Evolutionary computation technique combined with ensemble model for classification of diabetes

K Satish Kumar*, G V Suryakanth, K Giridhar

Abstract

Machine learning (ML) was a rapidly advancing technology in the modern world. It had a wide variety of applications such as medical diagnosis, stock market trading, email spam, and malware filtering, etc., ML algorithms train the computer to learn from the past data and make predictions on the unknown samples. This research mainly focuses on the prediction of the PIMA Indian diabetes disease. The diabetes dataset was taken from the UCI machine learning repository. The research work was broken down into three stages. The AdaBoost technique was applied to all the features of the PIMA Indian Diabetes dataset. The correlation technique was applied for feature selection and the selected features were trained and tested with AdaBoost. A novel Hybrid Genetic Algorithm (HGA) was designed and developed for feature selection and the selected features were trained and tested with AdaBoost. Even though the correlation identifies the feature subsets based on statistical relevance but it fails in providing optimal feature subset. This drawback was overcome by the proposed novel HGA by selecting an optimal feature subset that can improve the performance of the AdaBoost model. A comparison of correlation and HGA was performed. The HGA with AdaBoost outperformed when compared with correlation with AdaBoost and AdaBoost models in terms of accuracy. The proposed methods were also applied to other datasets like the Wisconsin breast cancer diagnostic and Cleveland heart disease datasets to show its broader applications. The HGA with AdaBoost outperformed other reported techniques for the PIMA Indian diabetes.

Keywords: Genetic algorithm; Correlation; AdaBoost; Pima Indian diabetes

Introduction

Diabetes is a disease caused due to rise in the sugar levels in the human body. At present more than 400 million people are affected with type 2 diabetes. Diabetes if left untreated it results in serious side effects. There is no cure for diabetes; if a patient is affected with diabetes then he has to follow a regular diet and medicine to keep diabetes in control. Diabetes is a serious health concern, only early prediction can save the patients. Generally, doctors go through a series of tests, and based on the test reports a decision is made whether the patient is having diabetes or not. The patient after the medical tests has to wait for a long time to get the report from the doctor stating whether the patient is diabetic or normal. The presence of decision support systems bridges the gap between patients and doctors, by making informed decisions quickly and taking action. Thus the presence of decision support systems benefits society by efficiently addressing growing information and make informed decisions. Once the patient confirms diabetes then only he can book an appointment and get a necessary diagnosis, medication from the doctors. If the patient is normal then there is no need to go to the doctor and he can follow some preventive measures like regular exercise and diet to avoid diabetes in the future.

CSE Department, Koneru Lakshmaiah Education Foundation
Green Fields, India

Corresponding author: Satish Kumar Kalagotla,
e-mail: satish7433@gmail.com

Pima Indian diabetes dataset was a collection of type 2 diabetic samples on several independent variables like pregnancies, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, glucose, and target variable outcome.1 All these observations were made on females of PIMA Indian origin with at least 21 years of age. Blood pressure was the force applied to the muscular walls of the blood vessels. It raises and falls with the phases of the heartbeat. It was measured in mm Hg. skinfold thickness: In females, skinfold thickness was measured at triceps with skinfold caliper; it was the measure of fat situated under the skin. It was measured in mm (millimeter). Insulin: when glucose level increases in the blood, the pancreas secretes insulin; insulin helps in stabilizing glucose in the body. It is measured as 2-Hour serum insulin (mu U/ml). Body mass index (BMI) is the ratio of weight to the height measured in kg/m2. Diabetes pedigree function was the diabetes status of the patient's family history and predicts how likely the patient to be affected by diabetes was. Age was measured in years, in the PIMA Indian diabetes dataset the samples were collected from subjects who are at least 21 years old. Pregnant attributes gave information about how many times a patient became pregnant. Glucose was a measure of plasma glucose concentration for 2 hours in an oral glucose tolerance test. The outcome variable said whether or not a person had diabetes, label 1 indicates a person had diabetes, label 0 indicates a person had no diabetes. The diabetes data set was preprocessed, and it is free from outliers. Inconsistent values were imputed with mean. Pregnancies attribute can hold zero values, zero value means there was no pregnancy for female patients. So zero values of pregnancies attribute were not imputed. The dataset was scaled using a Min-Max scaler, for efficient processing.

Literature survey

Proposed genetic algorithm with Naviebayes (GA_NBs) model that minimizes the computational cost, computational time and maximizes the receiver operating characteristics and accuracy than several existing methods. [2] Designed Genetic Algorithm and Multilayer Perceptron Neural Network (GA_MLP NN) for the classification task.[3] The proposed model showed improvement in classification accuracy and receiver operating characteristics for the PIMA Indian diabetic dataset. The results were compared with other methods reported accuracy for PIMA Indian diabetes dataset. The proposed model reduces the computational cost, time and improves accuracy and ROC.[4] Designed a safe system using Principal Component Analysis (PCA) and Adaptive Neuro-Fuzzy Inference System (ANFIS) for medical diagnostic decision making. PCA was used for dimensionality reduction; the reduced dimensions make the ANFIS classifier more effective. To optimize the parameters in ANFIS a hybrid algorithm that combines least square error (Forward pass) and gradient descent (backward pass) was adopted. Optimization of parameters aims to reduce search space and to achieve a good convergence rate. The proposed technique was also compared with the reported accuracy of other methods for the PIMA Indian diabetes problem.[5] Designed a novel hybrid algorithm based on Artificial Neural Network (ANN) and Fuzzy Neural Network (FNN). The algorithm aims at improving the performance of the classifier by effectively handling both fuzzy and crisp values in the datasets. The proposed algorithm was applied on the PIMA Indian diabetes and Cleveland heart disease datasets. The results obtained were the best when compared with the reported results in the previous studies.[6] Applied six machine learning algorithms like logistic regression, k nearest neighbor, support vector machines, naïve Bayes, decision tree, random forest on new diabetes data set, and the PIMA Indian diabetes dataset. The new diabetes dataset was formed by collecting the data from 952 participants based on the designed questionnaire on various parameters of diabetes. The proposed methods were evaluated on various performance measures like accuracy rate, error rate, sensitivity, specificity, precision, F-measure, Matthews correlation coefficient, and area under the curve. Among the six implemented algorithms, the random forest algorithm showed greater accuracy on both datasets.[7] Designed GA with RBF NN, In the designed model GA, was used for optimal feature subset selection and RBF NN was used as a classification model on the selected features. The model constructed was proven to be efficient in terms of accuracy and ROC with minimized computational cost and complexity. The proposed model was evaluated on various performance metrics. The model performance was compared with other methods reported results for the PIMA Indian diabetes.[8] designed a novel hybrid model for the classification of a real-life type 2 diabetes dataset. The class imbalance problem of the dataset was addressed using the K-means clustering algorithm. SVM was employed for the classification task, besides, a rule-based explanation component like SQRex-SVM, Eclectic was added. The proposed model was proven to be a promising tool for the diagnosis of diabetes.[9] Proposed a model that combines Genetic Algorithm (GA) and fuzzy logic rule-based classifiers. GA was used for selecting good features; the fuzzy logic rule-based classifier was used for the classification of diabetes. The main goal of the proposed approach was to minimize the cost and improve the efficiency of the model. GA tool was employed from MATLAB R2006b for GA implementation. The fuzzy toolbox was also obtained for the classification task. The model achieved optimal results with 3 attributes[.10] Proposed a Kernel Principal Component Analysis (KPCA) Genetic Algorithm (GA) and Support Vector Machine (SVM) as KPCA GA SVM for the classification of diabetes. KPCA was used for feature reduction, GA was used for feature selection, and SVM was used for classification. The proposed model performed better when compared with PCA SVM, GA SVM. The model was also compared with other existing techniques in terms of accuracy, specificity, sensitivity, and Matthews correlation coefficients, the results showed that KPCA GA SVM showed optimal results[.11] Proposed Differential Evolution (DE) Glowworm Swarm Optimization (GSO) Adaptive Neuro-Fuzzy Inference System (ANFIS) as DE GSO ANFIS for achieving better results on real-time data sets like Parkinson and RIM-ONE dataset. ANFIS was good at handling nonlinear datasets. Fuzzy C means (FCM) clustering was used to calculate the membership function in ANFIS to obtain a smaller number of fuzzy rules. GSO suffers from getting stuck at local optima to overcome this limitation DE was coupled with GSO. The modified GSO was then applied to compute the parameters of ANFIS to enhance the classification rate. The results of the proposed model are compared with ANFIS, GA_ANFIS, PSO_ANFIS, LOA_ANFIS, DE_ANFIS, GSO. The proposed DE_GSO_ANFIS had recorded the lowest MSE, RMSE, MAE, and highest R2 measure on both datasets. The developed model can be formed as a second opinion expert to predict suspect cases. This model is highly useful when domain experts were in demand.

From the literature survey, there are a variety of approaches that were followed to improve the classification model. The authors contributed a novel hybrid genetic algorithm (HGA) as a feature selection technique. The AdaBoost classifier with the proposed HGA aims at improving the accuracy with optimal memory use and low computational complexity.

Exploratory data analysis

The type-2 PIMA Indian diabetes dataset having 768 samples of which 500 samples belong to the normal class, 268 samples belong to the diabetes class. The PIMA Indian diabetes dataset samples count for each feature was given in Table 1.

The PIMA Indian diabetes statistics were calculated before pre-processing and presented in (Table 2).

Table 1: *PIMA Dataset information*

Feature Names	Total number of instances
Pregnancies	768
Glucose	768
Blood Pressure	768
Skin Thickness	768
Insulin	768
BMI	768
Diabetes Pedigree Function	768
Age	768
Outcome	768

Table 2: *PIMA Dataset information*

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
count	768	768	768	768	768	768	768	768	768
mean	3.84	120.89	69.1	20.53	79.79	31.99	0.47	33.24	0.34
std	3.36	31.97	19.35	15.95	115.24	7.88	0.33	11.76	0.47
min	0	0	0	0	0	0	0.078	21	0
25%	1	99	62	0	0	27.3	0.24	24	0
50%	3	117	72	23	30.5	32	0.37	29	0
75%	6	140.25	80	32	127.25	36.6	0.62	41	1
max	17	199	122	99	846	67.1	2.42	81	1

The PIMA data has no missing values but it has zero values The below Table 3 shows how many zero counts a feature had, these zero recorded values have to be treated to make the dataset standardized.

Table 3: *PIMA data with zero counts*

Attributes	Zero counts
Pregnancies	111
Glucose	5
Blood Pressure	35
Skin Thickness	227
Insulin	374
BMI	11
Diabetes Pedigree Function	0
Age	0

Outlier analysis was performed so that while imputing the wrongly recorded values with mean will not affect the distribution.[12] Table 4 describes the statistics of the PIMA Indian diabetes data after performing outlier analysis. The outliers were removed from the dataset to make the dataset standard.

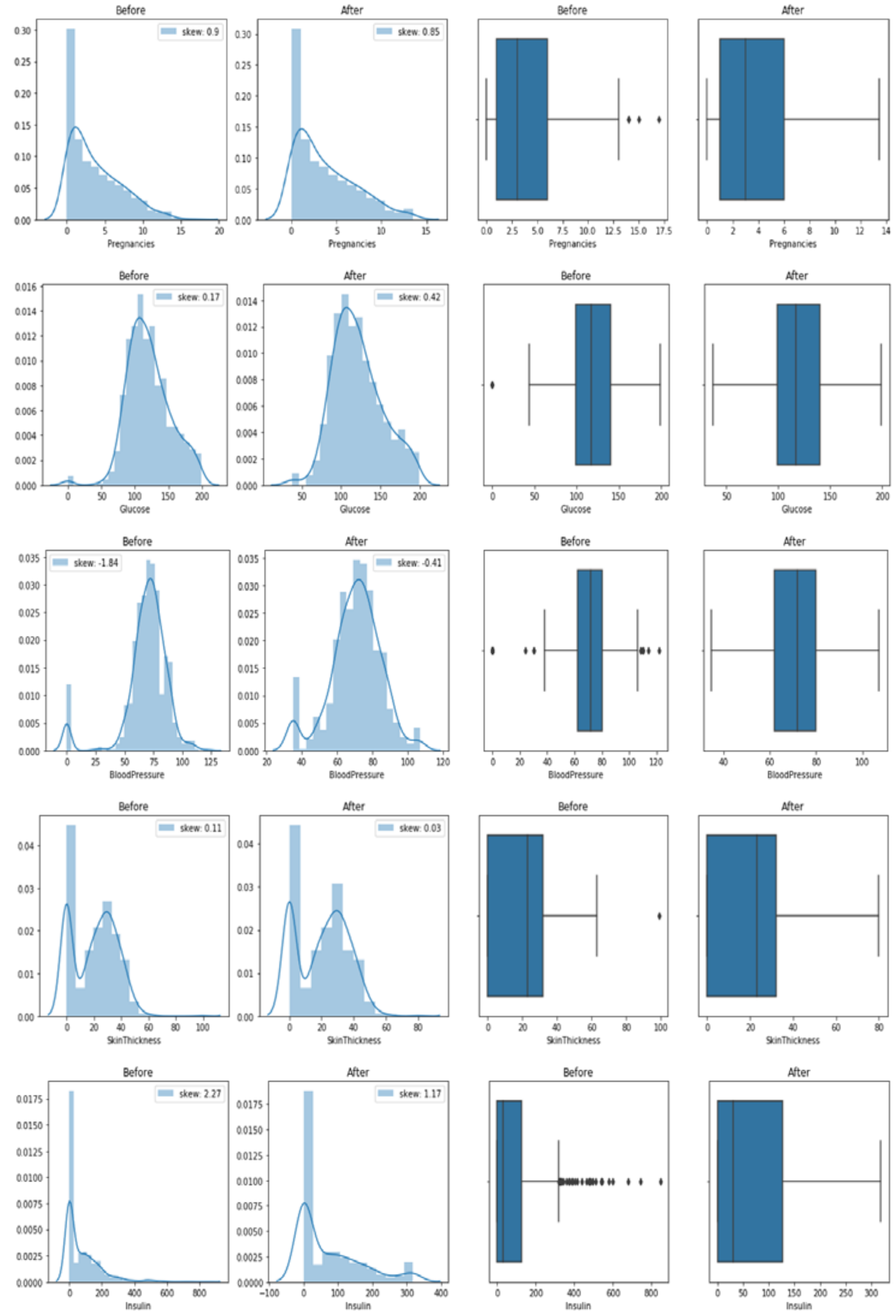
Table 4: *The PIMA Indian diabetes data statistics after performing outlier analysis*

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Out come
count	639.00	639.00	639.00	639.00	639.00	639.00	639.00	639.00	639.00
mean	3.80	119.11	72.12	20.56	65.93	32.00	0.42	32.71	0.31
std	3.26	29.16	11.34	15.33	79.56	6.43	0.25	11.08	0.46
min	0.00	44.00	38.00	0.00	0.00	18.20	0.078	21.00	0.00
25%	1.00	99.00	64.00	0.00	0.00	27.30	0.24	24.00	0.00
50%	3.00	114.00	72.00	23.00	37.00	32.00	0.35	29.00	0.00
75%	6.00	137.00	80.00	32.00	120.00	35.95	0.58	40.00	1.00
max	13.00	198.00	106.00	60.00	318.00	50.00	1.191	66.00	1.00

The below Figure 1 visualize the data before the outlier how the data distribution graph and after the outlier analysis how the data distribution graph was. It was noticed that there was no change in distribution after performing outlier removal. Box plot visualizes the presence of outliers before and after outlier analysis. The dots in the box plot showed the presence of outliers. Outliers affect the mean value, so the elimination of outliers resulted in an unbiased mean. The plots presented in Figure 1 illustrate that the dataset had the same distribution after performing outlier removal when compared with the data distribution before outlier removal. This showed that the dataset is standardized without losing its distribution property after outlier removal.

Table 5 explains the statistics of data after performing outlier analysis. From Table 5, it can be noticed that there were zero recorded values with pregnancies, skin thickness, and insulin.

Figure 1: Data distribution and box plot of the PIMA Indian diabetes data before and after outlier analysis.



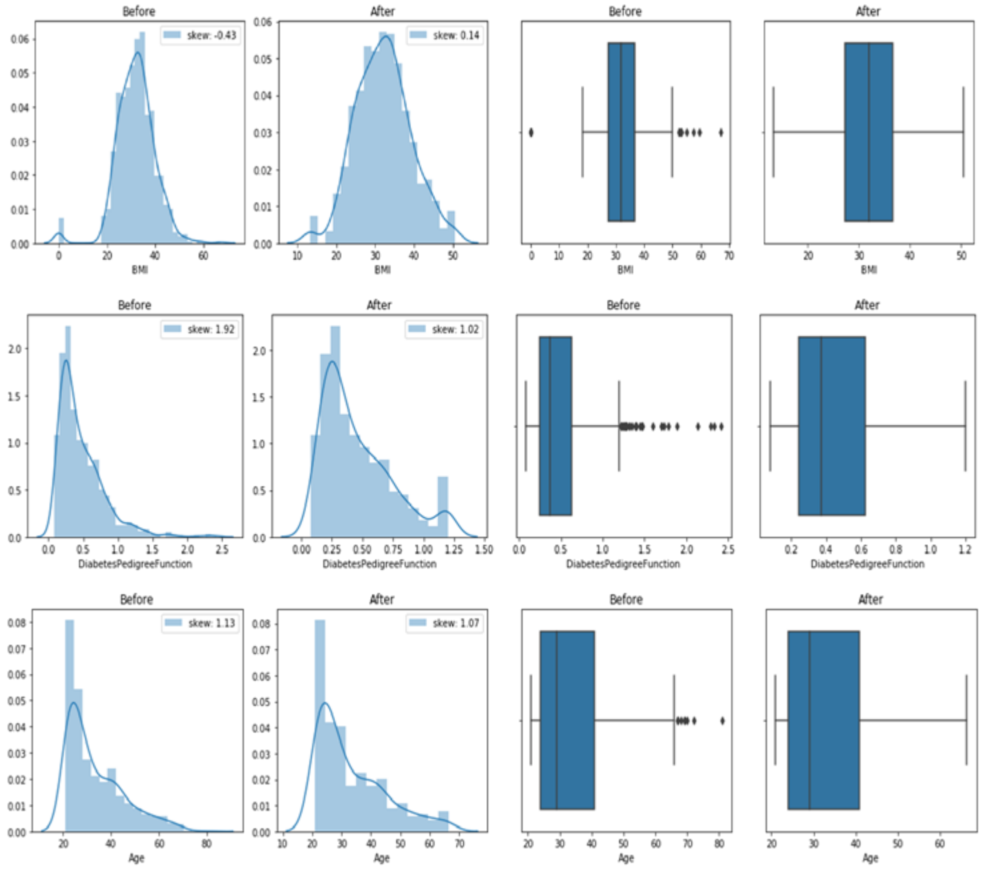


Table 5: Attributes that contain zero values after performing outlier analysis

Attributes	Zero Counts
Pregnancies	88
Glucose	0
Blood Pressure	0
Skin Thickness	179
Insulin	307
BMI	0
Diabetes Pedigree Function	0
Age	0

After outlier analysis, the dataset contains a total of 639 samples out of which 439 were class 0 (normal) and 200 were class 1 (diabetes). From the above data, it was observed that pregnancies, skin thickness, and insulin contain zero values. Pregnancies feature can have zero values which say that the female patient has not been pregnant at the time of sample collection. Skin thickness and insulin had zero values which say that the patient had no skin thickness and insulin in the body which cannot be true they may be missed during the data collection stage. In the pre-processing stage, skin thickness and insulin were replaced with the mean values. Insulin zero recorded values were filled with mean values of insulin that were grouped with glucose because glucose and insulin are related to each other. Skin thickness zero recorded values were filled with mean values of skin thickness that were grouped with BMI because skin thickness and BMI were related to each other. Below Table 6 shows the statistics of the data after imputing with mean values. Now, from the data, it is evident that there were no zero recorded values with skin thickness and insulin.

Table 6: After imputing the skin thickness and insulin the Pima Statistics

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
count	639	639	639	639	639	639	639	639	639
mean	3.80	119.11	72.12	28.26	127.27	32.00	0.42	32.71	0.31
std	3.26	29.16	11.34	9.154	58.88	6.43	0.25	11.08	0.46
min	0	44	38	7	15	18.2	0.078	21	0
25%	1	99	64	22	82	27.3	0.242	24	0
50%	3	114	72	28.26	121	32	0.358	29	0
75%	6	137	80	34	158	35.95	0.586	40	1
max	13	198	106	60	318	50	1.191	66	1
max	13.00	198.00	106.00	60.00	318.00	50.00	1.191	66.00	1.00

Scaling the data using min-max Scaling. 12 Age attribute was scaled to 100 years and the remaining features were scaled in the range of 0 to 1. Scaling was performed for efficient processing of the data. Below Table 7 showed min-max scaled data statistics. From the below table it is evident that data was in the range of 0 to 1.

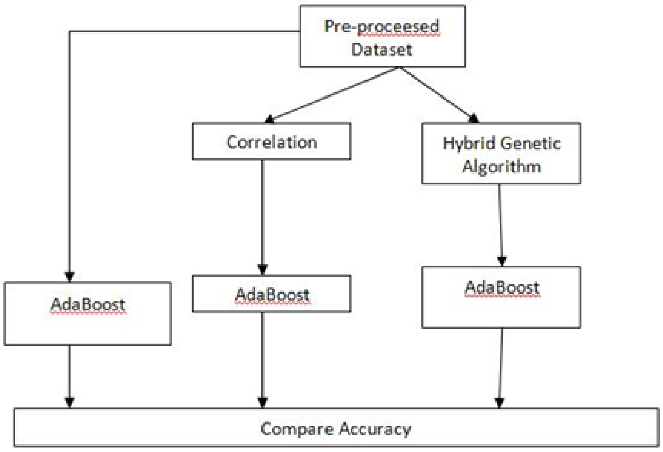
Table 7: *Min Max scaling applied to Pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, Diabetes pedigree function*

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
count	639	639	639	639	639	639	639	639	639
mean	0.29	0.48	0.5	0.4	0.37	0.43	0.31	0.32	0.31
std	0.25	0.18	0.16	0.17	0.19	0.2	0.22	0.11	0.46
min	0	0	0	0	0	0	0	0.21	0
25%	0.076	0.35	0.38	0.28	0.22	0.28	0.14	0.24	0
50%	0.23	0.45	0.5	0.4	0.34	0.43	0.25	0.29	0
75%	0.46	0.61	0.61	0.5	0.47	0.55	0.45	0.4	1
max	1	1	1	1	1	1	1	0.66	1

System architecture

Figure 2 describes the system architecture, here the pre-processed PIMA Indian diabetes dataset was processed in three stages and the outcomes of each stage were compared. In stage 1, the dataset was trained and tested with AdaBoost. In stage 2, the features were selected using correlation, and the selected features were trained and tested by AdaBoost. In stage 3, the features were selected using HGA, and the selected features were trained and tested by AdaBoost. Finally, all the outcomes of the three stages were compared in terms of accuracy.

Figure 2: *System architecture*



Methodology

AdaBoost

AdaBoost was a homogeneous ensemble technique it was used to get unbiased results. The major advantages of AdaBoost were, AdaBoost was easy to implement and have a high classification rate. They were flexible and not prone to overfitting. The major disadvantages of AdaBoost were, AdaBoost was sensitive to noise and outliers. The AdaBoost algorithms were applied in a variety of applications like disease prediction, text, and image classifications.

Algorithm: AdaBoost. A boosting algorithm aggregates classifiers. Each classifier gave a weighted vote in the final prediction.

Input:

G: represents a training data of size d;

M: represents how many models to be constructed;

A learning model.

Output: An Ensemble classifier.

Method: Initialize the sample weight to 1/d for each sample in the G.

For i=1 to M do:

Generate a new training data G^i by selecting the data samples with replacement from G concerning the weights associated with it.

Use training set Gi, to generate a classifier H^i;

Calculate the error of H^i; error rate of Hi is given by

$$error(H^i)=\sum_k^d w_k \times err(G') \qquad (5.1.1)$$

If error $(H^i)>0.5$ then reassign the weights to 1/d.

Go to step c; end if for correctly classified samples in Gi do update weights as error (H^i)=error (H^i)//1-error (H^i);

Normalize the weight of each instance.

End for test: Ensemble classifier was used to predict the class label of new sample X,

Initialize the weight to zero for each class category

For i=1 to M do:

w=log1[1-error(H^i)/ error(H^i)]// voting power of classifier.

Class label=H^i (X)//prediction made by H^i

Add wi to weight for the class label.

End for return the class label that had more weight.

Description: The AdaBoost classifier was used for improving the performance of the classifier. The AdaBoost was a homogeneous ensemble classifier that strengthens the weak classifiers collectively. The Dataset was a collection of samples, where each sample was associated with initial weight the associated weights were normalized so that the sum of all the weights of all samples is 1. The decision stump was constructed and its error rate was calculated. If the error rate was greater than or equal to 0.5 then reassign of weights was done. The classified sample weight was decreased and in the misclassified samples, weights were increased. Again the weights were normalized, based on the updated weights new dataset was generated, again decision stump was generated with an error rate; the process was repeated until the maximum estimators were reached. The generated decision function determines the result of the test samples. The results thus obtained were unbiased and efficient.

Correlation

Correlation was used to remove redundant features. It tells how one attribute was correlated with another attribute.

Input: D was a feature matrix of size mxm; rows and column represent same feature names; O was a null matrix of size mxm.

Output: One dimensional L array of selected features.

```
Method: For i:1 to m:
For j:1 to m:
If i!=j:
Calculate:
$r_{i,j} = \frac{\sum_{i=1}^{N} (a_i - \overline{A})( b_i - \overline{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N} (a_i b_i) - N\overline{A}\overline{B}}{N\sigma_A\sigma_B}$
N: no of samples
$\overline{A}$ and $\overline{B}$ were the mean of A and B features
$a_i$ and $b_i$ were the instances of A and B features
$\sigma_{A,}$ $\sigma_B$ The standard deviation of A and B features.
O (i,j) ri,j
End if
End for j
End for i
i,j max (O (i,j)) i{0,...m-2}; j{0,...m-2}
K Null (1xm)
if O (i,m-1)>O (j,m-1):
Update K with i th feature else update K with j th feature end if.
L Unique (K)
```

Description: The correlation coefficient was used to measure the dependency of both features. The correlation value lies between -1 to 1. If the Correlation coefficient r was zero then there was no correlation between features. If correlation coefficient r <0, then there was a negative correlation this means both features were discouraging each other. If r>0, then there was a positive correlation this means both features have raising nature for each other. The correlation did not guarantee causality; this means the presence of one feature does not influence the presence of another feature.

Hybrid Genetic Algorithm (HGA)

A genetic algorithm (GA) was used for optimization problems. The major advantages of GA were they are robust; they perform well on discrete or continuous problems. They were good at handling noisy data. GA was stochastic in nature; it can approximate solutions to specific problems with limited search space. The major disadvantages of GA were it was very difficult to design a fitness function, bit encodings, and operators. GA was computationally expensive. The major applications of GA were, it was used in optimization problems, neural networks, parallelization, image processing, vehicle routing problems, scheduling applications, machine learning, traveling salesman problems, and DNA analysis.

The correlation technique identifies the features based on the statistical relevance. It identifies positively correlated features to select the feature subset. Even though the features were selected based on statistical relevance, the selected feature subset cannot guarantee optimal performance. The selected correlated features when trained with AdaBoost showed lower performance when compared to AdaBoost performance on the PIMA Indian diabetes dataset. This clearly says that the correlation technique cannot guarantee an optimal feature subset. This drawback of correlation motivated the authors to design an HGA for the selection of feature subset which can achieve optimal performance.

Algorithm for the proposed Hybrid Genetic Algorithm (HGA)

The proposed HGA integrates genetic algorithm (GA) and AdaBoost. Here, AdaBoost was considered as an objective function to evaluate the fitness of the feature subsets Hybrid genetic algorithm (HGA)

1. Read the pre-processed PIMA Indian diabetes dataset.
2. Generate a population with a fixed size.
3. Calculate the fitness of the individuals in the population.
4. In the selection phase, select the parents that have good fitness values.
5. In the crossover phase, the selected parents in the mating pool were crossover and new offsprings were generated.
6. In the mutation phase, randomly selected individuals were mutated to bring diversity to the feature subset.
7. The fitness was calculated for the mutated individuals.
8. The individual with maximum fitness value was selected as the best.

9. The process is repeated from step 4 to step 8 until the maximum generations were completed.

Description: The initial population of fixed size was randomly selected. The population consists of individuals; each individual was a linear representation of bit strings. The data set has eight features so the linear bit string had eight bits. Each bit taken a binary value 1/0, 1 represents the presence of a feature, 0 represents the absence of a feature. The fitness of the individuals was evaluated by the AdaBoost classifier. In the selection phase the top best individuals that had good fitness scores were chosen as parents. The parents were treated with uniform one-point crossover to obtain new individuals. The new individuals may perform better or may underperform, if the new individual was not performing well then in the next generation at the selection phase, it gets eliminated. In the mutation phase, a random bit of the individual was flipped to obtain a new variety of feature subsets, and their fitness was calculated. The process was continued until all the generations were completed. Finally, at the end of generations, it reports the best individual that contributed more to the outcome.

In the medical domain, the proposed HGA is applied to the PIMA Indian diabetes dataset. The diabetes dataset has eight features each feature indicates the test result of the patient. For a common man to undergo all these tests is very expensive and troublesome. Once the tests are performed based on the test result the experts have to give their decision in time accurately this may not be feasible since it is a tedious process and also depends on the availability of the experts. Machine learning algorithms can offer the best solution for these types of situations. The proposed method reduces the dimensions of the dataset by identifying the relevant features. In this way, the number of tests to be performed on patients can be reduced. This helps the common man to undergo few tests and get benefitted financially. The reduced features also reduce the storage space, computational time, and cost. Once the features are selected they have to be trained by the classifier. Even though they are a variety of classification algorithms available, there is no guarantee that a specific algorithm works well on all the problems. To enhance the predictive rates of the disease a homogenous ensemble classifier AdaBoost is used. The AdaBoost classifier constructs multiple weak decision trees to get an unbiased result. The classifier performance shows that the proposed model can be considered as a second opinion when the medical experts are in demand. Thus the proposed model can help in reducing the computation costs, storage space and maximizing the accuracy of the model. The decisions made by the model are valid and are very helpful to the common man.

The proposed hybrid genetic algorithm (HGA) was successful because the model uses efficient bit encoding mechanisms to represent feature subsets, it had limited search space, and it used the homogeneous ensemble technique the AdaBoost as a fitness evaluator to evaluate the fitness of the feature subsets. The adoption of AdaBoost as a fitness function provides unbiased results and less prone to overfitting. The HGA had a linear bit string encoding which means each feature was represented as a bit of 1/0. 1 means the presence of feature and 0 means the absence of a feature. The HGA had an initial population. The population consists of a set of individuals. Each individual was a vector with bit encodings. The HGA was designed with genetic operators. They were selection, crossover, and mutation. Each individual was evaluated by AdaBoost to get its fitness value before the selection operator applied. The selection operator selects individuals (parents) based on the fitness scores above the determined threshold. For selection, roulette wheel selection was used. Next, in the matting pool, the selected parents were crossover to get new offspring. Next in mutation, random individuals were taken and mutated with a probability of 0.5, which means it inverts 1 to 0 and 0 to 1. The flipping of bits by the mutation operator generated new feature subsets in the population. After a fixed number of generations, the best individual was identified. The resultant individual was the final feature subset. The fitness function AdaBoost was fed with feature subsets in the population one by one to calculate its accuracy scores (fitness score). AdaBoost creates multiple decision stumps. Since AdaBoost learns progressively, the decision made by one stump influence the decision made by another stump. So finally the majority say determines the final prediction. The design of AdaBoost as the fitness function for the genetic algorithm plays a major role in selecting the optimal feature subset. The advantage of HGA here was optimizing the search space and introducing the diversity in the population to extract feature subset that has a major impact on the classifier performance. The advantage of AdaBoost here was the ensemble voting of decision stumps. These decision stumps collectively determine the final decision without bias and overfitting.

The proposed HGA was successful for the PIMA Indian diabetes dataset because it generates a random initial population of feature subsets. In the population, new feature subsets were generated with the help of selection, crossover, and mutation operators. The limited search space and generation of diversified features are the major strengths of the proposed algorithm. The proposed approach avoids the brute force approach of exploring all the possible feature subsets. The proposed model guarantees an optimal feature subset that can achieve global optima.

K-fold

In this work, 10-fold cross-validation was used for comparing the methods. The stratified sampling strategy was applied for 10-fold cross-validation. In stratified sampling, two-class labeled PIMA data was divided in equal proportions in each fold. For each ith fold of 10-fold, ith fold was used for testing and the remaining folds were used for training. The average of 10-test fold accuracy was used for comparing the model performance.

Performance metrics

Performance metrics were a measure to show how good a classifier was performing.[12] The precision determines the confidence of the model; recall determines the completeness of the model. F-measure tells how reliable our model was. Table 8 presents the performance metric formulae.[13-14]

tp: true positives

tn: true negatives

fp: false positives

fn: false negatives

Table 8: *Performance metrics*

Performance metrics	Formula
Accuracy	tp+tn/(tp+tn+fp+fn)
Precision	tp/(tp+fp)
Recall (sensitivity)	tp/(tp+fn)
F-measure	2*(recall*presicion)/(recall+precision)

Results and Discussion

The experiments were carried out in a Python programming environment. This section gives insight into the experimental results.

From the Table 9, Correlation technique selected Age, BMI, and Glucose as a feature subset. The genetic algorithm selected glucose, skin thickness, and age as a feature subset. Even though both feature selection methods extracted the same length feature subset but they have different feature names in the feature subset.

Table 9: *Selected features with feature selection techniques*

Method	No. of features	Features
Orignal features without feature selection algorithm	8	1.Pregnancies
		2. Glucose
		3. Blood pressure
		4. Skin thickness
		5. Insulin
		6. BMI
		7. Diabetes pedigree function
		8. Age
Features selected using Correlation	3	1.Age
		2.BMI
		3.Glucose
Features selected using Hybrid Genetic Algorithm	3	1.Glucose
		2.Skin thickness
		3.Age

Predictive functions were listed in Tables 10. Table 11 presents the scores of performance metrics with the proposed classifiers. For each classifier, this shows the boundary conditions to classify the data samples. The decision stumps were grouped according to their decisions, for each group the voting power of each decision stump is added. The final prediction was determined by the group with the largest sum.

Table 10: *Predictive function of proposed classifiers for single fold*

Classifier	The predictive function of AdaBoost for fold1 with 10 AdaBoost decision stump
AdaBoost	H(x)=0.2278 (Glucose<=0.65)+
	0.3091(Glucose<=0.52)+
	0.3497 (Age<=0.28)+
	0.4003 (Glucose<=0.37)+
	0.4251 (BMI<=0.26)+
	0.3848 (Glucose<=0.72)+
	0.4437 (Glucose<=0.31)+
	0.3950 (BMI<=0.38)+
	0.3902(Age<=0.31)+
	0.3950 (DiabetesPedigreeFunction<=0.41)

Correlation-based AdaBoost	H(x)=0.2278(Glucose<=143.50)+
	0.3737(Glucose<=101.50)+
	0.3963 (BMI<=26.35)+
	0.3711 (Age<=0.31)+
	0.4606 (Glucose<=89.50)+
	0.4208 (Glucose<=140.50)+
	0.4687 (Glucose<=179.50)+
	0.4757 (Age<=0.62)+
	0.4306 (BMI<=31.45)+
	0.4472 (Age<=0.41)
Hybrid Genetic Algorithm with AdaBoost	H(x)=0.2278(Glucose<=0.65)+
	0.3737 (Glucose<=0.37)+
	0.3633 (Age<=0.28)+
	0.4119 (SkinThickness<=0.29)+
	0.4626 (Glucose<=0.30)+
	0.4556 (Age<=0.56)+
	0.4408 (Glucose<=0.63)+
	0.4547 (Glucose<=0.80)+
	0.4398 (Age<=0.41)+0.4890 (Glucose<=0.26)

Table 11: PIMA Indian diabetes performance metrics results

	Accuracy	Precision	Recall	F_measure
AdaBoost classifier (All features)	0.77621	0.77385	0.41	0.534
Correlation with AdaBoost	0.7654	0.6472	0.5449	0.5859
HGA with AdaBoost	0.781	0.6849	0.56	0.6117

Figure 3 presented the accuracy scores of the developed methods. From Figure 3, the hybrid genetic algorithm (HGA) with AdaBoost scored maximum compared to AdaBoost and correlation with AdaBoost.

Figure 3: *Comparison of classifiers performance using the PIMA Indian diabetes dataset.*

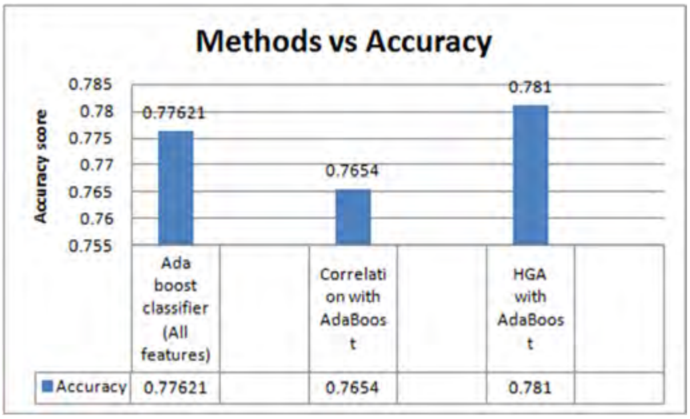


Figure 4 presented the comparison of precision scores of the developed methods. From figure 4, it was identified that AdaBoost scored has the maximum precision score.

Figure 4: Comparison of classifiers precision scores using the PIMA Indian diabetes dataset.

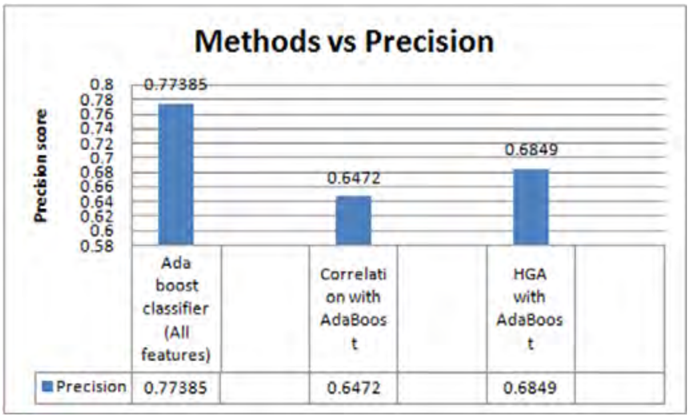


Figure 5 presented the comparison of recall scores of the developed methods. From Figure 5, it was identified that the hybrid genetic algorithm (HGA) with AdaBoost had a maximum recall score Figures 6 and 7.

Figure 6 represents the F-measure score comparisons of developed methods; Figure 7 represents performance comparison of HGA with AdaBoost with earlier reported techniques for PIMA Indian diabetes.

Figure 5: Comparison of classifiers recalls scores using the PIMA Indian diabetes dataset.

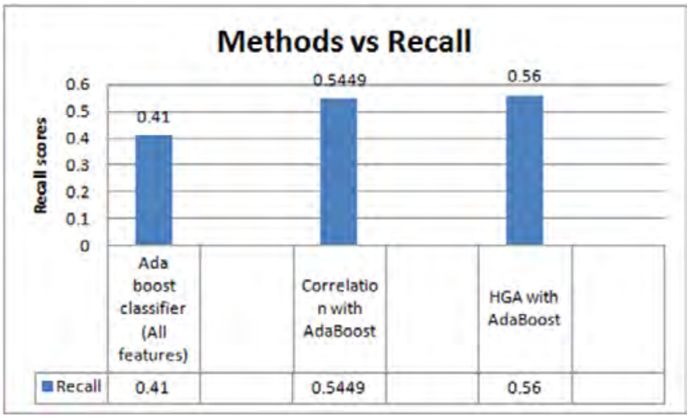


Figure 6: Comparison of classifiers F_measure scores using the PIMA Indian diabetes datasets.

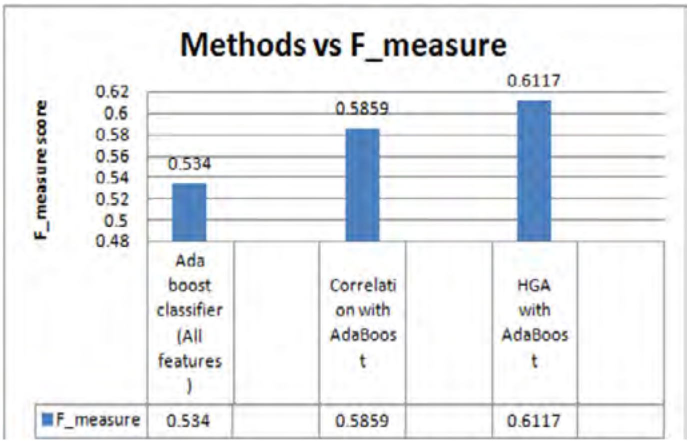


Figure 7: Comparison of existing methods accuracy with proposed method for the PIMA Indian diabetic dataset.

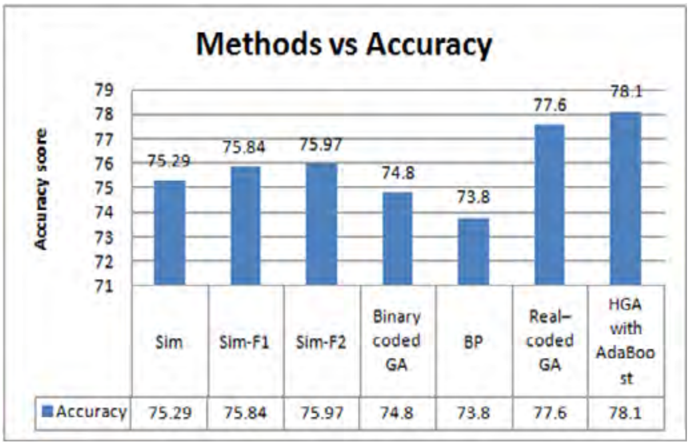


Table 12 compares the results with the performance of state-of-the-art techniques using PIMA Indian diabetic datasets.

Tables 13-15 presented selected features using proposed methods for Wisconsin breast cancer diagnostic and Cleveland heart disease datasets. Table 14 and 16 presented performance metric scores of experimental methods for Wisconsin breast cancer diagnostic and Cleveland heart disease datasets.

Table 12: Compares the results with the performance of state-of-the-art techniques using PIMA Indian diabetic datasets

S.No.	Source	Method	Accuracy
1		Sim	75.29%
2	(Luukka. 2011)	Sim-F1	75.84%
3		Sim-F2	75.97%
4		Binary coded GA	74.80%
5	(Orkcu and Bal, 2011)	BP	73.80%
6		Real–coded GA	77.60%
7	Our proposed method	HGA with AdaBoost	78.10%

__Table 13:__ Features selected using feature selection technique on Wisconsin breast cancer diagnostic dataset

Original Features	Correlation Selected Feature Subset	HGA Selected Feature Subset
1.radius_mean	1.area_se	1. radius_mean
2.texture_mean	2. area_worst	2. perimeter_mean
3.perimeter_mean	3. compactness_mean	3. symmetry_mean
4. area_mean	4. compactness_se	4. texture_se
5. smoothness_mean	5. compactness_worst	5. area_se
6. compactness_mean	6. concave points_worst	6. smoothness_se
7. concavity_mean	7. concavity_mean	7. concave points_se
8. concave points_mean	8. concavity_se	
9. symmetry_mean	9. fractal_dimension_mean	
10.fractal_dimension_mean	10. perimeter_mean	
11.radius_se	11. perimeter_se	
12.texture_se	12. perimeter_worst	
13.perimeter_se	13. radius_worst	
14. area_se	14. smoothness_se	
15. smoothness_se	15. smoothness_worst	
16. compactness_se	16. symmetry_worst	
17. concavity_se	17. texture_mean	
18. concave points_se	18. texture_worst	
19. symmetry_se		
20. fractal_dimension_se		
21.radius_worst		
22. texture_worst		
23.perimeter_worst		
24. area_worst		
25. smoothness_worst		
26. compactness_worst		
27.concavity_worst		
28. concave points_worst		
29. symmetry_worst		
30. fractal_dimension_worst		
Classlabels=[0,1]		

Table 14: Performance metric scores of experimental methods using the Wisconsin breast cancer diagnostic dataset

PerformanceMetric Methods	Accuracy (%)	Precision (%)	Recall (%)	F_measure (%)
AdaBoost with all features	95.2	94.6	85.7	89.6
Correlation with AdaBoost	95.1	92.3	86.3	88.9
HGA with AdaBoost	95.6	93.7	89.3	90.8

Table 15: Features selected using feature selection techniques for the Cleveland heart disease dataset

Original Features	Correlation Selected Feature Subset	HGA Selected Feature Subset
1. age	1. age	1.thalach
2. sex	2. ca	2. old peak
3. cp (chest pain)	3. exang	3. slope
4. trestbps	4. restecg	4. ca
5. chol	5. thal	5. thal
6. fbs	6. thalach	
7. restecg	7. trestbps	
8. thalach		
9. exang		
10. old peak		
11.slope		
12. ca		
13. thal		
Class labels=[0,1,2,3,4]		
[0] absence of heart disease		
[1,2,3,4] presence of heart disease.		

Table 16: Performance metric scores of experimental methods based on the Cleveland heart disease dataset

PerformanceMetric Methods	Accuracy (%)	Precision (%)	Recall (%)	F_measure (%)
Accuracy (%)	Precision (%)	Recall (%)	F_measure (%)	89.6
AdaBoost with all features	78.1	83.5	66.9	72.9
Correlation with AdaBoost	76.6	75.6	74.0	73.8
HGA with AdaBoost	79.8	80.0	74.9	77.1

the neighborhood. The proposed model by selecting the optimal feature subset saved a lot of storage space, reduced computational time and cost. The proposed model scored an accuracy of (78.1%), precision (68.4%), recall (56%), and F-measure (61.1%). The AdaBoost classifier with all features scored an accuracy of (77.6%), precision (77.3%), recall (41%), and F-measure (53.4%). The correlation with AdaBoost scored an accuracy of (76.5%), precision

(64.7%), recall (54.4%), and F-measure (58.5%). The proposed method outperformed AdaBoost when trained with all features. The correlation technique selected age, BMI, Glucose as features from original features, but this feature subset did not score well compared to AdaBoost with all features. As correlation technique selects features by statistical relevance so it will not guarantee an optimal solution. This drawback of correlation was overcome by HGA by selecting glucose, skin thickness, age as features from original features, and these selected features when trained with AdaBoost provided better accuracy scores. The genetic operators in the HGA generated diversified feature subsets in the population. The genetic algorithm with AdaBoost as a fitness function selected the best feature subset among the population of feature subsets. The proposed HGA with AdaBoost outperformed AdaBoost with all features and correlation with AdaBoost. This says that HGA with AdaBoost selected an optimal feature subset that was capable of achieving a global solution. The F-measure of the proposed model is above 50% which ensures the reliability of the model. The proposed model also performed well in terms of performance when applied on the Wisconsin breast cancer diagnostic and Cleveland heart disease datasets (95.6% and 79.8% respectively). The proposed model outperformed when compared with other reported algorithm's performance for the PIMA Indian Diabetic dataset. Hence the proposed model can be considered as a second opinion expert in medical diagnosis when the medical experts are in demand. The proposed model benefits the common man in reduced medical costs by limiting the tests that one has to go through. It also helps the common man by providing informed decisions accurately. The research efforts made can be used in various applications other than the medical domain. The designed model can be further enhanced to improve efficiency by adopting various embedded feature selection techniques and designing hybrid classifiers.

Acknowledgement

No Acknowledgement.

Conflict of Interest

Author have No Conflict of Interest.

References

1. Dheeru D, Casey G. Machine learning repository. University of California, Irvine, School of Information and Computer Sciences. 2017.
2. Choubey DK, Paul S, Kumar S, et al. Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection. ICCCS.2017;451-455.
3. Choubey DK, Paul S. GA_MLP NN: A hybrid intelligent system for diabetes disease diagnosis', International Journal of Intelligent Systems and Applications. 2016;8(1):49-59.
4. Polat K, Gunes S. An expert system approach based on principal component analysis and adaptive neurofuzzy inference system to diagnosis of diabetes disease. Digital Signal Processing, 2007;17(4):702-710.
5. Kahramanli H, Allahverdi N. Design of a hybrid system for the diabetes and heart diseases. Expert systems with applications. 2008;35(1-2):82-89.
6. Tigga NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods', Procedia Computer Science. 2020;167:706-716.
7. Choubey DK, Paul S. GA_RBF NN: A classification system for diabetes. Int J Biomed Eng Technol. 2017;23(1):71-93.
8. Barakat N, Bradley AP, Barakat MNH. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE T INF TECHNOL B. 2010;14(4):1114-1120.
9. Ephzibah EP. Cost effective approach on feature selection using genetic algorithms and fuzzy logic for diabetes diagnosis. IJSC. 2011;2(1):1-10.
10. Dinesh MG, Prabha, D. Diabetes Mellitus Prediction System Using Hybrid KPCA-GA-SVM Feature Selection Techniques. J Phys Conf Ser. 2021;1767:1.
11. Balasubramanian K, Ananthamoorthy NP. Improved adaptive neuro-fuzzy inference system based on modified glowworm swarm and differential evolution optimization algorithm for medical diagnosis. NEURAL COMPUT APPL. 2020;33:7649-7660.
12. Han J, Kamber M, Pei J. Data mining concepts and techniques (3rd edn). The Morgan Kaufmann Series in Data Management Systems, 5. Hardcover 2011.
13. Luukka P. Feature selection using fuzzy entropy measures with similarity classifier. Expert Systems with Applications. 2011;38(4):4600-4607.
14. Orkcu HH, Bal H. Comparing performances of backpropagation and genetic algorithms in the data classification.', Expert systems with applications, 2011;38(4):3703-3709.