JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

An Efficient prediction of Diabetes Using Machine Learning Algorithms

[1] Prof Sonia Bajaj, [2]Deepakkumar Ishwar, [3]Roshan Jogi, [4]Hritvik Gayaki

[1]Head of Department,CSE,G H Raisoni University, [2]Student,CSE,GHRU, [3] Student,CSE,GHRU, [4] Student,CSE,GHRU
[1]Computer Science Enginnering,
[1]G H Raisoni University Saikheda , Chhindwara,Madhya Pradesh, India

Abstract : Diabetes in the population is one of the world's main diseases. Diabetes is achronic condition if there is either insufficient insulin in the pancreas or insulin cannot be used properly in the body. Long-term risks such as cardiac disease and liver cancer could lead to diabetes if not treated. Therefore, a timely prevalence of diabetes is very important to people around the world. In particular, the challenges to newer people and the workforce were identified as diabetes. Diabetes is tracked if patient changes in diet and lifestyle in the early days could be identified. Type 1 and type 2 diabetes were one of the most common forms of diseases; however, other types of diabetes, such as gestational diabetes pregnancy and other forms of diabetes. This paper outlines a method for predicting early diabetes, taking important risk effects into consideration. This approach uses the strategy to display a feature selection procedure. In Order to construct a good subset to improve predictive performance, the selected features are assessed through the rating of design models. Ultimately, these characteristics are learned from the neural network classifier and defined.

I. INTRODUCTION

Diabetes is the best - known chronic condition. Persons with type 2 diabetes typically undergo several types of complications, which in turn lead to death. Even so, glucose disappearance regulates sugar and is responsible for diabetes. Diabetes of type 2 develops if the body does not use the insulin of the pancreas effectively[1]. The chances of both companies having type-2 diabetes are prevention steps and foreign food patterns. The photoplethysmograph (PPG) waveform review extracts features that can predict type 2 diabetes, based on variations in the size of the curves of the PPG [2]. The PPG signal is monitoring blood tissue and vessel volumetric changes. Since the origin of a PPG waveform has not been established by agreement, it is intended to reflect the optically obtained plethysmograph. Changes in microvascular tissue bed blood volume can be identified by PPG. The pulsatile components of PPG are associated with increases in blood flow in the artery[3].

The popular oximeter of the pulse is used to highlight the skin and then calculate the changes of light absorption (index finger, ear, or forehead). There has been a range of current clinical applications, including pulse oximeters, the vascular system for measuring the blood pressure, and HR monitoring programs, using PPG technology in commercial medical devices available [5]. A reflection on citizens' quality of life, growth, and development is good physical, mental and social health. A poor individual's working productivity is higher than that of a healthy person. Good people also have a positive role to play in society[4]. However, increased contamination of the atmosphere, the use of fertilizers and pesticides to boost crop yields, a disparity in physical and mental activities, and a rise in the use of packaged food can cause overweight and impeccable health problems[6]. Some of the biggest risks to life include fluctuating blood pressure (BP), diabetes, cancer, cardiovascular diseases, and renal failure. Diabetes, which affects a significant number of people worldwide, is one of the most common disorders[3]. It is a condition that causes blood glucose levels to rise. It can be classified into five groups, specifically

- Type-1
- Type-2

- Gestational diabetes
- Impaired glucose tolerance (IGT)
- Impaired fasting glycemia (IFG)

Type 1: While only about 10% of patients with diabetes have this form of diabetes, the number of cases of this type has been rising recently. The condition arises as an autoimmune disease and is thus also known as juvenile diabetes at a very young age of fewer than 20 years. The pancreatic cells producing insulin are destroyed in this type of diabetes by the body's defense system. Patients with Type 1 diabetes should be followed by insulin injections, frequent blood tests, and nutritional constraints[4].

Type 2: Almost 90% of cases of diabetes of this form are known as diabetes that is adult and diabetes without insulin. In this situation, the different organs of the body become insulin resistant, increasing insulin demand. The amount of insulin required at this point is not provided by the pancreas. Patients need to follow a strict diet, practice exercise and monitor blood glucose to keep this form of diabetes at bay[6]. Obesity can contribute to type 2 diabetes by becoming overweight and physically inactive. The risk of diabetes is often known as being higher when aging.

Gestational diabetes: It is diabetes that occurs in pregnant women when there is insufficient insulin in the pancreas as a result of the elevated sugar levels.No treatment can lead to childbirth complications. Diabetes can be treated by tracking the diet and taking insulin[2].

IGT and IFG: The transition states from mild to diabetes are the IGT and IFG. The individual is vulnerable to the type-2 disorder by IGT and IFG.

2. ARTIFICIAL NEURAL NETWORK

The Artificial Neural Network (ANN) is a model for computer machinery training based on a biological neural network structure and function. Input and output are changed as the network knowledge flows across the network affects ANN structure[4]. The ANN is considered a nonlinear data modeling method that models complex input-output relations. Three basic layers are contained in a neural network as shown in Figure 1.

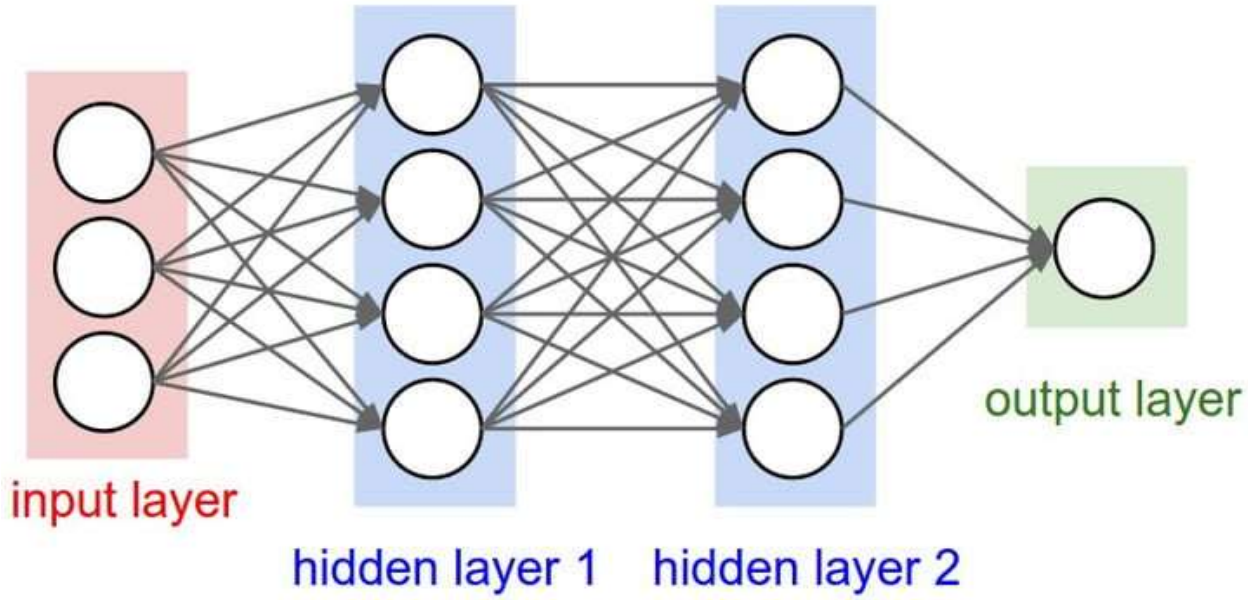

Fig 1: Artificial neural network
https://www.researchgate.net/figure

Input Layers: Elementary level of an ANN which contains data in the form of different texts, numbers, audio files, pixels etc.

Hidden Layers: There are hidden layers in the center of the ANN model. One secret layer can be hidden like a perceptron or many layers. These secret layers perform different types of arithmetic operation and describe input data characteristics.

Output Layer: It compares the accuracy by means of stringent center layer calculations.

There are several parameters and hyper parameters in a neural network that influence the model output. The ANN performance depends mainly on these parameters. Parameters include weights, biases, learning rate, batch size, etc as illustrated in figure 2. Per node weights the ANN. There are certain weights allocated to each node in the network. For the calculation of the weighted sum and the bias, a transmission function is used.

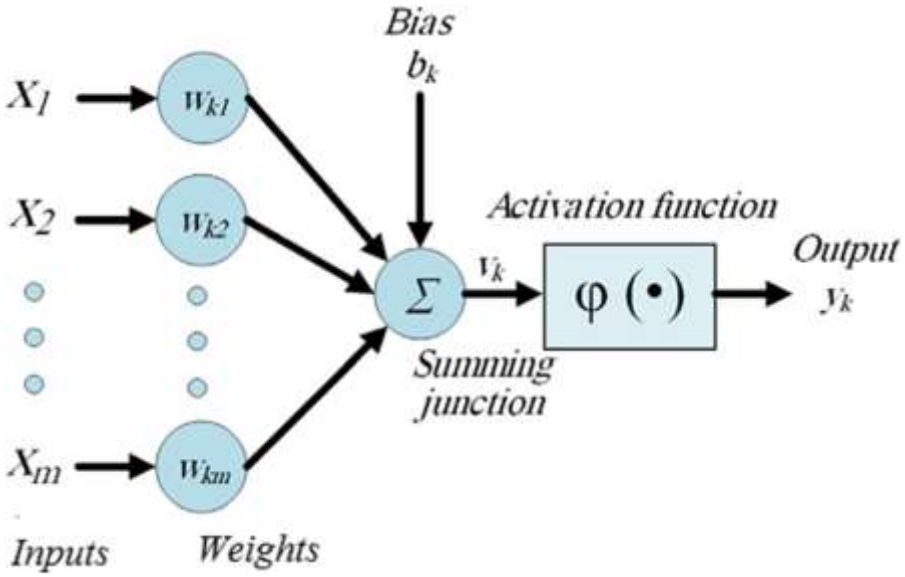

Fig 2: Simple structure of ANN
https://joems.springeropen.com

In other words, one secret layer MLP is assigned the corresponding variable $f_u(x)$

$$f_u(x) = A\left(b_2 + w_2\left(s(b_1 + w_{1.x})\right)\right) \quad 1$$

'W2' and 'W1' are the matrix weight, and the kernel feature is 'A' with the bias objects 'b2' and 'b1.'In addition, the variable h is specified as the hidden state

$$h(x) = s(b_1 + w_{1.x}) \quad 2$$

Multilayer perceptrons can be learned through experience. This approach uses iterations to ensure the lowest possible number of possible errors before the required input-output mapping has been achieved, which involves the collection of training data, including some input and related output vectors[5]. All model parameters for MLP training are learned. Let W_2 ,b_2, W_1, B_1 be the set of learning parameters.

I. RESEARCH METHODOLOGY

3.System Architecture

ANN preparation is an iterative process that begins with data collection. The pre-processing of the data then makes the data ready and the workout more successful[7]. The data must be separated into three separate sets during this phase of data pre-processing, primarily for training purposes, validation, and testing. On preliminary processing, the information must be chosen, and the network architecture must be configured in terms of several layers of neurons and the required type of Network such as multi-layer, competitive and dynamic[4,7]. The next step is to pick the algorithm for the preparation. An algorithm for training should be selected to deal with the network and the problem. After the network has been trained, the performance of the network is evaluated and any required improvements to the data, network design, and training algorithms can be established as illustrated in figure 3. The subsequent iteration renders these improvements. This is accomplished before the success of the network is reached. The entire process is iterated.

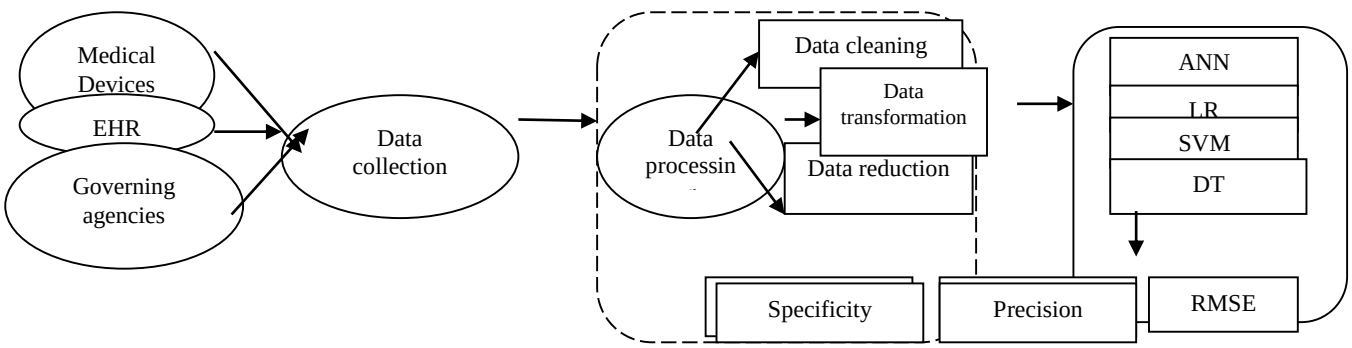

Fig 3: Architecture of the predictive analysis system Health care
Data Collection

The input for the device is the raw diabetic large data or data collection [5]. The unstructured broad input data can be accessed in different formats (flat files, .csv, tables, ASCII, text, etc.) from various electronic health records (EHR), clinical networks, and external sources (government sources, hospitals, pharmacies, insurance companies, etc.)

Data Preprocessing

The most important method for the process of data mining is data pre-processing. There are several missing values, null values, and defective values in data obtained from different sources. This leads to additional issues during data processing [7]. If there are more insignificant and noisy data, and mining data or artificial intelligence algorithms would be difficult to implement on such datasets. Different tasks are carried out for preprocessing data, including cleaning, incorporation, transformation, reduction, and discretion.

- Data cleaning: Data cleaning involves the completion of the missing qualities and the expulsion of noisy data. The high-level information contains exceptions expelled for irregularities to be determined. Different parameters are contained as zero in the data set such as weight, glucose, BMI, etc. It is important to substitute these values. Cleaning the data supersedes certain values to any of the data set's median values [8,9].
- Data reduction: Data reduction gets a reduced picture of the dataset. The volume is a wide liter but the equivalent results (or almost the equivalent) are given. Dimensionality reduction is used to decrease data set characteristics [3]. This preprocessing task is used to extract required or correct device attributes.
- Data transformation: The transformation of data involves knowledge smoothing, standardization, and aggregation. The dataset we use has several non-numerical values that must be numeric during training and modeling. Conversion to numeric of non-numeric functions. It is impossible to multiply the matrix in a string so that we must convert the string to a number[3]. Inputs to a fixed size are resized. Neural and linear models feed-forward have a fixed number of input nodes such that the input data is still the same size.

Machine learning is a set of methods that can automatically recognize data patterns and use them to predict future results or to take other decision-making forms under certain conditions. Machine learner introduces different algorithms, so machines can understand the current circumstances and can make reasonable decisions based on these machines. Machine learning runs autonomously and makes its own decisions [2].

Machine learning are two primary types: supervised learning and unsupervised learning

Unsupervised learning: A sample of the data set is given without marking class in unsupervised learning. Examples of unsupervised learning can be found in algorithms like K-means, self-organizing maps, etc [8].

Supervised learning: The method offers sample input and tagged classes in supervised learning. The model learns of the exercise data set sample and provides an etiquette class as a test data set output. Subcategories such as grouping and regression are covered in supervised learning. Examples of supervised learning methods include SVM, NB, KNN, RF, DT, and ANN. supervised learning is helpful in the field of healthcare for making predictions [8]. A short overview of the widely used supervised diabetes prediction learning algorithms is discussed below

- Logistic regression

Logistic regression (LR) is the most common model for understanding the correlation between dependent and independent variables. LR is often used when customers are looking to predict sickness or state of health. The LR model assesses the probability of a relevant educational T2DM [7], based on the feedback from risk factors. If an issue has T2DM, Y will be1; otherwise Y will be 0.The probability of generating T2DM in a person has been described as p(Y=1 b| Z) = p (Z). The formulation of the LR model is then shown below

$$logit(p) = ln\left[\frac{p(Z)}{1-p(Z)}\right] = \gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2 + \cdots \ldots \ldots + \gamma_k Z_k \qquad 1$$

After exponentiating both sides

$$\frac{p(Z)}{1-p(Z)} = e^{\gamma_0+\gamma_1 Z_1+\cdots\ldots\ldots+\gamma_k Z} \qquad 2$$

The probability of an individual developing T2DM is

$$p(X) = \frac{e^{\gamma_0+\gamma_1 Z_1+\cdots\ldots\ldots+\gamma_k Z_k}}{1-e^{\gamma_0+\gamma_1 Z_1+\cdots\ldots\ldots+\gamma_k Z}} \qquad 3$$

Where Z= (Z_1, Z_2…… Z_k) represents a risk factor

Where $\gamma = (\gamma_0 + \gamma_1 + \gamma_2 + \cdots \ldots \ldots + \gamma_k)$ are the coefficients estimated by using the method of maximum likelihood

- Support Vector Machine (SVM)

Using hyperplanes, the supporting vector machine classifies groups of data. It is better for x versus other variables for binary classification. SVM is a collection of supervised models for the classification and regression of computers. For instance, if two courses are given, the SVM classifier better divides two clusters[10]. One cluster does not sit closer to the data points of the other cluster for generalization purposes. It should be far from each other's clusters. This point is closest to the spectrum of the classification known as support vectors.

- Random Forest Algorithm

The most well-known and fundamental computer algorithms for learning are the Random Forest Algorithms. Random forest technology provides greater accuracy and strength. In most cases, it is difficult to develop the exhibition and, also, difficult to deal with various forms of information like numerical, apparent, and dual information [10,11]. Forests of numerous choices grow random trees. Random Forest is a kind of directed learning technique that uses a learning selection approach to classify and rebound. The trees in the random forest are identical, so no cooperation occurs between these trees when the trees are established. In the Random Forest Technique, the results of various assumptions are consolidated by the multiple trees of choice, which is why it is known as the meta-estimator.

Performance Measurement

In order to evaluate the results, we consider applying the uncertainty matrix [9]. We initially acquired a misunderstanding matrix in which the right and wrong classifier is determined in Table 1.

Table 1: Confusion Matrix

Actual class	Predicted class	
	Healthy	Diabetes
Healthy	True positive	False-negative
Diabetes	False-positive	True negative

This uncertainty matrix allows maximum accuracy, precision, specificity, sensitivity and F calculation to be quantified.

Accuracy is a performance calculation of the correct grade rate as defined by ACC. The ratio of accurate prediction and total prediction is determined in Eq 4.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad 4$$

The sensitivity analysis is to measure the real positive percentage of non-diabetes modules, properly defined. As indicated in Eq5, this can be determined as

$$Sensitivity = \frac{TP}{TP + TN} \quad 5$$

The specificity is measured as a genuinely negative rate, which shows the amount of right classified software modules and can be expressed in Eq 6.

$$Specificity = \frac{TN}{TN + FP} \quad 6$$

Precision is defined as the ratio of True Positive and (true and false) positives. It can be expressed as in Eq 7.

$$Precision = \frac{TP}{TP + FP} \quad 7$$

F-measure is defined as the mean of accuracy and performance in sensitivity. It can be calculated as shown in Eq 8

$$F = \frac{2 * P * Sensitivity}{P + Sensitivity} \quad 8$$

To minimize the difference between the expected value and the requested response, the creation of an ANN model involves a phased modifying of the vector. The difference is referred to as the cost function, calculated according to different parameters.

The performance index in terms of root mean squared error (RMSE) is used, which is formulated as in Eq 9.

$$RMSE = \sqrt{\frac{I}{k}\sum_{i=1}^{k}\left[[h(b_i) - y_i]\right]^2} \quad 9$$

Where k is several testing samples, $h(b_i)$ is the desired response and y_i is the predicted output of ANN.

IV. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statics of Study Variables

4. Results and discussion

This section presents the results between precision and accuracy Table 2 demonstrates the accuracy of the various algorithms and training times used, i.e. vector machine supports, random forest, and logistic regression algorithms.

Table 2: Accuracy level of the algorithms

Algorithms	Correctly classified instances	Incorrectly classified instances
LR	88.78%	17.67%
RF	99.89%	1.23%
SVM	85.78%	7.89%

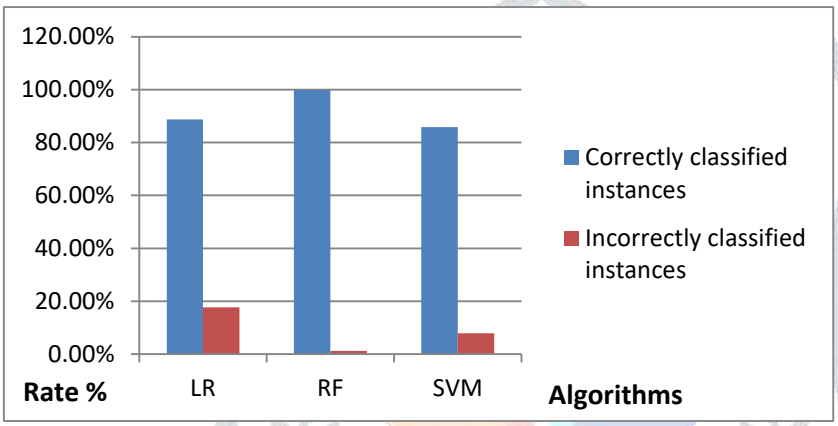

Fig 5: Rate of algorithms categorized correctly and incorrectly.

Figure 5 shows comparative rates for properly classified and incorrectly classifiable instances between Randomforest, Simple logistic, and SVM.

Table 3: RMSE of LR, RF, and SVM algorithm- percentage split and cross-validation

Parameters	LR		RF		SVM	
	percentage split	cross-validation	percentage split	cross-validation	percentage split	cross-validation
Root mean squared error	0.35	0.53	0.45	0.56	0.489	0.62

Root Mean Squared Error: Calculate the difference in values between a design or an optimization algorithm and the parameters are measured. Table 4 shows the values of specificity, sensitivity, accuracy, and precision for the algorithms used.

Table 4: Values of specificity, sensitivity, accuracy, and precision for algorithms

Algorithms	Specificity	Sensitivity	Accuracy	Precision
LR	93.45%	99.89%	99.67%	99.54%
RF	52.78%	99.02%	84.89%	99.21%
SVM	86.78%	89.98%	96.89%	99.35%

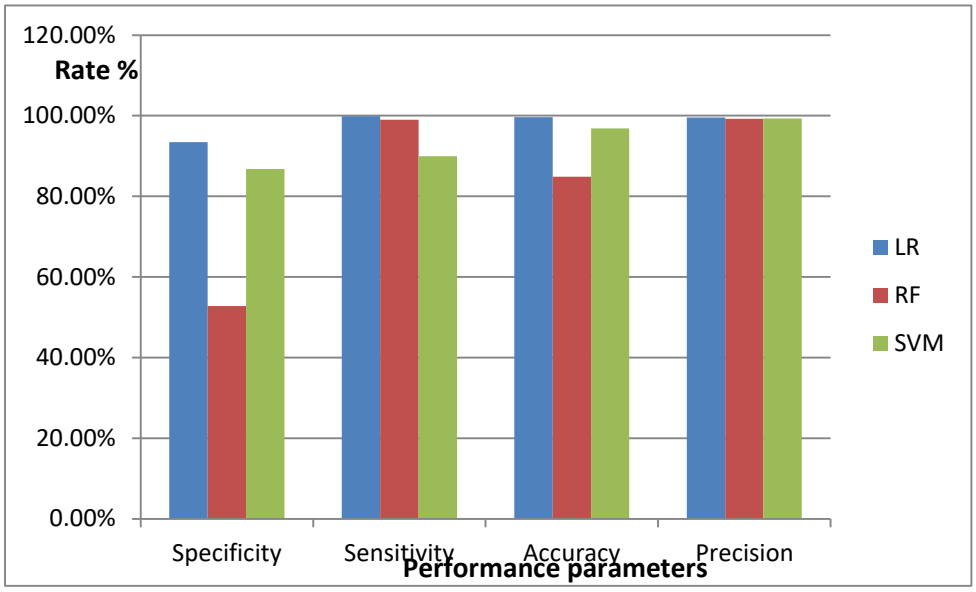

Fig 6: Rate Vs Performance parameters of the different algorithms

Figure 6 shows the specificities and sensitivities, precision and accuracy of the various algorithms. The best performance is given by the LR algorithms in relation to specificity, sensitivity, precision and precision.

II. Conclusion

III. Medical predictive analysis can allow physicians and health scientists to gather information and make sensible and effective choices from medical information. The neural network has shown greater prediction accuracy than other approaches like logistic regression, feature selection, decision tree, and so on as a data mining technique. The prediction of diabetes disease survival using different learning algorithms from the controlled neural network learning algorithms in this article. Based on the numerous algorithms presented and analyses carried out, the simple algorithm for logistic learning showed excellent classification with a sensitivity of 99.89% and a precision of 99.54% in diabetes prediction compared to other algorithms.

REFERENCES

[1] Yousef K Qawqzeh "*Neural Network-based Diabetic Type II High-Risk Prediction using Photoplethysmogram Waveform Analysis*" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 12, 2019,pp 89-92.

[2] P. Manaswini and S. Ranjit, "*Predict the onset of diabetes disease using Artificial Neural Network (ANN)*," International Journal of Computer Science & Emerging Technologies, vol. 2, I 2, 2011, pp303-311

[3] Aiswarya Iyer "*Diagnosis of diabetes using classification mining techniques*" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015,pp1-14.

[4] S. Kumari and A. Singh, "*A Data Mining Approach for the Diagnosis of Diabetes Mellitus*", Proceedings of Seventh international Conference on Intelligent Systems and Control, 2013, pp. 373-375

[5] C. W. Ting and C. Quek, "*A novel blood glucose regulation using TSK-FCMAC:afuzzy CMAC based on the zero-ordered TSK fuzzy inference scheme*," IEEE Transactions on Neural Networks, vol. 20, no. 5, pp. 856–871, 2009.

[6] Chiara Zecchin, Andrea Facchinetti, Giovanni Sparacino, Giuseppe De Nicolao, " *Neural Network Incorporating Meal Information Improves Accuracy of Short-Time Prediction of Glucose Concentration*" IEEE Transactions on Biomedical Engineering,vol.59 (6), pp:1-10,2012.