Diabetes Classification Using Cascaded Data Mining Technique

[1]J. N. Mamman [2]M. B. Abdullahi [3]A. M. Aibinu [4]I. M. Abdullahi
[1,2]*Department of Computer Science, Federal University of Technology, Minna. Nigeria*
[3]*Department of Mechatronic Engineering, Federal University of Technology Minna, Nigeria*
[4]*Department of Computer Engineering, Federal University of Technology Minna, Nigeria*

Abstract— **Clustering plays a major role in data mining for: building models from an input data set; predicting future data trends for further decision making; simulating and analysing model; and diagnosing of healthcare diseases. Currently, in diagnosis of healthcare diseases such as diabetes, the initial knowledge of the clustered data is required in the use of Artificial intelligence (AI) technique as data pre-processing and classification technique. However, the inability to have such a prior knowledge had led to less classification accuracy. In this work, a cascade of K-Means clustering algorithm and Artificial Neural Network (ANN) was proposed for clustering of diabetes dataset. The proposed model was implemented in two stages. In the first stage, a K-Means clustering was used to pre-process the dataset after the initial filtering operation. In the second stage, the ANN was used to classify the result obtained from the pre-processed dataset. The proposed cascaded model was applied on Pima Indian diabetes dataset (PIDD) obtained from one of the public repository. Experimental results shows that accuracy of 99.2% was obtained from the K-Means-ANN model. Further analysis also revealed that the cascade of K-means-ANN model outperformed the cascade of ANN-K-means model, thus establishing that the two cascaded models are not commutative.**

Keywords— **Data mining, diabetes disease, Pima Indian Diabetes Dataset, ANN, K-means clustering, Pre-Processed Data, ClassificationPut your keywords here, keywords are separated by comma.**

I. INTRODUCTION

Data Mining involves a set of techniques that can be used to create useful and necessary knowledge from data and to view unseen patterns in a voluminous dataset. Data Mining, this extraction regarding hidden predictive information from big databases, is an effective new technological know-how with fantastic potential to assist companies to pay attention to the most crucial information in their data warehouses. It also involves the analysis of large volumes of data from different views and summarizing them into information that will be useful.Data mining proffer ways to clearly see patterns which otherwise would have been quite difficult in predicting future behaviour [13]. The branch of science called data mining is vital in the extraction of non-trivial, useful information that was unknown. This is made possible through knowledge acquiring methods by the use of analytic methods. Also referred to as information harvesting, this technique makes use of pre-processed data as input and returns knowledge as output.

Data mining methods or techniques are tools that are often implemented at advanced institutions today for analysing available data and extracting information and knowledge to support decision-making [6]. The data mining models are of two types [22], Predictive and Descriptive. The predictive model makes forecast about unknown data values by utilizing the known qualities. Example include; Classification, Time series analysis among others. The descriptive model recognizes the patterns or interactions in data and explores this properties in the data studied. Example include; Clustering, Association rule, Sequence discovery, Summarization among others.

The application of data mining in the area of organic compound analysis, medical diagnosis, and design of products, automatic abstraction and prediction of the shares of audiences is increasingly growing [16]. It is also a veritable tool in scientific innovation, intelligence gathering, monitoring, marketing, fraud detection and it is gradually creeping into other fields [9].
Classification is one of the generally utilized data mining technique in classifying objects into distinctive classes focused around the class characteristics [20].

A. *DataMining Classification Techniques*

Data mining classification techniques are machine learning technique where each class contains an instance that is being distinguished because of its unique characteristics. It plays a major role in data mining and a useful predictive tool for building models from an input data set. It is also a tool used for predicting the class of labels which have never been discovered and are used in the classification of future data trends [21]. The field of healthcare is seen as "rich in information" yet "poor in knowledge" [12]. There are sufficient data available in healthcare sector and the tool needed for effective analysis for the discovery of hidden relationship and data trends is deficient. The field of science and business has effectively applied the concept of Knowledge discovery and data mining. Data mining if applied in the system of healthcare can lead to the discovery of valuable knowledge. The technology of data mining has provided approaches that are user-friendly to new and

undiscovered patterns in data. When knowledge from this is discovered it will play a very significant role in the administration of improved healthcare service. Knowledge discovery will also find application in the reduction of the occurrence of adverse effect in drugs when medical practitioner takes advantage of it and help in the suggestion of alternative therapy that are less expensive. Being able to predict the behaviour of patients based on the given history of the patient is one of the vital areas of application of data mining techniques in healthcare management. Very good health service means efficient diagnosis and correct administration of treatment. Inefficient clinical decisions can result in unacceptable consequences that are catastrophic [23].

B. *Diabetes*

Diabetes mellitus is a disease that occurs when a body is not able to deliver or react legitimately to insulin which is required to control glucose. Insulin is a standout amongst the most critical hormones in the body [7]. It supports the body in changing over sugar, starches and other sustenance things into the vitality required for everyday life. In any case, if the body does not create or appropriately utilize insulin, the repetitive measure of sugar will be determined out by urination. This ailment is called diabetes. The reason for diabetes is still a puzzle, despite the fact that obesity and absence of exercise seem to conceivably assume significant parts [17].

Glucose moves down in the circulatory system bringing about one's blood glucose or "sugar" to climb excessively high. There are two main types of diabetes. In type1 diabetes, the body totally quits creating any insulin, a hormone that empowers the body to utilize glucose found as a part of nourishments for vitality. Individuals with type1 diabetes must take day by day insulin infusions to survive. This manifestation of diabetes typically creates in kids or youthful grown-ups, yet can happen at any age. Type2 (likewise called grown-up onset or non-insulin-dependent) diabetes results when the body doesn't create enough insulin and/or is not able to utilize insulin legitimately. This type of diabetes typically happens in individuals who are above 40 years of age, overweight, and have a family history of diabetes, in spite of the fact that today it is progressively happening in more youthful individuals, especially young people. Diabetes not just is a helping component to coronary illness, additionally expands the dangers of creating Kidney infection, Blindness, Nerve damage, and vein damage. Measurements demonstrate that more than 80 percent of individuals with Diabetes bite the dust from some type of heart or vein maladies. Presently there is no cure for Diabetes; on the other hand, it can be controlled by infusing insulin, changing dietary patterns, and doing physical activities [18]. Diabetes is a significant wellbeing issue in both developing and developed nations and its frequency is climbing drastically [4].

Presence of outliers in a dataset affect the quality of the decision that can be made from such dataset. Hence, removing the outliers through data pre-processing is paramount for effective decision making. Further analysis has also shown that the accuracy results obtained from a single data mining technique without pre-processing is usually low due to the presence of outliers (noise) from the data. Hence, the need for a hybrid pre-processing and clustering data mining technique for effective decision making. Though, efforts have been made by researchers to use Artificial Intelligence (AI) hybridization techniques as data pre-processing and data clustering techniques. However, some of the existing and well known AI techniques need an initial knowledge of the clustered data, therefore the use of K-means is being envisaged as a good technique for generating initial cluster. Subsequent application of AI in a cascaded approach will therefore lead to increase in classification accuracy. Hence, in this paper the aim is to develop a cascaded model of data mining classification techniques for diabetes disease classification. In section II related literatures were reviewed in relation to the performance of data mining classification techniques using diabetes dataset. Section III includes the adopted methodology for the study, analysis and limitations of available techniques and the use of MATLAB tool for the analysis. Section IV presents the result obtained from the various stages of diabetes data classification using ANN in MATLAB. The study ends with section V and VI, which includes conclusion and recommendation respectively.

II. Review of Related Works in Data Mining Classification on Diabetes Dataset

Data mining has been applied by experts in virtually all areas of medicine. One of the challenging aspects in healthcare is the extraction of hidden knowledge from large diagnostic dataset. Machine learning (ML) offers techniques and tools that can help to solve diagnostic problems in a variety of medical fields.

Karegowda, Jayaram and Manjunath, [7] proposed a hybrid model for the classification of Pima Indian Diabetes Dataset on WEKA machine learning tool. The model used K-means and K-nearest neighbour to identify and eliminate wrongly classified instances. From the total of 768 instances available in PIDD, there are 376 cases with missing values leaving a total of 392 samples after removing the missing values. Performance analysis for a 70-30 ratio was performed for training and testing respectively. The experimental result showed that the proposed hybrid model when compared with k-NN improved the performance of k-NN with classification accuracy of 96.68%. However, k-NN classification accuracy is low, Computational cost for k-NN is high, finding k in very large dataset is computationally expensive and K-NN classification time is slow.

Gupta, Kumar and Sharma [5] conducted performance analysis of different data mining classification techniques involving Bayes Net, k- NN, SVM, RBFNN, MLP, LDA and Decision Tree on Pima dataset using Weka, Tanagra and Clementine tools. Knowledge discovery in database process was used in transforming the data for decision making. The relevant dataset to be used for the analysis was taken from the database while the target dataset that will undergo the analysis was also created. Noisy and inconsistent data were removed

and the target dataset which was created earlier was pre-processed to handle the noise. The analysis showed that SVM on Weka tool shows the most promising result for Pima Indian Dataset with 96.76% percentage accuracy. However, this study used two criteria measures for the comparison, Decision trees are prone to error with too many classes and can be complex and time consuming and SVM is slow in training and lacks transparency in result.

Arora and Suman [3] performed a comparative evaluation analysis of J48 (Decision Tree) and Multilayer perceptron (Neural Network) on five datasets. The two techniques used were passed directly to the Weka data mining tool for analysis. The evaluation analysis result shows that the neural network has a greater accuracy in diabetes and glass dataset with 75.39% accuracy. However, High processing time is required for large datasets and Decision trees are not good for predicting the value of a continuous class attribute.

Also Koklu and Unal [11] performed a comparative analysis of 768 diabetic patients using MLP, Naïve Bayes classifier and J48 as artificial intelligence classifiers performed on Pima Indian Diabetes Dataset using Weka tool. Each layer of MLP composed of neurons that are interconnected to each other by weights and the activation function accepts input from previous layers and generates output for the next layer. In the experiment, hyperbolic tangent sigmoid transfer function was used as the activation function. The analysis result compared shows that Naïve Bayes outperformed the other two classifiers with 76.30% percentage accuracy. However, only two criteria measure were used for the comparison and Classification accuracy is low.

Karegowda et al, [8] proposed a Cascaded K-Means and Decision Tree C4.5model using Rule based Classification for Diabetic Patients. Pima Indian Diabetes Dataset on WEKA machine learning tool was used for the analysis. The model used K-means to identify and eliminate wrongly classified instances. The correctly classified instances were used as input to Decision tree C4.5 after conversion of continuous data to categorical data for classification. Performance analysis for a 60-40 ratio was performed for training and testing respectively. The experimental result showed that the proposed hybrid model when compared with Decision tree C4.5 improved the performance of Decision tree with classification accuracy of 93.33%. However, Classification time is slow and Decision trees are prone to error with too many classes and can be complex and time consuming

Aibinu, Salami and Shafie, [2] proposed a biomedical signal classification using Pseudo complex- valued autoregressive (CAR) approach of a feedforward multilayer Complex valued Neural Network (CVNN). Complex data type 1 and complex data type 2 were the two RVD-CVD data normalization method that was evaluated. The complex-valued autoregressive coefficient was obtained from the weights and adaptive coefficients of a properly trained network. Complex data normalization technique within the range of 0 and π on Pima diabetes Indian dataset was used. Also monotonically increasing the hidden layers neuron after convergence was adopted. The training algorithm was used in order to find the set of parameters that minimize the mean of the sum of the squared error. The proposed model was compared with other neural Networks using Pima Indian Diabetes Dataset. Cross validation was used to validate the result. The experimental result shows that the proposed CVNN-CAR model outperform all other neural networks with 81.28% accuracy. However, No pre-processing of wrongly classified data and Classification accuracy is low

Khyati K. Gandhi and Nilesh B.Prajapati [10] proposed a feature selection and classification model for predicting the prevalence of diabetes. Feature selection such as K-Means clustering and F-score was used as a pre-processing technique for selecting optimal feature subset to increase the predictive accuracy. Z-score normalization was also used as a data pre-processing step which scales data in (-1, 1) range. The classification accuracy of SVM on Pima Indian Diabetes Dataset shows 98%. However, SVM slow in training lack of transparency in result.

Pandeeswari and Rajeswari [15] proposed the development of a hybrid model using K-Means clustering and Naïve Bayes classifier for the categorization of diabetes patients. The model consists of two stages, stage one used K-Means clustering to identify and remove wrongly classified instances. In the second stage, the correctly clustered instances from stage one was used as input to Naïve Bayes using Weka tool. The experimental result of the cascaded K-means clustering and Naïve Bayes has enhances the classification accuracy of a single classification technique (Naïve Bayes). The result shows that even with the missing values contain in the dataset, it still gives promising result with utmost accuracy rate. However, Naïve Bayes assumes independence of features.

III. CASCADED K-MEANS AND ARTIFICIAL NEURAL NETWORK

A. *Architectural Framework of the System*

Fig 1 and Fig 2 presents the system block diagrams for the development of two cascaded model of K-means clustering for data mining classification techniques. The block diagram contains different stages and processes involved in the cascaded classification techniques.

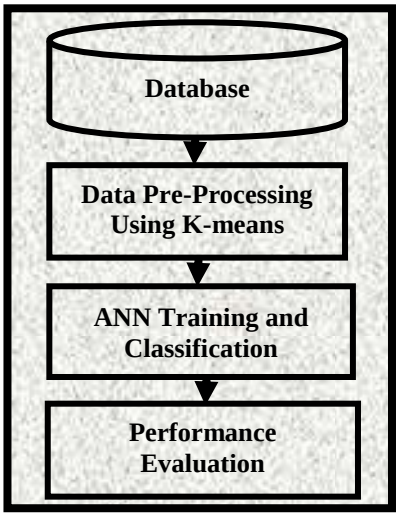

Fig 1: Block diagramof architectural Framework of model 1

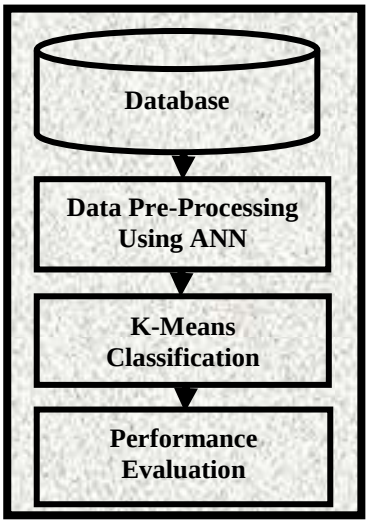

Fig 2: Block diagram of architectural Framework of model 2

B. *Database*

There are total of 768 samples of diabetic patients in the database. 5 patients had a 2-hour Oral Glucose Tolerance Test (OGTT) plasma glucose of 0, 28 patients had a diastolic blood pressure of 0, 192 patients had triceps skin fold thickness of 0, 11 of the patients had a BMI of 0, and 140 other patients had a 2-hour serum insulin readings of 0. The features with the values of zeros were removed during pre-processing. The data was clustered in order to obtain correct classification. Therefore, the total PIDD data for this study analysis is 392 un-processed samples, 8 attributes and 2 number of classes. Out of the 392 samples, 130 patients were positive (with diabetes) and 262 were negative (without diabetes) (Breault, Goodall and Fos, 2002).

In this study, diabetic patient data was obtained from Pima Indians Diabetes Dataset (PIDD) from machine learning repository of University of California, Irvine (UCI). A database comprised of training and testing data was created. The database to be used for the analysis comprised of 392 samples, 8 attributes and two classes (Tested positive and Tested Negative) of Pima Indian heritage above 21 years living in the United State. Table 1 presents the features of Pima Indians Diabetes Dataset.

TABLE 1 FEATURES OF PIMA INDIANS DIABETES DATASETS

S/N	Attributes	Mean	S. Deviation
1	Number of times pregnant,	3.8	3.4
2	2-hour OGTT plasma glucose,	120.9	32
3	Diastolic blood pressure (mm Hg),	69.1	19.4
4	triceps skin fold thickness (mm),	20.5	16.0
5	2-hour serum insulin (mu U/ml),	79.8	115.2
6	Body Mass Index (BMI) $(kg/m)^2$,	32.0	7.9
7	diabetes pedigree function,	0.5	0.3
8	Age (years),	33.2	11.8

C. *Data Pre-processing*

In order to improve the quality of the results obtained after mining and the complete mining process effectiveness, data pre-processing is implemented. Researchers and practitioners realize that in order to use data mining tools on the database effectively, data pre-processing is essential for successful data mining.

Intuitively, quality decision depends on quality data. Data pre-processing is an important step in the knowledge discovery

process. Preparing the data is very important as its quality influences the result from the analysis. This is because data are prone to noise due to huge size of databases, complexity and from multiple heterogeneous sources of data. The preparation and development of a quality data is achieved by pre- processing the data so as to eliminate the noise, irrelevant feature, redundant features and wrongly classified samples. These irrelevant features reduce classification accuracy, so data pre-processing is needed to prepare the data for mining tasks.

These pre-processing was implemented to remove noise (samples with missing attributes). The input data is the 768 samples with 8 attributes representing a 768x8 dataset. After the initial noise removal, the remaining noise free data were 392x8 samples.

D. *K-means Clustering (Filtering)*

K-means clustering is an algorithm that is used to classify or group objects based on attributes into K number of groups. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. The purpose of K-means clustering is to classify data. In this study, K-means clustering was employed to detect and remove incorrectly classified instances. The correctly classified samples obtained after the pre-processing using K-Means is 265 samples. This was done in order to improve the quality of mining result and efficient mining process. The classification technique that shows the highest accuracy and least error rate will be selected and recommended as the suitable technique for the classification of diabetes dataset.

E. *Cascaded Training and Testing for K-Means ANN*

The first cascaded classification model was implemented in two stages, the first stage used a simple K-Means clustering on the 392 pre-processed data as explained in section B. The wrongly classified data were removed, leaving correctly classified data. In the second stage, the correctly classified data was used as input to ANN for data classification.

1) K-Means-ANN Training

The Network training comprised of two stages (stage one and stage two). In stage one, the 392 pre-processed datasets obtained in section B was clustered using K-means clustering (with K=2) algorithm. This was done to remove irrelevant features and wrongly classified samples. After removing the wrongly classified samples, the remaining correctly classified samples were 265 samples (processed data) as explain in section D.

Finally, in stage two, the processed data from stage one was given as input to ANN. Stage two comprises of ANN training with a two layer, 8 input, 2 hidden layer and 20 hidden neurons feed forward back propagation ANN which was designed for the training of diabetes dataset using the processed dataset. The sample used for training in step two is 55.5% of the processed samples which is 147 samples. The above two stages were done in order to evaluate the effect of K-means clustering on the overall classification. Fig.3 is the flowchart for network training for the two stages in K-Means-ANN.

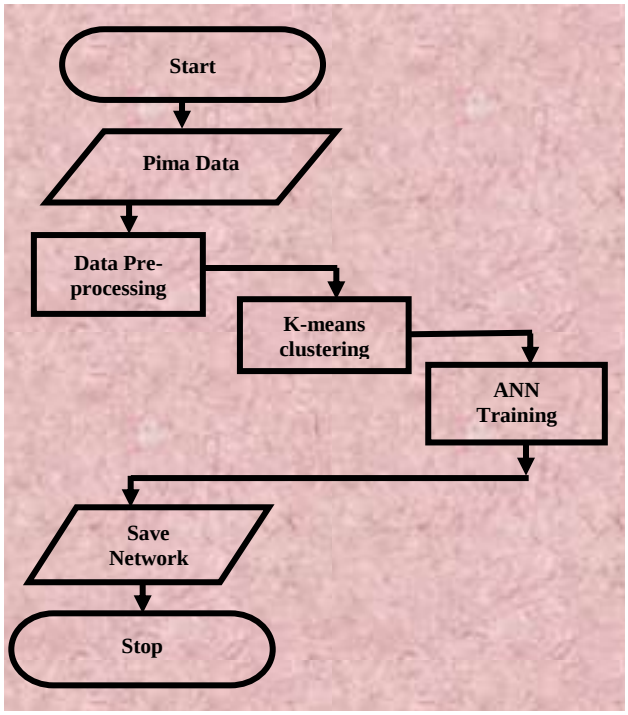

Fig 3: Network training flowchart for K-means- ANN

2) K-Means-ANN Testing

As explained in sub-section E1, network testing comprises of two stages, in the first stage, the processed data obtained from sub-section E1 was compared with the original class variable (Target) from PIDD. In stage two, 44.5% of data which has not be used for training was used for testing the already trained and saved network. In this stage, one hundred and eighteen (118) test samples comprising one hundred and nine (109) positive samples and nine (9) negative samples were used for testing. The result of the test samples obtained was compared with the original target data in PIDD. The network was tested to ascertain its performance by obtaining the number of samples that were correctly and incorrectly classified when compared with the target data. Fig.4 presents the flowchart of the network testing for the two stages.

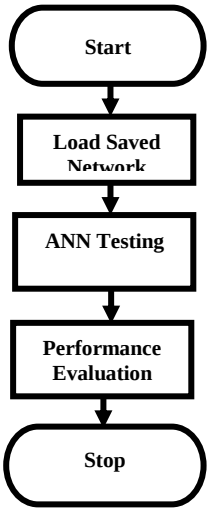

Fig.4: Network testing flowchart for K-means-ANN

F. *Cascaded Training and Testing for ANN- K-Means Classification*

Section E explained the cascaded K-means-ANN training and testing where K-means algorithm was used to cluster the data before classifying with ANN. However, the effect of reversing the two algorithm, which is removing noise, irrelevant features and wrongly classified samples using ANN first before classifying the data was investigated in this section.

This second cascaded classification model was also implemented in two stages, the first stage used ANN to cluster the 392 un-processed data as explained in sub-section E2. The wrongly classified data were removed, leaving correct classified data while in the second stage, the correctly classified data from first stage was used as input to K-Means for classification.

1) ANN –K-Means Training

As shown in Fig. 5, in stage one, the dataset after pre-processing was passed through a two layer, 8 input, 2 hidden layer and 20 hidden neurons feed forward back propagation ANN to classify the data as positive or negative using the target variable class. The output of the ANN was compared with the original target to obtain the correctly and incorrectly classified samples. After removing the wrongly classified samples, the remaining correctly classified samples (processed samples) was three hundred and thirty five (335) sample which was then used as input to the K-means clustering algorithm for final classification. In this stage, ANN was used for filtering (clustering). The ANN used here was to remove the noisy, irrelevant features and wrongly classified samples.

In stage two, sixty four point seven percent (64.7%) of the 335 processed samples obtained in stage one was clustered which is two hundred and seventeen (217) samples using K-means. The cluster mean of the result was saved which will be used for testing the remaining thirty five point three percent (35.3%) of the clustered data for training. In this stage, K-means clustering (with K=2) was used for classification on the 335 processed samples. The processed samples was given as input to K-means for classification. The above two stages were done in order to evaluate the effect of ANN clustering. Fig. 5 is the flowchart for the ANN-Kmeans network training.

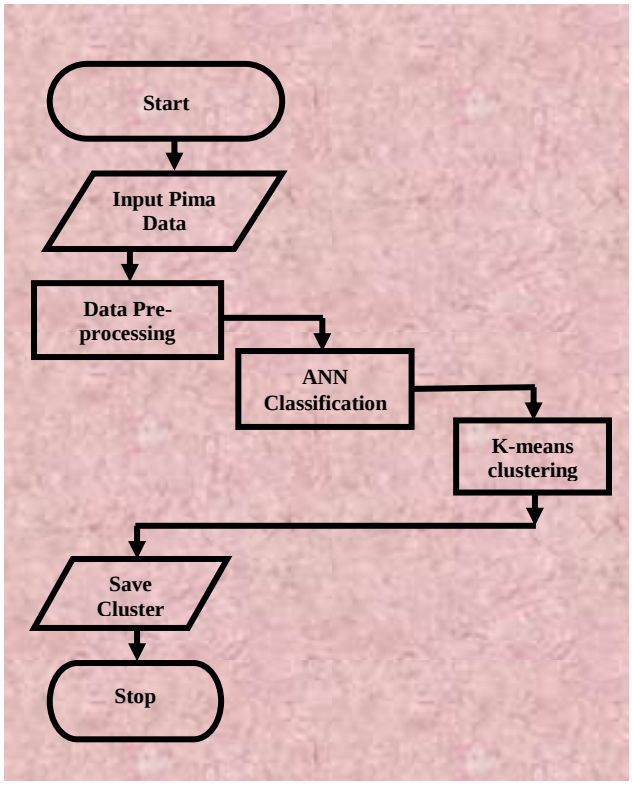

Fig. 5: Flowchart for ANN- K-means classification

2) ANN-K-means Testing

One hundred and eighteen (118) samples, which is 35.3% of the processed data from sub-section F1 was tested by comparing the samples with the cluster mean. The samples that belong to positive class and negative class were separated using the cluster mean. The testing was done to ascertain the performance of the cascaded ANN-K-means classification. The number of correctly classified and wrongly classified were recorded for performance evaluation. Fig. 6 is the flowchart for the testing of ANN-Kmeans classification.

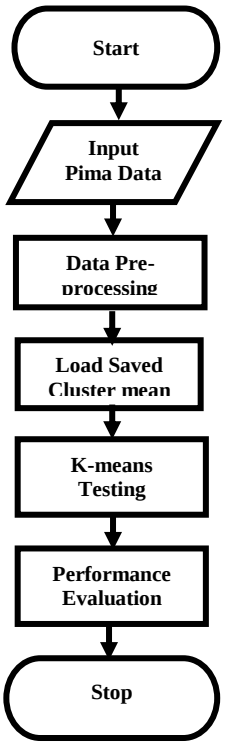

Fig. 6: Testing Flowchart for ANN- K-means classification

G. *Performance Evaluation*

Performance evaluation measures how well a data mining algorithm is performing on a given dataset. The Evaluation of data mining technique performance is a fundamental area of data mining. The evaluation is important for understanding the quality of the model and to examine the efficiency and performance of such model. In this study, classification accuracy, specificity, sensitivity and precision were used for the performance of the data mining techniques used for carrying out classification task. Mathematically, sensitivity, specificity, precision and accuracy are represented in equations 3.1, 3.2, 3.3 and 3.4 respectively [19]; [1]. Confusion matrix which shows a specific table layout of the performance of an algorithm that is used in evaluating the quality of an output of a classifier on a dataset was used for the analysis. This also represents the classification results.
Accuracy is the degree of closeness of a measured quantity to its actual value (Fawcett, 2004). Sensitivity is the fraction of positive examples predicted correctly by the model.

Specificity is the fraction of negative examples predicted as a positive class. Precision is the fraction of records that actually turns out to be positive in the group the classifier has declared as a positive class.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad 3.1$$

$$Specificity = \frac{TN}{\mathrm{TN} + \mathrm{FP}} \quad 3.2$$

$$Sensitivity(\mathrm{Recall}) = \frac{TP}{\mathrm{TP} + \mathrm{FN}} \quad 3.3$$

$$\mathrm{Pr}\,ecision = \frac{TP}{\mathrm{TP} + \mathrm{FP}} \quad 3.4$$

True Positive (TP) corresponds to the number of correct predictions that a patient is actually diagnosed positive. False Positive (FP) corresponds to the number of wrong predictions with the classifier that an instance is actually diagnosed positive.False Negative (FN) corresponds to the number of wrong predictions by the classifier that an instance is actually diagnosed negative. True Negative (TN) corresponds to the number of correct predictions that a patient is diagnosed negative.

IV. RESULTS AND ANALYSIS

This chapter presents the result obtained from the various stages of diabetes data classification using ANN. The Pima Indian Diabetes Dataset used for this classification was obtained from University of California machine learning repository. The data was pre-processed before training by removing the noise, irrelevant features and incorrectly classified samples.

A. Data Pre-Processing Results

Out of the 768 dataset in the database, a total of 392 samples was obtained after the pre-processing. Table 2 shows some of the achieved result of the pre-processed data in the dataset as explained in section III.

TABLE 2: PRE-PROCESSED DATA RESULT

S/N	Number of times pregnant	2-hour OGTT plasma glucose	Diastolic blood pressure (mm Hg)	Triceps skin fold thickness (mm)	2-Hour serum insulin (mu U/ml)	Body mass index (weight in kg/(height in m)^2)	Diabetes pedigree function	Age (years)	Class variable
1	1	89	66	23	94	28.1	0.167	21	0
2	0	137	40	35	168	43.1	2.288	33	1
3	3	78	50	32	88	31	0.248	26	1
4	2	197	70	45	543	30.5	0.158	53	1

B. K-means Filtering result

The second pre-process stage involves removing incorrectly classified instances from the pre-processed data obtained in section A shown in Table 2. The result shows that out of 392 samples, 265 were correctly classified while 127 samples which were wrongly classified were removed as shown in Table 3.

TABLE 3: CLASS VARIABLE AND K-MEANS FILTERING RESULTS

S/N	Class variable (Target)	K-means clustering Output
1	0	0
2	1	0
3	1	0
4	1	0
5	0	0
6	0	0
7	1	0
8	1	0
9	0	0
10	0	0
11	0	0
12	0	0
13	0	0
14	0	0
15	1	1

After applying K-means clustering, it was observed that out of the displayed fifteen results on Table 3, column two values (class variables) and column three values (K-means output) of serial numbers 2, 3, 4, 7 and 8 are wrongly classified. Hence, do not correspond to the target. Therefore, the application of K-means clustering on the dataset have eliminated the wrongly classified samples. Leaving the correctly classified samples to 265.

C. K-means -ANN Test Results on Processed Data

Table 4shows some of K-means-ANN test results. Fig. 7 shows confusion matrix while Table 5 shows the summary of the confusion matrix results for PIDD data indicating the TP, TN, FP and FN.

TABLE 4: ANN, ANN-K-MEANS AND K-MEANS ANN TEST RESULT

	ANN-K-MEANS RESULT		K-MEANS –ANN RESULT		
S/N	Output	Target	Output	Thresholded output (0.44)	Target
1	1	1	0.0000	0	0
2	0	0	0.0000	0	0
3	0	0	0.0000	0	0
4	0	0	0.0000	0	0
5	0	0	0.0002	0	0
6	1	1	0.0000	0	0
7	0	0	0.0000	0	0

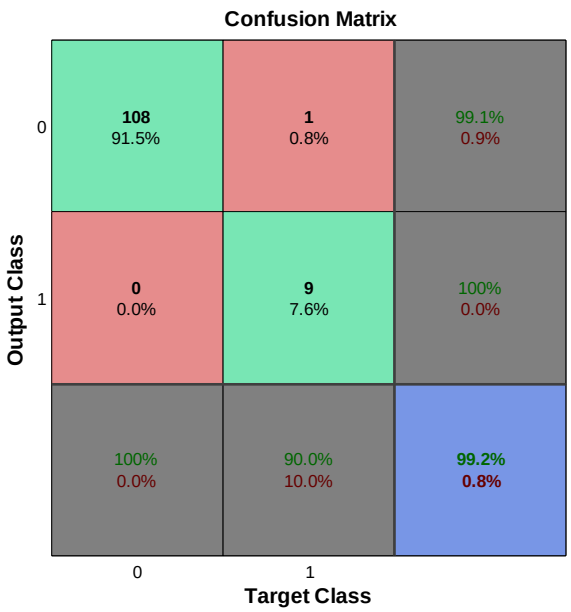

Fig. 7: Confusion Matrix for K-Means-ANN on PIDD test data

TABLE 5: CONFUSION MATRIX SUMMARY FOR K-MEANS-ANN

TP	FP	TN	FN	Total
108	1	9	0	118

The confusion matrix obtained using the K-means-ANN are given in Table 5. Accordingly, out of 108 data that the physician considered as positive, the K-means-ANN found that 108 were positive and 0 were negative. Moreover, out of the 10 data that the physician considered negative, the K-means-ANN found that 9 were negative and 1 were positive. Therefore the K-means-ANN gave values of 99.20% accuracy, 90% specificity, 100% sensitivity and 99.08% precision as shown in Table 6.

TABLE 6: KMEANS-ANN TEST RESULTS

Performance measurement	**K-means ANN Performance**
Accuracy	99.20%
Specificity	90%
Sensitivity	100%
Precision	99.08%

D. ANN- Kmeans Test Results on Processed data

Table 4 shows some of ANN-K-means test results. Fig. 8 showsANN-K-means confusion matrix while Table 7 shows the summary of the confusion matrix results for PIDD data indicating the TP, TN, FP and FN.

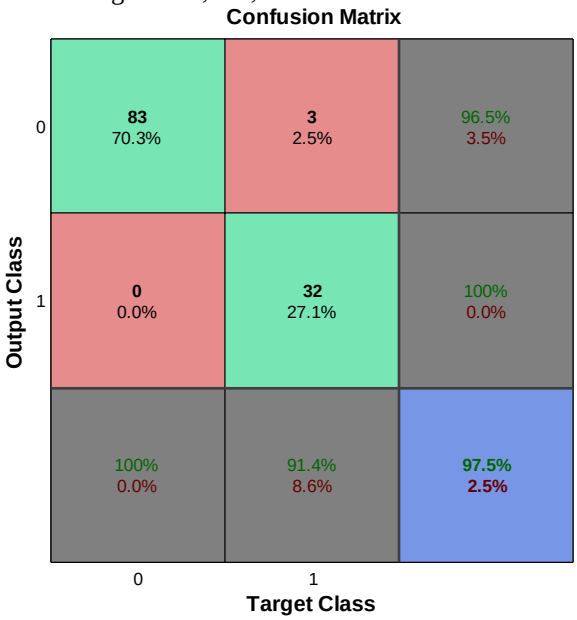

Fig. 8: ANN-K-means Confusion Matrix

TABLE 7: CONFUSION MATRIX SUMMARY FOR ANN-K-MEANS

TP	FP	TN	FN	Total
83	3	32	0	118

The confusion matrix obtained using the ANN-K-means are given in Table 7. Accordingly, out of 83 data that the physician considered as positive, the ANN-K-means found that 83 were positive and 0 were negative. Moreover, out of the 35 data that the physician considered negative, the ANN-K-means found that 32 were negative and 3 were positive. Therefore the ANN-K-means gave values of 97.45% accuracy, 91.43% specificity, 100% sensitivity and 96.51% precision as shown in Table 8. These results were obtained after several testing was performed on the K-means classifier.

TABLE 8: ANN-K-MEANS TEST RESULTS

Performance measurement	**ANN-K-means Performance**
Accuracy	97.45%
Specificity	91.43%
Sensitivity	100%
Precision	96.51%

E. Performance Evaluation Results

The performances of the cascaded techniques, K-Means-ANN and ANN-K-Means were individually computed using equations (3.1), (3.2), (3.3) and (3.4) from the values of TP, TN, FP, FN obtained and summarized in Table 9. From the study result on the Table 9, it can be deduced that K-means-ANN result is not the same as ANN-K-means result which implies that the system is not commutative. The results of the cascaded techniques on the basis of comparison shows that, K-means-ANN gave higher result than ANN-K-means with classification accuracy of 99.20%. The accuracy of K-means-ANN was better because the K-means clustering filtered more noise before the classification. That is, 265 samples filtered data was obtained unlike ANN-K-means filtering where 335 correctly classified samples were obtained. This is because K-means is better in filtering than classifying while ANN is better in classifying than filtering.

TABLE 9: PERFORMANCE EVALUATION RESULTS

Performance Metrics	**ANN- K-Means**	**K-Means – ANN**	**Best Performance**
True positive (TP)	83	109	
False negative (FN)	0	0	
True negative (TN)	32	9	
False positive (FP)	3	0	
Accuracy	97.45%	99.20%	K-Means – ANN
Specificity	91.43%	90%	ANN- K-Means
Sensitivity	100%	100%	K-Means –ANN and ANN-K-means
Precision	96.51%	99.08%	K-Means – ANN

Method	Accuracy %	Reference
Proposed model: K-Means+ANN, ANN+K-means, k=2	99.20 97.45	This study
k-means+DT continuous data)	93.33	Karegowda *et al* (2012b)
K-means + KNN,k=5	96.68	Karegowda *et al* (2012a)
ARTMAP-IC	81.0	Carpenter and Markuzon (1998)
GRNN	80.21	Kayaer and Yildirim (2003)
MLNN-GDA	77.60	Kayaer and Yildirim (2003)
MLNN-GDA-MM	76.56	Kayaer and Yildirim (2003)
MLNN-GDA + ADLR	77.60	Kayaer and Yildirim (2003)
MLNN-LM	77.08	Kayaer and Yildirim (2003)
MLNN-LM	82.37	Temurtas, Yumusak and Termutas (2009)
PNN	78.13	Temurtas *et al*, (2009)
CVNN-GDA	81.00	Aibinu, Salami and Shafie, (2010)
CVNN-CAR	81.28	Aibinu, Salami and Shafie, (2011)
PCA-ANFIS	89.47	Polat and Gunes 2007
LS-SNM	78.21	Polat, Gunes and Aslan (2008)
GDA-LS-SNM	79.16	Polat, Gunes and Aslan (2008)

Looking at the best performance, it is observed that the K-means-ANN is quite higher than the ANN-K-Means. It has been also observed that the two cascaded techniques shows higher classification accuracy of 97.45% and 99.20% for ANN-
K-means and K-means-ANN, respectively. It was found that the two techniques separately and the two cascaded models were compared according to performance parameters. To further verify the effectiveness of these techniques and models, they were tested using data of the same size. It was observed that the two cascaded models showed comparably good accuracies compared with single test data sets. Also it is recommended that the models will give optimal results of classification for any data set, provided it is trained with its previous data. Therefore, for better accuracy, specificity, sensitivity and precision of diabetic classification, K-means clustering with ANN is recommended.

F. Comparison with other published work

Results and methods adopted from various reported work were compared with different data mining techniques using Pima Indian diabetes dataset is shown in Table 10. From existing reported cases, this is the first application of K-means- ANN model on PIMA diabetes data. Also noticeable is the high accuracy value obtained in this study for K-means-ANN compared to earlier reported cases. This work is significantly different from any of the earlier work. From the table, it is also observed that most of the methods used were hybridized and the classification result were within the ranges of 76.56% – 96.68%.

TABLE 10. THE CLASSIFICATION ACCURACY WITH DIFFERENT DATA MINING TECHNIQUES ON PIMA INDIAN DIABETES DATASET

V. CONCLUSIONS

The study designed a two stage data clustering technique involving a K-Means based data pre-processing stage and ANN clustering stage for diabetes diagnosis was achieved by first filtering the 768 sample instances and secondly pre-process the filtered data using K-means clustering. The accuracy, specificity sensitivity and precision metrics are very important to better evaluate the performance of a technique which enables the researcher to achieve the third objective. When the classification technique and the two cascaded models were compared, the best accuracy, sensitivity and precision values were achieved with K-means-ANN, with classification accuracy of 99.20%. Since the correct classification of the patients is associated with sensitivity, the two cascaded techniques have better classification of diabetes than a single technique without pre-processing.

The implication of the results obtained when the cascaded techniques were implemented shows that the techniques when first clustered have better classification of diabetes cases (100% sensitivity) than when classified first before pre-processing. Therefore, the K-means clustering with the classification technique is adopted for the classification of diabetes disease.

VI. RECOMMENDATION

This study used a data mining classification technique for the classification of diabetes disease on Pima Indian Diabetes Datasets. Areas of improvements can be in the use of different dataset for the classification of other diseases.

REFERENCES

[1] Aibinu, A. M., Salami, M. J. E. and Shafie, A. A. "Application of modelling techniques to diabetes diagnosis". *In IEEE conference on biomedical engineering and sciences, Malaysia*. 2010.

[2] Aibinu, A. M., Salami, M. J. E., and Shafie, A. A. "A novel signal diagnosis technique using pseudo complex-valued autoregressive technique". *Expert Systems with Applications*; 2011; *38*(8), 9063-9069.

[3] Arora, R., and Suman, S. (2012). "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA". *International Journal of Computer Applications*, *54*(13), 21-25.

[4] Giveki, Davar, Hamid Salimi, GholamReza Bahmanyar, and Younes Khademian. "Automatic detection of diabetes diagnosis using feature weighted support vector machines based on mutual information and modified cuckoo search". *arXiv preprint arXiv: 2012;1201.2173*

[5] Gupta, S., Kumar, D., and Sharma, A. "Performance analysis of various data mining classification techniques on healthcare data". *International journal of computer science and Information Technology (IJCSIT);* 2011: *3*(4).

[6] Kabakchieva, Dorina. "Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies* 13, no. 1; 2013; p. 61-72.

[7] Karegowda, Asha Gowda, M. A. Jayaram, and A. S. Manjunath.

"Cascading K-means clustering and K-nearest neighbor classifier for categorization of diabetic patients". *International Journal of Engineering and Advanced Technonlogy* 1; 2012; p. 147-151.

[8] Karegowda, Asha Gowda, V. Punya, M. A. Jayaram, and A. S. Manjunath. "Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4. 5." *International Journal of Computer Applications;* 2012; 45

[9] Kaushik H. Raviya and Biren Gajja,"Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA tool". *Indian Journal of Research.* (Volume: 2, Issue: 1, January 2013 ISSN - 2250-1991.

[10] Khyati K. Gandhi and Nilesh B.Prajapati. "Diabetes prediction using feature selection and classification". International Journal of Advance Engineering and Research Development (IJAERD) Volume 1, Issue 5, (2014). e-ISSN: 2348 - 4470 , print-ISSN:2348-6406.

[11] Koklu, M., and Unal, Y. "Analysis of a Population of Diabetic Patients Databases with Classifiers". *human resources, 1*, 2; 2013.

[12] Krati Saxena and Shefali Singh. "Diabetes Mellitus Forecast Using Artificial Intelligence Techniques".International Conference of Advance Research and Innovation (ICARI-2014) 544 ICARI, ISBN 978-93-5156-328-0

[13] Neelamegam S. and Ramaraj E. "Classification algorithm in Data mining": An Overview. International Journal of P2P Network Trends and Technology *(IJPTT)* 4; 8; 2013; ISSN: 2249-2615.

[14] Pal, Jiban K. "Usefulness and applications of data mining in extracting information from different perspectives". 2011;

[15] Pandeeswari, L., & Rajeswari, K. "K-Means Clustering and Naive Bayes Classifier For Categorization Of Diabetes Patients*". International Journal of Innovative Science, Engineering & Technology (IJISET),* Vol. 2 Issue 1, 2015.

[16] Patil, Tina R., and Mrs SS Sherekar. "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification". *International Journal of Computer Science and Applications* 6; 2; 2013.

[17] Pham, Huy Nguyen Anh, and Evangelos Triantaphyllou. "Prediction of diabetes by employing a new data mining approach which balances fitting and generalization". *Computer and Information Science*. Springer Berlin Heidelberg; 2008; p. 11-26.

[18] Polat, Kemal, and Salih Güneş. "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease". *Digital Signal Processing* 17.4; 2007: p. 702-710.

[19] Raschka, S. "An Overview of General Performance Metrics of Binary Classifier Systems". *arXiv preprint arXiv:2014; 1410.5330*.

[20] Ravinder Reddy R., Padmalatha E., Ramadevi Y. and K.V.N Sunitha. "Performance Analysis of Classifiers for Intrusive Data and Rough Sets Reducts". IJCSNS *International Journal of Computer Science and Network Security*, VOL.14 No.8. 2014

[21] Shazmeen, S. F., Baig, M. M. A., and Pawar, M. R. "Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis". *Journal of Computer Engineering*, 10(6), (2013) 01-06.

[22] Tan Pang-Ning, Steinbach, M., Vipin Kumar. "Introduction to Data Mining". Pearson Education, New Delhi, ISBN: 978-81-317-1472-0, 3rd Edition, 2009.

[23] Ray, Kisor, Santanu Ghosh, Mridul Das, and Bhaswati Ray. "Design & Implementation Approach for Error Free Clinical Data Repository for the Medical Practitioners." arXiv preprint arXiv:1503.08636 (2015).