Comparison of Noninvasive Blood Glucose Estimation Using Various Regression Models

Shraddha K. Habbu[1,2,3](✉), Manisha P. Dale[4], Rajesh B. Ghongade[5], and Shrikant S. Joshi[2]

[1] Department of E&TC Engineering, All India Shri Shivaji Memorial Society's Institute of Information Technology, Pune, India
shraddha.habbu@viit.ac.in

[2] Department of E&TC Engineering, Vishwakarma Institute of Information Technology, Pune, India
shrikant.joshi@viit.ac.in

[3] Department of E&TC Engineering, Savitribai Phule Pune University, Pune, India

[4] Department of E&TC Engineering, Modern Education Society's College of Engineering, Pune, India
mpdale@mescoepune.org

[5] Department of Electronics Engineering, Bharati Vidyapeeth's College of Engineering, Pune, India
rbghongade@gmail.com

Abstract. In this work we aim to evaluate the performance of three types of machine learning models implemented for blood glucose level (BGL) estimation. Pulse photoplethysmography signal is acquired from 611 human subjects and used for the analysis. Time and frequency domain features are extracted using (1) Frame based and (2) Single Pulse Analysis technique. These two features are used as input to train neural network, Support vector machines and Random forest models. These trained models are used for the estimation of BGL values. The BGL estimation performance of these models, (i) neural network, (ii) SVM, (iii) RF and (iv) K-fold RF are compared based on two feature sets i.e. Frame based time and frequency domain features and Single Pulse Analysis based time and frequency domain features. The performance of each system model is evaluated on the basis of, (i) Coefficient of determination i.e. R^2, (ii) Spearman's coefficient of correlation, (iii) Pearson's coefficient of correlation and (iv) Clarke error grid analysis. We observed that Single Pulse Analysis technique shows better performance as compared to Frame based technique. The highest R^2 value (0.95) for Single Pulse analysis is obtained for K-fold RF network. For Single Pulse Analysis, all other models also show comparable BGL estimation accuracy with R^2 values ranging from 0.91 to 0.94. According to Clarke error grid analysis the values that lie in class A and class B are clinically accepted. We obtained highest prediction accuracy for Single Pulse analysis with K-fold random forest with 93.2% (class A) and 6.8% (class B).

Keywords: Photoplethysmograph (PPG) · Neural network · Noninvasive blood glucose measurement NIBGM · Blood glucose level (BGL) · Support vector machines (SVM) · Random forest (RF) · Single Pulse Analysis (SPA)

S. M. Thampi et al. (Eds.): SIRS 2019, CCIS 1209, pp. 306–318, 2020.
https://doi.org/10.1007/978-981-15-4828-4_25

1 Introduction

The need of self-monitoring of blood glucose levels is very crucial in diabetics. Currently glucometers that are available in market are either invasive or minimally invasive. The diabetic patients are in need of frequent monitoring of their blood glucose levels to detect the hyper or hypoglycemia events [1]. The glucometers available in market require blood sample which is obtained through puncture of the skin by a lancet. Frequent piercing causes discomfort to the patients. According to 2016 WHO survey report 422 million people were estimated infected by diabetes by 2014 which is 8.5% rise compared to 2012 survey [2]. This shows that mortality rate is increasing rapidly. With the increasing number of infected people by diabetes there is a need to develop a noninvasive method of blood glucose level measurement. However, the last couple of decades, work towards the development of a non-invasive glucose monitor has increased significantly among research groups with promising results. Many noninvasive techniques of blood glucose estimation have been studied and implemented, stating their particular benefits and drawbacks [3]. Noninvasive glucose determinations can be classified into optical and non-optical techniques. The optical properties of glucose are rather specific and these methods have shown better results and a better correlation with blood glucose content. With the advancements in technology, optical based techniques have gained lot of importance in development of biomedical instruments. Photoplethysmography (PPG) is an optical based technique which is used in the development of advanced health care. Various parameters like blood pressure (BP), respiratory rate, stroke volume, pulse transit time (PTT), heart rate variability (HRV), arterial stiffness, and blood glucose levels can be analyzed using PPG technique [1, 3–6].

Moreno et al. [6] presented a system for a simultaneous non-invasive estimate of the blood glucose level (BGL) using a Photoplethysmograph (PPG) and machine learning techniques. A comparative results of blood glucose estimation using various regression technique such as Linear Regression, Neural network, Support vector machines and Random forest are presented by the author. The coefficient of determination (R^2) obtained using linear regression (0.52), neural network is 0.54, Support Vector Machines (0.64) and 0.88 using Random forest regression technique.

Moreno et al. [6] uses 410 subjects PPG data and computed frequency domain features of fixed duration windowed PPG signal for the estimation of BGL values. The importance of the use of Single Pulse Analysis technique is motivated by the fact that the PPG signal looks similar to arterial pressure pulse but the wave contour does not remain the same [7]. In this technique we separate pulses from the entire signal and then extract time domain features of each single pulse. It captures the time varying nature of the pulse. In the current work, we used Single Pulse based time domain features.

In this paper, we present photoplethysmography based BGL estimation using two signal analysis techniques as (i) Frame based and (ii) Single Pulse Analysis. Time and Frequency domain features are computed and are used as input to train neural network, Support Vector Machine and Random Forest models and comparison of BGL estimation is performed using these models. A part of this work is funded by Board of Colleges and University Development (BCUD). Savitribai Phule Pune University, Pune. Development of data acquisition system, data collection work and preliminary experimentation work is carried out under this funding.

2 Methodology

2.1 System Implementation

Figure 1 shows the block schematic of implementation of BGL estimation system based on Frame based and Single Pulse Analysis technique. The PPG acquisition system is built using an SPO2 sensor which is operated in near infrared region [8]. This sensor consists of an optical transmitter (LED) and a receiver (pin photodiode) and is placed along the finger tip. The reflected signal is resultant of volumetric change observed in blood during systole. This change is detected by the detector and is amplified by signal conditioning circuitry. The amplified signal is PPG signal which is then interfaced to a processor to generate digitized pulsed data for recording and storing the data. The details of pulse data acquisition is explained in detail in our earlier research work reference paper [9]. The PPG data of 611individuals is recorded over a 3 min duration each. Using this system the PPG signal of the individuals are recorded along with their BGL values on the Accu-Check® [10] machine simultaneously. The range of BGL values varied from 70 to 450 mg/dl. The database details can be referred in reference [9]. From the collected data we created training dataset and testing dataset which could be used for experimentation. We extracted multiple instances of one minute window out of three minute recorded PPG signal with suitable time overlap. With this process we obtained around 1900 window segments of 1 min duration. By randomly observing the window segments we selected 1000 window segments which are noise, power line interference free and useful for training purpose and 500 samples for testing purpose. The details of training and testing dataset are provided in Table 1. DB3 is a combined training set formed by adding DB1 and DB2 consisting of 1000 samples and DBtest is the testing set consisting of 500 different samples. Preliminary experimentation were carried out using the three datasets and performance evaluation of baseline testing were carried out.

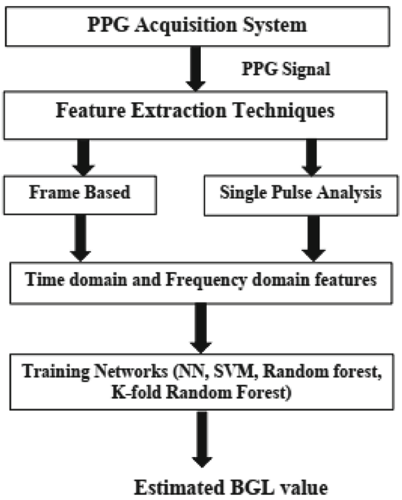


Fig. 1. System implementation

The best results were obtained using DB3 [9]. Hence DB3 dataset is used as a training dataset for further experimentation and DBtest is the testing dataset.

Table 1. Training and testing dataset

Training dataset	Testing dataset	
DB1 (500)	DB2 (500)	DBtest
DB2 (500)	DB1 (500)	DBtest
DB3 (1000)	DBtest	

2.2 Feature Extraction Techniques

The extracted PPG signal is a digitized signal of 1 min duration with a sampling rate of 100 samples per second giving a total of 6000 samples. To remove the noise components generated through power line interference, motion artifacts, muscle artifacts, high frequency artifacts and low amplitude the preprocessing of the signal is carried out. The base line interference is removed using moving average filter and the PPG signal is smoothed using Savitzky Golay filter [11]. The feature vector consists of features computed at window level and local features computed at frame level. The feature set in [6] consists of 33 dimensional features. We excluded age, weight, body mass index and oxygen saturation features to eliminate the person specific dependency. This reduced feature set with 29 features is called baseline feature set.

Frame Based Feature Set: We extracted features based on fixed duration frame with 500 samples per frame. Computed 29 baseline features as mentioned above for each frame and additional features such as pulse transit time, pulse interval and pulse amplitude of the PPG signal along with their statistical measures like mean, variance standard deviation and interquartile range are added to it. This results into 35 dimensional feature set. The details about the computation of Autoregressive coefficients, Kaiser Teager Energy, Heart rate, Spectral Entropy, Log energy profile, Pulse interval, Pulse transit time, and Pulse amplitude can be obtained in our previous research work presented in [9]. We represent this feature vector as FV and is given by

$$FV = \{FVAR_{PPG},\ FVKTE,\ FVHr,\ FVHs, FVLogE, VPTT, FVP\text{int}, FVPamp\}$$

Single Pulse Analysis Feature Set: As mentioned earlier, the PPG signal looks similar to arterial pressure but wave contour of pulse does not remain same hence to capture the variations in pulse wave shape of the entire signal we implemented single pulse technique. We split the pulses by finding the local minima and extracted features of each pulse based on energy, pulse interval, pulse transit time, pulse start time, pulse end time, peak amplitude along with its statistical measures like mean, variance, standard deviation and interquartile range. This feature set consists of 28 features and is given as

$$FV_{new} = \{PFV_n^{\mu}, PFV_n^{\sigma}, PFV_n^{skew}, PFV_n^{iqr}\}$$

The details about the computation of the above feature set is explained in our research work presented in [9].

2.3 BGL Estimation

We implemented BGL estimation using time domain feature set with 35 features and Single pulse feature set with 28 features using various regression techniques. As explained in Sect. 2.1 we created training dataset consisting of 1000 samples and testing dataset of 500 samples useful for training purpose and for testing purpose. The features of training dataset along with their BGL values are then given as input to various training models like neural network, support vector machine, random forest and k-fold random forest generating the training model as explained in further sections. A nonlinear regression model is trained by each network. These networks are then used for testing dataset to estimate their BGL values. Also the performance of each network is further analyzed based on Clarke error grid analysis and coefficient of determination.

Neural Network: For neural network we used Multilayer perceptron trained by Levenberg Marquart algorithm. The Multilayer perceptron is able to approximate nonlinear relationships. We used 3 hidden layers neural network topology with 30, 20, 10 neurons in layer 1, 2 and 3 respectively to train the network using nonlinear regression. Matlab 2017b neural network toolbox is used for neural network training and testing purpose.

Support Vector Machines: SVM regression is implemented using Matlab 2017 SVM tool box. We used Gaussian Kernel based on the earlier results reported in [6] which had shown best results. The specifications of kernel (kernel type, soft margin, and kernel parameters) were set based on the quality criterion i.e. R^2 value on the testing database.

Random Forest: Random forests are a ML technique based on a Set (forest) of classification and regression trees [12], each trained in a different way. The output of the system is an aggregation of the outputs of the trees of the forest. The methodology is justified by the fact that the mathematical expression of the error of a classifier depends on the bias and variance of each of the trees of the forest. In order to take advantage of the fact that averages reduce the bias and variance, the algorithm introduces a systematic controlled variability and bias [12]. The advantages of this technique is (i) low global error rate as the mean output reduces the variance and bias of the estimate, (ii) the execution time is low as the base classifier is a tree and (iii) computations are based on comparisons at each node. Also each node compares only one feature, making the system robust with respect to the scale of the inputs and the correlations between them [6]. The best configuration (number of trees in the forest, number of features tested at each node and maxi mum number of samples at a given node) was determined based on the quality criterion i.e. R^2 value on the testing database. The algorithm was programmed in Matlab using *'fit-r-ensemble'* function. We also implemented cross validation on testing dataset using K-fold random forest technique with 10 folds and maximum decision tree number 50. The same was also implemented in Matlab.

3 Experimentation and Results

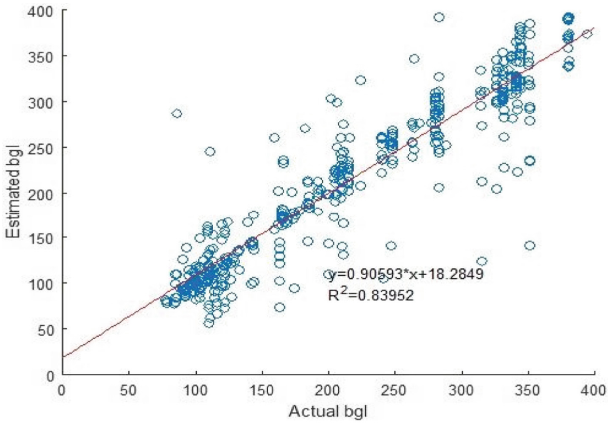
In this work we performed BGL estimation using two feature extraction techniques as (1) Frame based and (2) Single pulse analysis. The performance of BGL estimation is evaluated using coefficient of determination (i.e. R^2), computed via four different regression models namely neural network, SVM, Random forest and K-fold Random Forest. We also analyze the performance based on Clarke Error Grid Analysis. Clarke Error Grid Analysis is clinically accepted for validation of BGL estimation [13]. The grid is divided into 5 regions. Clarke grid scatter is plotted using Matlab function [14]. Region '**A**' represents prediction within 20% of the actual BGL value. Region '**B**' represents prediction more than 20% away from actual BGL but do not give false predictions. Region '**C**' represents false positives of either cases of hypoglycemia or hyperglycemia. Region '**D**' represents predictions that fail to detect cases of hypoglycemia or hyperglycemia. Region '**E**' represents prediction errors which could wrongly classify cases of hypo or hyperglycemia. Following section gives the details about the results of BGL estimation using four different regression models for 2 features sets i.e. (1) Framed based time and frequency features and (2) Single Pulse Analysis based time and frequency features. Table 2 gives the comparative performance measures of all the four regression models for baseline feature set with the reference [6]. Table 3 gives the comparative performance measures of all the 4 regression models for time and frequency domain and Single Pulse Analysis feature set. Figures 2, 3, 4 and 5 show the computation of R^2 and Clarke Error grid Analysis using 4 different regression models for Single Pulse analysis technique.

Table 2. Comparative BGL estimation performance of 3 different regression models NN, SVM and RF of baseline testing

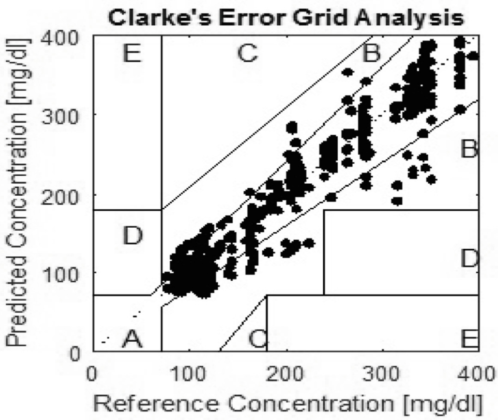
	NN	SVM	RF	K-fold RF
Monte Moreno [6] (33 features) R^2	0.54	0.64	0.88	–
Baseline system (29 features) R^2	0.71	0.78	0.83	0.93

Table 3. Comparative BGL estimation performance of 4 different regression models NN, SVM RF and K-fold RF for Frame based and Single Pulse Analysis.

	NN frame based	NN Single Pulse Analysis	SVM Frame Based	SVM Single Pulse Based
R^2	0.83	0.91	0.74	0.94
A	403 (80.6)	415 (83)	340 (68)	434 (86.8)
B	87 (17.4)	85 (17)	150 (30)	66 (12.2)
C	5 (1)	0	10 (2)	0
D	5 (1)	0	0	0
E	0	0	0	0
Spearman coefficient	0.886	0.948	0.909	0.970
Pearson coefficient	0.8794	0.954	0.905	0.977
	RF Frame based	RF Single Pulse Based	K-fold Frame based	K-fold Single Pulse Based
R^2	0.84	0.91	0.91	0.95
A	395(79.4)	424 (84.8)	447 (89.4)	466 (93.2)
B	96 (19.2)	75 (15)	53 (10.6)	34 (6.8)
C	3 (0.6)	0	0	0
D	6 (1.2)	1 (0.2)	0	0
E	0	0	0	0
Spearman coefficient	0.895	0.928	0.944	0.957
Pearson coefficient	0.893	0.937	0.947	0.965

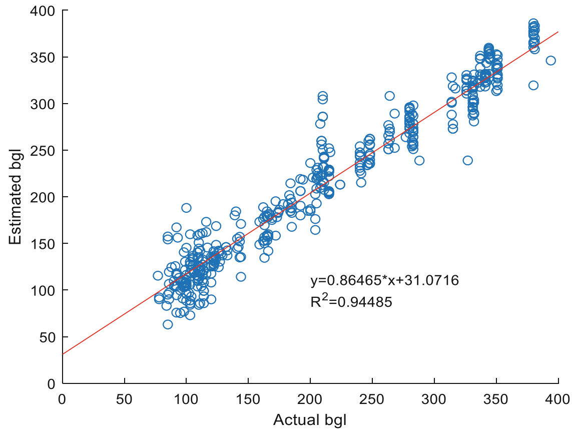


(a)

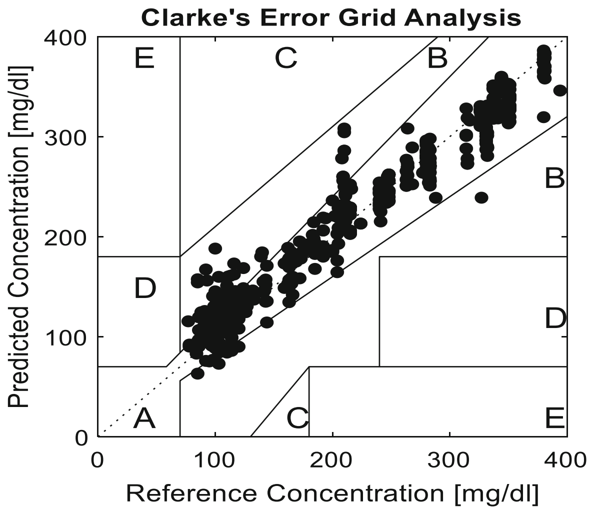


(b)

Fig. 2. Performance metric of Single Pulse analysis for nonlinear regression technique using NN (a) Coefficient of determination (i.e. R^2) and (b) Clarke error grid.

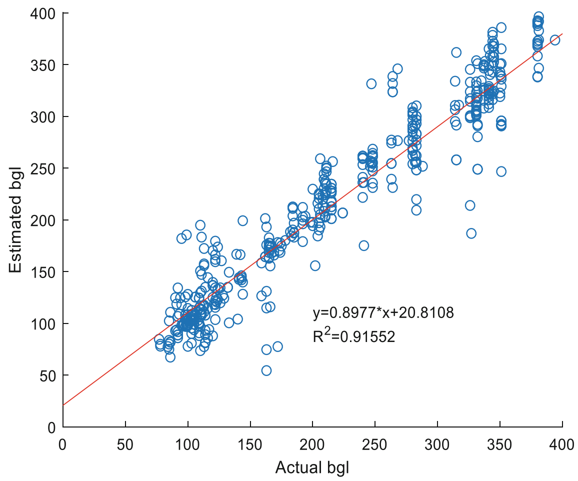


(a)

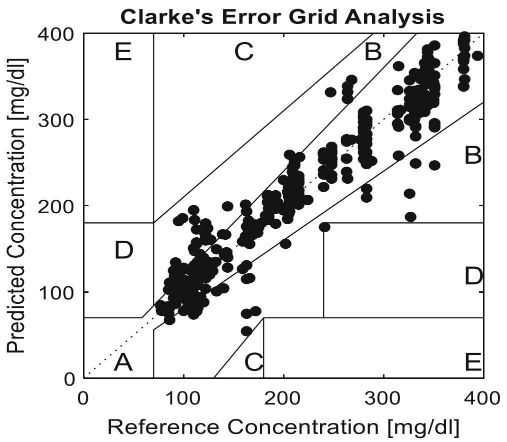


(b)

Fig. 3. Performance metric of Single Pulse analysis for nonlinear regression technique using SVM (a) Coefficient of determination (i.e. R^2) and (b) Clarke error grid.

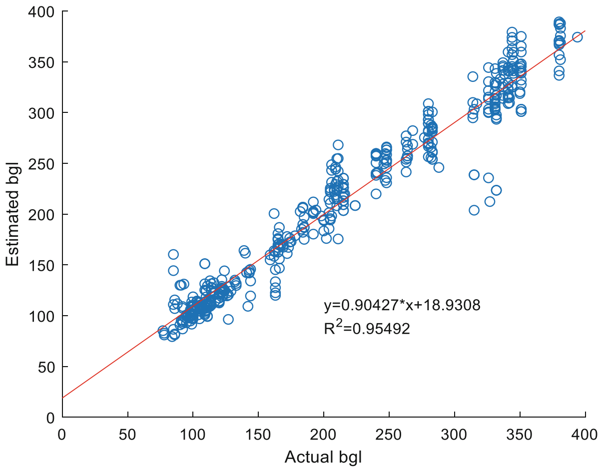


(a)

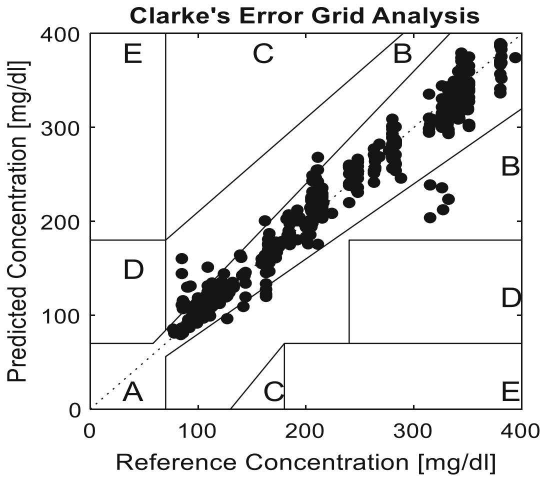


(b)

Fig. 4. Performance of Single Pulse analysis for nonlinear regression technique using RF (a) Coefficient of determination (i.e. R^2) and (b) Clarke error grid



(a)



(b)

Fig. 5. Performance metric of Single Pulse analysis for nonlinear regression technique using K-fold (a) Coefficient of determination (i.e. R^2) and (b) Clarke error grid.

4 Observations and Discussion

From Table 2, it is observed that we obtained comparable baseline results with the results presented in the literature [6]. By eliminating the person specific dependent features and reducing the feature size there is improvement in R^2.

Table 3 represents comparative performance evaluation of machine learning models based on two PPG signal analysis techniques i.e. (i) Frame based and (ii) Single Pulse Analysis. Single Pulse Analysis technique outperforms Frame based technique for all performance evaluation metrics and this is also true for all machine learning models i.e. NN, SVM, RF and K-fold RF. The performance of Single Pulse Analysis shows highest accuracy via R^2, Spearman and Pearson correlation coefficient performance metrics. High value of Spearman's and Pearson's correlation coefficient shows that the estimated and actual BGL values show monotonic and linear relationship. The results obtained for Single Pulse Analysis with neural network shows improvement in BGL value estimation ($R^2 = 0.84$ for Frame based and 0.91 for Single Pulse technique) as compared with reference result ($R^2 = 0.54$) obtained by Monte et al. [6].

The performance of machine learning models (NN, SVM, RF, K- fold RF) for Single Pulse analysis via R^2 are comparable and lie in range 0.91 to 0.95. According to Clarke error grid analysis we obtained the highest estimation accuracy for K-fold random forest with 93.2% prediction cases lying in class A and 6.8% cases in class B. The Random forest technique uses multiple trees which reduces the risk of over fitting. For the system performance validation purpose cross validation is very important as it ensures that each test sample occur in training and testing phase. This eliminates the chance of obtaining very good or bad results. With 10 fold cross validation we obtained R^2 value of 0.95. The accuracy obtained using SVM is also comparable with K-fold Random forest.

5 Conclusion

We investigated the performance of different machine learning models which includes (1) Neural network, (2) SVM, (3) RF and (4) K-fold RF in BGL estimation. We explored two feature extraction techniques as (1) Frame based and (2) Single Pulse Analysis. From the results it is concluded that Single Pulse Analysis technique shows high accuracy as compared with Frame based technique. Using Single Pulse Analysis all the trained models have shown comparable results. Random forest has shown highest accuracy as this method operates by constructing multiple decision trees during training phase. As SVM and RF models shows comparable results, these models can be used for noninvasive BGL real time system implementation.

References

1. So, H.-F., Choi, K.-S.: Recent advances in Noninvasive glucose monitoring. Med. Devices (Auckl) **5**, 45–52 (2012)
2. World Health Organization: Global status report on diabetes 2016. WHO Press, Geneva, Switzerland (2016)

3. Ciudin, A., Hernandrez, C., Simo, R.: Non-invasive Methods of Glucose measurement: Current status and future Perspectives. Curr Diabetes Rev. **8**(1), 48–54 (2012)
4. Fortino, G., Giampa, V.: PPG based methods for noninvasive and continuous blood pressure measurement: an overview and development issues in body sensor networks. In: IEEE Transaction (2010)
5. Poon, C., Zhang, Y.: Cuffless and noninvasive measurements of arterial blood pressure by pulse transit time. In: 27th Annual International Conference on IEEE EMBS (Engineering in Medicine and Biology Society), Shanghai, China, 17–18 January 2005
6. Moreno, E., et al.: Noninvasive estimation of blood glucose and blood pressure from PPG by means of machine learning techniques. J. Artif. Intell. Med. **53**(2), 127–138 (2011)
7. Millasseau, S., Guigui, F., Kelly, R., Prasad, K., Cockcroft, J., Ritter, J.: Noninvasive assessment of the digital volume pulse. Comparison with the peripheral pressure pulse. Hypertension **36**, 952–956 (2000)
8. Jeon, K.J., Hwang, I.D., Hahn, S., Yoon, G.: Comparison between transmittance and reflectance measurements in glucose determination using near infrared spectroscopy. J. Biomed. Opt. **11**, 014022 (2006)
9. Habbu, S., Dale, M., Ghongade, R.: Estimation of blood glucose by noninvasive method using photoplethysmography. Sādhanā **44**, 135 (2019). https://doi.org/10.1007/s12046-019-1118-9
10. Accu-chek Aviva. https://www.accuchek.com/us/glucose-meters/aviva.html. Accessed 2 May 2011
11. Schafer, R.W.: What is a Savitzky-Golay filter? (Lecture notes). IEEE Signal Process. Mag. **28**(4), 111–117 (2011). https://doi.org/10.1109/msp.2011.941097
12. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
13. Clarke, W., Cox, D., Gonder-Frederick, L.A., Carter, W., Pohl, S.: Evaluating clinical accuracy of systems for self-monitoring of blood glucose. Diabetes Care **10**(5), 622–628 (1987)
14. Codina, E.G.: Clarke error grid analysis. Matlab central file exchange (2008). http://www.mathworks.com/matlabcentral/fileexchange/20545-clarke-error-grid-analysis. Accessed 2 May 2011