International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-7, May 2020

Detection of Diabetes By Machine Learning Technique

Vandana Bavkar, Arundhati A. Shinde

Abstract*: Diabetes is a most important health dispute that has reached distressing levels; today approximately half a billion individuals are living with diabetes universal. Diabetes is a state that damages the body's capability to process glucose in blood, otherwise known as blood sugar. It is a metabolic disease that reasons high blood sugar. The hormone insulin transfers sugar from the blood into your cells to be stored for energy. With diabetes, your body either doesn't make sufficient insulin or can't efficiently use the insulin it does makes. The motive of this research is to design a method or prototype which can detect or predict the diabetes in patients with high precision. Therefore different machine learning classification algorithms namely decision tree, support vector machine, Naïve Bayes and k-NN are used in this research work for prediction of the diabetes. Two databases are used for experimentation. The first one is created from hospital with 82 patients and second one is readily available Pima Indian Diabetes database. The performances of different machine learning algorithms are estimated on different measures like Precision, Recall, F-measure and accuracy. The objective of this research is to study the accuracy of different machine learning algorithms and hence identify set of suitable algorithms for prediction of diabetes for further research work.

***Keywords*: Blood Glucose, Diabetes, NIR Spectroscopy, Machine Learning Algorithms.**

I. INTRODUCTION

Diabetes is a disease that arises when the proportion of glucose in your blood is very high. According to International Diabetes Federation (IDF), nearly 463 million peoples were living with diabetes worldwide [1], [2]. Glucose is the main source of energy. We get energy from the food we eat. This food is converted into energy by insulin which is naturally occurring hormone developed by beta cells of pancreas. At times, the situation arises that body doesn't produce sufficient amount of insulin or it is not properly used by body to get energy. In this case glucose remains there in your blood and doesn't spread in your cells. So the amount of unused glucose in blood increases with increases the sugar level. This condition is very dangerous for health. Even though diabetes has no cure, we can manage this with regular exercise and diet.

Revised Manuscript Received on May 30, 2020.
***** Correspondence Author
Vandana Bavkar*, Department of Electronics, Bharati Vidyapeeth (Deemed to be University), Pune, India. Email: cvaidehi@gmail.com
Arundhati A. Shinde, Department of Electronics, Bharati Vidyapeeth (Deemed to be University), College of Engineering,Pune, India. Email: aashinde@bvucoep.edu.in

Hence, detection at an early stage is very important. There are mainly three types of diabetes namely Type1, Type2 and Gestational diabetes. Type1 condition occurs when your body attacks your pancreas with antibodies and body doesn't make insulin. So it is called as insulin dependent diabetes. Type2 diabetes is more common in adults and teens. So it is also called as adult onset diabetes or non-insulin dependent diabetes. In this case pancreas creates some insulin, but it is not enough or body doesn't use it properly. Third type, Gestational diabetes normally occurs in pregnant women. Pregnancy usually causes some form of insulin resistance. It is very important to control gestational diabetes to protect baby's growth and development. All these type of diabetes need treatment and if they are detected at an early stage one can avoid the complications associated with them. In recent years NIRS (Near Infrared Spectroscopy) has come up as a prospective method for noninvasive glucose monitoring because of little immersion and extra diffusion of NIR light into the skin [3]-[5]. NIR spectroscopy is situated in the wavelength region of 730-2500nm. Glucose is an optically active substance. When NIR light falls on body part (finger, palm, ear) some part of light radiations is absorbed by glucose molecules and remaining part is transmitted to other side. Therefore, blood glucose measurement is possible by measuring transmitted and reflected light. Many researchers have tried several body places such as tongue (saliva), forearm, palm, ear lobe, finger etc.to measure blood glucose noninvasively [6]. In this research work, for data collection NIR sensors are used. In recent years, different machine learning techniques are used for implementation of algorithms of automated identification system for diabetes. Recently, different algorithms are used to analyze patient datasets for secreted information such as Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (K-NN), Artificial Neural Networks (ANN), Linear Regression etc. [7]. The remaining sections are organized as follows: Section II describes related work of NIR spectroscopy and different machine learning algorithms used for prediction of diabetes and blood glucose concentration.

II. RELATED WORK

Sauad Larabi-Marie-Sainte et al describes many machine learning techniques for prediction of diabetes. Different classifiers are used on PIMA Indian database with evaluation metrics like accuracy, precision, Recall, F-measure and ROC. The highest accuracy is obtained by implementing REPTree i.e. Reduced Error Purning Tree [8].Deepti Sisodia et al describe three machine learning classification algorithms namely Decision tree,

Support vector machine (SVM) and Naïve Bayes for detection of diabetes at an early stage. Highest accuracy is achieved with Naïve Bayes on PIMA Indian dataset [9].

Marjan Gusev et al reviewed different noninvasive blood glucose measurement methodologies and explain different machine learning and neural network method for glucose measurement [10].

CF So et al describes deep learning method in their research work for blood glucose monitoring. For data collection NIR wavelength range is used. An improved method based on Monte-Carlo approach for partial least squares is proposed and got the optimal accuracy. They have used in vitro dataset with 512 patients [11].

A.Thammi Reddy et al describes minimal rule based classification using principle component analysis (PCA). Naïve bayes classifier shows better performance with PCA getting optimal accuracy [12].

Chang Sheng Zhu et al describe logistic regression model to estimate the diabetes by combining k-means clustering and Principal Component Analysis (PCA). Readily available PIMA Indian database is used here. Firstly, Principle Component Analysis is used for dimensionality decrease. After this K-means clustering is used to remove incorrectly classified instances. With correct classification and clustering, data is fed to logistic regression model. They achieved better accuracy with this model [13].

M.S. Barale. et al describe cascaded model for prediction of diabetes. They have cascaded k-means clustering with different machine learning algorithms like SVM (Support Vector Machine), LR(Logistic regression) , K-NN . With K-means and SVM grouping they got highest accuracy [14].

N. Sneha et al describes analysis of diabetes mellitus using different machine learning techniques. Attribute selection is done using different classifiers. With decision tree (DT) and random forest (RF) method better accuracy is achieved [15].

Jyoti Yadav et al describe NIR blood glucose measurement technique with different body sites (earlobe finger). With variation in angles three different polarizers were used at 45º, 90º, and 180 º. Two NIR transmitters are used of 640nm and 740nm. Aqueous glucose solutions are used with different glucose concentrations. With different angles of polarization at 180º, the deviation in output voltage is highest [6].

Hussan Ali et al describe a method for glucose concentration measurement. The research work is based on photo acoustic spectroscopy means effect of light energy on a substance by acoustic detection. The researchers found the linear relation between detected acoustic signal measured at body site and actual blood glucose concentration. [17].

Abishek Thekkeyil Kunnath et al describe a method which is voltage intensity based non-invasive blood glucose monitoring system by occlusion method. At the detector side after receiving reflected light, different voltage levels are received for individuals. The researcher found correlation between variation in light intensity and glucose concentration. Prediction of diabetic condition is done and approximate glucose value is predicted [18].

Ola S.Abadalsalam et al proposed a model created using Rayleigh Scattering Spectroscopy. The relation between scattering angle and the glucose concentration is found out. The performance measure used here are Mean Squared Error and Clarke Error Grid Analysis [19].

Praful P. Pai et al study focuses on photo acoustic (PA) spectroscopy. In vitro experimentation was done with glucose solutions at different NIR wavelengths. Time and frequency domain feature extraction is done and various peak amplitude and area based features were extracted. They found correlation between these features and glucose concentration. Regression analysis was used to find the relation. The performance of the system is calculated using clark error grid analysis [20].

Masab Ahmad et al proposed model consisting of transmittance spectroscopy on the body site ear lobe. Different body parameters like tissue thickness, blood oxygen saturation are considered and combined with a linear regression analysis method to propose real time architecture for glucose measurement [4].

Shraddha Haboo et al describe a method based on photoplethysmography (PPG) and neural networks. Different frequency and time domain features are extracted from PPG signal and single pulse analysis is also done in this work. The extracted features are fed to the neural network as inputs and blood glucose estimation is performed. Clark Error Grid analysis is used for estimation and obtained good accuracy with80.6% samples in class A. [16]

III. METHODOLOGY

Glucose is optically active substance. Many investigators have exploited this property for glucose measurement without puncturing the skin.

Electromagnetic spectrum defines total range of light according to its wavelength. Infrared region is divided into three types namely Near Infrared (NIR), Mid Infrared (MIR) and Far Infrared (FIR). The blood has several components like water, hormones, lipids etc. When NIR light (750-2500nm) passes through any body part , the absorption of light by water, lipid is less as compared to MIR region. So the light can penetrate deeper in the body part [21]-[23]. Therefore, NIR spectroscopy is used for this research work. The proposed system is based on the photoplethysmography (PPG). Fig.1 below shows block diagram of the system.

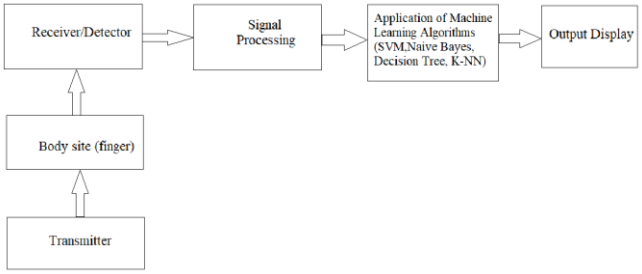

Fig.1. Block Diagram of System

The projected system uses NIR sensors for transmitting and receiving the light rays within the body site. The PPG signal is recorded using NIR sensor attached to fingertip. The output of detector is given to the analog input of arduino board for further analysis.The NIR light is transmitted from NIR transmitter through body site and reflected light is received at the detector output. The output of detector is photoplethysmography (PPG) signal. This signal requires filtering and amplification done by signal processing stage. The data collection is done with 82 patients within age group of 30 to 85. According to diabetic and blood pressure condition patients are classified.

Different frequency and time domain features are extracted from this PPG signal [16].

These features are fed to neural networks for training and prediction of diabetes is done using different machine learning algorithms.

3.1 Different machine learning techniques used:

1. Support Vector Machine (SVM)

Machine Learning algorithms are basically classified into two types, supervised learning and unsupervised learning. SVM is a type of supervised machine learning algorithm. It is widely used for regression analysis and classification. Supervised learning means training the data which is well labeled and with this labeled data predict the output for new data. The algorithm learns from previous inputs. The SVM model represents various classes of input data in a hyper plane in multidimensional space. The hyper plane is generated by algorithm to reduce the error. For two dimensional space the hyper plane is a straight line dividing the plane in two parts.

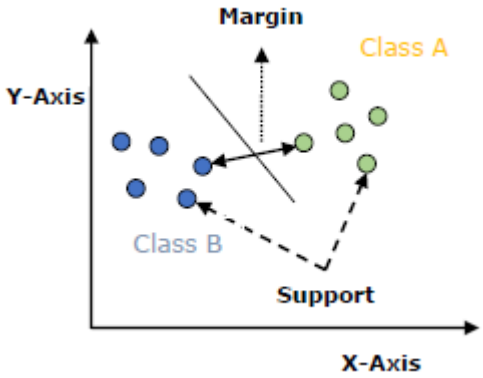

Fig.2. Support Vector Machine (SVM)

The decision plane which separates different data points is the hyper plane. Support vectors are the points very nearer to the line. The distance between two closest points of various classes is given by margin.

2. Naïve Bayes Classifier

Naïve Bayes classifier is another type of machine learning algorithm which is probabilistic and statistics based model. It is purely based on Bayes theorem.

Bayes Theorem:

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

Where,

$P(A|B)$= Posterior Probability

$P(B|A)$= Likelihood Probability

$P(A)$= Prior Probability

$P(B)$= Marginal Probability

The grouping of two words Naïve and Bayes comes out as Naïve Bayes algorithm. Naïve means the existence of any feature is independent on any other feature. As this algorithm is based on Bayes theorem, it is called Naïve Bayes algorithm. The following steps are used for implementation of this algorithm.

1. Frequency Table Creation.
2. Likelihood Table Creation. (P(B|A))
3. Posterior Probability Calculation. (P(A|B)

3. Decision tree algorithm:

A Decision Tree is a simple representation for classifying examples. It is a Supervised Machine Learning where the data is continuously split according to a certain parameter. Decision tree algorithm is designed for regression analysis and classification.

The tree involves of two main entities i.e. leaves and decision nodes. Decision nodes are the points where data splitting is done and the levels are the ultimate outputs. Two main types of decision tree are classification tree and regression tree. In classification tree, output is categorical and in regression tree, output variable is continuous in nature. Entropy calculation and information gain, these two important parameters are considered for implementation of decision tree algorithm.

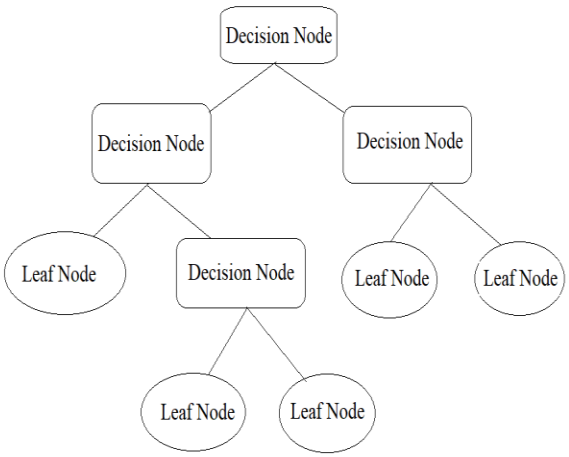

Fig.3. Decision Tree

4. KNN classifier

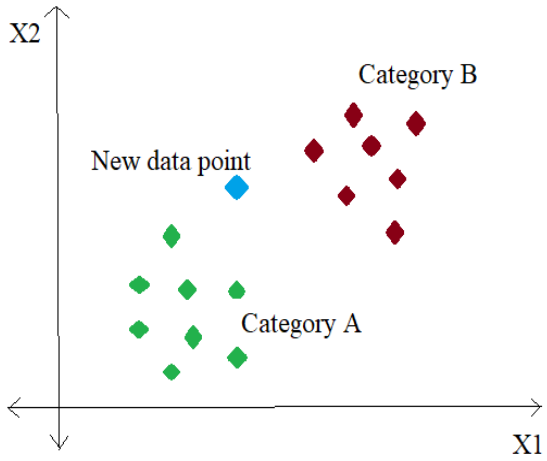

Fig.4. K-Nearest Neighbor (KNN)

One more supervised machine learning algorithm is k-nearest neighbor (KNN) algorithm. It is easy to implement and have simple construction. As it is supervised machine learning algorithm, it depends on

labeled input to learn a function that generate or predict accurate output for new sample of data. This algorithm stores all input data and classify new data point based on similarity. The new data point is assigned to class which has the nearest neighbor (k). As value of k increases, prediction accuracy might increase.

3.2. Database used:

Two databases are used for this research work. First one is created with patients from hospital and another is PIMA Indian diabetes dataset.

Database1:

Database is created with 82 patients having different conditions like non-diabetic, diabetic, with BP within age group of 30 to 85. The NIR sensor is clipped to finger-tip to measure the PPG signal at detector side. The output of detector is given to the analog input of arduino board for further analysis.

Database2:

PIMA Indian diabetes dataset is eventually from the National Institute of Diabetes and Digestive and Kidney Diseases. It is mainly used for prediction of diabetes based on different measures. Different attributes of this database are Number of times pregnant, Plasma glucose concentration, Diastolic blood pressure, Triceps skin fold thickness , 2-Hour serum insulin, Body mass index, Diabetes pedigree function, Age (years) and the last one is output variable i.e. Class variable (0 or 1) [24].

3.3. Accuracy measures:

Different machine learning techniques are used in this research work. All the algorithm implementation is done in Matlab software. Accuracy, F-measure, Precision and Recall are used as performance parameters for different classifiers. All above measures are calculated from confusion metrics data. Different accuracy measures described as follows:

Accuracy: It determines the accuracy of the algorithm for prediction of instances.

Precision: Precision estimates the fraction of correct classified occurrences among the ones classified as positive. It is the ratio of accurate positive results to the number of positive results predicted by the classifier.

Recall: Recall is a metric that enumerates the number of accurate positive predictions made out of all positive predictions that could have been made. It is the ratio of accurate positive results to the number of all relevant samples.

F-measure: F-Measure offers a way to combine together precision and recall into a single measure that captures both properties.

IV. RESULTS AND DISCUSSION

Table I below shows accuracy measures for different classifiers on dataset with 82 patients.

Table- I: Accuracy measures for classifiers (Database1)

Classifier	Precision	Recall	F-measure	Accuracy
SVM	0.8636	0.4634	0.6032	68.29%
Naïve Bayes	0.3636	0.5714	0.4444	75.61%
Decision Tree	0.7727	0.8947	0.8293	91.46%
KNN	0.3636	0.7273	0.4848	78.04%

Table II shows accuracy measures for PIMA Indian diabetes database using different classifiers.

Table- II: Accuracy measures for classifiers (PIMA Indian database)

Classifier	Precision	Recall	F-measure	Accuracy
SVM	0.6700	0.8656	0.7554	71.61%
Naïve Bayes	0.8420	0.8019	0.8215	76.17%
Decision Tree	0.9540	0.9427	0.9483	93.23%
KNN	0.8720	0.8134	0.8417	78.51%

Table-I and Table II represents different performance values of all classification methods calculated on several measures.

From Table I & Table II it is analyzed that Decision tree algorithm is giving the maximum accuracy. Hence the Decision tree machine learning classifier can predict the probabilities of diabetes with more accuracy as compared to other classifiers. Performances of all classifiers based on several measures are plotted via a graph in Fig. 5 and Fig. 6.

Table- III: Classifiers Performance (Database1)

No. of patients	Classifier	Correctly classified	Incorrectly classified
82	SVM	57	25
	Naïve Bayes	62	20
	Decision Tree	75	07
	KNN	65	17

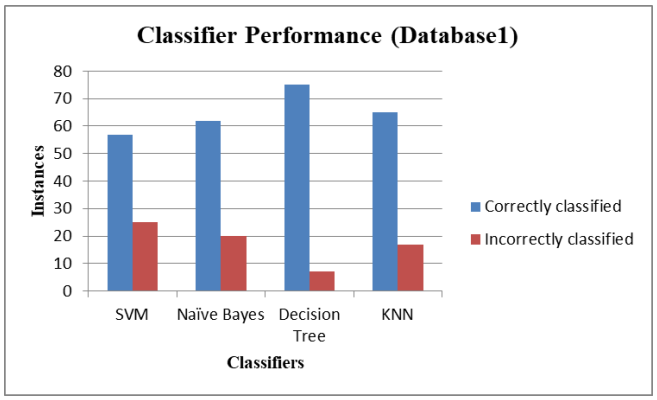

Fig. 5 Classified Instances

Table- III: Classifier's Performance (PIMA Indian Database)

No. of patients	Classifier	Correctly classified	Incorrectly classified
768	SVM	551	217
	Naïve Bayes	585	183
	Decision Tree	716	52
	KNN	604	164

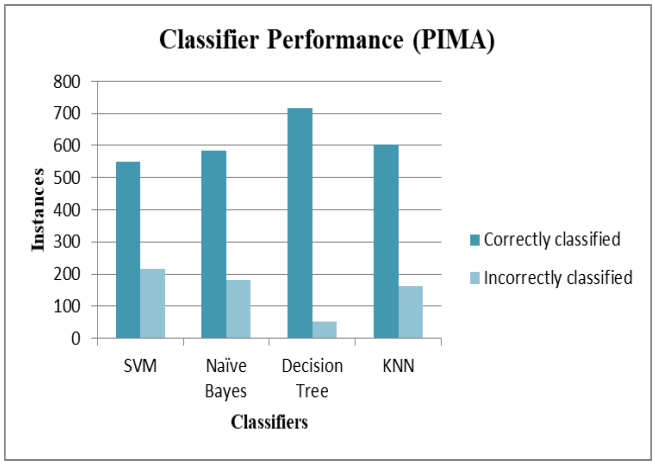

Fig. 6 Classified Instances for Pima database

V. CONCLUSION

Untreated high blood sugar from diabetes can harm your nerves, eyes, kidneys and other organs. Therefore it is extremely important to detect diabetes at an initial stage. 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 91.46% for database1 and 93.23% for database2 (PIMA). The research work can be extended for prediction of actual glucose concentration value using different machine learning algorithms.

ACKNOWLEDGMENT

We would like to thank Bharati Hospital and Research Center, Pune, India for allowing us to collect data in pathology laboratory section.

REFERENCES

1. International Diabetes Federation, About Diabetes, http://www.idf.org/about-diabetes
2. International Diabetes Federation – Facts & Figures, https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html
3. Praful P. Pai, Pradut Kumar Sanki, Arjit De, Swapna Banerjee, "NIR photoacoustic spectroscopy for Non-invasive Glucose Measurement" 37th Annual International Conference of IEEE Engineering in Medicine and Biology Society, 25-29 Aug 2015, Milan, pp 7978-7981
4. Masab Ahmad, AwaisKamboh, Ahmed Khan, "Non-invasive blood glucose monitoring using Near Infrared Spectroscopy", EDN Network, Oct16, 2013.
5. Kiseok Song, Unsoo Ha, Seangwook Park, JoonsungBae, Hui-Jun, "An impedance and multi-wavelength Near-Infrared Spectroscopy IC for Non-invasive Blood Glucose Estimation", IEEE Journal of Solid-State Circuits,Vol 50, No.4, April 2015, pp 1025-1037.
6. JyotiYadav, Asha Rani, Vaijnder Singh, BhaskarMoahnMurari, "Comparative study of Different Measurement Sites using NIR Based Non-invasive Glucose Measurement system",4th ICECCS, Procedia computer Science 70,,pp469-475,2015.
7. Marjan Gusev et al, "Noninvasive Glucose Measurement Using Machine Learning and Neural Network Methods and Correlation with Heart Rate variability" Hindawi, Journal of Sensors Volume 2020.
8. Souad Larabi-Marie-Sainte et al," Current Techniques for Diabetes Prediction: Review and Case Study", MDPI, Applied Science Open Access Journal, October 2019.
9. Deepti Sisodia et al, "Prediction of Diabetes using Classification Algorithms", International Conference on Computational Intelligence and Data Science, 2018.
10. Marjan Gusev et al," Noninvasive Glucose Measurement Using Machine Learning and Neural Network Methods and Correlation with Heart Rate Variability" Hindawi Journal of Sensors, vol 2020.
11. CF SO et al, "Deep Learning Analysis for Blood Glucose Monitoring Using Near Infrared Spectroscopy" Biomedical Journal of Scientific & Technical Research, vol 21, issue 3, 2019.
12. A. Thammi Reddy et al, "Minimal Rule-Based Classifiers using PCA on Pima-Indians-Diabetes-Dataset", International Journal of Innovative Technology and Exploring Engineering, Volume-8 Issue-12, October 2019.
13. Changsheng Zhua," Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques", Informatics in Medicine Unlocked, Science Direct, April 2019.
14. M.S. Barale et al," Cascaded Modeling for PIMA Indian Diabetes Data", International Journal of Computer Applications, vol 139, April 2016.
15. N. Sneha et al,"Analysis of diabetes mellitus for early prediction using optimal features selection", Sringer open Journal of Big data, 2019.
16. Shraddha Habbu et al," Estimation of blood glucose by non-invasive method using photoplethysmography" Sadhana, Indian Academy of Science, May 2019.
17. Hussan Ali Al naam, Mohamed Osman, Idress, AbdalsalamAwad, Ola S. Abdalsalam,Frangoon Mohamed, "Non-invasive Blood Glucose Measurement Based on Photo-acoustic Spectroscopy", International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering,7-9 sep 2015,khartoum,pp-1-4.
18. Abishek Thekkeyil Kunnath et al "Voltage intensity based noninvasive blood glucose monitoring", ICCCNT, 2013.
19. Ola S. Abdalsalam et al "Design of simple noninvasive glucose measuring device", International conference on computing, electrical and electronics engineering, 2013.
20. Praful P. Pai et al, "Accuracy enhancement for noninvasive glucose estimation using dual-wavelength photoacoustic measurements and kernel based calibration", IEEE transaction on measurement, vol 67, no 1, January 2018.
21. Jonas Kottman, JulienM.Ray, Markus W. Sigrist "Mid-infrared Photoacoustic detection of glucose in human skin: Towards noninvasive diagnostics", sensors 2016, MDPI, Switzerland.
22. O.S.Khalil, "Non-invasive Glucose measurement Techniques: An update from 1999 to the Dawn of the new millenium", Diabetes Technology Therapeatics, Vol.6, no.5, 2004, pp 660-697
23. Alexandra Werth et al, "Noninvasive glucose measurements in skin using Mid-IR quantum cascade laser spectroscopy" IEEE 2017.
24. PIMA Indian Database, http://networkrepository.com/pima-indians-diabetes.php
25. A.A.Shinde, R.K.Prasad "Non Invasive Blood Glucose Measurement using NIR technique based on occlusion spectroscopy " International Journal of Engineering Science and Technology (IJEST) Vol. 3 No.12 December 2011.
26. Pamshangphy Raikham et al, "Noninvasive blood components measurement using optical sensor system interface", 3rd international conference on microwave and photonics, 9-11 Feb 2018.
27.] Aishwarya Lakshmi K et al "Studies on relating to monitoring blood glucose levels using noninvasive optical methods"IEEE conference on recent trends in electronics information& communication technology, May 19-20, 2017.
28. Heungjae Choi et al, "Recent developments in minimally and truly noninvasive blood glucose monitoring techniques" IEEE 2017.
29. JyotiYadav, Asha Rani, Vijander Singh, "Near-Infrared LED based Non-Invasive Blood Glucose Sensor", International conference on Signal Processing and Integrated Network (SPIN),20-21 feb 2014

AUTHORS PROFILE

Vandana Bavkar is currently working as research scholar in Bharati Vidyapeeth (Deemed to be University), Pune, India. She has obtained Master of Engineering from MIT College of Engineering Pune in 2013. She has published research paper on Spectral Analysis of Blood Glucose in 2018. She is member of IEEE. Her area of research includes Biomedical Engineering, Signal Processing, and Neural Networks.

Dr. Arundhati A. Shinde is currently working as Head of Department in Bharati Vidyapeeth (Deemed to be University), College of Engineering, Pune, India. She has obtained Ph. D from BVDU. She has published 18 research papers in national and international journals. Her research interest includes Instrumentation & Control, Digital Signal Processing, Embedded system and Biomedical Signal Processing.