Grenze International Journal of Engineering and Technology, Jan Issue

Smart Diabetic Prediction: A Machine Learning Approach

Sangeetha C P[1], Dhanya S[2], Anjali S V[3] and Anna Sajan[4]
[1]KMEA Engineering College, Edathala, Cochin, Kerala
[2-4]Muthoot Institute of Technology and Science, Varikoli, Cochin, Kerala

***Abstract*—With the tremendous advancement in information and communication techniques, the health care system is being transformed into smart healthcare systems. The large volume of patient data, such as Electronic Health Records, has shown the importance of new technologies like machine learning and artificial intelligence in advanced disease prediction. Diabetes Mellitus is a worldwide health disorder which needs to be addressed seriously, especially for elderly society. This is because it can be the first phase of many other serious diseases like heart attack, kidney disorder, etc. In this article, a smart diabetic monitoring/prediction system is proposed with different machine learning algorithms. To be precise, the risk of diabetes can be predicted well in advance on the basis of the patient health data and genetic conditions. This is where the role of machine learning and artificial intelligence comes in predicting the diseases. Diabetes diagnosis is usually done with patient data, such as blood glucose and random blood glucose from the daily examination. To predict degenerative diseases, machine learning methods such as random forest and support vector machines are used. Accuracy is one of the essential variables that should be examined during the disease diagnosis. In this article, we use a machine learning algorithm that can more accurately diagnose diabetes than the current ones. Also the efficiency of different existing algorithms will be compared with the proposed methodology.**

***Index Terms*— Diabetic prediction, machine learning, neuro-fuzzy, accuracy.**

I. INTRODUCTION

Diabetes Mellitus is a chronic disorder which causes the blood glucose also called blood sugar to rise [1]. The human body in its normal functioning maintains the necessary level of the glucose by producing the protein called insulin which acts as a hypoglycemic agent. The insulin acts as the agent for moving the glucose around the tissue, thus decreasing the level of glucose in the blood. Diabetes may be caused by the absence of this agent or that the hypoglycemic agent is not properly used up in the human body. The disease is categorized into different types.

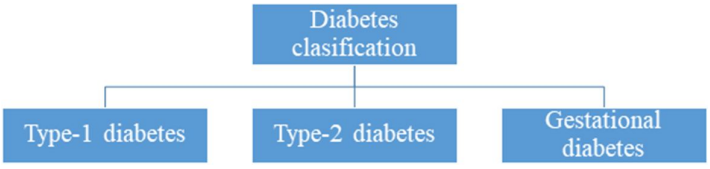

Figure 1 Classification of diabetes

Grenze ID: 01.GIJET.7.1.608_1

Type-1 diabetes is also known as the insulin-dependent polygenic diabetes-mellitus. This is caused due to the inefficiency of the pancreas to produce enough beta cells, which is the insulin producing cell. The type-1 is generally seen in children under the age of 20. If not taken care of, the disease can lead to many more problems like cardiovascular diseases, retinopathy, kidney failure, etc. There is also a hereditary factor associated with its occurrence [2]. Type-2 is a lifelong disease which is caused by the ineffective use of insulin in the human body. The body gains resistance to insulin. It is also known as non-insulin-dependent polygenic diabetes mellitus.
Obesity in children also causes type-2 diabetes. The other type of diabetes is caused in women during their gestation period and is called gestational diabetes. Gestational diabetes includes two classes, one which is controlled by the diet and exercise and the other which require one to take insulin and other medication. There is a 40% chance that the women who had gestational diabetes will have type-2 diabetes in the future [3].
Any disease will have a set of common symptoms that occur. These symptoms allow the doctors to predict whether they have the disease or not. In case of diabetes, the symptoms may be increased thirst and urination, increased hunger, feeling tired, blurred vision, numbness or tingling in the feet or hands, sores that do not heal, unexplained weight loss etc. there are other factors like hereditary factor, age etc. which also contribute to the disease. These factors are the dataset associated with diabetes.
With the emergence of artificial intelligence, machine learning has gained application in every field of study because it provides learning capabilities in the workplace. For the detection of diabetes, the factors which contribute to the occurrence of the disease and also the external symptoms are taken as the input to the model to predict whether a person has diabetes or not. The aim is to build a machine learning model which predicts the risk level of diabetes in patients with better accuracy. There are many algorithms which are used to build the same. Random forest, support vector machine, Naive Bayes to name a few. In this paper a neuro-fuzzy approach is used to predict the disease. The neuro-fuzzy approach combines together the merits of both neural networks and fuzzy inference systems.

II. Literature Survey

Priyanka Sonar and Prof. K. Jaya Malini discusses in their paper the different approaches to predicting diabetes using machine learning. On the basis of accuracy, different algorithms like naive Bayes, SVM, decision tree and ANN are compared and briefly explained. The dataset features used for the model are glucose level, blood pressure,Body Mass Index (BMI), Skin fold thickness in mm, Insulin value in 2 hours, Hereditary factor-Pedigree function, Age of patient in years. Out of 768 instances they collected about 75% is used to train the model and the rest 25%is used to test it.The machine learning matrices used to compare the different algorithms are precision, FI score and recall [1].
Diabetes can lead to many other dangerous diseases, depression being one of them. Raid M. Khalil, Adel Al-Jumaily aims in the paper to predict the depression associated with diabetes in patients using machine learning algorithms. The paper identifies and compares four different machine learning techniques to check which one gave the best result with the given dataset size. The study used both sociodemographic factors and clinical factors to build up the model. The paper briefly explains the four mentioned algorithms and comparing the accuracy , finds that the support vector machine has the highest accuracy of all [4].
Hospital charges for the regular admission to take tests and checking that the blood glucose is maintainable is costly for chronic diabetes persons. Juan Camilo Ram´ırez, David Herrera in their paper discusses how they can predict whether a readmission is required for the patient having diabetes within 30 days with the help of his/her medical records. Different algorithms like logical regression, support vector machines, random forest, neural networks are compared . The performance analysis was based on different matrices like FI score ROC AUC and they concluded that the random forest algorithm outperforms all others. The paper is concluded by giving a comparison result between the algorithms [2].
Chinmayi Chitnis, Roger Lee discusses the model which predicts the type -2 diabetes among people. The model uses 9 parameters each with 768 instances. The paper also discusses the different machine learning algorithms, including Artificial Neural Networks, Logistic Regression, K Nearest Neighbors, Decision Tree and Random Forest algorithm. According to the analysis, the neural network performs the best out of it. The paper aims at finding the simple, efficient most accurate algorithm by adjusting the training methods, hyper-parameter setting etc. the paper also includes a complete comparison of above mentioned algorithms based on the accuracy of its strengths and weaknesses [3].

Diabetes retinopathy is the eye disease caused due to diabetes, which damages the retina of the eye, eventually leading to complete blindness. The test for the disease include a visual acuity test, dilation of pupils, optical consistency tomography which cost a lot of time and is not good for the people. So S M Asiful Huda , Ishrat Jahan Ila , Shahrier Sarder, Md. Shamsuzzoha, Md. Nawab Yousuf Ali proposed to use machine learning algorithms to predict the chance of getting the diabetic retinopathy. The study was using Logistic Regression, Support Vector Machine, K- Nearest Neighbor and Decision Tree algorithms. 80% of the dataset was used to train the model and rest for testing. The data set was normalized and standardized, and the paper used many matrices accuracy, precision and recall to evaluate their study [5].

Machine learning applications in the medical field are very vast. Cancer is one of the deadliest diseases in the world. Different studies are going on to implement machine learning to predict cancerous cells with high precision. Kaan Uyar, Umit Ilhan, Ahmet Ilhan, Erkut Inan Iseri in their paper explains how to detect breast cancer cells using different approaches. They used the UCI repository dataset to train and test the theory. The different approaches used genetic algorithm bases recurrent fuzzy neural networks and adaptive neuro-fuzzy inference systems. For different variables different combinations of both the algorithms are designed. The study used 81 instances for training and 35 instances for testing. The comparison of different combinations of algorithms based on sensitivity, specificity, F-score, precision PME and accuracy [7].

Rahib H. Abiyev and Sanan Abizade discuss how Parkinson's disease can be predicted using a neuro-fuzzy approach. The training and testing dataset was used from the UCI repository. Dysphonia and dysarthria are recognized as the main symptoms for the disease. These symptoms were observed in almost 90% of the patients. The paper explains about fuzzy classification and learning through gradient descent methods. The paper concludes by giving the comparison of different algorithms based on accuracy [9].

The application of machine learning in the educational field is explained in the work done by Quang Hung Do and Jeng-Fung Chen. The study applies a neuro-fuzzy approach to accurately calculate the performance of students under different contexts. The data was taken from University of transport technology. Along with academics, socio-economic factors and other related factors are considered for implementing the model. The paper also compares the neuro- fuzzy approach with other classifier algorithms in terms of accuracy. The aim is to classify the students into classes poor, average and good based on the parameters. The neuro-fuzzy architecture for the implementation is also explained in the paper. The paper concludes by saying that the neuro-fuzzy algorithm is better in accuracy than other algorithms and it is greater than 90% [10].

Md. Faisal, Faruque Asaduzzaman, Iqbal H. Sarker gives the performance analysis of different machine learning algorithms for detecting diabetes. The paper compares support vector machine, k nearest neighbor algorithm, Naive Bayes, C4.5 decision tree. The comparison was done on the basis of accuracy, precision, recall, FI score. As the result of the analysis the C4.5 decision tree was found to be a better algorithm. The data set was from 200 patients, which had almost 16 attributes [12].

III. Machine Learning Based Diabetic Prediction

Machine learning is the way by which the machine learns to adapt to the particular situation to take optimal decisions. There is a tremendous study going on so that machine learning can be applied in every field possible. Diabetes is a chronic disease which requires clinical attention and regular check up. The application of machine learning in this can reduce a lot of time, money and can increase the ease in detection. Different symptoms and parameters are used as input to the machine which is trained to detect the disease like weight, blood pressure, body mass, skinfold thickness, age, hereditary factor, etc. The necessary data are taken from patient records who have been detected with diabetes. These data sets are used to train the machine and later test it. The data are available as raw data. In the preprocessing stage the raw data is transformed so that they can be applied to the algorithm. There are different algorithms being implemented and analyzed in the field for detecting diabetes, like SVM, decision tree, naive Bayes, logical regression, etc. There is no single algorithm that can be said which outperforms all the other, because different studies done in the years have proven different algorithms better than the other. There are different learning methods in machine learning like supervised learning, unsupervised learning, reinforced learning.

A. Supervised learning

The supervised learning will have input data completely labelled with the output. The learning process finds the mapping between the input and the output. The supervised learning as the name suggests has a supervisor like behavior. The test results depend on the mapping that was obtained from the training phase. There are two main

types, they are classification and regression. The classification tries to classify a fixed number of output values while the regression model works with any range of output values. Major examples of supervised learning algorithms are the support vector machine, artificial neural network, linear regression, decision tree. The computational complexity of this type of learning is relatively simpler.

B. UnSupervised learning

The unsupervised learning will have un-labelled input data, that is the dataset will only have inputs and no output. The algorithm tries to classify the input into groups having some similarities. The aim of the algorithm is to find a distribution in this input data so that it can decide upon an output. This can be further classified as clustering and association. Clustering is used when subsets have to be formed out of a given dataset and association is used when a decision rule has to be drawn out of the inputs. K-means for clustering problems, Apriori algorithm for association rule learning problems are some of the examples for the unsupervised learning. The computational complexity is relatively higher.

C. Reinforcement learning

It is a learning method by which the machine needs to take the best decision for a particular situation. The learning decisions are dependent on one another. There are two different types of reinforcement learning, positive and negative. The positive learning increases the strength of the model when it takes a decision leading to a model with maximum performance, whereas the negative learning strengthens by avoiding or stopping a situation thus leading to an average performance.

D. Support vector machine

SVM is used for classification and regression application. This is a type of supervised learning. The algorithm groups the data points which are similar into classes. The decision making line or space which separates the classes is called the hyperplane. This process reduces the chance of errors. The aim of the algorithm is to divide the data set into different classes to get the maximum marginal hyperplane. SVM transforms input into a more suitable form so that the non separable data can be converted into separable input.

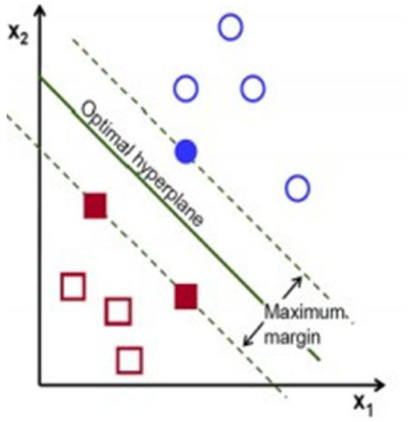

Fig 2: Finding the optimal hyperplane [4]

E. Decision tree algorithm

Decision tree is a supervised learning algorithm which divides the data set into different subsets according to different conditions and representation is like a tree. The decision process of the algorithm generally is in if-then-else conditions. We have different parameters which can be formed as the true or false condition (like Is the person below 20 years old) which can be applied and distinct decisions can be made with all the parameters included. This algorithm has relatively less computational complexity. Generally, this algorithm has less accuracy than other algorithms.

F. Random Forest algorithm

The algorithm is a supervised learning algorithm that can be used for classification as well as regression. The structure consists of a number of trees. These trees can be considered as decision trees. The algorithm actually

creates different trees for different data classes and combines the decision by voting. Random forest has comparatively less variance and high accuracy. The computation is comparatively complex and time consuming.

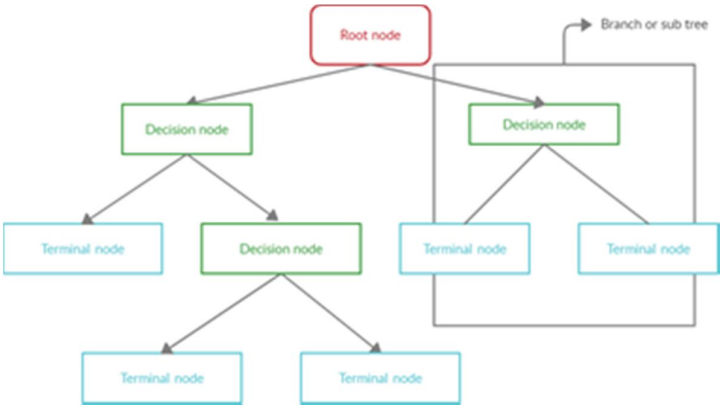

Fig 3: Decision Tree algorithm

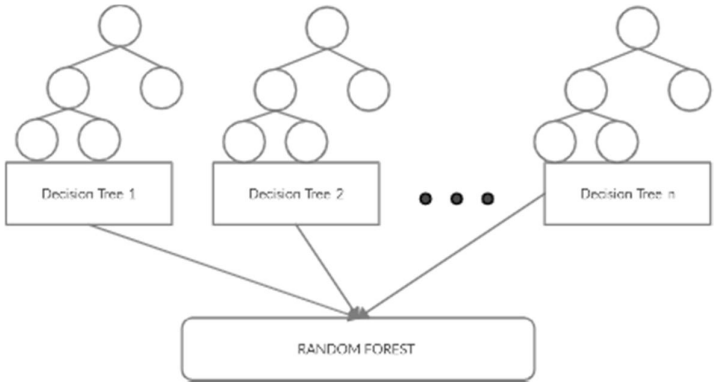

Fig 4: Random forest algorithm

G. Naive Bayes algorithm

The probability of an event lies in the relative knowledge about all the other events that are associated with the event. Bayes theorem explains how a probability of unknown classes can be found out. This algorithm is more useful to classify large random data. The probability can be expressed as[1]

$$P(H|E) = (P(E|H) * P(H)) / P(E)$$

Where,

- P(H|E) is the posterior probability of a hypothesis in which H gives the event E.
- P(E|H) given that the hypothesis H is true, when the probability of an event is E.
- P(H) the probability of hypothesis where the H is true, a preceding probability.
- P(E) states the probability of the event that is occurring.

IV. Proposed Methodology

As mentioned above, there are different types of algorithms that can be used to implement machine learning to detect diabetes or may be many other diseases. The difference lies in how much accuracy, precision the application requires. Here, a neuro-fuzzy approach in detecting diabetes in patients is considered. There are two components in this approach, i.e. it combines the effects of both neural networks as well as the fuzzy inference engine. FIS is a rule based classification algorithm which does not learn on its own. Neural networks on the other hand have learning capabilities and can self organize. Hence, both the methods are combined to form the neuro-fuzzy system. This method resembles the use decision making approach in humans.

Fuzzy inference system consists of a fuzzification layer, knowledge base, inference engine, and defuzzification module. In the fuzzification layer, the crisp input value or data is converted into a fuzzy set. These fuzzy values are then given to the inference engine. The knowledge base contains all the conditions that are to be applied to the input in the form of if else statement. In the fuzzy inference engine the appropriate decision is made by analyzing and applying the rules to the input fuzzy set. In the end the defuzzification is done on the output fuzzy set to get the crisp data.

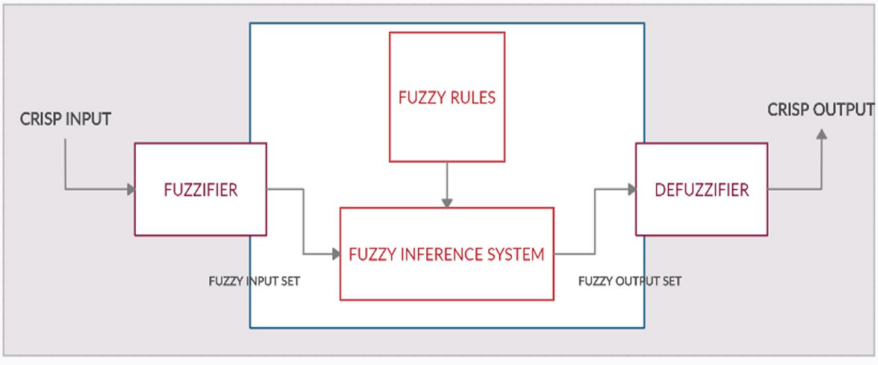

Fig 5: Fuzzy inference system

Neural networks represent the neural network in the brain. There are basically three layers in the network. First one is the input layer which is exposed to the external signal or data. The last layer is the output layer. The middle layer is a hidden layer, which extracts relevant features from the first layer. According to the output required different activation functions are applied into the input layer.

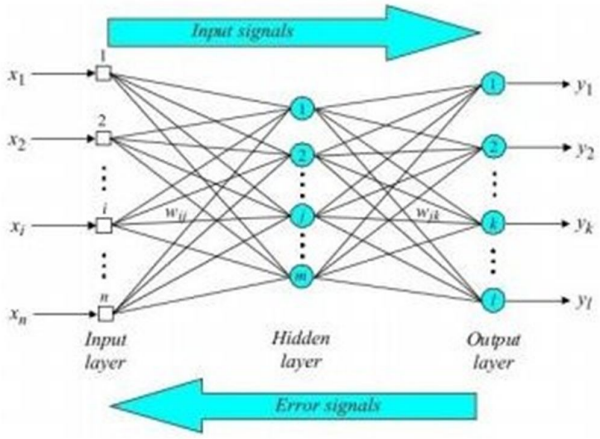

Fig 6:Neural Network[3]

In a neuro-fuzzy system, merits of both the approaches are utilized. It is a fuzzy system trained using a neural system. The first layer is the fuzzification layer which converts the crisp value to fuzzy values. The fuzzified values are then applied to the neural network as the input. The fuzzy rules are also applied to the neural network layer. Each neuron, thus combines the input using the fuzzy operations.The last layer is the defuzzification layer which converts the single output fuzzy value to a crisp output thus giving the necessary output for the system with a particular set input.

The membership function in the membership layer of each input is identified in this layer. There are different types of membership functions, a Gaussian function is utilized in this study, since this function has fewer parameters and smoother partial derivatives for parameters.

There are different types of membership functions, a Gaussian function is utilized in this study, since this function has fewer parameters and smoother partial derivatives for parameters. The Gaussian membership function is defined as[10]:

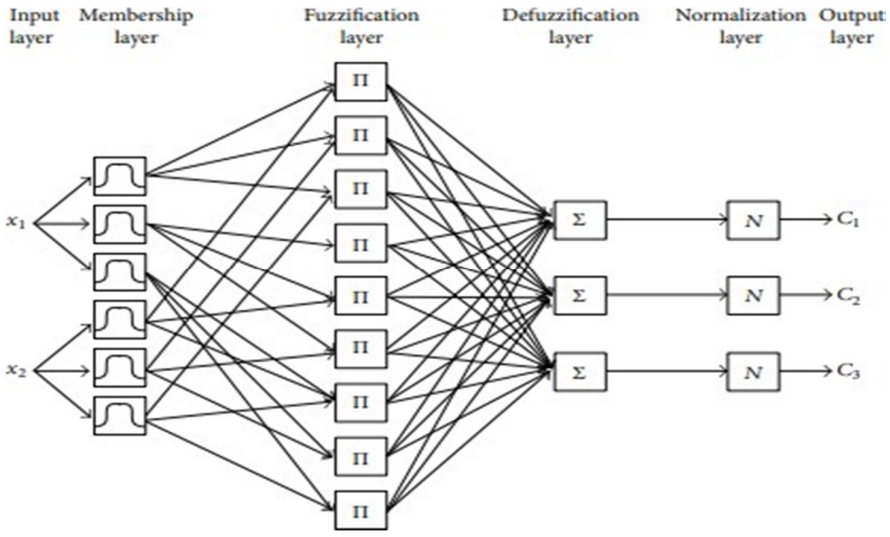

Fig 7: Neuro-fuzzy classifier[10]

$$\mu_{ij}(x_{sj}) = \exp\left(-\frac{(x_{sj} - C_{ij})^2}{2\sigma_{ij}^2}\right) \quad (1)$$

In the Fuzzification layer, each node generates a signal corresponding to how the fuzzy rule is applied to each sample. It is called the firing strength of a fuzzy rule. The firing strength of the rule is as follows[10]: Each class is affected according to their weight. So, at the defuzzification layer the weighted sum is calculated as the output. The weighted output can be calculated as[10]:

$$a_{is} = \prod_{j=1}^{N} \mu_{ij}(x_{sj})$$

$$\alpha_{is} w_{ik} \quad (2)$$

where N is the number of features.

$$\beta_{sk} = \sum_{i=1}^{M} \quad (3)$$

Sometimes the sum output may not be less than one. So these output values need to be normalized using different normalizing functions. This is done in the normalization layer. The normalized value of the sample at the Kth class is given by[10]:

$$O_{sk} = \frac{\beta_{sk}}{\sum_{l=1}^{K} \beta_s} \quad (4)$$

The model has different outputs for different sets of sample inputs.

V. Training The Model

In order to train the model, the if-else statements need to be obtained according to the input parameter. This may be obtained from the k-mean clustering algorithm. The algorithm groups the input data according to the similarities into different clusters and finds the mean value of each cluster. For training any model we can use different training algorithms. Here the scaled conjugate gradient method is used. In the conjugate algorithm method the weights are adjusted in the conjugate direction. TheCost function is calculated from the difference in the target function and class mean value and taking its least square mean. SCG finds the near-optimal parameter from the cost function.

VI. DATASETS

The data set for training and testing the model is the general symptoms obtained from the diabetes patients. The UCI machine learning repository has datasets for diabetes prediction. The data set for early prediction contained 520 instances with 17 attributes. Along with those some other attributes from another dataset were also added to increase the accuracy of the model. The whole dataset was classified into a number of classes based on common factors. Out of these dataset about 75% is used to train the model and rest for testing it.

VII. RESULT AND DISCUSSION

There are different matrices by which machine learning models can be compared. They are accuracy, precision, FI score, recall. Different outputs that are obtained can be taken in 4 ways, True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN). These values determine the different matrices for an algorithm.
Accuracy is the percentage of true prediction out of the total prediction done by the model Precision is the ratio of correctly obtained positive predictions to the total number of positive predictions which include the true positive and false positive Recall is the ratio of correct positive prediction to all the observations that were made. To actually assess the performance of the model it has to be compared with other algorithms such as SVM, Naïve Bayes.
The accuracy of the SVM classifying approach is approximately 83.6% and that of Random Forest approach is 78.4%. The neuro-fuzzy approach used in this paper has about 91.7% accuracy. So it is seen that the neuro-fuzzy method actually has better accuracy and can be increased further.

TABLE I: TABULAR REPRESENTATION OF COMPARISON DIFFERENT ALGORITHM

Comparison parameters	Support vector machine	Random forest algorithm	Neuro-fuzzy network
Accuracy(%)	86.3	78.4	91.7
Precision	0.75	0.70	.78
Recall	0.99	.81	.98

VIII. CONCLUSION

Machine learning has a vast area of application. Different algorithms like decision tree, neural network, random forest, etc. have different uses and comparatively different accuracy and precision. Here the diabetes detection is done using the combination of neural networks and fuzzy logic called the neuro-fuzzy approach. There were different attributes used for learning. These attributes are which helps the machine to learn from the data. Larger the features in the dataset better will be the output. Training the model with the dataset is the other important factor which increases the efficiency. According to the application required, compare different algorithms and use the most suitable one. In this study it is seen that opting the neuro-fuzzy over the other algorithms have improved the prediction ability of the model.

REFERENCES

[1] P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 367-371, doi: 10.1109/ICCMC.2019.8819841.

[2] J. C. Ramírez and D. Herrera, "Prediction of diabetic patient readmission using machine learning," 2019 IEEE Colombian Conference on Applications in Computational Intelligence (ColCACI), Barranquilla, Colombia, 2019, pp. 1-4, doi: 10.1109/ColCACI.2019.8781796.

[3] R. Lee and C. Chitnis, "Improving Health-Care Systems by Disease Prediction," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2018, pp. 726-731, doi: 10.1109/CSCI46756.2018.00145.

[4] R. M. Khalil and A. Al-Jumaily, "Machine learning based prediction of depression among type 2 diabetic patients," 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nanjing, 2017, pp. 1-5, doi: 10.1109/ISKE.2017.8258766.

[5] S. M. A. Huda, I. J. Ila, S. Sarder, M. Shamsujjoha and M. N. Y. Ali, "An Improved Approach for Detection of Diabetic Retinopathy Using Feature Importance and Machine Learning Algorithms," 2019 7th International Conference on Smart

Computing & Communications (ICSCC), Sarawak, Malaysia, Malaysia, 2019, pp. 1-5, doi: 10.1109/ICSCC.2019.8843676.
[6] K. VijiyaKumar, B. Lavanya, I. Nirmala and S. S. Caroline, "Random Forest Algorithm for the Prediction of Diabetes," 2019 IEEE International Conference on System,
[7] Computation, Automation and Networking (ICSCAN), Pondicherry, India, 2019, pp. 1-5, doi: 10.1109/ICSCAN.2019.8878802.
[8] K. Uyar, U. Ilhan, A. Ilhan and E. I. Iseri, "Breast Cancer Prediction Using Neuro-Fuzzy Systems," 2020 7th International Conference on Electrical and Electronics Engineering (ICEEE), Antalya, Turkey, 2020, pp. 328-332, doi: 10.1109/ICEEE49618.2020.9102476.
[9] Y. Dong, R. Wen, Z. Li, K. Zhang and L. Zhang, "Clu-RNN: A New RNN Based Approach to Diabetic Blood Glucose Prediction," 2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB), Hangzhou, China, 2019, pp. 50-55, doi: 10.1109/ICBCB.2019.8854670.
[10] Abiyev, Rahib & Abizade, Sanan. 2016. Diagnosing Parkinson's Diseases Using Fuzzy Neural System. Computational and Mathematical Methods in Medicine. 2016. 1-9. 10.1155/2016/1267919.
[11] Do, Quang Hung & Chen, Jeng-Fung. 2013. A Neuro-Fuzzy Approach in the Classification of Students' Academic Performance. Computational intelligence and neuroscience. 2013. 179097. 10.1155/2013/179097.
[12] R. Deo and S. Panigrahi, "Performance Assessment of Machine Learning Based Models for Diabetes Prediction," 2019 IEEE Healthcare Innovations and Point of Care Technologies, (HI-POCT), Bethesda, MD, USA, 2019, pp. 147-150, doi: 10.1109/HI- POCT45284.2019.8962811.
[13] Faruque, M. F., Asaduzzaman, & Sarker, I. H. 2019. Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus. 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). doi:10.1109/ecace.2019.8679365.