommendation for next researches.

A. Motivation

Computer-based automated disease diagnosis system will be most useful in medical sectors.

Failure to study the data and show its effect on the accuracy of classification, how this data is processed and improve its quality, selection of important and relevant characteristics and ability to distinguish leads to misdiagnosis, lack of accuracy and time spent on trying more than one method to determine which methods are best for this data which were supposed to improve before the beginning of algorithm training

B. Contribution

1- Datasets are the core of any analytical system and their characteristics have high impact on ML techniques used to extract knowledge
2- Applying five machine learning and deep neural network on five different structure datasets with different characteristic.
3- Analysis and comparative study of multiple models to identify the most robust method that give high accuracy with different datasets and different complex characteristic of data.

C. Challenges

1- Finding and acquiring heart disease datasets with large size (sufficient number of samples), their features characterized by high mutual information, less sparsity, and independent
2- Communicating with cardiologists to make sure of correctness and quality of selected features of datasets

D. *Paper layout*

The rest of this paper is organized as follows. First, the related work is described in Section II. Section III describes the methodology of work. Sections IV, explain the proposed framework. Section V discusses the results and analysis of the proposed work. Finally, the conclusions and future work are presented in Section VI.

II. RELATED WORKS

Different most common data complexity measures were produced and described in [2][5][6][7][8][9] with the aim of determined the effect of data complexity on the selection and constructed classifier. In [5][2] studies, number of features (dataset dimension), size of dataset (number of samples), number of labels, data sparsity, correlation and other factors were described with analysis study of their effect on classification model in term of classification accuracy.

In [2]the effect of different characteristics of seven different datasets on five techniques of feature selection and three classification models were examined in terms of time complexity and classification accuracy. This study produced decision tree to decide the performance of each feature selection method according to the complexity of specific dataset.

Theoretical Complexity Score (TCS) was a complexity measure produced in [5] aimed to measure the complexity of multilabel datasets(MLDs) and its effect on accuracy of classifier.

Another data complexity measures were produced in [8] such as the inter-class, Fisher's Discriminant Ratio, the largest fraction of points denoted by F2, F1, and F3 respectively, the largest fraction of points denoted by L1, L2, and L3. In addition, nearest neighbor measures denoted by N1, N2, and N3 to estimate overlap of inter-class. Also, there are T1 and T2 measures, where T1 refers to total number of hyperspheres measurement while T2 measure the ratio of datasets size to datasets dimension.

In [10] all above measures in [8] were used to evaluate the performance of Synthetic minority over-sampling technique (SMOTE) method of solving imbalance datasets problem.

In [1] the performance of different machine learning methods were evaluated in terms of accuracy and time complexity. Two standard datasets, Hungarian and Statlog were used for examination and analyzing of machine learning techniques such as Linear Regression, Naive Bayes, REP Tree, M5P Tree, Random Tree, JRIP, and J48. This study aimed to produce recommendation of the best classification model-based machine learning, where the random tree achieved high classification accuracy about of 100% and less time of prediction about 0.01 seconds. Using different machine learning methods (KNN, SVM, DT, NB, LR, and ANN), heart diseases prediction system was designed and testing using Cleveland heart disease dataset in [11]. This dataset contains 303 samples with 14 features, so in this study four features selection methods (Relief, MRMR, LASSO, and LLBFS) in addition to proposed features selection methods called fast conditional mutual information feature selection algorithm (FCMIM) were used to improve the classification accuracy and reduce processing time. As a result, the SVM with proposed FCMIM features selection method achieve higher accuracy of 92.37% as compare with other

classifier and features selection methods as well as deep learning model.

By using different methods in [12] a proposed prediction model was produced to heart disease prediction. the first step in this study was combination of five datasets: Statlog, Cleveland, VA, Switzerland, Hungarian and Long Beach to construct one large dataset then Least Absolute Shrinkage and Selection Operator (LASSO) and Relief methods were used as feature selection methods to overcome the overfitting problem and enhance the classification accuracy. The hybrid classifiers were designed by making use from Boosting and Bagging such as: Gradient Boosting Method (GBBM), Decision Tree Bagging Method (DTBM), AdaBoost Boosting Method (ABBM), K-Nearest Neighbors Bagging Method (KNNBM), and Random Forest Bagging Method (RFBM). Analysis results-based comparison study for all models indicated that highest accuracy of 99.05 was achieved using RFBM and Relief feature selection methods.

III. METHODOLOGY

A. Datasets Description

Five standard datasets are used in our study. The summary explanation about these datasets with website link corresponding to each one:

- The first heart disease Cleveland database contains 303 samples and 76 attributes but only 14 of them are used in our study, including the predicted attribute. It available at (https://archive.ics.uci.edu/ml/datasets/heart+disease)
- The heart statlog Cleveland hungary database is combined from different datasets as: Cleveland: 303, Hungarian: 294, Switzerland: 123, Long Beach VA: 200, Stalog (Heart) Data Set: 270 thus, it consists of 1190 instances (patients from US, UK, Switzerland and Hungary) and 11 common attributes. It available at (https://www.kaggle.com/datasets/sid321axn/heart-statlog-cleveland-hungary-final)
- The heart failure prediction dataset consists of 918 observations and 11 attributes, this dataset was created by combining five datasets already available independently. It available at (https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction)

Fourth and fifth datasets was created by The Behavioral Risk Factor Surveillance System (BRFSS: a health-related telephone survey that is collected annually by the Centers for Disease Control (CDC)).

- The heart_disease_health_indicators_BRFSS2015 it contains 253,680 survey responses from cleaned BRFSS 2015 and 21 attributes. It available at (https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset)
- The fifth dataset also come from CDC The dataset contains 18 variables (9 Booleans, 5 strings and 4 decimals) and 319,795 samples. It available at (https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease)

B. Preprocessing

Quality of methods and results depend on the good quality of data. Data in reality should be cleaned since it contains miss, noisy, outlier and inconsist data.

all of the datasets used in this study was preprocessed previously in the source they available in it. In spite of that we make sure that all datasets have no missing values, outliers, or noisy data as it seen in Figure (2).

Some of attributes of datasets have been encoded, so the categorical data convert from nominal values to numerical data to be suitable with some classification methods. The factorize method was used for this purpose which is available in (*pandas)* library in python. This method encodes the object as enumerate value, where the attribute in datasets is factorized to its categoric distinct values and then each distinct value encodes by giving unique integer value to it.

Most of dataset was imbalanced so we processed them to be balanced using down sampling technique by select randomly some of normal instances. All used dataset was normalized using Z-score normalization to make every data point has the same scale so each feature is equally important.

C. Feature Selection and Feature Reduction

In this study we test the correlation between the features before using features selection to see if the number of features is suitable or need to be reduce as well as the feature that high correlated with another feature can be dropped from dataset. This technique-based correlation is used as feature reduction method. In addition, the zero variance features which denoted by constant column has been dropped also since it considers irrelevant feature.

The dimension (number of features) of data set has important impact upon the classification process. So, the correct chose of number and type of feature that are more relevant is essential matter in classification method. Sometime using too much number of features leads to overfitting problem while using small number of features less than required leads to underfitting problem.

The feature selection method used in this work called Mutual information (MI) based feature selection. This method returns mutual information with positive value that describe the dependency between each feature value and the target class, zero value refer to independent variables whereas the high value refers to higher dependency between them[13].

The MI calculated as in (1):

$$I(X;Y) = H(X) - H(X|Y) \dots \dots (1)$$

Where:

$I(X;Y)$ is the mutual information between feature X and target class Y

$H(X)$ is entropy for X and $H(X|Y)$ is the conditional entropy for X given Y

D. Learning Methods

Five machine learning are used in this study: Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), logistic regression, naïve bayes (NB), KNN and single layer neural network (ANN)[14][15][16].

In the other hand, deep neural network is used as deep learning method with four hidden layers[17].

IV. Proposed Framework

Five classifiers-based machine learning and deep neural network were applied on different five datasets with different characteristics (data complexity measures) as it clear in framework in Figure (1). After preparing preprocessing of datasets, four data complexity measures were evaluated for each dataset, then each dataset was divided to 70% of training and 30% of testing data. The results of each classifier for each dataset were analyzed according to data complexity of each dataset in terms of four evaluation metrices (accuracy, precision, recall, and F1_score). The recommendation of which selection methods and which classifier is suitable for each dataset according to its characteristics was produced to be developed as future works.

Bellow is overall program pseudocode of proposed work:

```
# This program load and analysis the datasets then classify it to healthy and patient, finally it predicts the class labels to test samples
Data=Load (dataset path)
Data.shape # return size and dimension of dataset
Data.hist( ) # draw data histogram
#Call data_ preprocessing function(data)
Data_preprocess(data)
# Down sampling for imbalanced data
Normal data=data[target==0]
Abnormal data=data[target==1]
L=Len(Abnormal)
New_Normal = Normal data. Sample (L, random)
New_data= New_Normal. Append (Abnormal data)
#Call correlation_function
corr_features =correlation_function (dataset, threshold=0.7)
New_data.drop(corr_features)
#Call Gini function ()
 impurty_values=(New_data)
# Split data to features and target label)
X=new_data[:,:target]
Y=new_data[:,target]
#Split x to train and test data
X_train, x_test,y_train,y_test(x,y, test ratio=0.3)
#Call mutual_information(new_data) as in aquation(1)
Mut_val=mutual_info_classif(x_train, y_train)
sel_seven_cols = SelectKBest(mutual_info_classif, k=7)
sel_seven_cols.fit(x_train, y_train)
x_train.columns[sel_five_cols.get_support()]
x_test.columns[sel_five_cols.get_support()]
for cloumn in x_train.columns:
if column not
in(x_test.columns[sel_five_cols.get_support()]):
 dropped_col=column

end
end
x_train.drop(dropped_col)
x_test.drop(dropped_col)
model1=logistic Regression
model2= K_Nearest Neighbors
model3= Decision Tree
model4=RandomForest
model5=support vector machine
for each model:
model. Fit(x_train,y_train)
pred=model. Predict(x_test)
#Call evaluation metrics
accuracy_score= metrics.accuracy_score(y_test,pred)
recall_score= metrics.recall_score(y_test,pred)
precision_score = metrics.precision_score(y_test,pred)
f1_score= metrics.f1_score(y_test,pred)
```

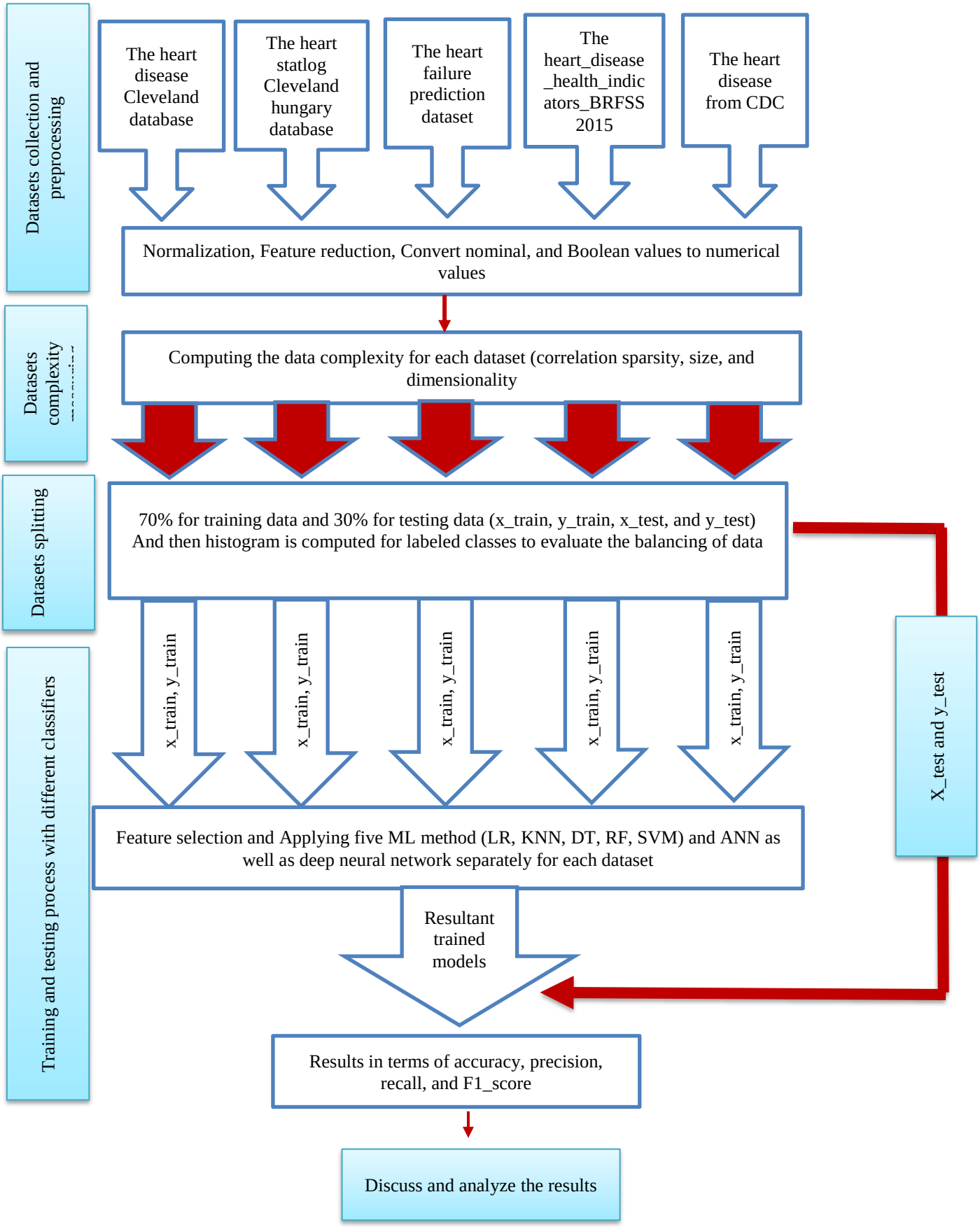


Fig.1. General proposed framework

V. Results and Discussion

A. Dataset Analysis

Each dataset has been analyzed in terms of its characteristics with aims of studying them impact on the accuracy of classification. Four characteristics are considered: dataset dimensions (number of features), size of datasets (number of instances), statistical measure (correlation of features) and data sparsity measure[8][2].

Number of samples (size of datasets) related strongly with selecting the suitable classifier where it affects the performance of classifier by increase the classification accuracy and decrease the underfitting problem due to increasing the variety of information[18]. Table (1) shows the size and dimension correspond for the code of each dataset.

TABLE 1: TWO CHARACTERISTICS AND CODE FOR EACH DATASET

Dataset	Dataset code	Number of samples	Number of features	Number of samples in each class
heart disease Cleveland database	A	303	14	patient 165 normal 138
The heart Statlog Cleveland Hungary database	B	1190	11	patient 629 normal 561
The heart failure prediction dataset	C	918	11	patient 508 normal 410
The heart disease health indicators BRFSS2015	D	253,680	21	Patient 23893 Normal 229787
The heart disease from CDC	E	319,795	18	Patient 27373 Normal 292422

As it can be seen in the table (1) that the two data sets D and E are imbalanced data sets, so we deal with this problem by selecting random samples from the healthy class (which containing a large number of samples) with the same number of samples in the patient class.

Figures (2a-2e) show the histogram of data which indicates the occurrence frequency of each distinct value within the one feature (attribute column). From these figures, the distribution and sparsity of distinct values within each feature can be visualized, as well as they help to make sure if there is an outlier value.

Correlation measure play important role in determining the redundant features that are correlated to each other's with linear relationship between them so it used as pre-step of dimension reduction by identify the feature with high correlation (high dependency) and ignore one of them[19][20]. The threshold value used for measure the degree of correlation is 0.7, where the two features with correlation more than or equal to 0.7 consider dependent features and one of them should be dropped. The results show that the features of all datasets are independent with the maximum value of the correlation between each of the two features being 0.6. Therefore, there is no frequent features that is dropped according to its high cross-correlation.

Data sparsity measures the distribution of significantly value among the distinct value for each feature. One of the sparsity measurements that used in our study is Gini index. Gini index refer to the distribution of distinct value of each feature across various classes, it ranges between 0 (each distinct value in specific feature belong to one class) and 1 (the distinct values distributed randomly across classes). The feature with less Gini index is the least sparse and it has most classification decision (classification purity)[20][21].

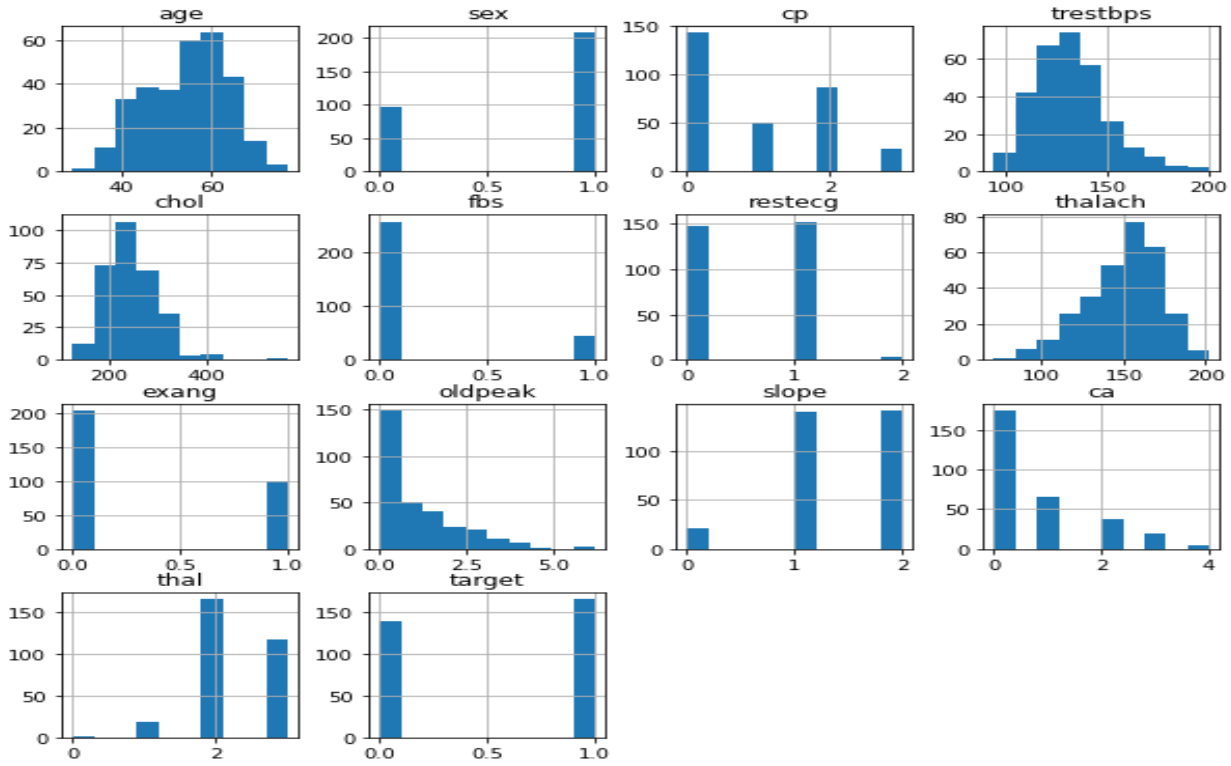
Table (2) shows the impurity of each attribute in each dataset. As it is clear the impurity of datasets is closed to 0.5 then the datasets are high sparsity as in:

$$Gini = 1 - \sum_j p_j^2 \dots\dots\dots\dots\dots (2)$$

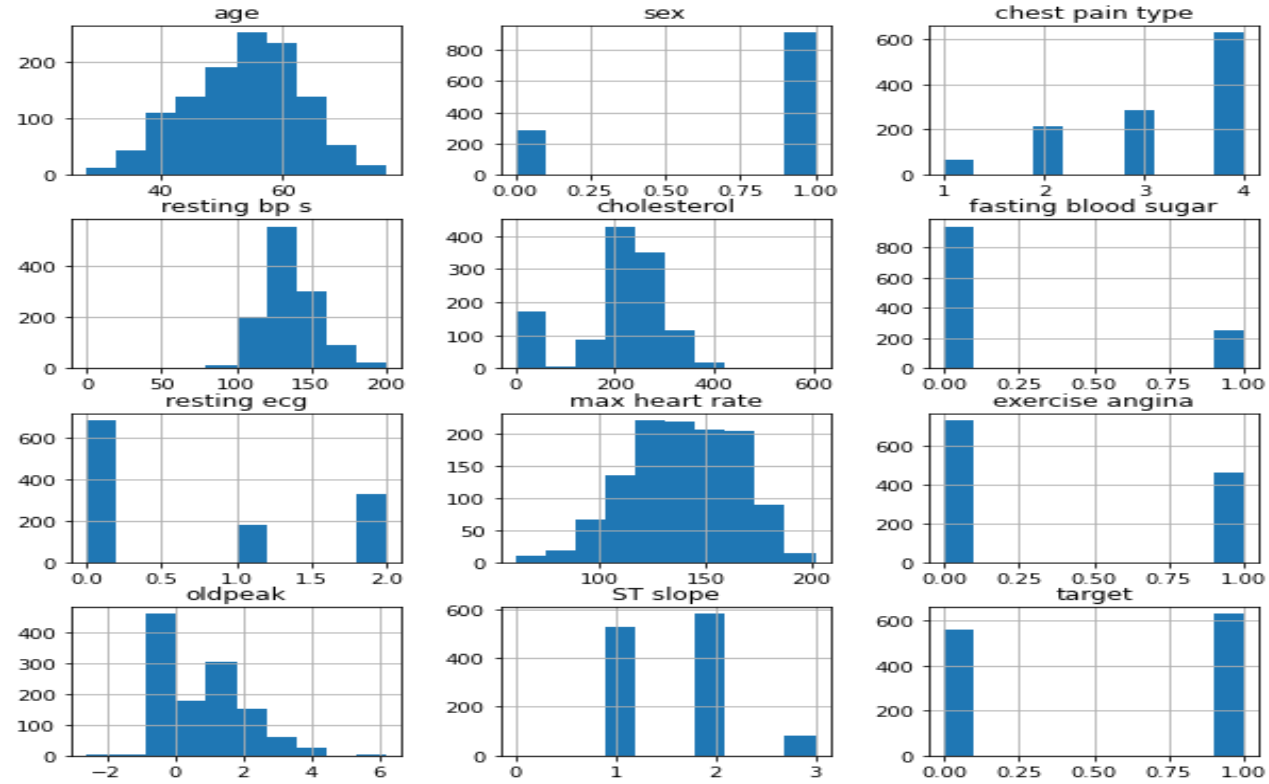
where the p_j^2 is the square of probability of j feature distinct value.

The mean value of impurity for the first three datasets A, B and C is about 0.3 which is less sparsity. In contrast, the results in table (2) for the last two data sets D and E have a high variance value of 0.5 indicating that this data set is high in impurities. So, as it will be clear in tables (3) and (4) this characteristic has high impact on the performance of all used classifier.

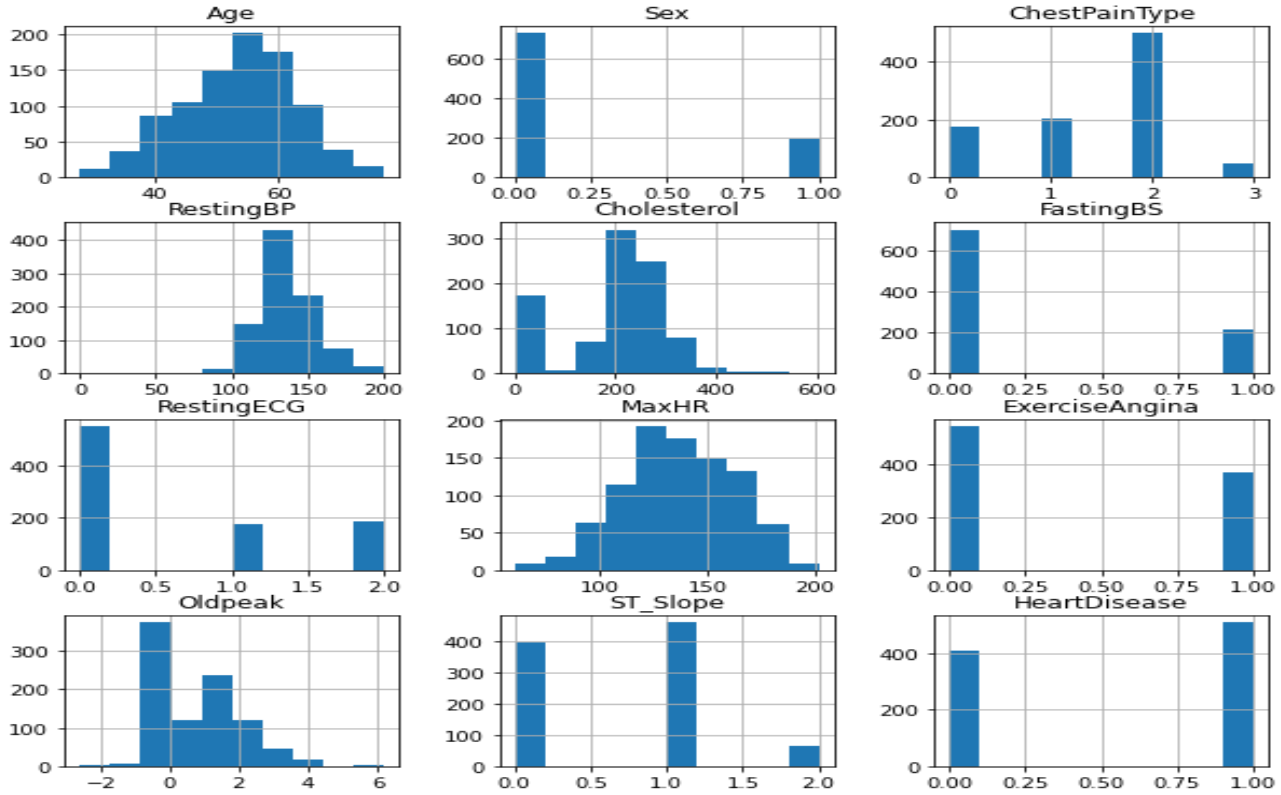
A


age
sex
cp
trestbps
chol
fbs
restecg
thalach
exang
oldpeak
slope
ca
thal
target

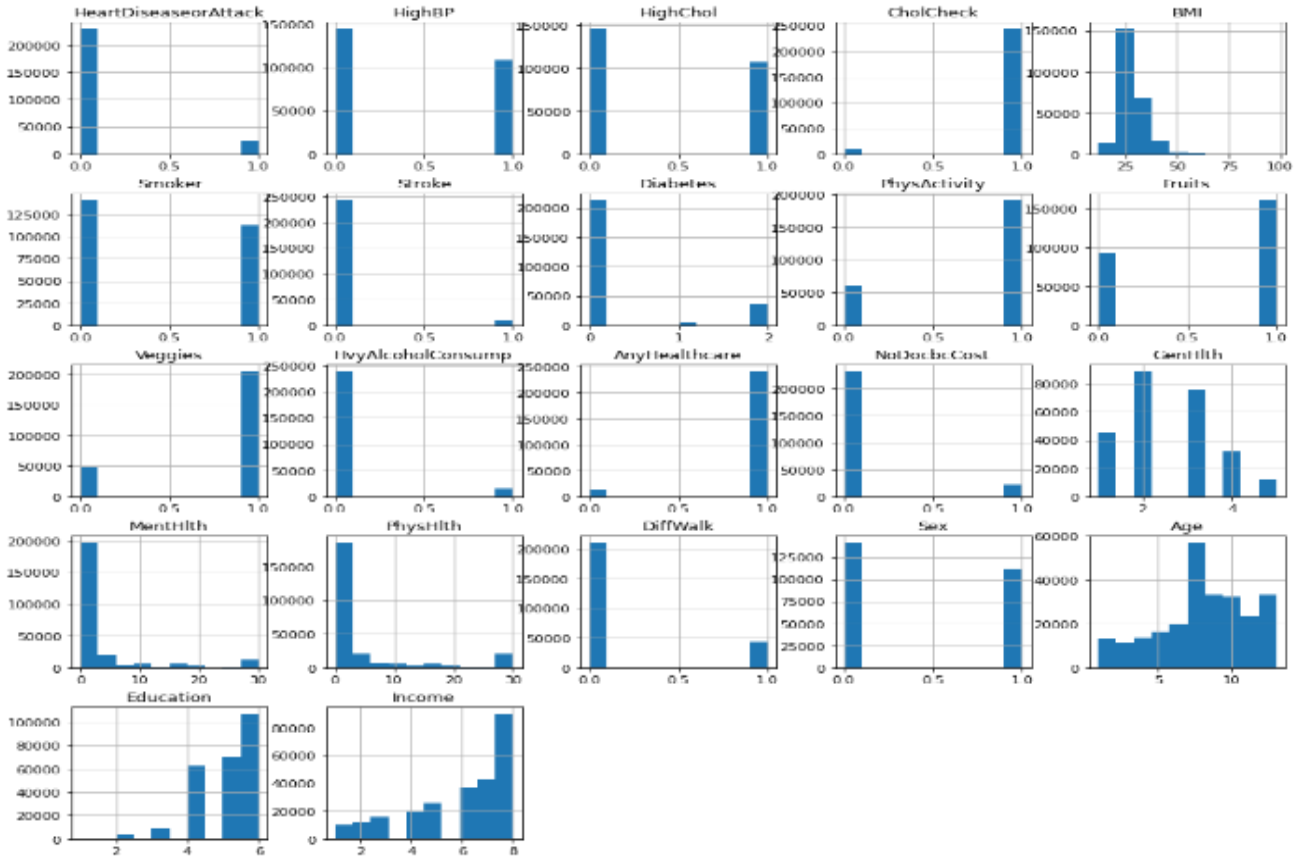
B


age
sex
chest pain type
resting bp s
cholesterol
fasting blood sugar
resting ecg
max heart rate
exercise angina
oldpeak
ST slope
target

C


Age
Sex
ChestPainType
RestingBP
Cholesterol
FastingBS
RestingECG
MaxHR
ExerciseAngina
Oldpeak
ST_Slope
HeartDisease

D


HeartDiseaseorAttack
HighBP
HighChol
CholCheck
BMI
Smoker
Stroke
Diabetes
PhysActivity
Fruits
Veggies
HvyAlcoholConsump
AnyHealthcare
NoDocbcCost
GenHlth
MentHlth
PhysHlth
DiffWalk
Sex
Age
Education
Income

E

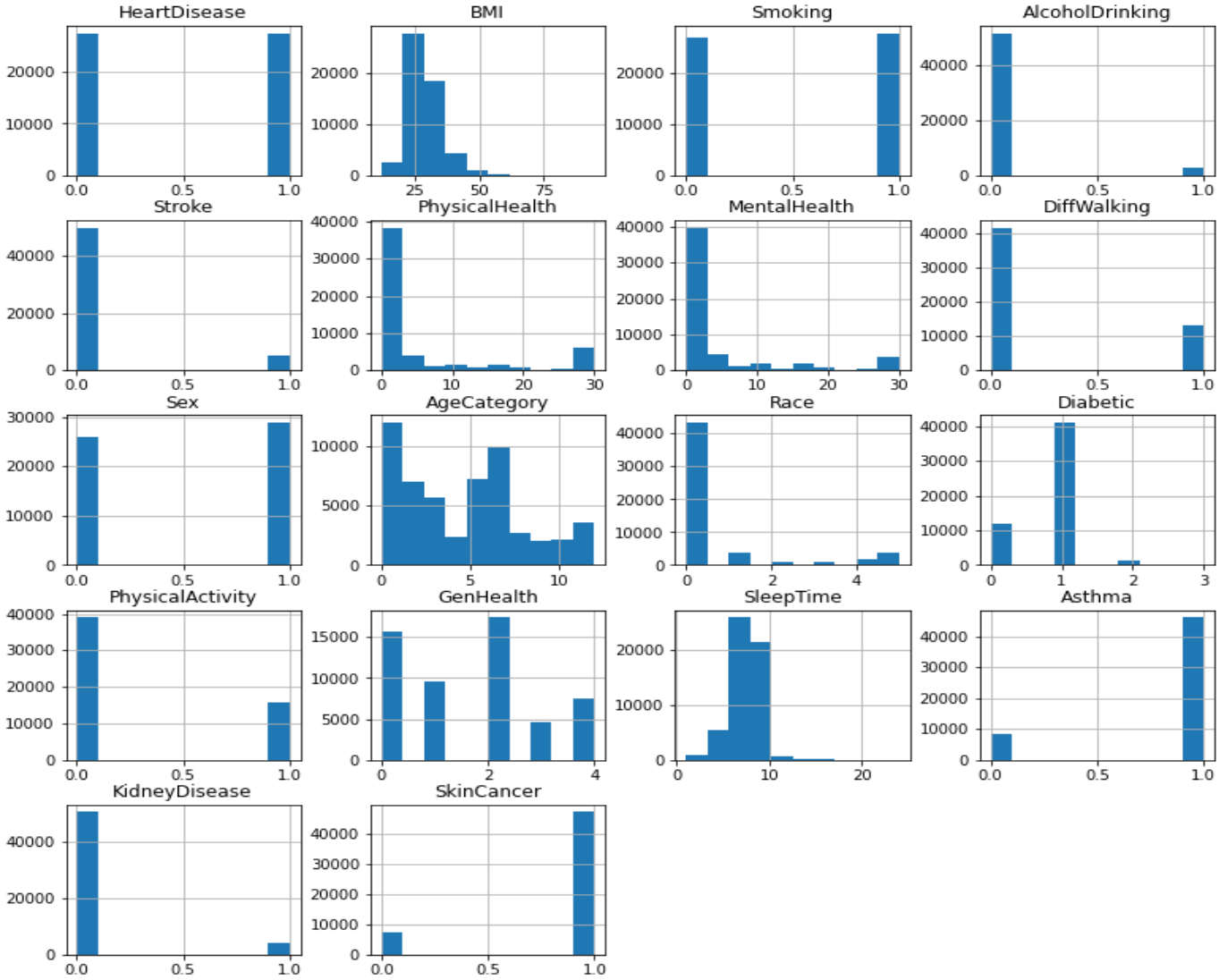


Fig.2. Histogram of each dataset according to its code

Table 2: The sparsity of each feature in each dataset

Data sets A		**Data sets B**		**Data sets C**		**Data sets D**		**Data sets E**	
Attribute	***Gini impurity***	***Attribute***	***Gini impurity***	***Attribute***	***Gini impurity***	***Attribute***	***Gini impurity***	***Attribute***	***Gini impurity***
Age	0.414	Age	0.43	Age	0.427	HighBP	0.434	BMI	0.468
Sex	0.457	Sex	0.45	Sex	0.448	HighChol	0.452	Smoking	0.482
Chest pain type	0.362	Chest pain type	0.35	Chest pain type	0.35	CholCheck	0.496	AlcoholDrinking	0.498
resting bps	0.418	resting bps	0.44	resting bps	0.44	BMI	0.492	Stroke	0.473
Cholesterol	0.213	cholesterol	0.31	cholesterol	0.315	Smoker	0.480	PhysicalHealth	0.467
fasting blood sugar	0.49	fasting blood sugar	0.47	fasting blood sugar	0.459	Stroke	0.473	MentalHealth	0.495
resting ECG	0.48	resting ECG	0.49	resting ECG	0.488	Diabetes	0.467	DiffWalking	0.457
Thalach	0.31	max heart rate	0.36	max heart rate	0.364	PhysActivity	0.489	Sex	0.492
exercise angina	0.4	exercise angina	0.38	exercise angina	0.374	Fruits	0.500	AgeCategory	0.402
Oldpeak	0.35	oldpeak	0.37	oldpeak	0.37	Veggies	0.498	Race	0.495
slope	0.41	ST slope	0.33	ST slope	0.303	HvyAlcoholConsump	0.498	Diabetic	0.464
Ca	0.37					AnyHealthcare	0.499	PhysicalActivity	0.486
thal	0.35					NoDocbcCost	0.499	GenHealth	0.420
						GenHlth	0.414	SleepTime	0.490
						MentHlth	0.494	Asthma	0.498
						PhysHlth	0.462	KidneyDisease	0.484
						DiffWalk	0.453	SkinCancer	0.490
						Sex	0.490		
						Age	0.420		
						Education	0.486		
						Income	0.470		

B. Feature Selection Analyses

Without using feature selection there is clear overfitting in accuracy results according to the characteristics of datasets. The importance of each feature in each dataset used in our study according to mutual information method are shown in figures (3-7). After using our feature selection framework there is clear reduction of overfitting with increasing in accuracy as illustrate in bellow tables. Where the features selected according to its importance were seven features.

Mutual information that measures the importance of features in a data set, is effective for data set A, B and C. From figures (3-5), these measurements make feature reduction method able to drop features of least importance and select only relevant features, as there is a discrepancy between the importance values of the features of these datasets. Thus, this reduction will increase accuracy and avoid overfitting because less important features have no role in decision making.

Based on figures (6) and (7), it is clear that high dimensionality (number of features) has no effect because the mutual information values of all features are few and some of them are close to zero, so they should be dropped because they are considered irrelevant features. Less mutual information values have a significant impact on the performance of all classifiers as will be seen in tables (3) and (4) even if there is an increase in the size of the datasets.

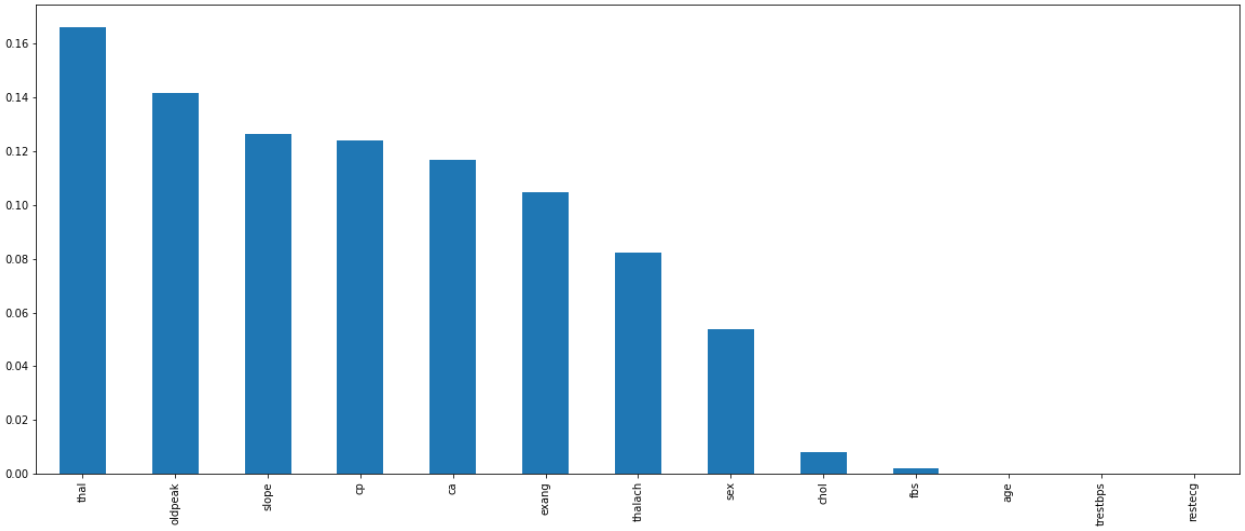


Fig.3. The mutual information value(importance) for the features of A dataset

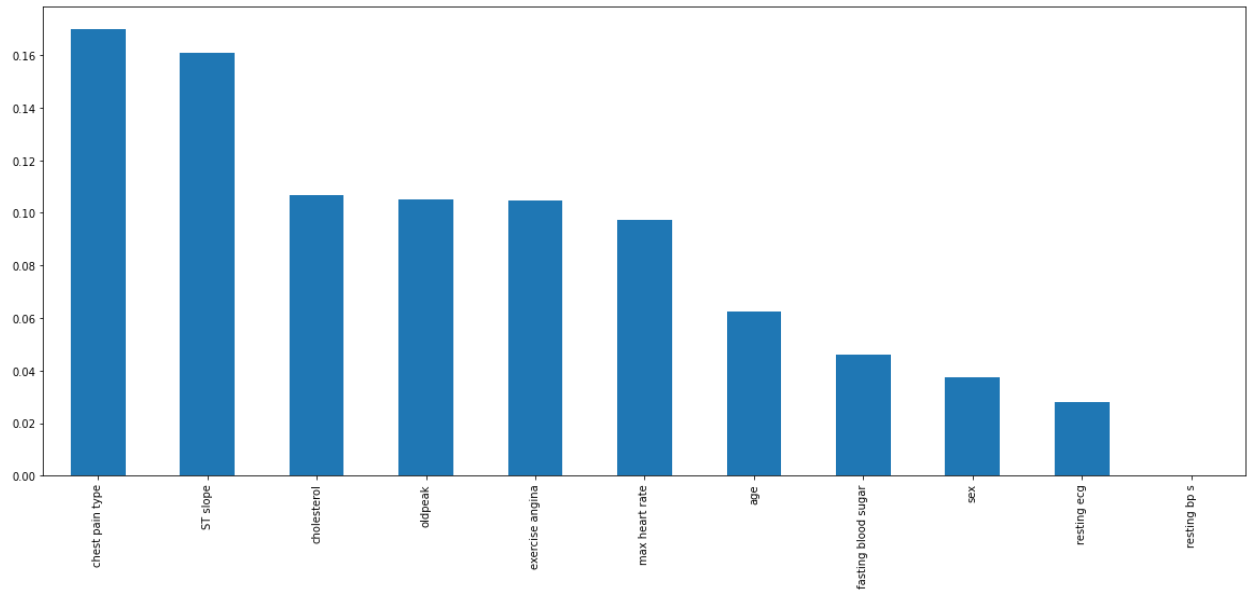


Fig.4. The mutual information value(importance) for the features of B dataset

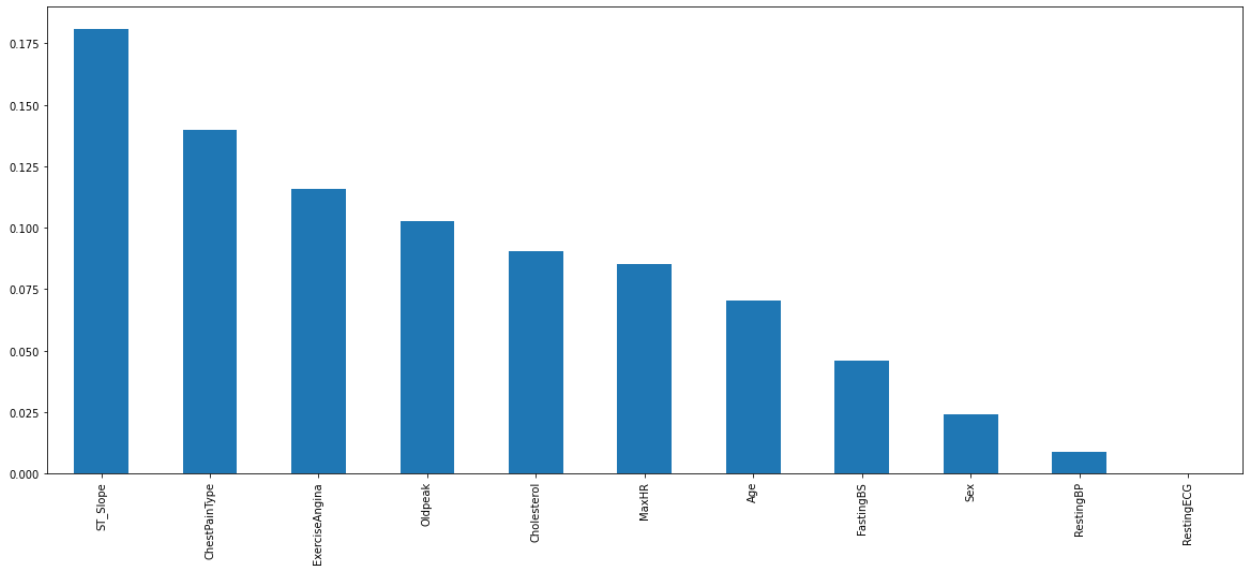


Fig.5. The mutual information value(importance) for the features of C dataset

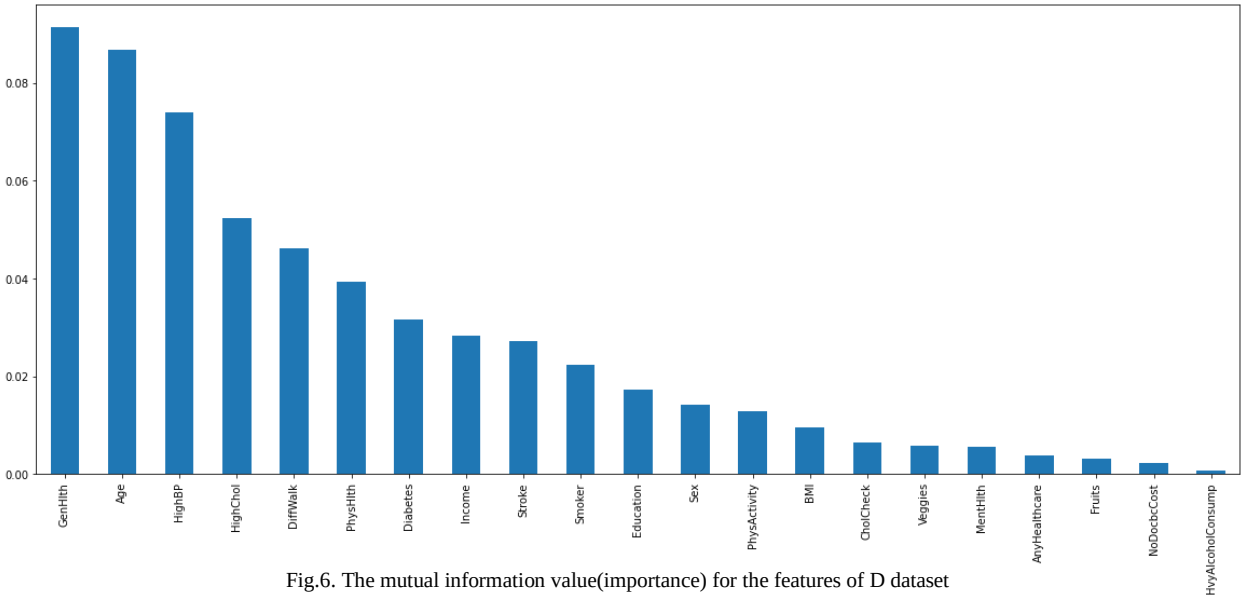


Fig.6. The mutual information value(importance) for the features of D dataset

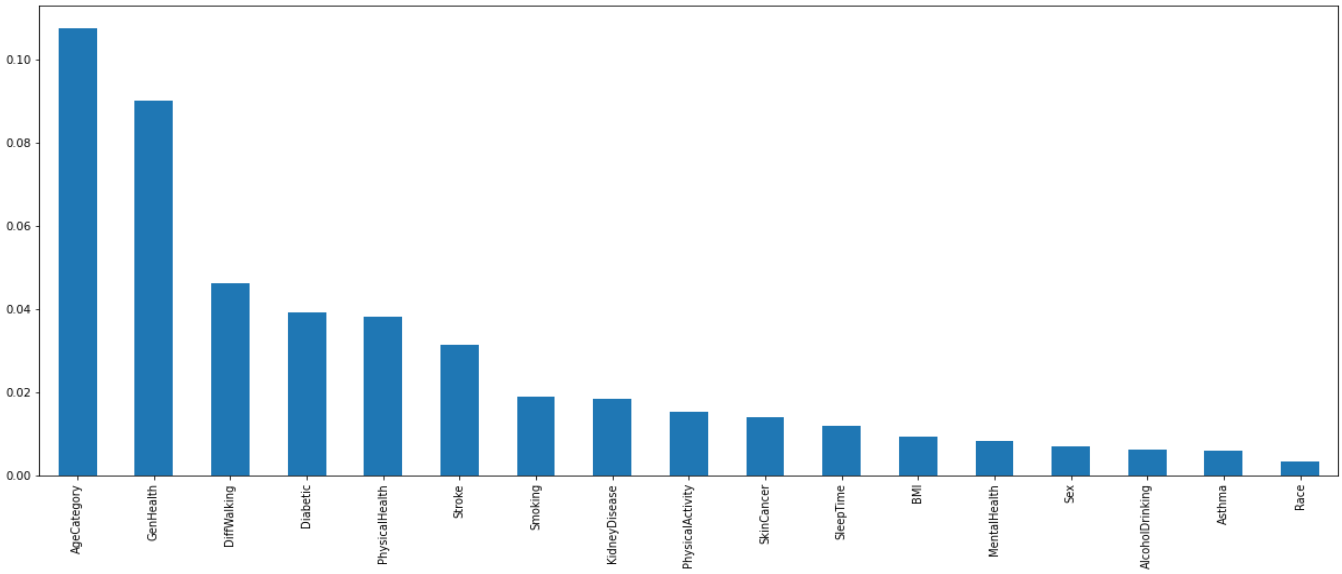


Fig.7. The mutual information value(importance) for the features of E dataset

According to if there is overfit or not, we decide using feature selection or not. As it clear in table (3) and table (4) the accuracy of some algorithms does not increase by using feature selection methods. The reason behind these results is the lack of overfitting so the reduction of dimension causes the underfitting of classification performance. In contrast, the rest algorithms achieve obvious improvement in term of accuracy where there is clear overfitting without using feature selection method as in table (3).

With dataset A there is overfitting problem (where there is clear difference between training accuracy and testing accuracy) with all machine learning therefore we should use feature selection to reduce the complexity of model and increase the accuracy as shown in table (3) with decreasing of overfitting as in table (4).

For dataset B there is an overfitting using DT, RF, SVM, whereas the reset algorithm has no overfit. Then we predict that the using of feature selection methods decrease the accuracy. The results in table (3) for dataset D, come true with our prediction. The overfitting problem DT, RF, SVM decreases with promise ratio using our feature selection methods.

In case of applying classification techniques on C dataset, DT, RF, and SVM have clear overfitting which need to using feature selection methods but our feature selection method has no good effect on the accuracy of these techniques. This problem was solved by increase the size of dataset with a greater number of samples as in dataset B.

There is no effect on performance of reset techniques when using feature selection method where there is no overfitting.

The three above datasets have common features so the effect of increasing number of samples are obvious in increasing the accuracy.
The results of applying DT, RF and KNN on dataset D have clear overfitting as shown in table (3) therefore, a feature reduction method had to be used to reduce the dimensionality of dataset then reduce the overfitting with increasing of accuracy as in table (4). Same thing when applying DT, RF, KNN, and SVM on dataset E.
From the results given in tables (3) and (4), it is clear that there is an obvious underfitting of classification performance for all classifier methods despite the large size of data sets D and E compared to the results in tables (3) and (4) for datasets A, B and C. The reason behind these results is lower Significance (mutual information values) and high impurity values for these datasets.
From all above results and analysis of datasets we can illustrate which machine learning is suitable according to complexity of data as in table (5).
As we show for same dimensionality of datasets the increase of datasets size increases the classification accuracy. Generally Random Forest algorithm achieves the high accuracy for all datasets with different characteristics. As a consequence, for that using machine learning for specific structure datasets used in this paper, can give significant results as compared with deep learning algorithm when there is limitation in size of dataset (limitation of acquires a greater number of samples) because the performance of deep learning increase with the increasing of dataset size. In other hand, the deep learning can find more deeper features than machine learning and achieve more classification accuracy with sufficient samples. The results in tables (3) and (4) gives argument to this outcome, where the deep learning give the same results of machine learning for the datasets D and E with a greater number of samples although these datasets are more complex than other. These results come true with the results achieved in [22] which show that XGBoost produced results and perform better than modern and recent deep learning algorithms on datasets with large size.
The maximum correlation value between features for all datasets not exceed the threshold value 0.7 so there is no feature dropped as redundant value.

The sparsity has clear impact on some ML algorithm especially the DT algorithm which based on information gain of each feature in dataset. As shown in table (3) and (4) for datasets D and E where the high impurity of these datasets is 0.74 closed to 0.5, the accuracy is 69% which is least accuracy as compare with another algorithm and accuracy achieved by DT for reset datasets A, B, and C. So, using our feature selection framework which based on the importance mutual information achieve promise improvement for this problem where the accuracy increases up to 74%. In the other hand, the mutual information values for datasets D and E as shown in figures (6) and (7) are close to 0 for all features in datasets with respect to target class, that mean there is high score of independencies between them and that interpret the less classification accuracy as compare with datasets A, B, and C which have less size and dimensions. This result is important argument of the challenge of extract and find correct and relevant feature.
Deep learning prevents this challenge but it more complex than machine learning. All these results come true with results produced by Oreski et al. (2017) [2] which concluded that the information gain based feature selection more suitable to datasets with low correlation and high sparsity.

Table 3: The accuracy result of all used algorithm for all datasets without using feature selection methods

Classifier model	**A**		**B**		**C**		**D**		**E**	
	Train accuracy	***Test accuracy***	***Train accuracy***	***Test accuracy***	***Train accuracy***	***Test accuracy***	***Train accuracy***	***Test accuracy***	***Train accuracy***	***Test accuracy***
Logistic regression	87%	74%	83%	84%	84%	87%	76%	76%	72%	72%
Decision tree	100%	73%	100%	83%	100%	81%	99%	67%	99%	67%
Random Forest	100%	75%	100%	91%	100%	89%	99%	75%	99%	74%
KNN	73%	62%	87%	87%	88%	88%	79%	72%	78%	71%
SVM	74%	68%	97%	87%	98%	83%	77%	77%	78%	73%
Single layer neural network	67%	67%	83%	83%	71%	71%	77%	77%	72%	72%
Deep neural network	74%	74%	90	90%	88%	88%	%77	%77	74%	74%

Table 4: The accuracy result of all used algorithm for all datasets with using feature selection methods

Classifier model	A		B		C		D		E	
	Train accuracy	*Test accuracy*	*Train accuracy*	*Test accuracy*	*Train accuracy*	*Test accuracy*	*Train accuracy*	*Test accuracy*	*Train accuracy*	*Test accuracy*
Logistic regression	87%	74%	83%	82%	85%	88%	76%	76%	69%	69%
Decision tree	100%	75%	99%	88%	100%	81%	79%	74%	77%	73%
Random Forest	100%	81%	99%	92%	100%	84%	79%	75%	77%	73%
KNN	87%	80%	86%	85%	88%	86%	76%	74%	74%	72%
SVM	87%	80%	92%	88%	92%	83%	76%	76%	73%	73%
Single layer neural network	79%	73%	84%	82%	81%	84%	76%	76%	69%	69%
Deep neural network	91%	80%	88%	88%	85%	87%	76%	76%	73%	73%

Table 5: Complexity measures and best machine learning method in term of accuracy for each dataset

Dataset code	Size (no. of samples)	dimensions	Max correlation value	Mean of sparsity	Feature selection	Best Method	Accuracy
A	303	14	0.5> <0.6	0.386	with	Random forest	81%
B	1190	11	0.5> <0.6	0.398	with	Random forest	92%
C	918	11	0.5> <0.6	0.394	without	Random forest, KNN, DNN	89%
D	47785 (after down sampling)	21	0.5> <0.6	0.474	without	SVM, Single layer NN, DNN	77%
E	54746 (after down sampling)	18	0.4> <0.5	0.474	without	Random forest, DNN	74%

I. Conclusion and Further work

In conclusion, this work aims to find best learning classification method for the heart disease dataset according to its data complexity and show the effect of data complexity on the performance of learning method. Five datasets of heart diseases are used and five machine learning, single layer neural network and deep neural network are applied. Based on the results one can conclude that for all small and medium size datasets including in this paper, with different data complexity measures the machine learning achieves similar or higher classification accuracy than deep neural network, where the Random Forest algorithm achieves higher classification accuracy for all datasets with different characteristics. In addition, SVM and DNN done well with datasets characterized by high dimension and size. In other hand, the less mutual information and high

impurity measures of data have large impact on the performance of classifier techniques even if it has large size and dimensions, where these measures have large impact on the decision-making process, means when these two dataset characteristics are efficient the performance of all classifier algorithms will be higher.

This work paved the way, in near future, to use ECG signal as tool to predict the heart diseases using both machine and deep learning to identify which one is the best in term of complexity and accuracy.

REFERENCES

[1] R. G. Nadakinamani *et al.*, "Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques," *Comput. Intell. Neurosci.*, vol. 2022, 2022.

[2] D. Oreski, S. Oreski, and B. Klicek, "Effects of dataset characteristics on the performance of feature selection techniques," *Appl. Soft Comput.*, vol. 52, pp. 109–119, 2017.

[3] A. Gacek, "An introduction to ECG signal processing and analysis," in *ECG Signal Processing, Classification and Interpretation*, Springer, 2012, pp. 21–46.

[4] K. H. Boon, M. Khalil-Hani, and M. B. Malarvili, "Paroxysmal atrial fibrillation prediction based on HRV analysis and non-dominated sorting genetic algorithm III," *Comput. Methods Programs Biomed.*, vol. 153, pp. 171–184, 2018.

[5] F. Charte, A. Rivera, M. J. del Jesus, and F. Herrera, "On the impact of dataset complexity and sampling strategy in multilabel classifiers performance," in *International conference on hybrid artificial intelligence systems*, 2016, pp. 500–511.

[6] J. Ribeiro, R. Silva, L. Cardoso, and R. Alves, "Does Dataset Complexity Matters for Model Explainers?," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 5257–5265.

[7] F. Branchaud-Charron, A. Achkar, and P.-M. Jodoin, "Spectral metric for dataset complexity assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3215–3224.

[8] N. Anwar, G. Jones, and S. Ganesh, "Measurement of data complexity for classification problems with unbalanced data," *Stat. Anal. Data Min. ASA Data Sci. J.*, vol. 7, no. 3, pp. 194–211, 2014.

[9] Y. Zhang, S. Wei, C. Di Maria, and C. Liu, "Using Lempel–Ziv complexity to assess ECG signal quality," *J. Med. Biol. Eng.*, vol. 36, no. 5, pp. 625–634, 2016.

[10] J. Luengo, A. Fernández, S. García, and F. Herrera, "Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling," *Soft Comput.*, vol. 15, no. 10, pp. 1909–1936, 2011.

[11] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in e-healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, 2020.

[12] P. Ghosh *et al.*, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021.

[13] J. Brownlee, "Information gain and mutual information for machine learning," *Preuzeto*, vol. 18, p. 2020, 2019.

[14] S. Marsland, *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC, 2011.

[15] B. Mahesh, "Machine learning algorithms-a review," *Int. J. Sci. Res. (IJSR).[Internet]*, vol. 9, pp. 381–386, 2020.

[16] J. Alzubi, A. Nayyar, and A. Kumar, "Machine learning from theory to algorithms: an overview," in *Journal of physics: conference series*, 2018, vol. 1142, no. 1, p. 12012.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] C. M. Van der Walt, "Data measures that characterise classification problems." University of Pretoria, 2008.

[19] Y.-H. Chen and S.-N. Yu, "Selection of effective features for ECG beat recognition based on nonlinear correlations," *Artif. Intell. Med.*, vol. 54, no. 1, pp. 43–52, 2012.

[20] M. S. Bin Sinal and E. Kamioka, "An Efficient Arrhythmia Detection Using Autocorrelation and Statistical Approach," *J. Comput. Commun.*, vol. 6, no. 10, pp. 63–81, 2018.

[21] S. Goswami, C. A. Murthy, and A. K. Das, "Sparsity measure of a network graph: Gini index," *Inf. Sci. (Ny).*, vol. 462, pp. 16–39, 2018.

[22] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Inf. Fusion*, vol. 81, pp. 84–90, 2022.