APPLICATIONS OF
MODELLING AND SIMULATION

http://arqiipubl.com/ams eISSN 2600-8084 VOL 4, 2020, 101-109

CSE-DT Features Selection Technique for Diabetes Classification

Matthew T. Ogedengbe* and Charity O. Egbunu

Department of Mathematics/Statistics/Computer Science, University of Agriculture, Makurdi, Nigeria

*Corresponding author: matt554real@gmail.com

Submitted 01 February 2020, Revised 28 February 2020, Accepted 06 March 2020.

Abstract: Diabetes has become one of the world deadliest disease. It is a sickness which occurs as a result of increase in blood sugar level in the body. Most people living with it encounter various complications in their body organs if it remain undetected and untreated at the early stage. Most literatures considered all features of a diabetes dataset as risk factors in diagnosing diabetes and this has resulted to low classification accuracy and longer execution time since all the features in the dataset are involved in the classification process. Selecting the most relevant features as the risk factors improves the performance of classifiers in term of classification accuracy and other performance measures. This paper presents feature selection technique called Classifier Subset Evaluator (CSE) which selects most relevant risk factors for the prevalence of diabetes in the body. The selected features (risk factors) were passed to J48 decision tree (DT) classifier for training and testing, and the DT classified all the instances of the dataset based on these selected features. The CSE and DT were hybridized as a proposed Classifier Subset Evaluator Decision Tree (CSE-DT). The CSE-DT was experimented on Pima Indian Diabetes dataset (PIDD) acquired from the UCI data repository and implemented on Waikato Experiment for Knowledge Analysis (WEKA). The CSE-DT was compared with Naïve-Bayes, Support vector machine (SVM) and Decision Tree for the evaluation measure in terms of F-Measure, Precision, ROC, Recall and Accuracy. The results show that the CSE-DT attained a better classification accuracy value of 81.64% among others.

Keywords: Accuracy; Classifier Subset Evaluator; Diabetes; Decision tree; Naïve-Bayes; Support vector machine.

1. INTRODUCTION

Diabetes disease has an adverse effect on human life if it is not detected early and treated at the immediately at the early stage. This causes unimaginable increase of glucose (blood sugar) in human's body. Rises in blood sugar result to inadequate reproduction of insulin in the body or failure of the body to respond to the produced insulin. The basic symptoms of diabetes are intensify thirst, hunger and frequent urination. Diabetes can occur due to several factors such as unhealthy consumption of food substance, heredity and obesity. However, anyone suffering from diabetes is prone to develop serious complications such as cardiovascular disease, heart attack, kidney disease and stroke [1-3]. Statistically, it has been estimated that 8.8% of global population has diabetes in 2017 [4]. Basically, there exist three types of diabetes. Type 1 diabetes is a condition which is caused as a result of insufficient insulin produced by pancreas and this mainly occur in children and adolescents as a result of genetic disorder. The Type 2 diabetes is mostly occurs in adults and it is caused by high sugar level in the body. On the other hand, Type 3 is known as Gestational diabetes which mostly common in pregnant women during pregnancy period without prior history of diabetes [3]. However, it has been observed and established by the researchers [5] that Type 2 diabetes has the highest occurrence of about 90% globally.

Moreover, due to the life threatening effect of diabetes and its complications, it is paramount to provide effective and precise predictive model for prediction/classification of diabetes. The predictive model can be developed since there is possibility to identifying people with risk of acquiring diabetes because of some similar risk factors, such as blood pressure, number of times pregnant plasma glucose concentration, family history of diabetes and body mass index (BMI). Numerous predictive models have been developed by researchers [6] for classification and prediction problems including logistic regression, decision tree, support vector machine (SVM), Naïve-Bayes, artificial neural network (ANN) and random forest.

In literatures, all features of a diabetes dataset are mostly considered as the risk factors in diagnosing diabetes, whereas some of these features are not responsible for the prevalence of diabetes in the body. The classifiers in [7] employed all the features in the dataset for classification and this resulted to low classification accuracy since some irrelevant features are included for the classification process. Selecting the most relevant features as the risk factors improves the performance of classifiers in term of classification accuracy and other performance measures. This research proposed a Classifier Subset Evaluator (CSE) as a feature selection technique for selecting the most relevant diabetic features/attributes with standard metaheuristic (J48 Decision Tree) classifier. The proposed CSE and DT are combined as CSE-DT to identify and select the most relevant features and to increase classification accuracy. The CSE-DT performance was compared with the existing Naïve-Bayes, decision tree and SVM in [7] on PIDD dataset for performance evaluation.

The contribution of this study are;

- Development of CSE as a feature selector for selection of most influential risk factor for predicting diabetes
- Hybridizing the CSE with J48 Decision Tree classifier for classification accuracy improvement

2. RELATED WORK

Researchers have done lots of researches on the classification accuracy for various illness encountered by human in the last decade. A deep machine learning algorithms in [4] was developed for the detection of diabetes. The authors used RR-interval signal which is known as heart-rate variability (HRV) signals which was obtained from electrocardiogram (ECG) signals. Convolutional neural network (CNN) and Long short-term memory (LSTM) were combined to extract features set of the input HRV data. The extracted features were loaded into support vector machine (SVM) for classification process. The results shows that the proposed classification system using ECG signals has an accuracy of 95.7% when compared to the CNN and CNN-LSTM without SVM. In [7], diabetes was predicted using three common classification algorithms namely DT, SVM and Naïve-Bayes. The experiments was performed on PIDD dataset of UCI data repository. These algorithms were evaluated on performance measure index in terms of Precision Accuracy, Receiver Operating Characteristic (ROC), Recall and F-Measure. Their results obtained shows that the performance of Naïve Bayes attained best accuracy of 76.30% among others. In [5] the Genetic programming (GP) approach was adopted for prediction of diabetes, by training the GP and testing it with Diabetes dataset obtained from UCI repository. Results obtained using GP shows optimal performance when compared to other existing techniques reviewed in their study. However, their GP takes more time during classification but the implementation of the GP is at low cost which is also a significant advantage in prediction of diabetes. An algorithm for the classification of the risk of diabetes was developed in [9], and their model adopted four renowned classifiers namely; Naïve-Bayes, Decision Tree, ANN and Logistic Regression. The robustness of their designed model was improved by adopting Bagging and Boosting techniques. Results of their experiment shows that Random Forest classification technique achieved optimal accuracy among all the algorithms adopted.

A hybrid prediction model (HPM) with the use of *k*-means clustering was developed in [10] for a selected class label validation in a dataset which employed C4.5 algorithm for developing classifier model, which achieved 92.38% accuracy. The performance multilayer perception (MLP) in terms of prediction accuracy against the decision tree (J48 and ID3) algorithms was evaluated in [11]. It was observed that J48 outperformed others with prediction accuracy of 89.3%. An artificial meta-plasticity using multilayer perceptron (AMMLP), which serves as prediction mechanism for diabetes was developed in [12]. This mechanism achieved accuracy of 89.93%. A critical information in medical domain was explored in [1] using data mining techniques, it was discovered that data mining can be a useful tools in minimizing risk of developing deadly diseases such as diabetes, heart disease and kidney diseases. Seven mining techniques were adopted for mining the dataset obtained namely; Regression, MLP, Bayes, ZeroR, Logistic Regression, J.48 and Random Forest. Results showed that MLP has a better performance of 81.8% among others. Development of an ensemble system using data mining techniques was proposed in [2]. Three classification algorithms were adopted for the prediction of diabetes mellitus in human including decision tree, weighted *k*-nearest neighbor and logistic regression. The proposed technique adopted votes which was provided by each classifier to obtain an optimal result. The output (majority vote) of the proposed system is the function of estimated value of each classifier supplied as input to the voting mechanism. The ensemble technique obtained the highest classification accuracy of 80.60% among others.

A detailed survey was performed in [13] which summarized the performance of various data mining algorithms often used in medical research namely, genetic algorithm, Decision Tree, Naïve-Bayes, *k*-NN. The algorithms were experimented on heart disease. In their implementation, a GUI was designed to input patient data for predicting the prevalence of the disease in such patient. They reduced attributes by the use of genetic search algorithm and the results showed that Naïve-Bayes obtained a better performance. Clustering is another data mining technique for mining complex task. In [14], *k*-means was employed along other clustering algorithms to identify different impacting factors of a disease without complexity. Results depicted that *k*-means performance was better than others. A rebalancing algorithm was developed in [3] for the prediction of diabetes of imbalanced dataset. Two phase predictive model was adopted thereby the first was preprocessing of data using Synthetic Minority Oversampling Technique (SMOTE), and the second phase supplies the preprocessed data to SVM, Bagging, Multi-Layer Perceptron (MLP), Decision Tree and Simple Logistic for the purpose of selecting an optimal classifier for diabetes prediction. A 94.7% accuracy was achieved with 10-fold cross validation. The study in [15] compared the performance of pre-processing and non-pre-processing data mining techniques in order to discover the importance of pre-processing data. After the experiments, pre-processing classifier acquired an increase in classification accuracy compare to non-pre-processing classifier.

Early detection of diabetes only feasible when appropriate assessment of both uncommon and common symptoms are carried out. Prediction of the likelihood of prevalence of diabetes was conducted in [16] with a dataset of 520 instances. The dataset was analyzed on Random Forest, Logistic Regression, and Naïve-Bayes by applying Percentage Split evaluation method and 10-fold cross validation with Random forest obtained the best accuracy. A predictive model was developed in [17] which narrowed down the risk factors into gender, blood pressure, age, BMI, blood glucose level of diabetes, duration of diabetes suffers and family history. C4.5 decision tree, Naïve-Bayes and *k*-means clustering technique were employed to analyze the dataset. It was discovered that in Retinopathy, the most significant risk factor is a female patient which once have a hypertension. Also for Nephropathy, the most influential risk factor is the diabetes duration which may be more than 4 years. An automated system was developed in [18] for gestational diagnoses using hybridized classifiers for the prediction of the most influential risk factors of diabetes Type 2. The data and its attributes were tested to obtain a new set of classified data.

The modified J48 decision tree and SVM classifier were employed to mine the clinical dataset obtained for the purpose of predicting gestational diabetes of Type 2 and it risk factors. The results reflected the performance of the modified J48 decision tree as a better classifier due to increase in its accuracy and minimal error rate against SVM and others.

2. MATERIALS AND METHODS

This section consists of dataset adopted for this study, the data preprocessing phase which consist of feature selection, and the proposed decision tree classifier model.

3.1 Dataset Used

The dataset adopted for this research was obtained from the UCI data mining repository which consists of 768 instances. Instances of 268 patients were tested positive while instances of 500 tested negative. Tested positive indicates ‘1’ which implies patient is diabetic and tested negative indicate ‘0’, implies patient not diabetic. The diabetes dataset consists of 8 features/attributes, and they are numeric data type. These dataset were generated as a result of medical examination of individual. Table 1 depicts the features of each instance as recorded in the dataset.

Table 1. Attributes/Features for diabetes

S/N	Features	Abbreviation
1.	Body mass index	bmi
2.	Plasma glucose concentration	plas
3.	Diastolic blood pressure (mm Hg.)	pres
4.	Diabetes pedigree function	pedi
5.	2-Hours serum insulin (mu U/ml)	insu
6.	Number of times pregnant	preg
7.	Triceps skin fold thickness (mm)	skin
8.	Age in years	age
9.	Class variable (‘0’ or ‘1’)	class

3.2 Classifier Subset Evaluator (CSE)

This evaluator consists of two major components i.e. the attributes/features evaluator, search method and attribute selection mode.

- **Attributes/features classifier:** It evaluates features/attributes subsets on training data and uses classifier (such decision tree, Naïve-Bayes, random forest and SVM to compute the merit of attributes set. This study adopted Naïve-Bayes as a features classifier for estimating the accuracy of the features subsets.
- **Search method:** The classifier subset evaluator also consists of a search method which search all the features set to determine the optimal node. The Best-First search method was adopted in this study. This method examines the range of attribute subsets using greedy hill climbing as improved with a backtracking facility. The search direction was set to “Forward” with lookup cache size 1 and search termination of non-improving nodes to 5.
- **Attribute Selection Mode:** The attributes using the “full training dataset” was selected.

After the CSE have been applied on the dataset, it was discovered that five out eight attributes were selected as the most influential features/attributes as shows in Figure 1. These are referred to as risk factors for the prevalence of diabetes in the body.

3.3 Steps in Pre-processing (CSE feature Selection) Phase

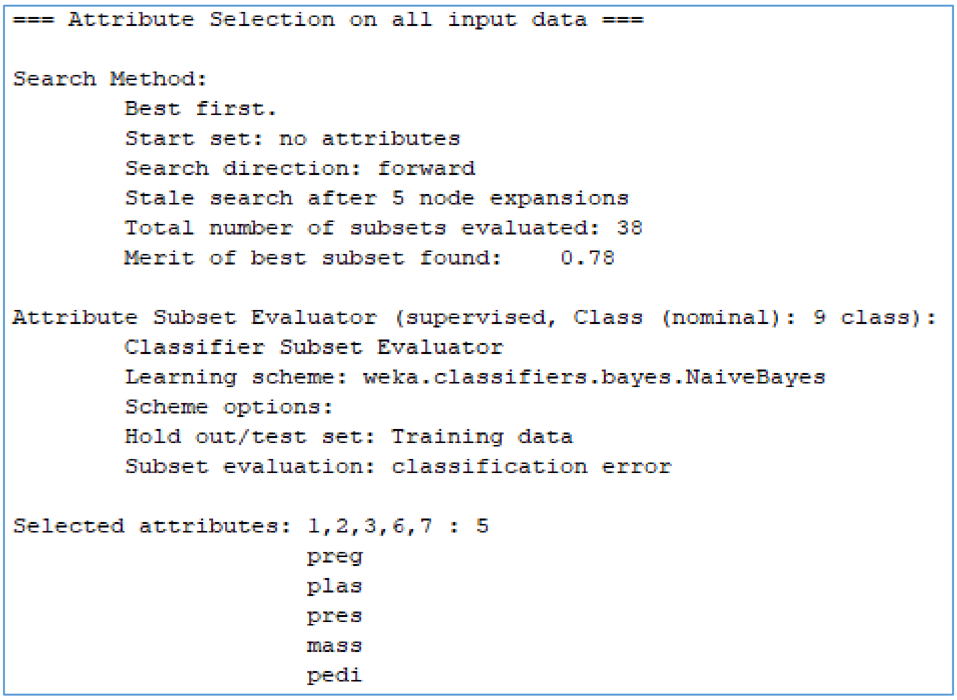
The following are steps required in selecting the most influential features/attributes using WEKA GUI.

Step 1: Uploading diabetes dataset

- From WEKA GUI select “Explorer” button
- From the preprocess menu select “open file” button and choose diabetes.arff file from the local filesystem and click on open.

Step 2: Attributes selection

- Select “Select attributes” button, the button comprises of *Attribute Evaluator*, *Search Method* and *Attribute Selection mode*.
- Select “ClassifierSubsetEval” from Attribute Evaluator.
- Select “BestFirst” from Search Method.
- Select “Start” button to perform the selection process.
- The selected attributes are displayed in “Attribute selection field” as shown in Figure 1.
- Click on “Preprocess” button to invert the selected attributes.
- Click on “Remove” button to discard the unselected attributes.



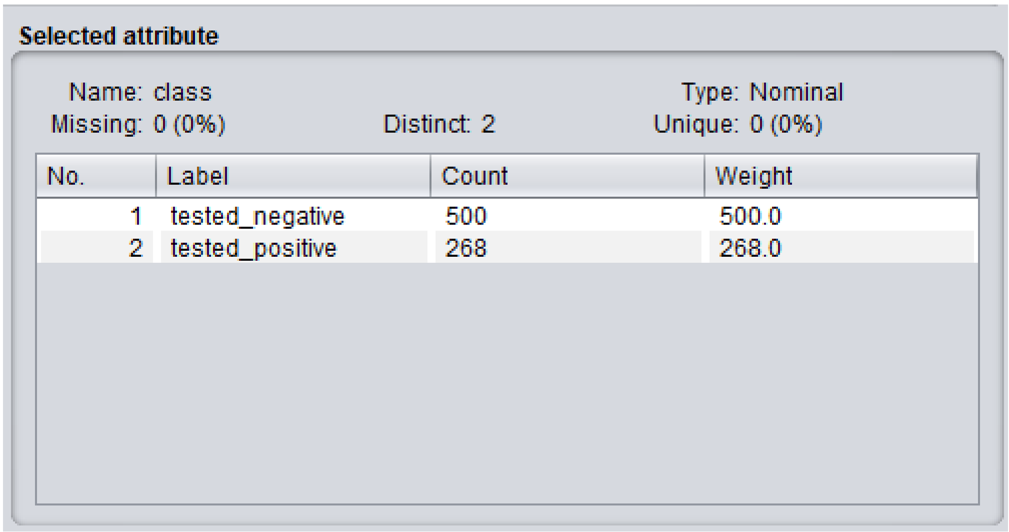
```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 38
        Merit of best subset found:    0.78

Attribute Subset Evaluator (supervised, Class (nominal): 9 class):
        Classifier Subset Evaluator
        Learning scheme: weka.classifiers.bayes.NaiveBayes
        Scheme options:
        Hold out/test set: Training data
        Subset evaluation: classification error

Selected attributes: 1,2,3,6,7 : 5
                     preg
                     plas
                     pres
                     mass
                     pedi
```

Figure 1. CSE feature (risk factors) selection summary



Selected attribute

Name: class
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	tested_negative	500	500.0
2	tested_positive	268	268.0

Figure 2. Selected attributes class labels

Triceps skin fold thickness, 2-Hours serum insulin and age were not selected as influential factors for the prediction of diabetes in Figure 1. Therefore, these three features were discarded and the remaining five relevant features were selected for the training and testing of the CSE-decision tree classifier for better prediction accuracy. The selected attributes consist of two label classes, tested_negative and tested_positive as shown in Figure 2. The tested_positive class has a weight of 268.0 and the tested_negative class is 500.0 and the feature/attribute data type is nominal. In Figure 2, Missing = 0 implies that the attribute is specified for all instances (no missing values), Distinct = 2 implies that the selected attributes have two different values: positive and negative and Unique = 0 implies that other instances do not have the same value.

The statistical frequency for the five selected (preg, plas, pres, mass, pedi and the class label) attributes are depicted in ‘Selected attributes window” as shown in Figure 3. The blue color signify tested_negative instances and the red color signify tested_positive instances of the class. Each of the attribute reflects the instances of both negative and positive class.

3.4 Decision Tree

Decision tree (J48) generates classified models which are in a tree structure form. It divides a data set into subsets and associated decision were created and incremented same time. The output result is a tree which consists of decision and leaf nodes. A decision node has at least two or more branches with a leaf node signifies a classification or decision. The uppermost decision node in a tree which matches to the best predictor is called root node. Decision trees can handle both categorical and numerical data [19].

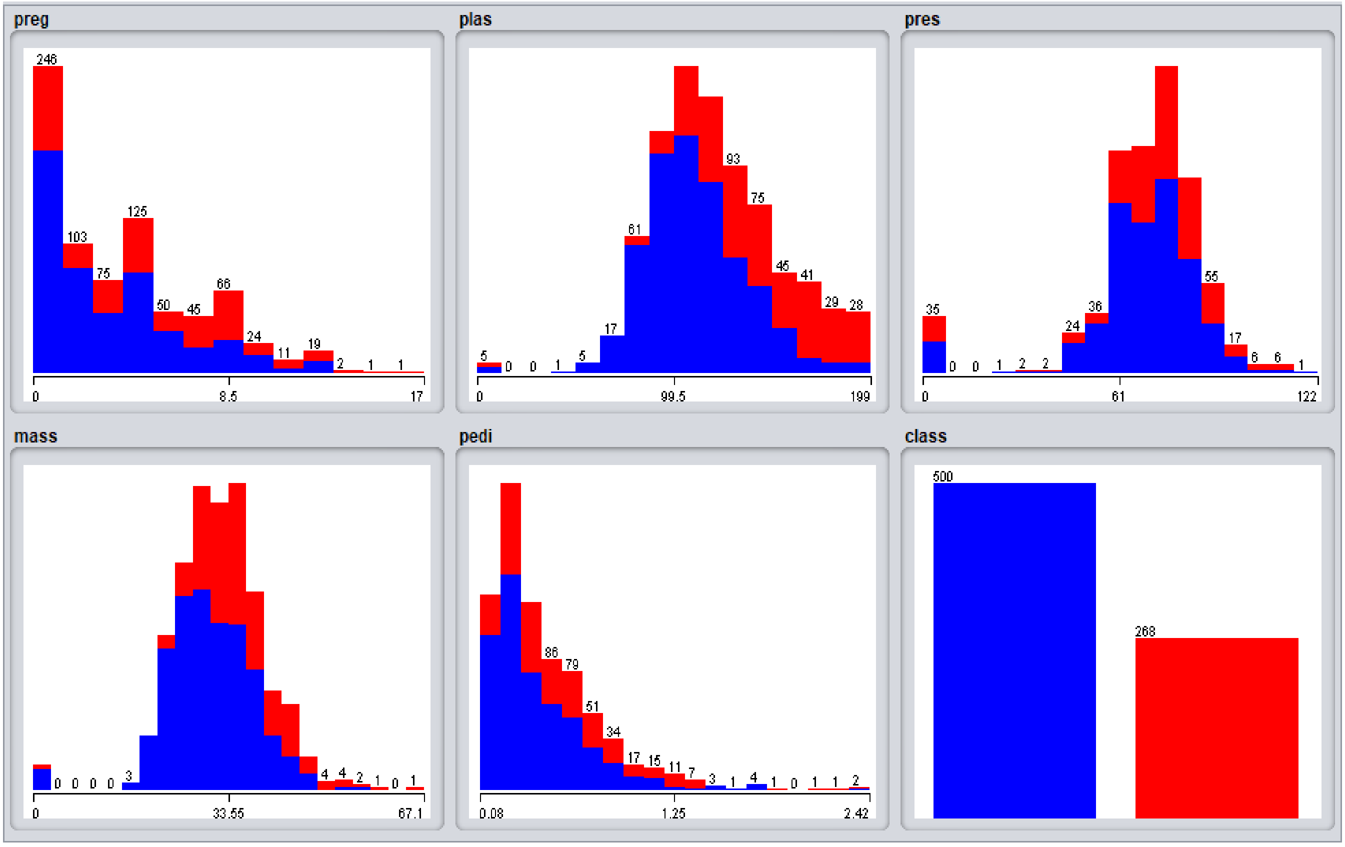


Figure 3. Number of instances in selected features/attributes window

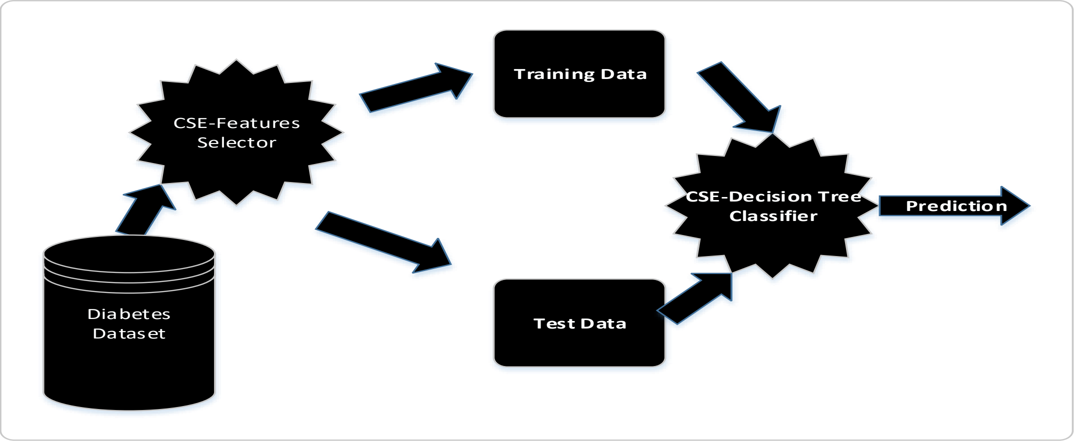


Figure 4. Proposed CSE-DT prediction frame work

3.5 Proposed CSE-DT Classifier

The proposed classifier in Figure 4 comprises of three major components namely the diabetes dataset repository, CSE-Feature selector and CSE-Decision Tree classifier. The diabetes dataset obtained from the UCI data repository was uploaded into CSE-Feature selector for selecting the determinant factors or the attributes most relevant for the diabetes prediction. The dataset was pruned and the less relevant attributes were discarded. The pruned dataset was partitioned into training and testing data for CSE-DT. The CSE-DT was trained and tested with the selected features/attributes for the prediction. The prediction shown in Figure 4 denotes the output of the CSE-DT which classifies the dataset into diabetes or not diabetes. It predicts patient base on the six features into diabetes or not diabetes.

4. RESULTS

In this study, CSE-DT classifier was experimented on diabetes dataset obtained from the UCI data mining repository and implemented on WEKA, a free machine learning toolkit comprises of standard machine learning algorithms. Figure 5 depicts the tree view of PIDD diabetes dataset with the most influential attributes. The summary of the proposed CSE-DT classifier was depicted in Figure 6, which consists of dataset with two classes namely tested_negative and tested_positive. The dataset consists of 768 instances of which 81.6% were correctly classified while 18.4% where misclassified. The confusion matrix reflects 36 instances of negative class were misclassified as tested_positive and 105 instances of positive class were misclassified as tested_negative.

4.1 CSE-DT Comparative Analysis

In this study, the performance of the proposed CSE-DT model was compared with that of Naïve-Bayes, SVM, and Decision Tree in [20] in terms of the following performance metrics: Precision, Recall, F-Measure, Prediction Accuracy and ROC. The diabetes dataset comprises of 768 instances. In Table 2, the performance of the listed classifiers were compared with the CSE-DT in term of instances that were classified correctly. It is shown that the CSE-DT has 627 instances correctly classified and this result is better than that of SVM, Naïve-Bayes and Decision Tree which were 586, 500 and 567 respectively.

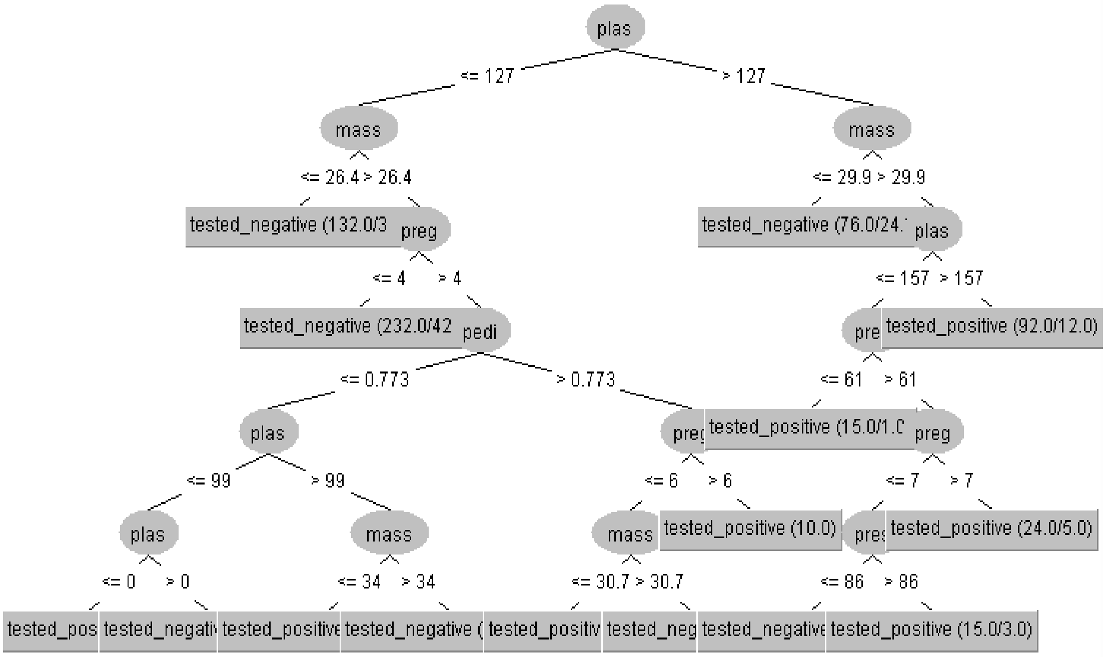
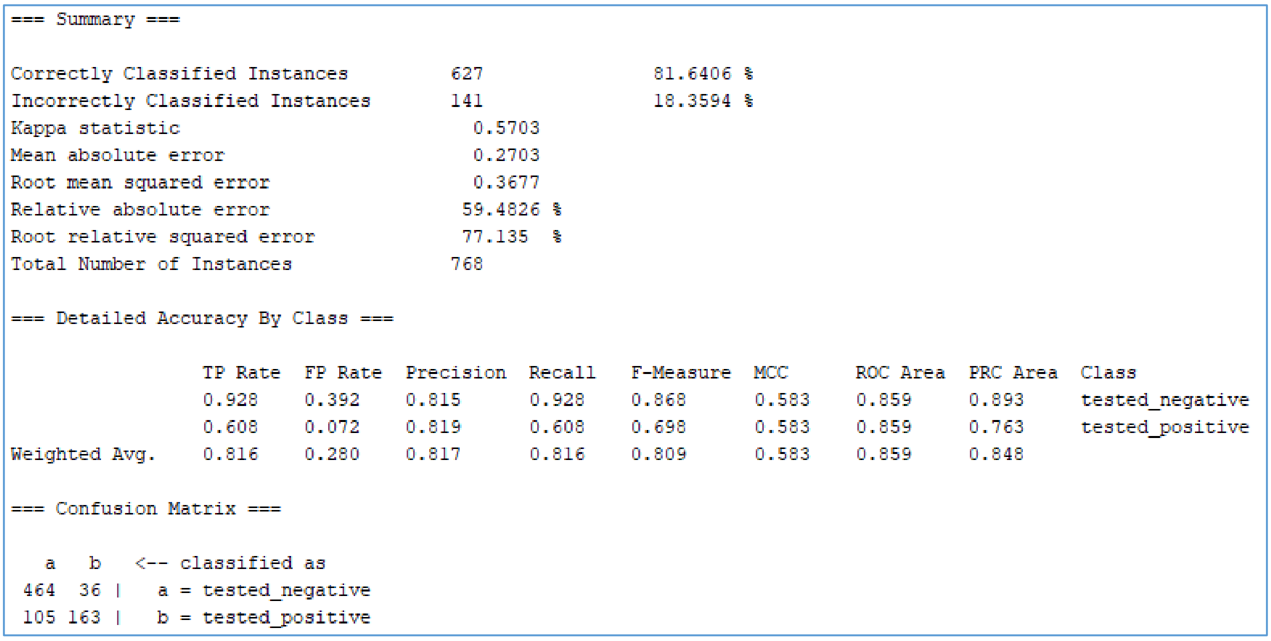


Figure 5. Decision tree view


```
=== Summary ===

Correctly Classified Instances         627               81.6406 %
Incorrectly Classified Instances       141               18.3594 %
Kappa statistic                          0.5703
Mean absolute error                      0.2703
Root mean squared error                  0.3677
Relative absolute error                 59.4826 %
Root relative squared error             77.135  %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.928    0.392    0.815      0.928    0.868      0.583    0.859     0.893     tested_negative
                 0.608    0.072    0.819      0.608    0.698      0.583    0.859     0.763     tested_positive
Weighted Avg.    0.816    0.280    0.817      0.816    0.809      0.583    0.859     0.848

=== Confusion Matrix ===

   a   b   <-- classified as
 464  36 |   a = tested_negative
 105 163 |   b = tested_positive
```

Figure 6. Summary of proposed CSE-DT classifier

Table 2. Classified Instances

Total No. of instances	Classification Algorithms	Correctly Classified	Incorrectly Classified
768	Naïve Bayes	586	182
	SVM	500	268
	Decision Tree	567	201
	CSE-DT	627	141

Table 3. Performance measures of CSE-DT with other classifiers

Classification Algorithms	Accuracy %	Precision %	Recall %	F-Measure %	ROC
Naïve-Bayes	76.30	75.90	76.30	76.00	0.819
SVM	65.10	42.40	65.10	51.30	0.500
Decision Tree	73.82	73.50	73.80	73.60	0.751
CSE-DT	81.64	81.70	81.60	80.90	0.859

The performance measures of CSE-DT were compared with SVM, Naïve-Bayes and the Decision Tree in terms of Recall, Precision, Prediction Accuracy, F-Measure and ROC values. Table 3 depicts that CSE-DT with the highest accuracy of 81.6%. This implies that CSE-DT classifier has better chances of predicting diabetes disease more accurately due to higher prediction accuracy value attained. Figure 7 and Table 3 depict the instance classification performance measure of all the classifiers including CSE-DT into correctly classified and incorrectly classified. It is shown that the CSE-DT obtained highest corrected classification performance.

Furthermore, Figure 8 depicts the performance measure of all the classifiers in terms of accuracy, precision, recall and F-Measure. Results show that CSE-DT outperformed among others in all the performance measures especially in classification accuracy. The proposed system obtained 81.64% accuracy against naïve Bayes of the highest performance in the existing system in [7]. Finally, the CSE-DT obtained the highest ROC value among all classifiers as depicted in Figure 9.

5. CONCLUSION

In machine learning, medical diagnoses has been the trending research areas. Awareness of the most relevant features (risk factors) of diabetes disease will help the researchers to focus on how to improve on classification accuracy of classifiers. This research has successfully hybridized CSE and DT as a single system CSE-DT for identifying, and selecting the most relevant features (risk factor) for diabetes classification. CSE-DT has been successfully applied on PIDD diabetes dataset which comprises of 768 instances and 8 features. The selected risk factors for the prevalence of diabetes by the proposed CSE-DT are number of times pregnant, plasma glucose concentration, blood pressure, diabetes pedigree function and mass. Meanwhile, the irrelevant features discarded are BMI, age and triceps skin fold thickness, and they are not considered as a risk factors. The proposed CSE-DT model has been compared with the existing algorithms in terms of classification accuracy, precision, recall, F-Measure and ROC. The results showed that the CSE-DT outperformed other algorithms with a classification accuracy of 81.64%.

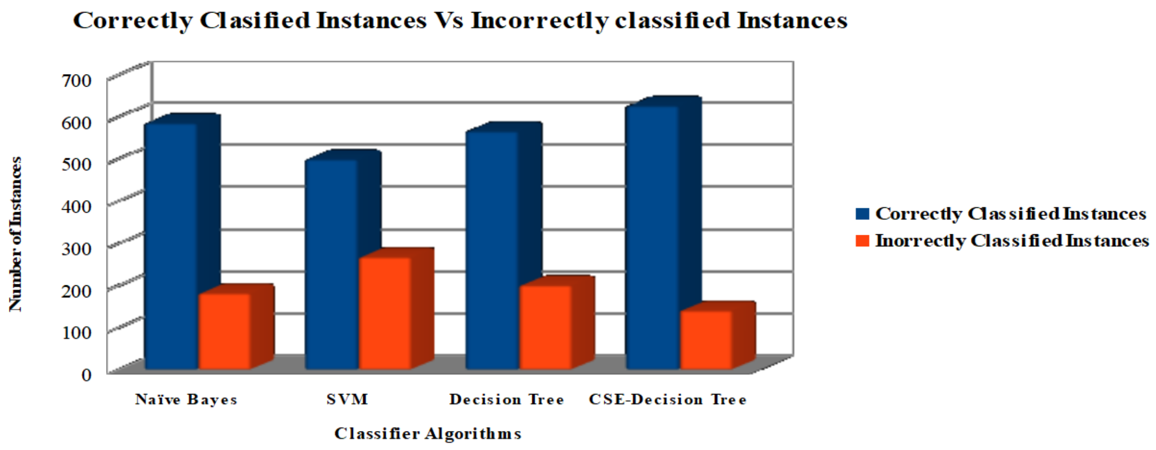


Figure 7. Instance classified performance measure

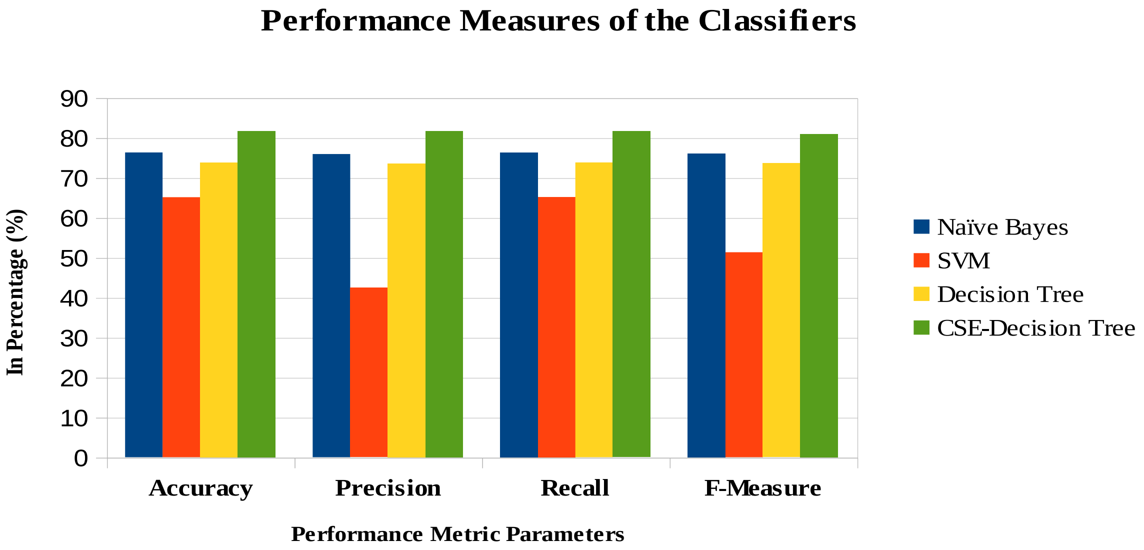


Figure 8. Performance measures of all the classifiers

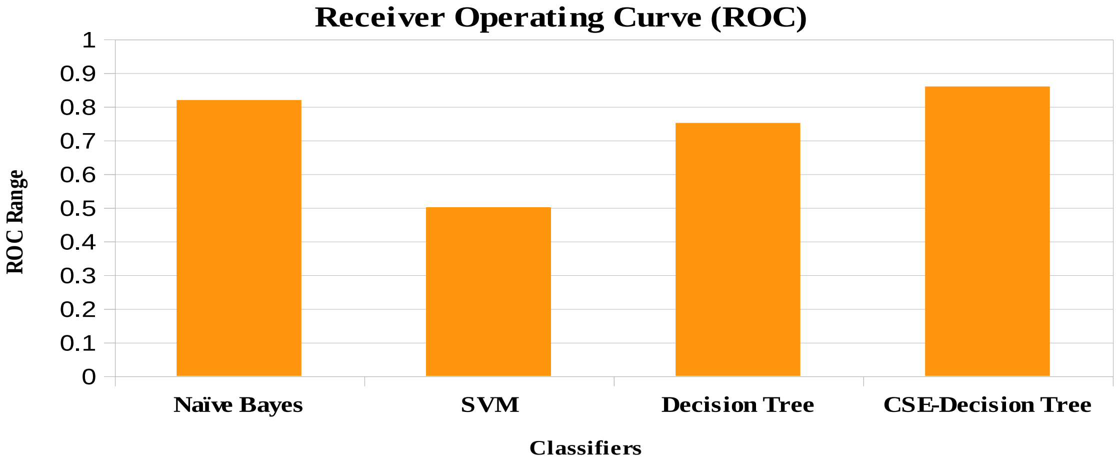


Figure 9. Receiver operating curve of all the classifiers

REFERENCES

[1] S. Hina, A. Shaikh and A. S. Sattar, Analyzing diabetes datasets using data mining, *Journal of Basic and Applied Sciences*, 13, 2017, 466-471.

[2] H. Wu, S. Yang, Z. Huang, J. He and X. Wang, Type 2 diabetes mellitus prediction model based on data mining, *Informatics in Medicine Unlocked*, 10, 2018, 100-107.

[3] M. Shuja, S. Mittal and M. Zaman, Effective prediction of type ii diabetes mellitus using data mining classifiers and SMOTE, *Advances in Computing and Intelligent Systems*, Springer: 2020, 195-211.

[4] G. Swapna, R. Vinayakumar, and K. P. Soman, Diabetes detection using deep learning algorithms, ICT Express, 4(4), 2018, 243-246.

[5] N. Sharma and A. Singh, Diabetes detection and prediction using machine learning/IoT: A survey, in *Advanced Informatics for Computing Research*, A. Luhach, D. Singh, P. A. Hsiung, K. Hawari, P. Lingras and P. Singh (eds). Communications in Computer and Information Science, 995, 2019, 471-479.

[6] L. Tapak, M. Hossein, H. Omid, and P. Jalal, Real-Data comparison of data mining methods in prediction of diabetes in Iran, *Healthcare informatics Research*, 19, 2013, 177-185.

[7] D. Sisodia, and D. Singh, Prediction of diabetes using classification algorithms, *Procedia Computer Science*, 132, 2018, 1578-1585.

[8] M. P. Bamnote, Design of classifier for detection of diabetes mellitus using genetic programming, *Advances in Intelligent Systems and Computing*, 1, 2014, 763-770.

[9] N. Nai-Arun, and R. Moungmai, Comparison of Classifiers for the Risk of Diabetes Prediction, *Procedia Computer Science*, 69, 2015, 132-142.

[10] B. M. Patil, Hybrid prediction model for type-2 diabetic patients, *Expert Systems with Applications*, 37, 2010, 8102-8108.

[11] A. Ahmad, A. Mustapha, E. D. Zahadi, N. Masah and N. Y. Yahaya, Comparison between neural networks against decision tree in improving prediction accuracy for diabetes mellitus, in *Digital Information Processing and Communications*, V. Snasel, J. Platos and E. El-Qawasmeh (eds), 188, 2011, 537-45.

[12] A. Marcano-Cedeño, J. Torres and D. Andina, A prediction model to diabetes using artificial metaplasticity, in *New Challenges on Bioinspired Applications*, J. M. Ferrández, J. R. Álvarez Sánchez, F. de la Paz and F. J. Toledo (eds), 6687, 2011, 418-425.

[13] K. Nandini and T. Deepa, A study on disease prediction using data mining technique, *Our Heritage,* 68(19), 2020, 100-107.

[14] T. Anand, R. Pal and S. K. Dubey, Cluster analysis for diabetic retinopathy prediction using data mining techniques, *International Journal of Business Information Systems*, 31(3), 2019, 372-390.

[15] G. Khurana and A. Kumar, Improving accuracy for diabetes mellitus prediction using data pre-processing and various new learning models, *International Journal of Scientific Research in Science and Technology*, 6(2), 2019, 502-515.

[16] M. M. F. Islam, R. Ferdousi, S. Rahman and S. Y. Bushra, Likelihood prediction of diabetes at early stage using data mining techniques, in *Computer Vision and Machine Intelligence in Medical Image Analysis*, M. Gupta, D. Konar, S. Bhattacharyya and S. Biswas (eds), 992, 2020, 113-125.

[17] C. Fiarni, E. M. Sipayung and S. Maemunah, Analysis and prediction of diabetes complication disease using data mining algorithm, *Procedia Computer Science*, 161, 2019, 449-457.

[18] S. Saradha and P. Sujatha, Prediction of gestational diabetes diagnosis using SVM and J48 classifier model, *International Journal of Engineering & Technology*, 7(2.21), 2018, 323-326.

[19] S. B. Kotsiantis, Supervised machine learning: A review of classification techniques, *Informatica*, 31, 2007, 249-268.

[20] D. Sisodia and D. S. Sisodia, Prediction of diabetes using classification algorithms, *Procedia Computer Science*, 132, 2018, 1578-1585.