European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 08, Issue 02 , 2021

Optimized Feed Forward Neural Network For Classification Of Diabetes In Big Data Environment

N.V.Poornima[1], Dr.B.Srinivasan[2] , Dr.P.Prabhusundhar[3]

[1,2,3] Department of computer science Gobi Arts & Science college Gobichettipalayam.

ABSTRACT:

Diabetes can be turn into life threaten diseases, if it is not treated at an early stage. Especially, in the women, the chances of diabetes is higher as compared to men due to the hormonal changes during pregnancies. Due to this, they suffer a long term diabetes as well as other diseases due to their tensions and regular life chores. This can be prevented if the diagnosis is determined at an early stage. Mostly, the trained doctors are required to confirm the diabetes. It requires manual work and complete knowledge in it. This problem is avoided by several research works using machine learning algorithm for the classification. Those algorithms process effectively for smaller dataset with smaller number of attributes. Hence, to overcome this shortcomings, in this an optimization based classifier is proposed to process on larger datasets like BIG DATA. In existing, the improved K-means and logistic regression algorithm is applied. It able to improve the classification rate but the features for training and computational time is high. To overcome this problem, an optimization based machine learning approach is used in this paper. Before the classification process, the data is pre-processed to remove the missing values and non-available values. In this, the K-means clustering is applied to the data to remove the outlier data and to reduce the training time period of the prediction process. The prediction of diabetic is done with the help of feed forward neural network. The inputs for the network is by selecting dominant attributes in the dataset using optimization process. The objective function is to reduce the misclassification rate of the classifier. In this, the cuckoo search optimization and feed forward neural network is used. This approach able to improve the accuracy as compared to the existing technique. The whole process is realized in MATLAB R 2018a environment and evaluated in terms of accuracy, precision, recall, F-measure and Matthew correlation coefficient.This approach outperforms all other existing technique with F-measure of 96.7%.

Keyword: Diabetic, dataset, pre-processing, characteristics, reduction using CSO, Feed forward neural network.

1. Introduction:

The term Big in “Big Data” itself defines it operations and size is large. Big indicates not only the size, it also denotes the nature and processing of the data. Big data doesn’t mean it refers to only large amount of data. But, the size of data can be small but it is complex in nature and requires tedious process to perform an operation.

It is a special type of data whose volume is in exponentially increasing manner with respect to time. It requires special type of processor to store, retrieve and perform operations on it. Moreover, it cannot be processed with the simple data mining algorithms. It requires special tool and algorithms for processing it. All operations can be performed on the big data as similar to the normal data processing methods.

Big data can also be called as a combination of three V's. The three V's are Variety, Velocity and Volume. These three V's are change in nature and it will be of higher volume for the big data.

Here, in variety, there are three types. They are structured, unstructured and semi-structured. The term structured indicates that data is arranged in an ordered format. The unstructured data denotes there is no definite format for the data. In Semi-structured, it is a hybrid of both structured and unstructured format.

Big data plays a major role in many applications like healthcare units, Educational oriented, Banking sector, Information technology sector, Manufacturing and retail sector. In health care unit, it plays a greater role by storing the information in a larger and continuous manner for saving the patient information. Not only for the storing purpose, it able to help the doctors by diagnosis the diseases by combining big data and artificial intelligence algorithms. In this, the role of big data in diagnosis the diabetics using machine learning technique. The remaining section describes about the techniques used in the big data for predicting the diabetes.

Diabetes will become a deadly diseases if the diseases is not treated at an early stage. Because, the diabetes leads to the new diseases like weight gain, blindness and hypertension and heart related problems [1]. This diseases will vary based on the time period of the diabetes without diagnosis. For short term, the weight gain and hypoglycaemia will act as the prediction model. The nerves and eye and heart related diseases will occur in the long term patients.

A scalable random forest model is proposed for predicting the diabetes in a patient using Hadoop environment [2]. Here, the records of both diabetic and non-diabetic patients is collected. Then, the attributes present in the records are selected. Based on the labelled information, the criteria for the diabetic records and other attributes are determined. This information is processed with other algorithms also like traditional random forest, CART model and scalable random forest. Among these, the scalable random forest able to identify the diabetic records based on its attributes accurately with a percentage of 87.5%.

Association rule based diabetic prediction is performed using Hadoop environment [3]. Here, a survey of Apriori algorithms which works based on the association rule is used for the prediction of diabetes in a patient. In this, seven Apriori algorithms using map reduce functions is implemented. Among these, the R-Apriori is the best for the prediction of diabetes and MR-Apriori is the next choice for the prediction.

A datamining and decision tree based approach is proposed for the classification of diabetic and non-diabetic records [4]. Here, the attributes in the dataset is ranked using information gain ratio. The highest ranked attributes were used for the classification. The classification is performed with the help of one of the decision tree classifier ID3. It able to predict the diabetes with accuracy of 94%.

A feature reduction based approach is proposed for the classification of diabetes and non-diabetes [5]. Here, the attributes of the dataset is reduced with the help of filtering process.

Four types of filtering is used to rank the features and get trained with the help of three classifiers. Four filter were clustering variation, correlation based, and Information gain and chi-square method. Feature selection based classification able to improve the classification as compared to the traditional algorithms.

Generally, the machine learning algorithms were used for the detection of diabetes. It able to predict the disease accurately but it suffer from minor problem to predict. This problem is overcome by the hybrid of those approaches [6]. Here, the traditional machine learning algorithms were combined based on its individual performance in classification. Due to this, the accuracy of the diabetic prediction is improved as compared to the individual algorithms.

Five machine learning algorithms like Gaussian Mixture model, artificial neural network, extreme learning machine, and logistic regression and support vector machine techniques were applied on the diabetic classification [7]. Among these, the artificial neural network outperformed the other classifiers with accuracy of 89%.

Diabetes is predicted using the basic machine learning algorithms like k-nearest neighbour, random forest, linear discriminant analysis, Classification and regression tree analysis and support vector machine is used on PIMA dataset [8]. Random forest is the best to predict the diabetes as compared to the other classifiers.

A detail survey about the prediction models and the corresponding techniques used in the diabetic classification [9]. Here, the PIMA and Koges dataset is discussed. Mostly prediction model is neural network and technique is the machine learning algorithm. Because, the prediction model only can improve the classification rate.

A clustering based feature reduction is used for the diabetic classification on PIMA dataset [10]. Here, the k-means cluster is used to reduce the attributes for the classification. A type of decision tree classifier named J48 is used for the classification process. Due to this reduced attributes, it able to achieve accuracy of 90.04% as compared to the agglomerative and hierarchical clustering with same classifier.

The organization of the paper is as follows: Recent techniques in classification of diabetes is discussed in section 2. Shortcomings in the existing and novelty in the proposed method is explained in section 3. The proposed clustered and optimized classifier is elaborated in section 4. Results and discussion based on the proposed method is given in section 5. Paper is concluded with summary and its extension in further sections 6 and 7

2. Related works:

In this section, the recent techniques used in the classification of the diabetics is discussed and it is shown in the table format in table 1.

Author	Technique	Dataset	observations	demerits
Sisodia&Sisodia (2018)	Decision tree Naïve bayes Support vector machine	PIMA	Naïve bayes is best	All attributes are used but accuracy is 76% only
Sarwar et al., (2018)	Decision tree Naïve bayes Support vector	PIMA	SVM and KNN gives best result	Accuracy is only 77% with all attributes

	machine K-nearest neighbour Random forest			
Mir and Dhage (2018)	Naïve bayes Support vector machine Random forest CART using WEKA	PIMA	SVM is best	No feature reduction process which can reduce the training time.
Liu et al., (2018)	Multi task learning on type 2 diabetes	Data collected from a study in Truven health	Association between the complications can be identified	Basic demographic variables only used. MTL is similar to Single task learning with lesser variation in overall results
Chen et al.,(2018)	Smart diabetes monitoring module with Bigdata and cloud	Real time data collection from Hubei province	SVM and ANN is used for the prediction	Larger records processed with no feature reduction results in higher training time
Saru and Subhashree (2019)	Bootstrapping With the following techniques Decision tree Logistic regression with SVM K-NN	PIMA	Decision tree	Limited Base classifier is only used
Parast et al (2019)	Dynamic prediction model	DPP data from NIDDK data centre	AUC is of about 0.6-0.7	Lack of clarity in the data and assumptions are made
Bai et al., (2019)	Diabetic prediction using clustering and classification method	PIMA	For clustering-OPTICs (ordered clustering) and BIRCH (hierarchical clustering), OPTICS is best. Gaussian Naïve Bayes is used for	Recent classifiers may improve the performance efficiency using clustering

			classification	
Makino et al., (2019)	Deep learning and natural language processing for progression in type 2 diabetes	Data collected from clinic	NLP is used for processing text and report Deep learning is used for forecasting and feature extraction in time series model	Accuracy is of 71% only even with larger records and processing time
Fiami et al., (2019)	Prediction of diabetes and correlation between the diabetes	Real time data collected from hospital in Indonesia	Naïve bayes and K-means clustering. Best features determined.	Manual analysis is performed for finding the best feature set.
Mosquera Lopez et al (2020)	Predicting nocturnal hypoglycaemia	Data collected from Tidepool in the format of days and nights	Mutual information is used for feature selection. Support vector regression is used for the prediction	Efficiency and training time period can be improved with the help of recent algorithms and optimization
Maniruzzaman et al., (2020)	Predicting diabetics	Data collected from National immunity department	Logisitic regression is used for feature selection. Many machine learning classifiers were used. Random forest is the best	Feature selection can be improved further to define automated optimal parameter
Subramaniyam et al., (2020)	A study of machine learning algorithms on diabetic prediction	Dataset collected from different forms	Supervised, semi-supervised and un-supervised machine learning algorithms were used	Does not define a clear idea about the algorithm used in it
Prasad et al., (2020)	Predicting diabetic with kidney diseases	Data collected in real time	Decision tree Naïve bayes Random forest J48	An accurate selection of classifier cannot be obtained due to difference in performance
SivaParthipan et al., (2020)	Predicting diabetes using	-	Map and reduce functions in	Size and dataset used is not

	statistical assessment		Hadoop system	defined to get a clear idea about the working

3. Existing method:

In existing, the improved K-means clustering and logistic regression algorithm is used for the diabetes classification [26]. Here, the analysis is carried out on the PIMA dataset. Here, the data is classified in two stages. In the first stage, the unwanted data are removed using K-means clustering. Then, the accurate dataset is trained with the help of Logistic regression algorithm. Then, the classified result is verified in terms of precision, recall, F-measure and Matthew correlation coefficient. Overall, this approach is able to achieve F-measure 95 % as compared to the basic classifiers. Yet, this method is able to improve the prediction and data reduction, it has the following drawbacks in it.

- The K-means clustered is used only to remove the outlier record.
- The total processing will perform on all the attributes for the classification process.
- Basic LR classifier can be modified to reduce its computational time for processing all the attributes.

3.1 Research gap and Novelty:

From the survey of various works, it is observed that the following process only performed in the prediction and it is shown in the table format

Table 2. Novelty in the proposed method

Existing techniques	**Shortcomings**	**Proposed method**	**Advantage of proposed**
Feature extraction	Directly performed on the attributes	Directly performed on the attributes	Able to predict diabetic accurately
Outlier removal	No	K-means cluster	Reduce the training time
Feature reduction	Only few papers used. Mostly K-means clustering	Optimization approach using CSO	Able to select correct attribute even in larger dataset using fitness function
Classifiers	Mostly binary or ANN	Feed forward neural network	Able to understand the pattern in the attributes
Overall merits	Classified better with larger training time for smaller dataset oriented	Able to classify larger dataset using optimal attributes alone	Reduces training time by outlier removal and optimal attribute selection

4. Proposed method:

This section is to describe briefly about the working and flow of the proposed method for the classification of diabetes. The flow diagram for the clustered and optimize based classification is given in figure 1.

Figure 1. Flow diagram for the proposed method

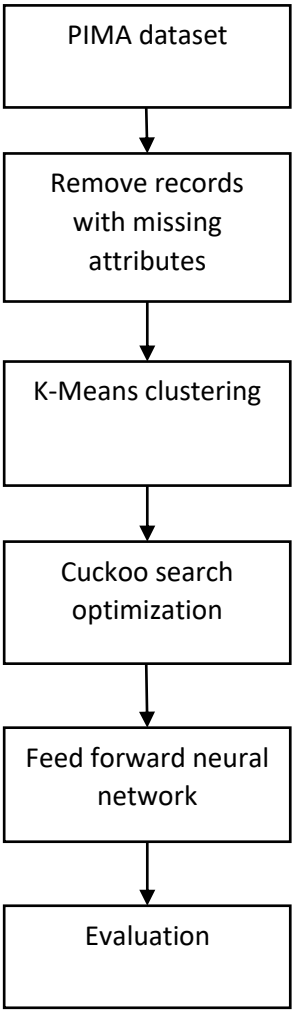

Figure 1 depicts the flow of implementation process in the proposed method. The individual process in the flow diagram is explained below:

4.1 Dataset:

In this, the PIMA diabetes dataset is downloaded from [27]. This dataset is about the diabetic patient with kidney diseases and simply it can termed as Type two diabetes.

In this dataset, all the records were belongs to women and the age were equal or above 21 years and eight attributes were collected from them as shown in the table 3.

Table 3. Attributes in the dataset

Attributes	**Use**	**Value range**
Pregnancy count	Number of times got pregnant	0-17
Glucose	Glucose level after two hours of oral consumption of glucose	0-199
Blood pressure	Diastolic blood pressure in mm/Hg	0-122
Skin thickness	Triceps skin fold thickness in mm	0-99
Insulin	2- hour Insulin test (muU/ml)	0-846
Body mass index	Weight in kg /height in cm	0-67.1
Diabetes pedigree function	Diabetes information	0.08-2.42
Age	Age of the women	21-81

4.2 Cleaning of data:

It is the second stage of process to remove the missing records in the data. In this, the missing records is removed instead of making assumption. Because the assumption leads to misclassification. In order to avoid that, here the missing attribute record will be removed completely. Even though, the records has zero values, those records are kept because for the complete analysis.

4.3 Outlier removal using K-means clustering:

The term outlier refers, some of the records can be synthesized one or repeated one. This information leads to the false classification rate and increase the processing time. To avoid the processing time and false rate, in this, the outliers in the dataset were removed using K-means clustering [28].

Let consider a dataset D with n data points and it is given with the following equation 1.

$D = \{D_1, D_2, \ldots, D_n\}$	(1)

In this, the dataset D is divided into only two groups. Because, the labels for the dataset is two namely diabetic and non-diabetic. Based on this, the membership functions for the cluster is identified by calculating the centroid for this two groups.

Then, the data points are assigned to each group by finding the distance between the point and centroid. The data point in which it has minimum distance cluster centroid and it will be associated to it. It is based on the following function in equation 2.

$$OF_{Kmeans} = \sum_{i=1}^{n}\sum_{k=1}^{K} R_{ik}\|D_i - \mu_k\|^2 \quad (2)$$

In the above equation 2, the term R indicates the belonging of the data point in the clusters. If the data point is belong to that cluster, then it will be 1 otherwise 0. Here, K =2 and n= 746. The centroid of the cluster is denoted using $\mu_k$and it is given in equation 3.

$$\mu_k = \frac{1}{n_k}\sum_{D_i \epsilon c_k} D_i \quad (3)$$

The points which are not belong to the cluster is denoted with the help of silhouette value in equation 4.

$$sil(D_i) = \frac{b(D_i) - a(D_i)}{\max\{a(D_i), b(D_i)\}} \quad (4)$$

The terms a and b indicates the average distance of the data point to other points in the same cluster and its neighbouring clustering respectively. The points which has minimum silhouette value will be considered as the outlier.

Based on this, the outlier in the dataset is determined and it will be removed. Then, the pre-processed dataset is used for the optimization process.

4.4 Attribute selection using Cuckoo search optimization:

This section is to describe the process of selecting the dominant attribute in the outlier removed dataset. The dominant attribute selection is performed to reduce the computational time for the calculation. In previous methods, it is performed manually by grouping attributes and validating it. But, in this it is performed through an automatic process called optimization.

The optimization process is to find the solution for a problem through iteration process. Here, the cuckoo search optimization is used. Because, the cuckoo lays it egg in other birds stronger nest and hatch its egg [29]. Due to this, the selection of nest is chosen by satisfying all the conditions. In this, the condition is to maximize the accuracy of the classifier. The objective function of the cuckoo search optimization is given in equation 5.

$$OF_{CSO} = 1 - Accuracy \quad (5)$$

The cuckoo search optimization follows the three conditions to find the solution for a problem. The three conditions are as follows.

- All cuckoos can lay egg only one at a time t and dump in any nest.

- The future generation of cuckoo is determined based on the nest which complete its hatching process.
- The nests are fixed and the host nest can demolish or abandon its nest when it finds the cuckoo egg. Its probability P_a and its values lies between 0 and 1.

Based on this, the Pseudocode for the CSO is given below.

start
objective function
initialization parameters like host nest, iterations boundary and number of cuckoo
while (iter < max(iterations))
select randomly the cuckoo and its solution using Levy flight equation 6
evaluate objective function
choose nest randomly
if (current fitness > previous fitness)
Replace the previous solution with current solution
End if
Remove worse nest based on P_a and find new solution using equation 6
Find local best solutions
Rank them and the top best solution is the global best solution.
End while
Output= dominant attribute and corresponding
Fitness function value
End

The new solution for a cuckoo is determined with the help of levy flight and it is given in equation 6.

$X_i^{t+1} = X_i^t + \alpha \oplus Levy\ (\lambda)$	(6)
$\alpha\ is\ step\ size, \alpha > 0$	(7)
$Levy = t^{-\lambda}, 0 < \lambda < 3$	(8)

Levy flight is used to provide random walk for the cuckoo and it follows power law step length distribution.

Based on the above Pseudocode, the optimization will determine the best attribute for the classification and it will be used for the training and testing.

4.5 Classification using feed forward neural network:

The role of feed forward neural network is to predict the diabetic based on the dominant attribute from the cuckoo search optimization. Here, the network will be trained with training data and tested with testing data.

The training and testing data is obtained by dividing the data into two sets using hold out approach with 0.3%. It divides the entire data as training with 70% of data and testing with the remaining 30 % of data.

The data is formed from the dominant attributes from the cuckoo search optimization and the corresponding labels are the targets for the network. The dominant attributes are Age, diabetic predictive value and insulin. The inputs for the neural network is given in equation 9.

$D_{neural} = [predictors\ responses]$	(9)
$predictors = dominant\ attributes = \{age.diabetic\ predictive\ value\ and\ Insulin\}$	(10)
$responses = class\ labels\ of\ data$	(11)

The equation 9 to 11 is common for both the training and testing data. Then, the network trained with the help of algorithm used in [30]. The network diagram for the feed forward neural network is shown in the figure 2.

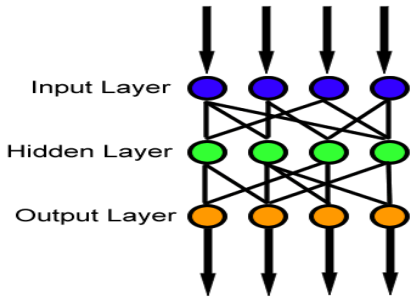

Figure 2. Layers in Feed forward neural network

Here, the input layers are predictors and the output layers are the responses. The hidden layers is to define the relation between the input and output. The hidden layers process the output with activation function to estimate the output. In this, ten hidden layers are used. The data travel in only one direction from input to output.

Based on this, the data will be trained and tested with the network and it is evaluated with the help of evaluation metrics in 4.6

4.6 Performance evaluation:

The proposed method optimized feed forward neural network is evaluated with the following metrics:

- Accuracy

- Precision
- Recall
- F-measure
- Matthew correlation coefficient

The above evaluation metrics and its formula for calculating them is given in the table format in table 4.

Table 4. Formula for evaluation metrics

Metric	**Formula**	**Eqn No**
Accuracy	$\frac{no\ of\ correctly\ identified\ classes}{total\ number\ of\ instances}$	(12)
Precision	$\frac{identified\ diabetic\ class}{actual\ diabetic\ class}$	(13)
Recall	$\frac{identified\ diabetic\ class}{actual\ diabetic\ class + wrongly\ identified\ diabetic\ class}$	(14)
F-measure	$2 * \frac{precision * Recall}{Precision + recall}$	(15)
Matthew correlation coefficient	$\frac{\frac{TP}{N} - S * P}{\sqrt{PS(1-S)(1-P)}}$	(16)

The terms used in the Matthew correlation coefficient is given in below equations

$N = total\ number\ of\ instances$	(17)
$TP = identified\ diabetic\ class$	(18)
$TN = identified\ non - diabetic\ class$	(19)
$FP = wrongly\ identified\ diabetic\ class$	(20)

$FN = wrongly\ identified\ non\ diabetic$	(21)
$S = \frac{TP + FN}{N}$	(22)
$P = \frac{TP + FP}{N}$	(23)

5. Implementation and discussion:

In this, the proposed method is implemented with the help of MATRIX Laboratory software R2018a version.

The input data for the processing which comprises of eight attributes

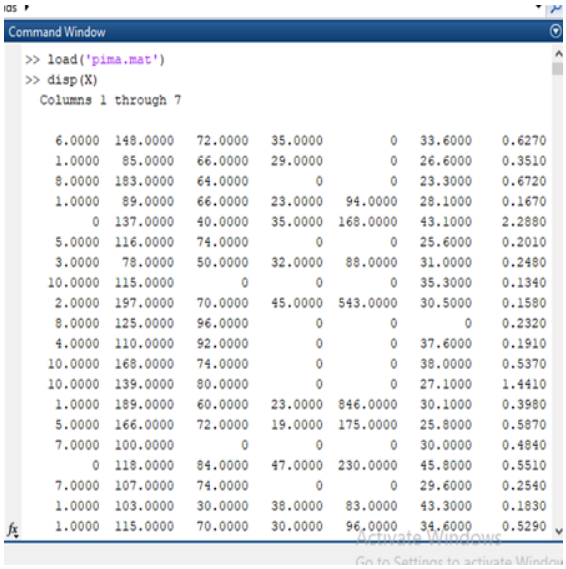

Figure 3. First seven columns of dataset

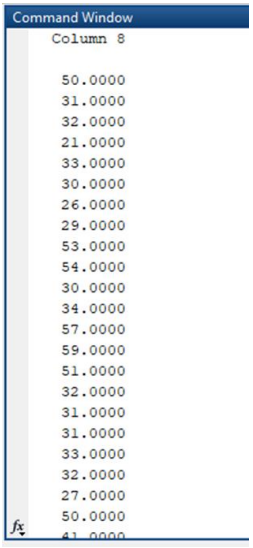

Figure 4. Eight attribute of dataset

The corresponding labels namely diabetic as 1 and non-diabetic as 0 for the above inputs is shown in figure 5.

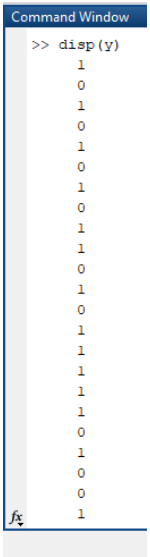

Figure 5. Classes for the dataset

In this, there is no missing records, but it has zero values in it. But, it is taken as the original dataset and then processed with K-means clustering for the outlier removal it is shown in the below figure 6.

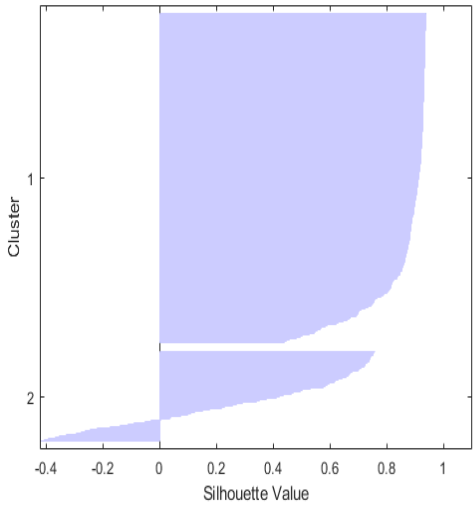

Figure 6. Outlier removal using K-means

Figure 6 shows that the values in the negative direction is the outlier data and it is removed from the dataset based on its centre values and the clean dataset is processed for the attribute selection process.

After the outlier removal process, the attributes of the records and its classes are passed to the optimization process to select the dominant attribute for the classification based on the fitness function. The corresponding convergence curve for the Cuckoo search optimization to reach maximization of fitness function is shown in the figure 7.

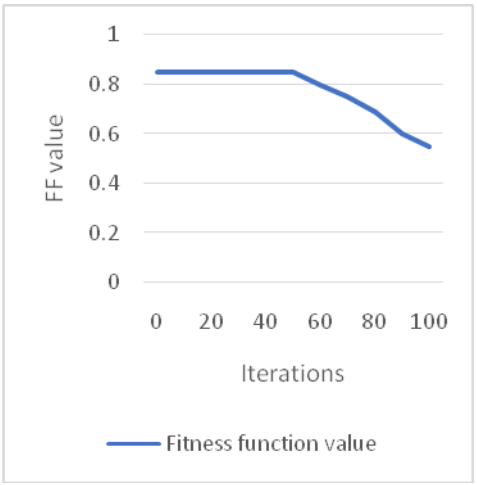

Figure 7. Convergence Curve for CSO

In the above figure 7, the curve is in decreasing manner. It is due to the minimization process is used in the optimization process. From the optimization, the three dominant attributes age, diabetic predictive and insulin level were used for the classification. Among these three, the two attributes has non-zero elements in it.

Then, the three attributes from the Cuckoo search optimization is used for the classification process using feed forward neural network. The training of the network is shown in figure 8

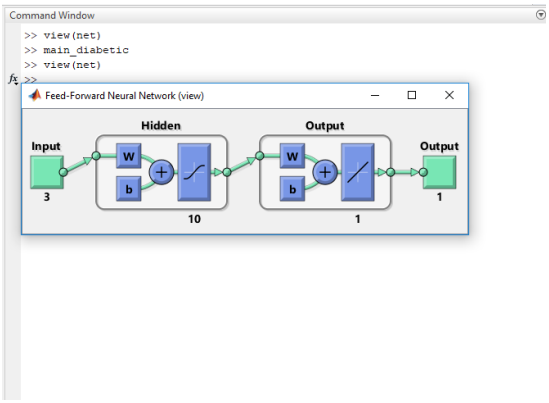

Figure 8. Training process of FFNN

From the figure 8, it is observed that the classifier uses only three attributes for the classification process. Based on the above training, the network is able to classify the tested dataset effectively and it is evaluated with the help of evaluation metrics and its comparison with the existing technique is shown in the below table 5.

Table 5. Performance comparison of CSO-FFNN versus K-means-LR

Method/ Metric	**Proposed CSO-FFNN**	**Existing K-means-LR**
Accuracy	92.3	90.7
Precision	93.4	91.6
Recall	97.2	96.4
F-measure	97.1	95.4
Matthew correlation	79.8	75.2

The diagrammatic comparison of the proposed and existing technique is shown in figure 9 using table 5 values.

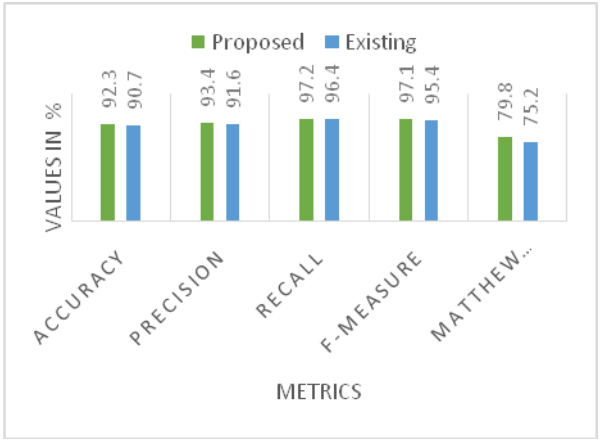

Figure 9. Comparison of results

Table 5 and figure 9 shows that the proposed method is able to achieve higher accuracy with minimal attributes as compared to the existing system. Hence, the proposed system is better as compared to existing in terms of performance as well as computing time.

6. Conclusion:

Diabetes can be turn into life threaten diseases, if it is not treated at an early stage. Especially, in the women, the chances of diabetes is higher as compared to men due to the hormonal changes during pregnancies. Due to this, they suffer a long term diabetes as well as other diseases due to their tensions and regular life chores. This can be prevented if the diagnosis is determined at an early stage. Mostly, the trained doctors are required to confirm the diabetes. It requires manual work and complete knowledge in it. This problem is avoided by several research works using machine learning algorithm for the classification. Those algorithms process effectively for smaller dataset with smaller number of attributes. Hence, to overcome this shortcomings, in this an optimization based classifier is proposed to process on larger datasets like BIG DATA. The proposed optimization helps to select an optimal attribute which is enough to predict diabetic or non-diabetic using feed forward neural network classifier. This approach outperforms all other existing technique with F-measure of 96.7%. Another advantage of the proposed approach is it able to convert into Chabot model for diabetic prediction.

7. Future work:

In future, the proposed method can be improved by using different optimization and classifier to reduce the training time further and improve the performance.

8. References:

1. Cichosz, S. L., Johansen, M. D., &Hejlesen, O. (2016). Toward big data analytics: review of predictive models in management of diabetes and its complications. Journal of diabetes science and technology, 10(1), 27-34.
2. Rallapalli, S., &Suryakanthi, T. (2016, November). Predicting the risk of diabetes in big data electronic health Records by using scalable random forest classification algorithm. In 2016 International Conference on Advances in Computing and Communication Engineering (ICACCE) (pp. 281-284). IEEE.
3. Muni Kumar, N. (2016). Survey on map reduce based apriori algorithms in medical field for the prediction of diabetes mellitus. RESEARCH JOURNAL OF FISHERIES AND HYDROBIOLOGY, 11(4), 13-18.
4. Shetty, S. P., & Joshi, S. (2016). A tool for diabetes prediction and monitoring using data mining technique. International Journal of Information Technology and Computer Science (IJITCS), 8(11), 26-32.
5. Mishra, S., Chaudhury, P., Mishra, B. K., &Tripathy, H. K. (2016, March). An implementation of Feature ranking using Machine learning techniques for Diabetes disease prediction. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (pp. 1-3).
6. Joshi, R., &Alehegn, M. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. International Research Journal of Engineering and Technology, 4(10), 426-435.
7. Komi, M., Li, J., Zhai, Y., & Zhang, X. (2017, June). Application of data mining methods in diabetes prediction. In 2017 2nd International Conference on Image, Vision and Computing (ICIVC) (pp. 1006-1010). IEEE.
8. Kumar, P. S., &Pranavi, S. (2017, December). Performance analysis of machine learning algorithms on diabetes dataset using big data analytics. In 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions)(ICTUS) (pp. 508-513). IEEE.

9. Jayanthi, N., Babu, B. V., & Rao, N. S. (2017). Survey on clinical prediction models for diabetes prediction. Journal of Big Data, 4(1), 26.
10. Chen, W., Chen, S., Zhang, H., & Wu, T. (2017, November). A hybrid prediction model for type 2 diabetes using K-means and decision tree. In 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS) (pp. 386-390). IEEE.
11. Sisodia, D., &Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585.
12. Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018, September). Prediction of diabetes using machine learning algorithms in healthcare. In 2018 24th International Conference on Automation and Computing (ICAC) (pp. 1-6). IEEE.
13. Mir, A., &Dhage, S. N. (2018, August). Diabetes disease prediction using machine learning on big data of healthcare. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-6). IEEE.
14. Liu, B., Li, Y., Sun, Z., Ghosh, S., & Ng, K. (2018, February). Early Prediction of Diabetes Complications from Electronic Health Records: A Multi-Task Survival Analysis Approach. In AAAI (pp. 101-108).
15. Chen, M., Yang, J., Zhou, J., Hao, Y., Zhang, J., &Youn, C. H. (2018). 5G-smart diabetes: Toward personalized diabetes diagnosis with healthcare big data clouds. IEEE Communications Magazine, 56(4), 16-23.
16. Saru, S., &Subashree, S. (2019). Analysis and prediction of diabetes using machine learning. International Journal of Emerging Technology and Innovative Engineering, 5(4).
17. Parast, L., Mathews, M., & Friedberg, M. W. (2019). Dynamic risk prediction for diabetes using biomarker change measurements. BMC medical research methodology, 19(1), 175.
18. Bai, B. M., Nalini, B. M., &Majumdar, J. (2019). Analysis and detection of diabetes using data mining techniques—a big data application in health care. In Emerging Research in Computing, Information, Communication and Applications (pp. 443-455). Springer, Singapore.
19. Makino, M., Yoshimoto, R., Ono, M., Itoko, T., Katsuki, T., Koseki, A., ...&Saitoh, E. (2019). Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. Scientific reports, 9(1), 1-9.
20. Fiarni, C., Sipayung, E. M., &Maemunah, S. (2019). Analysis and prediction of diabetes complication disease using data mining algorithm. Procedia Computer Science, 161, 449-457.
21. Mosquera-Lopez, C., Dodier, R., Tyler, N. S., Wilson, L. M., El Youssef, J., Castle, J. R., & Jacobs, P. G. (2020). Predicting and preventing nocturnal hypoglycemia in type 1 diabetes using big data analytics and decision theoretic analysis. Diabetes Technology & Therapeutics.
22. Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Abedin, M. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems, 8(1), 7.
23. Subramaniyan, S., Regan, R., Perumal, T., & Venkatachalam, K. (2020). Semi-Supervised Machine Learning Algorithm for Predicting Diabetes Using Big Data Analytics. In Business Intelligence for Enterprise Internet of Things (pp. 139-149). Springer, Cham.
24. Prasad, K. S., Reddy, N. C. S., &Puneeth, B. N. (2020). A Framework for Diagnosing Kidney Disease in Diabetes Patients Using Classification Algorithms. SN Computer Science, 1(2), 1-6.

25. Sivaparthipan, C. B., Karthikeyan, N., &Karthik, S. (2020). Designing statistical assessment healthcare information system for diabetics analysis using big data. Multimedia Tools and Applications, 79(13), 8431-8444.
26. Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10, 100-107.
27. https://www.dropbox.com/s/mvlwu7p0nyk2a2r/pima.mat?dl=0
28. Lei, D., Zhu, Q., Chen, J., Lin, H., & Yang, P. (2012). Automatic k-means clustering algorithm for outlier detection. In Information engineering and applications (pp. 363-372). Springer, London.
29. Gandomi, A. H., Yang, X. S., &Alavi, A. H. (2013). Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems. Engineering with computers, 29(1), 17-35.
30. Bebis, G., &Georgiopoulos, M. (1994). Feed-forward neural networks. IEEE Potentials, 13(4), 27-31.