INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

Check for updates

*Corresponding author.

cvaidehi@gmail.com

Machine learning algorithms for Diabetes prediction and neural network method for blood glucose measurement

Vandana C Bavkar[1]*, Arundhati A Shinde[2]

1 Research Scholar, Electronics Department, Bharati Vidyapeeth Deemed to be University, Katraj, Pune, 411043, India
2 Head of the department, Electronics Department, Bharati Vidyapeeth Deemed to be University, Katraj, Pune, 411043, India

Abstract

Objectives: To facilitate painless and easy method for prediction of diabetes with high accuracy and to measure blood glucose by noninvasive method using Photoplethysmography (PPG). **Method:** In this study, diabetes prediction is done using different machine learning algorithms on a dataset created by using samples from PIMA Indian Diabetes dataset and in vivo diabetes dataset. Machine learning algorithms used are Support Vector Machine (SVM), Decision Tree, Naïve Bayes Classifier and K Nearest Neighbor (KNN). A PPG data of 182 individuals is recorded over 1 minute duration each. Various frequency and time domain features of PPG single are extracted using single pulse wave analysis. Neural network is trained using extracted features and glucose measurement is performed. **Findings:** With decision tree algorithm we got highest accuracy of 89.97% for diabetes prediction and it proves to be good algorithm for consideration for measurement in diabetes treatment.Clarke Error Grid analysis for blood glucose prediction is clinically accepted, so we performed similar analysis. Using time and frequency domain features, we got 94.27 % data points in clinically accepted regions (Region A and Region B). **Novelty**: Based on the data collected or samples analyzed, and accuracy of our results, it is encouraging to see that further research may lead to affordable noninvasive method for detection of diabetes at early stage.

Keywords: Noninvasive; blood glucose; NIR; machine learning algorithms; photoplethysmography

1 Introduction

Diabetes is one of the recurrent diseases that target the elderly population worldwide.[1] This condition occurs when the blood glucose or blood sugar level is very high. Early detection and treatment of diabetes can avoid severe health problems. Presently available techniques for measurement of blood glucose are invasive, which lead to discomfort to patients. Some of the key facts are; the number of people with diabetes rose from 108 million in 1980 to 422 million in 2014. Global prevalence of diabetes

among adults over 18 years of age rose from 4.7% in 1980 to 8.5% in 2014. Between 2000 and 2016, there was a 5% increase in premature mortality from diabetes. [2] These key facts justify need of affordable and painless measurement of Blood Glucose measurement technique. This motivates researchers to find noninvasive methods for blood glucose monitoring. Within various options of measurements, optical measurement technique proves high potential based on the research work carried out until now. Many researchers have tried different spectroscopic techniques, including machine learning techniques for diabetes prediction. In non-invasive techniques, glucose concentration is measured through skin without extracting blood, interstitial fluid or without a needle penetrating through skin for reaching these fluids. [3] Since glucose is an optically active substance, by the non-invasive technique, the physical properties such as optical, acoustic and electrical properties of the fluid or underlying tissues can be measured. [4] There are many non-invasive techniques include mid-infrared (MIR) [5], near-infrared (NIR) spectroscopy [6], Raman spectroscopy [7], Impedance spectroscopy [8], Polarimetry [9]. In recent years NIRS (Near Infrared Spectroscopy) has come up as a potential technique for non-invasive glucose monitoring due to low absorption and more penetration of Near Infrared light into the skin.NIR spectroscopy uses the light in the 750-2500 nm region, which interrogates the tissue with low energy radiation. Radiation in NIR range can penetrate the skin much deeper than visible or Mid–infrared (MIR) radiation. For this research work NIR optical sensor is used. [10]

Many Machine Learning (ML) or Data Mining techniques have been utilized in predicting diabetes in the last couple of years. [11] Data mining is the process of extracting information or features from data and utilize them to create a decision making process with increased efficiency. [12] Different machine learning techniques are Support Vector Machine (SVM), Random Forest, Decision Tree (DT), K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Naïve Bays (NB) Classifier etc. [13,14]

Ioannis Kavakiotis reviewed the applications of machine learning, data mining techniques and tools in the field of diabetes research with respect to Prediction and Diagnosis, Diabetic Complications, Genetic Background and Environment, and Health Care and Management. [15] Anny Leema describes diabetes mellitus prediction using different machine learning classifiers. Classification techniques like KNN, SVM, Decision Tree are used and performance is measured using factors like precision, accuracy, sensitivity and specificity. SVM classifier performs better over other methods with the highest accuracy. For experimentation PIMA Indian dataset is used. [16] Harleen Kaur suggested model for prediction of diabetes using machine learning to approach. Supervised machine learning algorithms namely Radial Basis Functions (RBF) kernel, SVM, KNN, ANN and Multifactor Dimensionality Reduction (MDR) are used for experimentation. PIMA Indian diabetes dataset is used and feature selection is done with the help of Boruta wrapper algorithm which provides unbiased selection of important features. SVM linear model gives better accuracy among all models. [17] Naveen Kishor developed binary classifier model based on different machine learning approaches. Among all, random forest model showing the greatest accuracy with 75 %. [18] Neha prerna Tigga compared four machine learning classification techniques for diabetes mellitus prediction. They have created in vivo dataset with different attributes and used PIMA diabetes dataset also. Random forest technique shows better performance for both the datasets. [19]

2 Materials and Methods

2.1 Work Flow Diagram

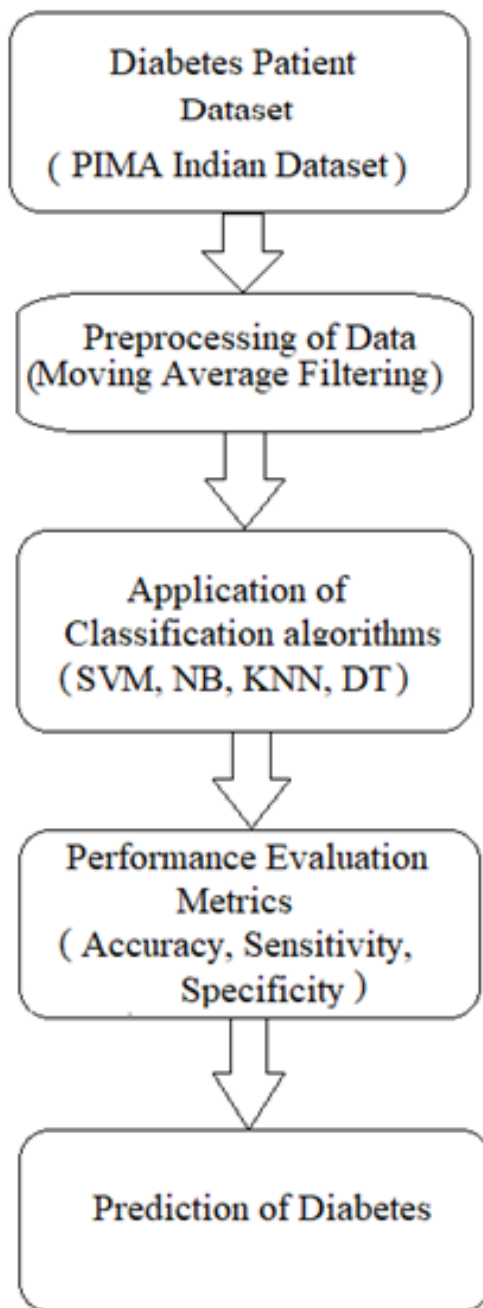The proposed procedure or work flow is as shown in Figure 1.

Fig 1. Work flow diagram

2.2 Dataset

PIMA Indian dataset (Dataset 1) is originally from the National Institute of Diabetes and Kidney diseases.[20] The objective of the dataset is to diagnostically predict whether or not a patient has diabetes based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. The dataset consist of several medical predictor (independent) variables and one target (dependent) variable i.e. outcome. Independent variables include the number of pregnancies of the patient, their Body Mass Index (BMI), age and so on. In particular, all patients here

are females at least 21 years old of PIMA Indian heritage. The different attributes of PIMA Indian database are listed below.

- Number of times pregnant
- Plasma Glucose concentration at 2 hours in an oral glucose tolerance test
- Diastolic blood pressure (mmHg)
- Triceps skin fold thickness (mm)
- Two-hour serum insulin (mu U/ml)
- Body Mass Index
- Diabetes Pedigree Function
- Age (years)
- Class variable (0 or 1)

The dataset has 768 observations with 8 attributes and one outcome. PIMA is a group of Native Americans living in Arizona.

The second dataset (Dataset 2) is in vivo dataset created with 82 patients chosen randomly. These patients are having different medical conditions like diabetic, non-diabetic, high or low blood pressure. All the patients are in the range of 30-85 age groups. The different attributes of vivo database are listed below.

- Gender
- Age
- Blood Pressure

2.3 Classification methods

Machine learning is the area in which machine learns from previous experience. This field is very similar to artificial intelligence. Basically, there are two types of machine learning algorithms, supervised and unsupervised. We have chosen to use supervised type of algorithm since we already know the output in the datasets we have. Supervised learning means mapping input to output based on labeled input output pairs. Labeled data consists of training examples. Each pair consists of input data and desirable output data. Different machine learning algorithms used here are Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbor (KNN) and Naïve Bays classifier. All the classification algorithms are implemented in MATLAB, which is a commercial mathematics software developed by MathWorks, Inc. It is used for algorithm development, data visualization, data analysis and provides interactive environment for numerical calculation. [13] [14]

2.3.1 Support Vector Machine

It is a type of supervised machine learning algorithm used for both regression and classification. In SVM, each data point is plotted on N-dimensional hyper plane (N- the number of attributes/features) that distinctly classifies the data points. To separate data points to two different classes, many possible hyper planes are there. We have to choose the one with maximum margin. Maximum margin is the distance between data points of two classes. Maximizing the distances between the nearest data points (either class) and hyper-plane will help us to decide the right hyper-plane. In the SVM classifier, it is easy to have a linear hyper-plane between these two classes. But, another burning question which arises is, should we need to add this feature manually to have a hyper-plane. No, the SVM algorithm has a technique called the kernel trick. The SVM kernel is a function that takes low dimensional input space and transforms it to a higher dimensional space i.e. it converts not separable problem to separable problem. It is mostly useful in non-linear separation problem. Simply put, it does some extremely complex data transformations, then finds out the process to separate the data based on the labels or outputs you've defined. The training data is represented as form of n data points.

$$M = \{(x_1, y_1), (x_2, y_2), \ldots \ldots (x_n, y_n)\}$$

Where y_n=1 or -1, a target or output variable that denotes the class to which the point x_n belongs where n is the number of data samples. The SVM classifier maps the input vectors into decision value and performs classification. The Hyperplane is defined as $w^T.x+b=0$ where w is p dimensional weight vector and b is scalar. The vector w is perpendicular to the separating hyperplane. Scalar parameter b is used to increase the margin. When training dataset is linearly separable, we select these hyperplanes in such a way that there are no points in between them and try on maximizing the distance between hyperplane. Mathematically, we will maximize the distance between the Hyperplane which is defined by $w^T x+b= -1$ and the hyperplane defined by $w^T x+b= 1$ as shown in Figure 2 . This distance is equal to $\frac{2}{|w|}$. Briefly, SVM works by identifying the optimal decision boundary that separates data points from different classes and predict the class of new data point based on this separation boundary. [21,22]
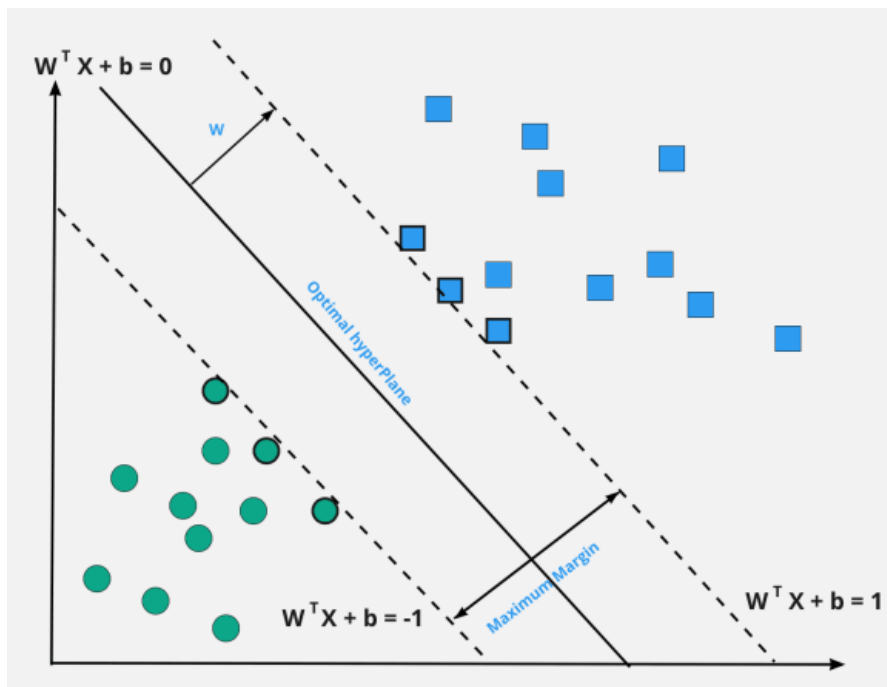
$W^T X + b = 0$

$W^T X + b = -1$

$W^T X + b = 1$

Optimal hyperPlane

Maximum Margin

w

Fig 2. Support vector machine

2.3.2 Decision tree

Decision tree is another type of supervised learning algorithm which has tree like structure. Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand. The internal node represents attributes/features of the dataset and branches represent the decision rule and leaf node represents the outcome or result. A decision tree makes decisions by splitting nodes into sub-nodes as shown in Figure 3. This process is performed multiple times during the training process until only homogenous nodes are left. It is the only reason why a decision tree can perform so well. [22]
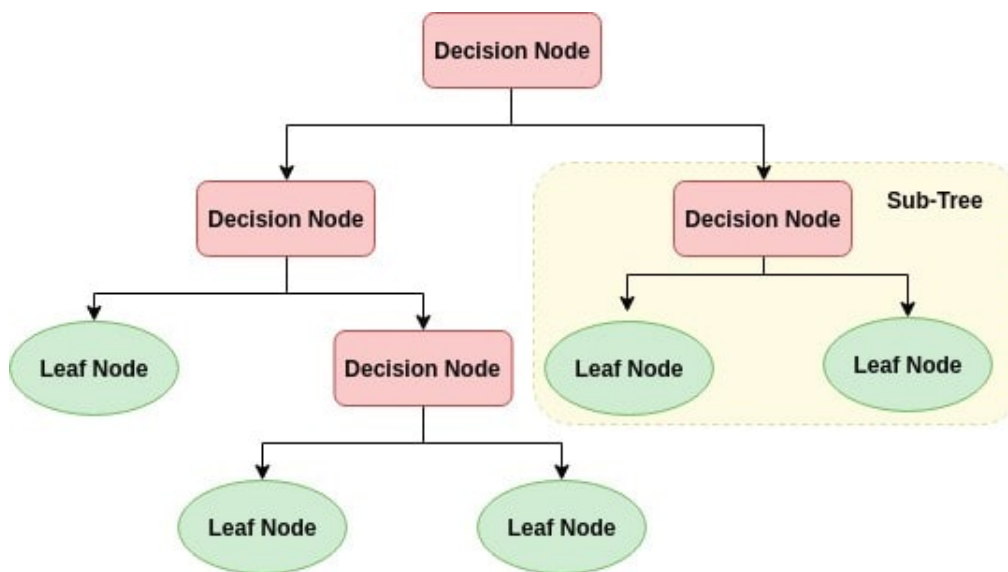

Fig 3. Decision tree

2.3.3 K-Nearest neighbor

KNN is one of the simplest supervised machine learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, and then it classifies that data into a category that is much similar to the new data. In the classification problem, algorithm will find the k-nearest neighbor of unseen data point and then it will assign the class to unseen data point by having the class which has the highest number of data points out of all classes of k neighbors as shown in Figure 4. For distance metrics, Euclidean metric is used. [19]

$$d\left(x, x'\right) = \sqrt{\left(x_1 - x_1'\right)^2 + \ldots + \left(x_n - x_n'\right)^2}$$
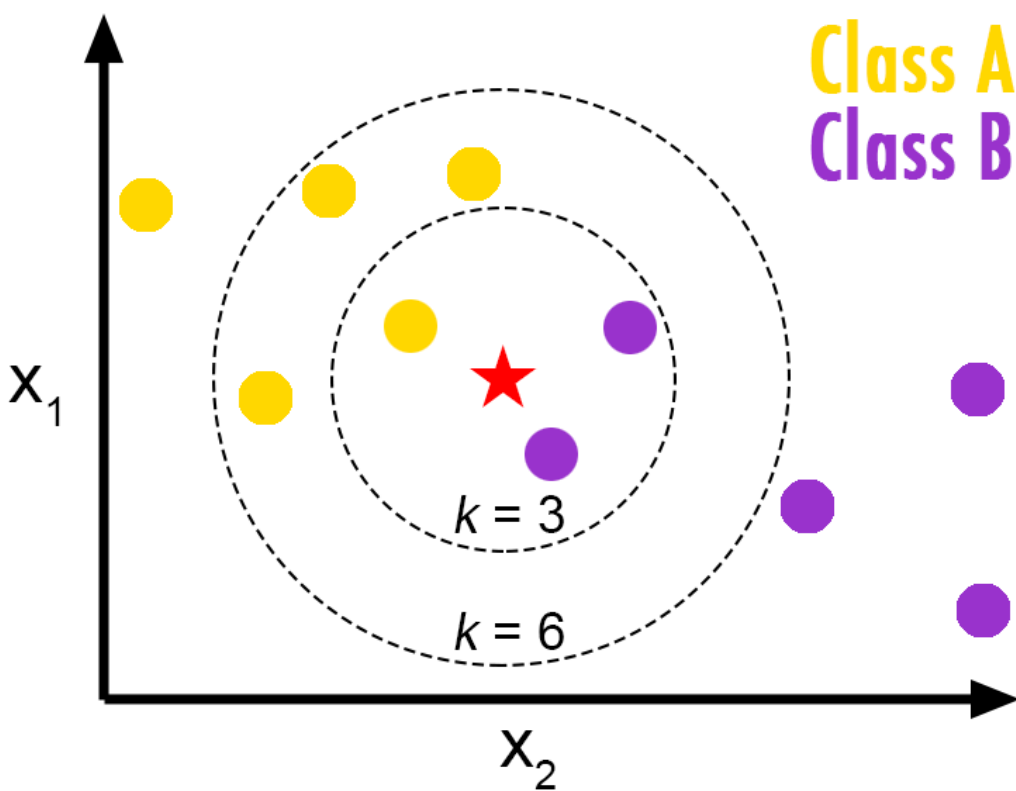

Fig 4. K-Nearest neighbor

2.3.4 Naïve Bays

It is a probabilistic classifier based on Bays theorem. Bayes' Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge. Naive Bayes is a machine learning model that is used for large volumes of data, even if you are working with data that has millions of data records the recommended approach is Naive Bayes. Prediction of membership probabilities is made for every class such as the probability of data points associated to a particular class. The class having maximum probability is appraised as the most suitable class. NB classifiers conclude that all the variables or features are not related to each other. [22]

Bayes' theorem:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Where,

$P(A \mid B)$= Posterior Probability

$P(B \mid A)$ = Likelihood Probability

$P(A)$ = Prior Probability

$P(B)$ = Marginal Probability

2.4 Feature selection

Feature selection methods can reduce the number of attributes, which can avoid the redundant features. In this research work we have used Principle Component Analysis (PCA) method for Dataset 1 to reduce the dimensionality. To make predictions of unknown, unseen data points, testing is used. In this research work 70 % data is used for training and 30% data is used for testing. [23]

2.5 Modified approach

Dataset 2 is created through in vivo experimentation on 182 patients having different conditions like diabetic, nondiabetic. Photoplethysmography (PPG) technique is used for measurement of blood glucose concentration. [24,25] Different time and frequency domain features like auto regression coefficient, peak distance, power spectral density, pulse transit time etc. are extracted from PPG signal recorded from these patients. [26,27] These feature vectors are given as input to artificial neural network and at the output we get predicted glucose concentration.

3 Results and Discussion

3.1 Performance evaluation metrics

Performance of different classifiers is compared using performance evaluation metrics such as sensitivity, specificity and accuracy. Table 1 shows performance metrics for PIMA Indian diabetes database using different classifiers.

Table 1. Performance metrics for classifiers (Dataset 1)

Classifier	Sensitivity	Specificity	Accuracy
SVM	0.7386	0.6795	0.7222
Naïve Bayes	0.7941	0.7667	0.7857
Decision Tree	0.9107	0.8775	0.8997
KNN	0.7402	0.6331	0.7035

Table 2 below shows performance metrics for different classifiers on vivo Dataset 2 with 82 patients.

Table 2. Performance metrics for classifiers (Database 2)

Classifier	Sensitivity	Specificity	Accuracy
SVM	0.6000	0.8148	0.75
Naïve Bayes	0.75	0.7222	0.7273
Decision Tree	0.8947	0.9206	0.9146
KNN	0	0.7500	0.7187

Tables 1 and 2 represent different performance values of all classification methods calculated on several measures. Decision tree algorithm is giving the maximum accuracy compared with other algorithms. Hence, it is proven that the Decision tree machine learning classifier can predict the probabilities of diabetes with higher accuracy as compared to other classifiers.

3.2 Classifiers performance

In this research work four different classifiers are used for diabetes prediction. They are compared with earlier researcher's work and tabulated in following Tables 3, 4, 5 and 6.

For Naïve Bays and Decision tree classifier our algorithm shows better performance as compared to SVM and KNN.

Table 3. SVM Classifier Performance (PIMA Dataset)

Sources	Accuracy (%)
Chang sheng Zhu [28]	58
Nitrsh Warke [29]	62
Deepti Sisodia [21]	65.10
A. Thammireddy [23]	68
Aishwarya Mujumdar [30]	68
Our algorithm	**72.22**
N.Sneha [22]	77.73
Neha Prerna T [19]	74.4
Muhammad Azeem Sarwar [31]	77

Table 4. Naïve Bays Classifier Performance (PIMA Dataset)

Sources	Accuracy (%)
Aishwarya Mujumdar [30]	67
Neha Prerna T [19]	68.9
Amina Azar [32]	71.4
Nitrsh Warke [29]	72
N. Sneha [22]	73.48
Muhammad Azeem Sarwar [31]	74
A. Thammireddy [23]	76
Getu Gamo Sagaro [33]	76
Our algorithm	**78.57**

Table 5. KNN Classifier Performance (PIMA Dataset)

Sources	Accuracy (%)
Amina Azar [32]	65.19
Nitrsh Warke [29]	66
Our Algorithm	**70.35**
Neha Prerna T [19]	70.8
Aishwarya Mujumdar [30]	72
Abdulhakim [16]	75.97
Muhammad Azeem Sarwar [31]	77
Chang sheng Zhu [28]	78

Table 6. Decision Tree Classifier Performance (PIMA Dataset)

Sources	Accuracy (%)
Nitrsh Warke [29]	68
Neha Prerna T [19]	69.7
Muhammad Azeem Sarwar [31]	71
Quan Zou [34]	72.75
Naveen Kishore G [18]	72.91
Getu Gamo Sagaro [33]	73
N.Sneha [22]	73.18
Aishwarya Mujumdar [30]	74
Amina Azar [32]	75.65
A. Thammareddy [23]	76.22
Seyede Somayeh [35]	80.2
Our Algorithm	**89.97**

3.3 Photoplethysmography

PPG (photoplethysmography) is a simple, optical based noninvasive technique used in the development of advanced health care.[24] PPG is a noninvasive optical technique for measuring changes in blood circulation that is mainly used for monitoring blood perfusion in the skin. The PPG signal is recorded with a sensor similar to pulse oximeter which works only in NIR spectrum. The PPG finger sensor consists of a light emitting diode (LED), which is often red or infrared, and a photo detector (PD). PD and LED are on the opposite side of the finger. The light is emitted from the LED to the skin and a small fraction of light intensity changes is received by the PD, which are related to blood flow, blood volume, blood vessel wall movement, and the orientation of red blood cells in the underlying tissue. Various parameters like blood pressure, heart rate, pulse transit time and blood glucose can be analyzed. Each recorded PPG pulse contains useful information for cardiovascular assessment. More detailed information can be obtained by analysis of the PPG signal sequences recorded over some time interval of about few minutes.[25–27]

A functional relationship exists between pulse signal and blood glucose concentration.[25] For measuring blood glucose concentration, time and frequency domain features are extracted from PPG waveform. We have done single pulse wave analysis for the purpose of concentration measurement. Single pulse wave is shown in Figure 5. In the following section we summarize different features extracted from PPG signal.

3.4 Time domain features of single pulse

Time domain features are over the X and Y axis of pulse as shown in Figure 5.
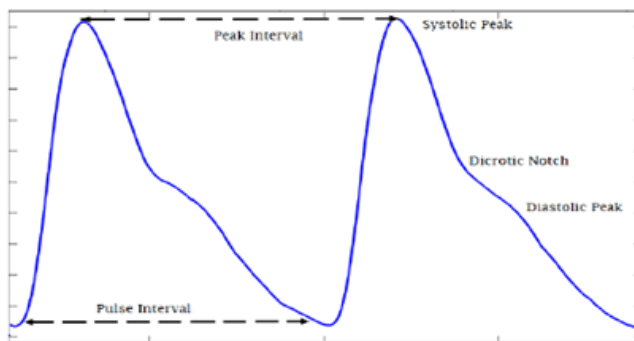

Fig 5. Single pulse wave from PPG signal

Below time domain features are extracted for our experimentation.

- Width period- Time taken for a single period
- Highest peak value- Maximum amplitude of the signal
- Time of the highest peak value- Time value when amplitude of a signal is maximum
- Diastolic peak amplitude- Amplitude of diastolic peak
- Time of Diastolic peak- Time at which diastolic peak occur
- Notch amplitude- Amplitude of notch
- Time of notch- Time at which notch occurs in a signal
- Time difference - Total time taken from start to peak, peak to notch, notch to diastolic peak and last diastolic peak to end
- Mean amplitude value- Mean amplitude value of a single period
- Standard deviation of single period- Standard deviation of amplitudes
- Mean amplitude- Mean of amplitudes from start to-peak, peak to notch, notch to diastolic peak and diastolic peak to end

Frequency Domain Features : Frequency domain features are related to the rate of change of signal. Below frequency domain features are extracted for our experiment.

- Auto Regression Coefficients
  These coefficients capture the change in the shape of the pulse occurring due to change in blood flowing through different veins and capillaries. These coefficients are also used to models the spectral envelope of PPG signals.[25]

- Kaiser Teager Energy
  It is well-known method used for finding energy profiles of signals with periodic signal components. This property indicates whether the signal is clean or noisy signal. [25]
- Power Spectral Density
  It measures damping of pulses, the spectral harmonic components and presence of noise. [25]

These features are given as input to neural network.

3.5 Neural network configuration

In this research work, we have used three hidden layer neural network topology. The neural network structure is shown in Figure 6. Neural network tool in MATLAB is used to build the structure. Preliminary experimentation is carried out with variations in neurons in each hidden layer to get maximum accuracy. We started with 2 hidden layer and (15, 10) neurons in layer 1 and 2 respectively. With this configuration we got optimum results. Therefore, we increased one more layer with (15, 10, 5) neurons in layer 1, 2, 3 respectively. Further experimentation with varying neurons is carried out, and we found that (20, 15, 10) neuron configuration gives maximum accuracy with maximum data points in clinically accepted regions (region A and B) of Clarke Error Grid Analysis as shown in Figure 7.
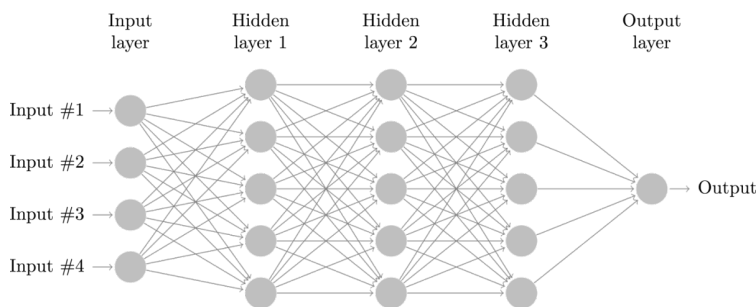

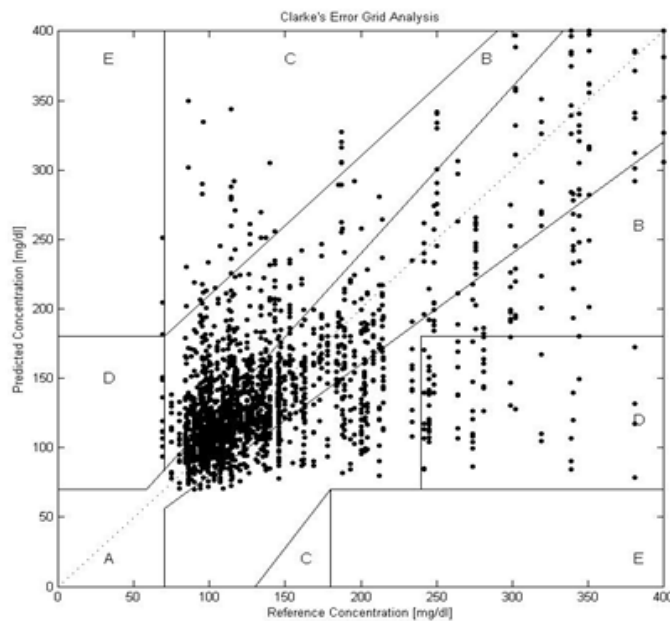**Fig 6.** Three layer neural network in MATLAB

Fig 7. Clarke error grid for dataset 2

The predicted results are compared with the actual glucose value using Clarke Error Grid Analysis. As shown in Figure 7, region A represents 59.66% data points within 20% of actual blood glucose value. Region B represents 34.61% data points, region C, D and E are negligible. Total 94.27% results are in clinically accepted region (region A and B) as shown in Figure 7.

4 Conclusion

The accuracy of proposed system comes to 89.97% for diabetes prediction, when compared with similar previous research work for different algorithms like KNN, Decision Tree and Naïve Bayes classifier. In this research work, two databases are used for experimentation. During this work, four machine learning algorithms are used for prediction of diabetes and performance of each type is measured with respect to different accuracy measures. The results of all four algorithms are compared with actual results of patients. Actual results are recorded using the traditional Invasive Method. After comparing each algorithm with actual result, it is found out that Decision tree algorithm shows better performance amongst all with accuracy of 89.97% for dataset1 for diabetes prediction. For dataset 2, using Clarke error grid analysis we got 94.27 % data points in clinically accepted region A and B. The research work can be extended for extraction of derivative features for better results of measurement of blood glucose concentration.

Acknowledgement

References

1) International Diabetes Federation, About Diabetes. . Available from: http://www.idf.org/about-diabetes.
2) Diabetes fact sheet details. . Available from: https://www.who.int/news-room/fact-sheets/detail/diabetes.
3) Chen C, Zhao XL, Li ZH, Zhu ZG, Qian SH, Flewitt A. Current and Emerging Technology for Continuous Glucose Monitoring. *Sensors*. 2017;17(12):182–182. Available from: https://dx.doi.org/10.3390/s17010182.
4) Larabi-Marie-Sainte, Aburahmah, Almohaini, Saba. Current Techniques for Diabetes Prediction: Review and Case Study. *Applied Sciences*. 2019;9(21):4604–4604. Available from: https://dx.doi.org/10.3390/app9214604.
5) Gusev M, Poposka L, Spasevski G, Kostoska M, Koteska B, Simjanoska M, et al. Noninvasive Glucose Measurement Using Machine Learning and Neural Network Methods and Correlation with Heart Rate Variability. *Journal of Sensors*. 2020;2020:1–13. Available from: https://dx.doi.org/10.1155/2020/9628281.
6) Narkhede P, Dhalwar S, Karthikeyan B. NIR Based Non-Invasive Blood Glucose Measurement. *Indian Journal of Science and Technology*. 2016;9(41). Available from: https://dx.doi.org/10.17485/ijst/2016/v9i41/98996.
7) Li N, Zang H, Sun H, Jiao X, Wang K, Liu TCY, et al. A Noninvasive Accurate Measurement of Blood Glucose Levels with Raman Spectroscopy of Blood in Microvessels. *Molecules*. 2019;24(8). Available from: https://dx.doi.org/10.3390/molecules24081500.
8) Kamat DK, Bagul D, Patil PM. Blood Glucose Measurement Using Bioimpedance Technique. *Advances in Electronics*. 2014;2014:1–5. Available from: https://dx.doi.org/10.1155/2014/406257.
9) Pirnstill CW, Malik BH, Gresham VC, Coté GL. In Vivo Glucose Monitoring Using Dual-Wavelength Polarimetry to Overcome Corneal Birefringence in the Presence of Motion. *Diabetes Technology & Therapeutics*. 2012;14(9):819–827. Available from: https://dx.doi.org/10.1089/dia.2012.0070.
10) Tang L, Chang SJ, Chen CJ, Liu JT. Non-Invasive Blood Glucose Monitoring Technology: A Review. *Sensors*. 2020;20(23):6925–6925. Available from: https://dx.doi.org/10.3390/s20236925.
11) Vizhi K, Dash A. Diabetes Prediction Using Machine Learning. *International Journal of Advanced Science and Technology*. 2020;29(06):2842–2852. Available from: https://sersc.org/journals/index.php/IJAST/article/view/13795.
12) Baig M, Nadeem M. Diabetes prediction using machine learning algorithms. 2020. Available from: https://doi.org/10.13140/RG.2.2.18158.64328.
13) Rout M, Kour A. Prediction of Diabetes Risk based on Machine Learning Techniques. *IEEE International Conference on Intelligent Engineering and Management*. 2020. Available from: https://doi.org/10.1109/ICIEM48762.2020.9160276.
14) Alic L, Abbas HT, Rios M, Abdulghani M, Qaraqe K. Predicting Diabetes in Healthy Population through Machine Learning. *IEEE 32nd International Symposium on Computer-Based Medical Systems*. 2019;p. 567–570. Available from: https://doi.org/10.1109/CBMS.2019.00117.
15) Kavakiotis I, Tsave O, Salifoglou A. Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural biotechnology Journal*. 2017. Available from: https://doi.org/10.1016/j.csbj.2016.12.005.
16) Hassan AS, Malaserene I, Leema AA. Diabetes Mellitus Prediction using Classification Techniques ". *International Journal of Innovative Technology and Exploring*. 2020. Available from: https://doi.org/10.35940/ijitee.E2692.039520.
17) Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. *Applied computing and Informatics*. 2018. Available from: https://doi.org/10.1016/j.aci.2018.12.004.
18) Kishore GN, Rajesh V, Reddy AVA, Sumedh K, Sai TR, Reddy. Prediction of Diabetes Using Machine Learning Classification Algorithms". *International Journal of Scientific and Technology Research*. 2020;9(1).
19) Tigga NP, Garg S. Prediction of Type 2 Diabetes using Machine Learning Classification Methods. *Procedia Computer Science*. 2020;167:706–716. Available from: https://dx.doi.org/10.1016/j.procs.2020.03.336.
20) Medical records for Pima Indians. . Available from: https://www.kaggle.com/kumargh/pimaindiansdiabetescsv.
21) Sisodia D, Sisodia DS. Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*. 2018;132:1578–1585. Available from: https://dx.doi.org/10.1016/j.procs.2018.05.122.

22) Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*. 2019;6(1). Available from: https://dx.doi.org/10.1186/s40537-019-0175-6.

23) Reddy AT, Nagendra M. Minimal Rule-Based Classifiers using PCA on Pima-Indians-Diabetes-Dataset". *International Journal of Innovative Technology and Exploring Engineering*. 2019;8(12):12–12. Available from: https://doi.org/10.35940/ijitee.L2476.1081219.

24) Moraes J, Rocha M, Vasconcelos G, Filho JV, de Albuquerque V, Alexandria A. Advances in Photopletysmography Signal Analysis for Biomedical Applications. *Sensors*. 2018;18(6):1894–1894. Available from: https://dx.doi.org/10.3390/s18061894.

25) Habbu KS, Joshi S, Dale M, Ghongade R. Noninvasive Blood Glucose Estimation Using Pulse Based Cepstral Coefficients. *IEEE International Conference on Signal Processing and Information Security (ICSPIS)*. 2019. Available from: https://doi.org/10.1109/ICSPIS48135.2019.9045897.

26) Habbu S, Dale M, Ghongade R. Estimation of blood glucose by non-invasive method using photoplethysmography. Indian Academy of Science. 2019. Available from: https://doi.org/10.1007/s12046-019-1118-9.

27) Zhang Y, Zhang Y, Siddiqui SA. Non-invasive Blood-Glucose Estimation Using Smartphone PPG Signals and Subspace KNN Classifier". *Electrotechnical Review*. 2019;86(1):68–74.

28) Sheng ZC. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques", Informatics in Medicine Unlocked. *Science Direct*. 2019. Available from: https://doi.org/10.1016/j.imu.2019.100179.

29) Warke M, Kumar V, Tarale S, Galgat D, Chaudhari. Diabetes Diagnosis using Machine Learning Algorithms. *International Research Journal of Engineering and Technology*. 2019;6(3). Available from: https://www.irjet.net/archives/V6/i3/IRJET-V6I3277.pdf.

30) Mujumdar A, Vaidehi V. Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*. 2019;165:292–299. Available from: https://dx.doi.org/10.1016/j.procs.2020.01.047.

31) Sarwar MA, Kamal N, Hamid W, Shah MA, Shah. Prediction of Diabetes Using Machine Learning Algorithms in Healthcare. In: and others, editor. IEEE 24th International Conference on Automation and Computing. 2018. Available from: https://doi.org/10.23919/IConAC.2018.8748992.

32) Azrar A, Ali Y, Awais M, Zaheer K. Data Mining Models Comparison for Diabetes Prediction. *International Journal of Advanced Computer Science and Applications*. 2018;9(8). Available from: https://dx.doi.org/10.14569/ijacsa.2018.090841.

33) Sagaro GGG, Battineni F, Amenta. Comparative Machine Learning Approach: A Follow- Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods. Multidisciplinary Digital Publishing Institute, Machines. 2019. Available from: https://doi.org/10.3390/machines7040074.

34) Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics*. 2018. Available from: https://doi.org/10.3389/fgene.2018.00515.

35) Mirzajani SS, Salimi S. Prediction and Diagnosis of Diabetes by Using Data Mining Techniques. *Avicenna Journal of Medical Biochemistry*. 2018;6(1):3–7. Available from: https://dx.doi.org/10.15171/ajmb.2018.02.