

Optimized Regression Neural Network For Classification Of Diabetes In Big Data Environment

N.V.Poornima¹, Dr.B.Srinivasan², Dr.P.Prabhusundhar³

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

ABSTRACT

One of the main problems in the human body is diabetes. Diabetes is one of the critical health issues in the human body it causes human life severely. In women, the diabetes rate is more and higher as compared with the others. It severely affects pregnant ladies while they are affected by diabetes. This can be affected the daily routine. The one and only method are to take dialysis at the beginning stage. The famous doctors check the patients and provide the precautions and the medicines for that. They handle the patients affected by diabetes. To solve this problem several methods and techniques are used in the machine learning algorithm under the big data environment. This algorithm provides good prediction results by taking the data from the dataset. In this diabetes prediction, the machine learning algorithms combine with the big data environment. In the previous methods, the k-means clustering and cuckoo search is used in optimization. In this existing system, its accuracy and performance are high but it takes a huge time to computational process. To solve this problem optimization based on machine learning is applied. Here the dataset is downloaded from the Indian diabetes dataset. For removing the unwanted and missing data in the dataset we use pre-processing and clustering. And the feature reduction is done in the use of glow-worm optimization. The glow-worm is the fastest method compared to the cuckoo search optimization. The classification of diabetes is done with the help of feed-forward neural networks. The aim of the classification is that the user misunderstanding the values after removing the unwanted data. The whole process is realized in the MATLAB R 2018a environment and evaluated in terms of accuracy, precision, recall, F-measure, and Matthew correlation coefficient. This approach outperforms all other existing techniques with an F-measure of 98.6%.

Keyword: Diabetic, dataset, pre-processing, characteristics, reduction using glow-worm, Feedforward neural network.

INTRODUCTION

As we, all know nowadays healthcare industries are producing huge amount of data regarding their patient's day by day. It is difficult to store this amount of data due to its structured and unstructured format. Now this

way of storing data can be used for the big data for classification and producing better output. This paper [1] shows the impacts of the diabetes.

In this they have used the simpler structure of GRN is Close classification Accuracy to get the high result. In this, they have used the multilayer perceptron (MLP) and radial basis function (RBF) for the classification and comparison of the disease using the specific manner to get the better result in the neural structure. Backpropagation training algorithms [2] have tested them. Therefore, this system increases its prediction accuracy for the diabetes.

As we all know diabetes is a spectacular disease caused for all age groups, it sometimes leads to heart related diseases and make the patient into an unconditional manner. This reason diabetic is made as the one of the killing diseases in the entire world. So, our researchers have used the technique of R programming correlated with Decision tree algorithm for the development and prediction of diabetes in the early manner [3].

This data prediction helps the modern doctors to predict their patient's disease in the early stage itself. This data mining techniques help the doctors to predict the disease in the high accurate and efficient manner in this paper they have used many recent algorithms for the early prediction of disease [4].

They have told that this cause of diabetes 2 is due to the improper way of insulin and unordered manner. In this they have, mention type 2 is the main reason for the increase of diabetic's patients. For this, they have used k-clustering algorithm for the classification and the process of data prediction. In this classification, they have used Pima Indian Diabetes dataset. They have rewarded that this classification should produce the high accuracy and less computational time. They obtained a classification accuracy of 86% with 83% sensitivity and 87% specificity. In addition, The Area under the Receiver Operating Characteristic Curve (ROC) is 87%. [5].

They have used the diabetes for the prediction of disease for the better result in diabetes. They have used the support vector regression model, which is the newly introduced method for which it helps to predict the value of diabetes at a different time like before sleeping and overnight glucose level. Diabetes can be caused by the high and low level of glucose in the body. So, in this system it helps to predict the glucose level of patient in the high and low manner. According to the 4 weeks value they have calculated the accuracy of diabetes and this help them to compare with the normal diabetic's level and particular glucose level. [6].

They have come with a system [7] to predict the disease of kidney failure using the modern technique of Artificial Intelligence. They have used the AI with machine learning algorithms and other new technologies to

predict the data with high accuracy and effectively. They have used Neural Network for the better time series. They have also used logistic regression and neural network for the better time series and result.

As we all know now our information system [8] has collecting the huge data and they are stored them in the better way. The data stored are used for the future researchers. And the accessing of health data is getting in various forms. So this helps to involve the data mining for the predicting and evaluating the disease in the system. Cloud helps to store the data and retrieve the data in efficient and effective manner. Cloud computing will help to get predictive results at a larger speed. This study shows that the diabetes prediction and future models with the big data techniques.

[10] Has explained the new methods in diabetic prediction and deliver this method to the users. The symptoms of diabetes include failures in nerve, heart, kidney, and low eye power. It is difficult to store this amount of data due to its structured and unstructured format. Now this way of storing data can be used for the big data for classification and producing better output. According to India's di-centric report, diabetes is one of the important and rising disease in the world. For analysing a new database and extract, the new information from the database is considered a difficult problem. This process is called mining of data and it has several techniques to be used for getting more accuracy in the diabetes prediction

LITERATURE REVIEW

[13] As we all know nowadays healthcare industries are producing huge amount of data regarding their patient's day by day. It is difficult to store this amount of data due to its structured and unstructured format. Now this way of storing data can be used for the Bigdata for classification and producing better output. According to India's dicentric report diabetes is one of the important and rising diseases in the world. So it is very difficult to analyse and found the disease in the earlier manner so, it leads to a long-term disease. In this paper, Researchers have used the modern technique of Hadoop/ MapReduce algorithms to find and predict the disease in the early stage. This help to give affordable and available disease prediction to the patients.

In this paper [14], they have mentioned the bigdata algorithms for the classification and analyse of disease. In this they have collected vast variety of dataset regarding the patients. They are difficult to collect, keep, format, evacuate, remedial, correlate, analyse and visualize using traditional methods and tools. So, they have used the modern technique of R for the data Analyzation and classification to predict the diabetic disease in early stage. As we all know diabetics is a spectaculars disease caused for all age groups, it sometimes leads to heart related diseases and make the patient into an unconditional manner. This reason diabetic is made as the one of the killing diseases in the entire world. So, our researchers have used the technique of R

programming correlated with Decision tree algorithm for the development and prediction of diabetics in the early manner.

The currently evolved technique [15] of neural network is used for the prediction of diabetics. In this paper they have used the GRN technique for the prediction of disease using the dataset. In this they have used the simpler structure of GRN is Close classification Accuracy to get the high result. In this they have used the multilayer perceptron (MLP) and radial basis function (RBF) for the classification and comparison of the disease using the specific manner to get the better result in the neural structure. They have been tested by backpropagation training algorithms.

In this paper [16] the authors have been explained how they have to predict the diabetics 2. They have told that this cause of diabetics 2 is due to the improper way of insulin and unordered manner. In this they have mention type 2 is the main reason for the increase of diabetic's patients. For this they have used k-clustering algorithm for the classification and the process of data prediction. In this classification they have used Pima Indian Diabetics dataset. They have rewarded that this classification should produce the high accuracy and less computational time. They obtained a classification accuracy of 86% with 83% sensitivity and 87% specificity. Also, The Area under the Receiver Operating Characteristic Curve (ROC) is 87%.

Now adays the disease diagnosis and prediction is one of the difficult and confused system. As we all know current emerging diseases diabetics is an extraordinary disease that couldn't be easy to predict it easily and effectively. So prediction of this disease can be made by currently used technique for the classification [17]. This way of efficient prediction can be made by this paper. For this they have used the two stages of classification they are: (a) Linear Discriminant Analysis (LDA) based genetic algorithm for feature selection, (b) Generalized Regression Neural Network (GRNN) for classification. From the first method they have reduce the cost of prediction but take more computational time and produce high accuracy result. The objective of this method can be evaluated by accuracy, confusion matrix and Receiver-Operating Characteristic (ROC). They have also evaluated the method with other data methods This method produces the accuracy of 80.2017% than a ROC of 0.875.

[18] They have expanded that data prediction has taken a huge role in the prediction of disease. For the prediction of disease in the data analysis data mining has taken a vital role in the process of classification and evaluation. This data prediction helps the modern doctors to predict their patient's disease in the early stage itself. This data mining techniques help the doctors to predict the disease in the high accurate and efficient manner in this paper they have used many recent algorithms for the early prediction of disease.

They have come with a system [19] to predict the disease of kidney failure using the modern technique of Artificial Intelligence. They have used the AI with machine learning algorithms and other new technologies to predict the data with high accuracy and effectively. They have used Neural Network for the better time series. They have also used logistic regression and neural network for the better time series and result.

This paper [20] explains the prediction model with three different metrics. The database consists of the patient's name, date of birth, blood group, blood pressure. The three algorithms are called naïve, decision tree, clustering algorithms. They have classified and find the differences between the disease and predict the disease evaluation methods for the classification of data and in which algorithm should produce the high accuracy result.

In this [21] they have used the diabetics for the prediction of disease for the better result in diabetics. They have used the support vector regression model which is the newly introduced method for which it helps to predict the value of diabetes at a different time like before sleeping and overnight glucose level. Diabetics can be caused by the high and low level of glucose in the body. So, in this system it helps to predict the glucose level of patient in the high and low manner. According to the 4 weeks value they have calculated the accuracy of diabetics and this help them to compare with the normal diabetic's level and particular glucose level.

They have used logistic regression [22] for the finding of Risk Factor in the diabetics. They have used four classification algorithms naïve, decision tree, AdaBoost, and random forest for finding the diabetics. In this system they have used certain rules for the prediction of disease and getting the better result. The k-protocol algorithm gives 70% accuracy in this system.

As we all know now our information system [23] has collecting the huge data and they are stored them in the better way. The data stored are used for the future researchers. And also, the accessing of health data is getting in various forms. Therefore, this helps to involve the data mining for the predicting and evaluating the disease in the system. Cloud helps to store the data and retrieve the data in efficient and effective manner. Cloud computing will help to get predictive results at a larger speed. This study shows that the diabetes prediction and future models with the big data techniques

[24] Has explained the new methods in diabetic prediction and deliver this method to the users. The symptoms of diabetes include failures in nerve, heart, kidney, and also low eye power. It is difficult to store this amount of data due to its structured and unstructured format. Now this way of storing data can be used for the Bigdata for classification and producing better output. According to India's dicentric report diabetes is one of the important and rising diseases in the world. For analysing a new database and extract the new

information from the database is considered a difficult problem. This process is called mining of data and it has several techniques to be used for getting more accuracy in the diabetes prediction and also extract the different data from the database. This study cooperates with the medical science field. For the mining tool, the special software is developed and used for the analysis. PIDD dataset is used for taking the diabetic data this was getting from UGC used in prediction methods. This is the most efficient dataset which can be used to predict the diabetic values. Increasing the accuracy of this system will be combined with the bootstrap technique and comparing the values and performance obtained from this system is used decision trees.

[25] Karthikeyan et.al in big data several categories used the machine learning concepts to perform and operate good prediction results. They have mentioned the bigdata algorithms for the classification and analyse of disease. In this they have collected vast variety of dataset regarding the patients. They are difficult to collect, keep, format, evacuate, remedial, correlate, analyse and visualize using traditional methods and tools. Guessing the values in a health area is a difficult task so the concept of big data helps the medical authorities for predicting the results. In this, the big data is cooperated with the machine learning algorithms and uses the six different types of processing. The dataset consists of all about the patient's details and this can be accessed with the help of the different machine learning algorithms. Comparing the results obtained from these algorithms provides the overall prediction of accurate results and those results also sensitive. For the prediction of disease in the data analysis data mining has taken a vital role in the process of classification and evaluation. This data prediction helps the modern doctors to predict their patient's disease in the early stage itself. This data mining techniques help the doctors to predict the disease in the high accurate and efficient manner in this paper they have used many recent algorithms for the early prediction of disease. Based on the result to be taken from these algorithms it will evaluate which algorithm is most suited for among the six algorithms. From this paper, the hospital authorities can easily predict the diabetics as soon.

EXISTING METHOD

At the earliest method, there is similarity in the dataset but after taking the dataset there are several steps makes a different. In the existing system. In the existing system pre-processing, take same steps used in the proposed system. But for the purpose of the optimization process it uses cuckoo search and after that, it uses glow-worm optimization .These two processes provides an accuracy of 95.5% and 96.3% accuracy. And classification used for the comparison of two methods is used the feed forward neural network. But this system is not providing high accuracy and also the performance. In proposed system regression neural network used in the classification.

PROPOSED SYSTEM

The proposed system explains the several features more than in the existing system. Here the block diagram explains the functionalities and operations in the proposed system.

The proposed system is very different from the existing model by making the tiny changes. Here initial step is that to access the diabetes data from the databases. Here they uses the same dataset used in the previous process. Diabetes data is taken from PIMA dataset. After receiving the data, the data enters into the pre-processing by removing the useless data has and removes the unused spaces. The pre-processing finished with the help of clustering methods. With the use of these methods, the data removes all the unwanted datas. After getting the pure diabetes data it goes into the feature extraction and feature selection methods. Dolphin glow-worm method help to optimize the data from the data bases. Classification and the comparison between the existing systems is done by a neural network called regression neural networks. From getting these metrics from the optimization and the feature, extraction process the evaluation completes with several metrics.

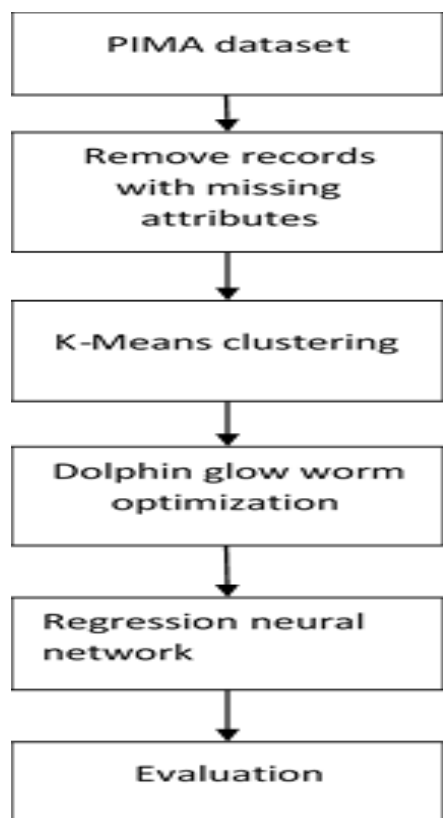


Fig 1: block diagram of proposed model

Dataset

PIMA dataset is used in the prediction of the diabetes. In this dataset is based on the Indian people lifestyle contains several parameters. This parametric can help the patient to identify their disease before fully affected. It specially designed for women's especially for above 21-year-old women's. Initially the dataset is downloaded from [8]. And there are around eight parameters in the diabetes database. It separated into two parts as trained and verified part. Most of them are in the trained part. In addition, it operates before train and tests the values in the dataset. To evaluate the accuracy to be used in the dataset we are used the learning data.

Several parameters like temperature, blood pressure, DBP, DPF, weight, insulin, glucose etc. In addition, the dataset contains two different type of data is diabetic and the non-diabetic data. The below table (1) shows the values in the dataset.

Parameters	Description	Value range
Temperature	Heat of the body	97°F-99°F
DBP	Contains the blood pressure	0-122
DPF	It is attractive attribute used in the dataset.	0.0675–2.53
Weight	Overall weight of the patient	0-65
Insulin level	Check insulin level at every two hours	0-846

Glucose level	Measures the plasma and oral glucose level.	4.0 to 5.4 mmol/L
---------------	---	-------------------

Table 1: dataset attributes

Removal of unwanted data

The data taken from the dataset not without errors. Some data in the dataset is not true. So the user cannot access and use all the data from the dataset. Here, we want to clarify the data whether it is correct or not. For assuring the quality of data, the data will enters and separates in to testing as well as training. Sometimes the datasets contains free spaces. This also an issue faced by the user. Because the free spacing makes confusion while taking data. For removing these unwanted things in the dataset, the pre-processing techniques are used.

Feature extraction using dolphin glow-worm

The DGO is an Updated version of GSO Algorithm. The GSO helps to solve multimodal optimization problems by finding multiple optima. In glow-worm optimization, the light emitted from it is proportional to the luciferin. It will interact with the neighbouring agent so it increases the luciferin level.

In this the initial stage of this algorithm is to initialization of population Z.

$$Z = \{Z_1, Z_2, \dots, Z_m\}; 1 \leq n \leq m$$

Where n is the size of the population and Z_m represents m^{th} solution in the population. Followed by population initialization, initial quantity of luciferin D_0 and neighbourhood range T_0 are also initialized. The luciferin update is based on the luciferin level at previous time instant and the fitness of the solution,

$$D_{c+1}^n = (1 - \rho) D_c^n + \gamma G(Y_{t+1}^c)$$

where ρ is the luciferin decay constant with value in the interval $[0,1]$, D_c^n is the luciferin level of n^{th} agent at the time c , γ is the luciferin enhancement constant, and $G(Y_{t+1}^c)$ is the fitness of n^{th} agent at the time $c+1$.

In the movement phase, the agent looks for higher luciferin and move to particular place. It is mainly due to the probability of the value. It can be found by

$$Y_c = \frac{M_c^z - M_t^z}{\dots}$$

$$\sum_{e \in O_c^n} M_t^e - M_c^n$$

Where O_t^q is the neighbour of z^{th} Glowworm at the time t , defined as where E^{zz} is the Euclidean distance between the search agents z and y at the time c and S_c^n is the neighbourhood range of n^{th} agent. Thus, based on the probability Y_T , a glow-worm r selects z^{th} Glowworm in the neighbourhood N_t^n . The movement of the agent is then represented as

$$Z_{c+1}^n = Y_c^n + \delta (Z_c^z - Z_c^n / |Z_c^z - Z_c^n|)$$

It calculates the difference between the previous and the current location at different times. And the new location is updated in the locations.

Comparison using Regression Neural Network

Regression Neural Network predicts an output variable as a function of the inputs. The input variable values may be independent variables. In this model we can perform the actions by:

$$G_1(n_1, n_2) = X_{11}n_1 + X_{12}n_2$$

Where X_{11} and X_{12} are just weights/coefficients. Before going any further, we should notice that, here, there is no constant term in the model. However, we could have introduced such term by setting

$$G_1(n_1, n_2) = X_{11} * n_1 + x_{12} * n_2 + d.$$

It can be represented as below:

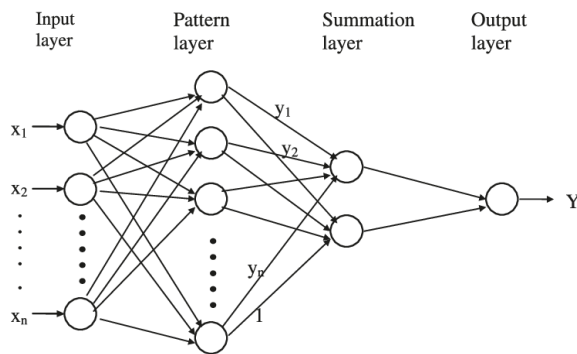


Fig 2: Regression neural network

This model cannot able to classify and find the output using the system in an easier manner so we have to use the non-simpler way for the classification. We can encourage and produce the output by the following method. First, we could consider that the quantity

$$\mathbf{b}(\mathbf{x}_{11} * \mathbf{n}_1 + \mathbf{x}_{12} * \mathbf{n}_2)$$

It is no longer give a final output but it produces an intermediate output, called **m1**

$$F_3(n_1, n_2) = b_2(w_1l_1 + w_2l_2 + w_3l_3)$$

$$= b_2(w_1 * b_{11}(x_{11}n_1 + x_{12}n_2) + w_2 * b_{12}(x_{21}n_1 + x_{22}n_2) + u_3 * b_{13}(x_{31}n_1 + x_{32}n_2))$$

Where we should mainly keep in mind that's are non-linear activation functions and **x**'s and **u**'s are weights. In the next figure, we give the network graphical representation of this model in the same way as we did for the previous ones.

PERFORMANCE EVALUATION

The performance evaluation of regression neural network can be obtained from the following metrics such as Precision, Recall, F-measure, Matthew correlation coefficient, Accuracy. The below table (1) shows that the formulas for calculating these metrics.

Metrics	Formula	Eq no
Precision	$\frac{\text{total selected diabetes}}{\text{actual diabetic class}}$	(1)
Recall	$\frac{\text{selected diabetes}}{\text{total diabetes + misidentify}}$	(2)
F-measure	$2 * \frac{\text{precision} * \text{Recall}}{\text{Precision} + \text{recall}}$	(3)
Mathew correlation coefficient	$\frac{\frac{TP}{n} - S * P}{\sqrt{PS(1 - S)(1 - P)}}$	(4)
Accuracy	$\frac{\text{no of perfect classes}}{\text{total number of classes}}$	(5)

Table 2: formulas used in evaluation

IMPLEMENTATION AND RESULTS

For the implementation of this system, we have used MATRIX Laboratory Software R2018a version. MATLAB is originated from the word “MATRIX Laboratory”. It is developed by MathWorks. It is a computing environment in multi paradigm programming languages. It helps to perform matrix manipulation, plotting of functions, and data implementation. It also helps to make interface between other languages.

For our system it takes 8 variable dataset and perform the classification and pre-processing of data.

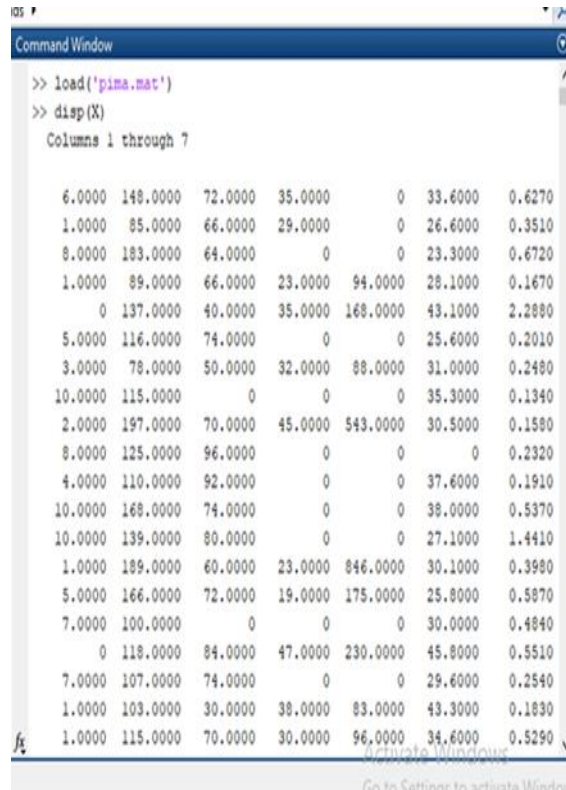


Fig 3. First seven columns of the dataset

It has explained the first seven columns of the dataset. It helps to make an efficient metric evaluation manner.

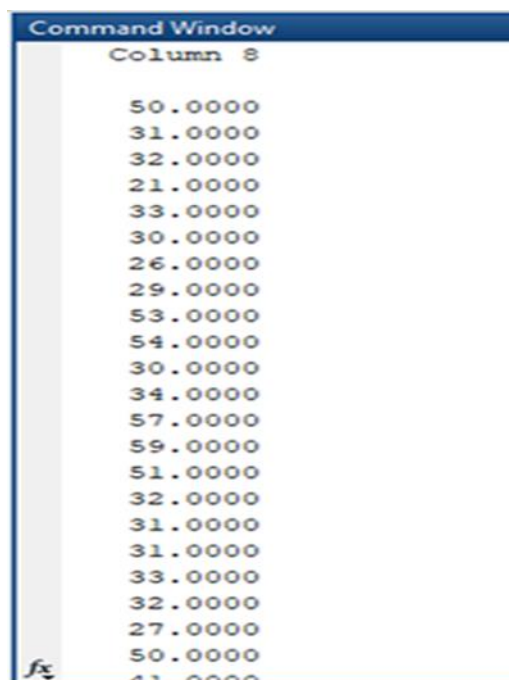


Fig 4. Eight attributes of the dataset

In the pre-processing time the dataset has been made into 1 for diabetic and 0 for non- diabetic. It can be easily understood from the below diagram.

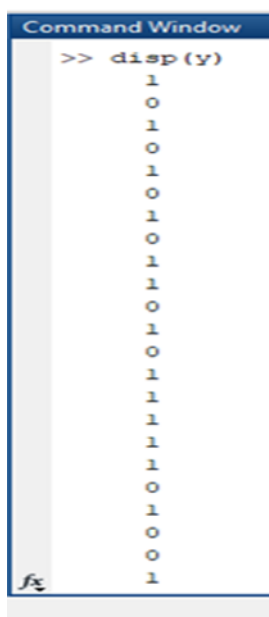


Fig 5. Classes for the dataset

In this, we have cleared all the unwanted and missing records there are no missing records, but it has zero values in it. It has taken it as original dataset and then it has to be pre-processed using the K-means clustering method. This processing is called data pre-processing. This way of Pre-processing helps to give an accurate output. We can understand it from figure 6.

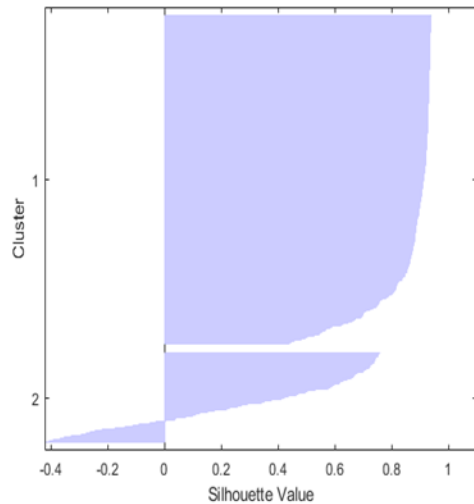


Fig 6. Outlier removal using K-means

It shows that the values the outlier data that present in the negative area shows the accuracy of the data pre-processing. It helps our system to get an accurate and efficient result.

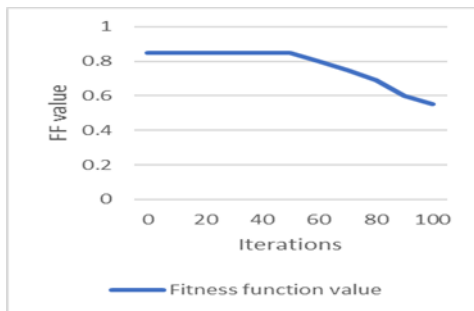


Fig 7. Convergence Curve for DWO

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.

Method / Metric	proposed DGO-RN	Existing GSO-FFNN	Existing CSO-FFNN	Existing K-means-LR
Accuracy	94.6	93.9	92.3	90.7
Precision	95.8	94.5	93.4	91.6
Recall	98.4	98	97.2	96.4
F-measure	98.6	97.9	97.1	95.4
Matthew correlation	81.5	80	79.8	75.2

Method/ Metric	Proposed GSO-FFNN	Existing CSO-FFNN	Existing K-means-LR
Accuracy	93.9	92.3	90.7
Precision	94.5	93.4	91.6
Recall	98	97.2	96.4
F-measure	97.9	97.1	95.4

Matthew correlation	80	79.8	75.2
---------------------	----	------	------

Table 3. Performance comparison of DWO-FFNN versus K-means-LR

Based on the training, the model has been evaluated and optimized using our DWO and Regression Network Model we have concluded with these factors and are classified and compared using the system.

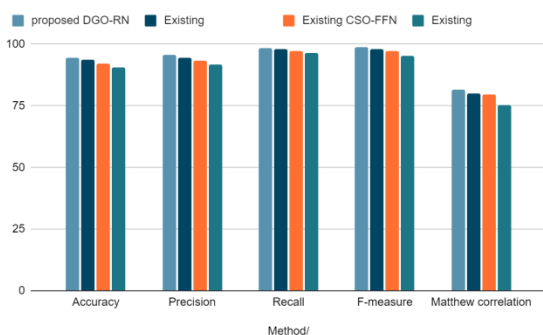


Fig 9. Comparison of results

From figure 9 and Table 5 we have concluded that our system should produce an accurate result to the entire system. It is better in computational time and performance.

7. CONCLUSION

This system avoids the rapid death due to the diabetics. It helps to get the patient about their disease in the early stage. For this problem, we have created a system on the machine learning algorithm for the prediction of diabetes. As part of this it first takes the patient dataset of PIMA about diabetics. Then it should be used for the machine learning algorithms and the clustering techniques are applied in this system for the pre-processing. And the optimization is done with the use of a dolphin glow worm. The optimization provides both for the diabetic as well as non-diabetics. From this optimization, we have concluded that our system should provide 97% of accuracy for the prediction. The classification under this prediction using the Regression neural networks. In the future to integrating more techniques, the system provides maximum accuracy and speed by implementing it in Artificial Intelligence.

References:

1. Cichosz, S. L., Johansen, M. D., &Hejlesen, O. (2016). Toward big data analytics: a review of predictive models in the 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 a 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 reduces based apriori algorithms in the 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). 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. Eswari, T., Sampath, P., & Lavanya, S. J. P. C. S. (2015). Predictive methodology for diabetic data analysis in big data. *Procedia Computer Science*, 50, 203-208.
14. Rumbold, J. M., O'Kane, M., Philip, N., & Pierscionek, B. K. (2020). Big Data and diabetes: the applications of Big Data for diabetes care now and in the future. *Diabetic Medicine*, 37(2), 187-193.
15. Sharmila, K., & Manickam, S. (2015). Efficient prediction and classification of diabetic patients from big data using R. *International Journal of Advanced Engineering Research and Science*, 2(9), 56-58.
16. Kayaer, K., & Yildirim, T. (2003, June). Medical diagnosis on Pima Indian diabetes using general regression neural networks. In *Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP)* (Vol. 181, p. 184).
17. Ndaba, M., Pillay, A. W., & Ezugwu, A. E. (2018, May). An improved generalized regression neural network for type II diabetes classification. In *International Conference on Computational Science and Its Applications* (pp. 659-671). Springer, Cham.
18. Bani-Hani, D., Patel, P., & Alshaikh, T. (2019). An optimized recursive general regression neural network oracle for the prediction and diagnosis of diabetes. *Global Journal of Computer Science and Technology*.
19. Jayashree, J., & Kumar, S. A. (2018). Linear Discriminant Analysis Based Genetic Algorithm with Generalized Regression Neural Network—A Hybrid Expert System for Diagnosis of Diabetes. *Programming and Computer Software*, 44(6), 417-427.
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/mvluw7p0nyk2a2r/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.
31. Karthikeyan, R., Alli, P. Feature Selection and Parameters Optimization of Support Vector Machines Based on Hybrid Glowworm Swarm Optimization for Classification of Diabetic Retinopathy. *J Med Syst* **42**, 195 (2018). <https://doi.org/10.1007/s10916-018-1055-x>