

Estimation of blood glucose by non-invasive method using photoplethysmography

SHRADDHA HABBU^{1,2,*}, MANISHA DALE³ and RAJESH GHONGADE⁴

¹Department of Electronics and Telecommunication, AISSMS-Institute of Information Technology, Pune, India

²Department of Electronics and Telecommunication, Vishwakarma Institute of Information Technology, Pune, India

³Department of Electronics and Telecommunication, Modern Education Societies College of Engineering, Pune, India

⁴Department of Electronics and Telecommunication, Bharati Vidyapeeth's Deemed University College of Engineering, Pune, India

e-mail: shraddha.habbu@viit.ac.in; mpdale@mescoepune.org; rbghongade@gmail.com

MS received 29 January 2019; revised 13 March 2019; accepted 16 March 2019; published online 6 May 2019

Abstract. This paper presents a system which estimates blood glucose level (BGL) by non-invasive method using Photoplethysmography (PPG). Previous studies have shown better estimation of blood glucose level using an optical sensor. An optical sensor based data acquisition system is built and the PPG signal of the subjects is recorded. The main contribution of this paper is exploring various features of a PPG signal using Single Pulse Analysis technique for effective estimation of BGL values. A PPG data of 611 individuals is recorded over duration of 3 minutes each. BGL value estimation is performed using two types of feature sets, (i) Time and frequency domain features and (ii) Single Pulse Analysis (SPA). Neural network is trained using above mentioned proposed feature sets and BGL value estimation is performed. First we validate our methodology using the same features used by Monte Moreno in his earlier work. The experimentation is performed on our own dataset. We obtained comparable results of BGL value estimation as compared with Monte Moreno, with maximum $R^2 = 0.81$. Further, BGL estimation using (i) Time and frequency domain features and (ii) Single Pulse Analysis (SPA) is performed and the resulting coefficient of determination (i.e., R^2) obtained for reference vs. prediction are 0.84 and 0.91, respectively. Clarke Error Grid analysis for BGL estimation is clinically accepted, so we performed similar analysis. Using Time and frequency domain feature set, the distributions of data samples is obtained as 80.6% in class A and 17.4% in class B. 1% samples in zone C and Zone D. For Single Pulse Analysis technique (SPA) the distribution of data samples are 83% in class A and 17% in class B. The proposed features in SPA have shown significant improvement in R^2 and Clarke Error grid analysis. SPA technique with the proposed feature set is a good choice for the implementation of system for measurement of non-invasive glucometer.

Keywords. Blood glucose measurement; non-invasive; blood glucose level (BGL); neural network; photoplethysmograph (PPG); single pulse analysis (SPA).

1. Introduction

Diabetes is the breakdown of body's ability to regulate amount of glucose (sugar) in the blood. It is a chronic disorder that currently has no cure. High glucose levels or hyperglycemia can lead to serious complications. Patients suffering from Type I and Type II diabetes need constant monitoring of their blood glucose levels so that proper insulin dosage can be given to them. Currently all glucometers available in market are of invasive type which are cumbersome and painful. Hence there is a need for coming

up with non-invasive techniques for evaluating the blood glucose levels which will ease the patient from discomfort due to frequent piercing. Due to its enormous potential to support the monitoring of patient's health at home and in the hospital the interest in computer based advanced health care system has significantly increased in the last few decades and to aid to this Monte Moreno proposed an alternative method of noninvasive measurement of blood pressure and blood glucose levels based on photoplethysmography [1, 2].

The research work on non-invasive automatic glucose estimation began in 1974 by March [3]. Since then many researchers have explored non-invasive techniques for blood

*For correspondence

glucose measurement. So *et al* [4] presented various non-invasive glucose monitoring techniques. Non-invasive 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. Optical based techniques include spectroscopy (thermal, fluorescence, Raman, Mid Infra-red and near Infra-red), optical rotation property and photoplethysmography [5, 6]. The limitation of thermal spectroscopy is that variations in tissue temperature could alter cutaneous vascular and refractive index responses and due to this fact, the method has an important limitation. The detection of fluorescence when exciting a glucose solution by an ultraviolet laser light at 308 nm depends on the glucose concentration in the solution. In humans, the fluorescence phenomenon depends not only on glucose concentration, but on skin pigmentations and epidermal thickness and so far its liability has been demonstrated only *in vitro* experiments. Raman spectroscopy technique is impractical to be used in the clinical practice for development of portable and handy glucose monitoring device. Mid infrared (MIR) light penetration in the skin is of only few micrometers which is a significant limiting factor. Also the MIR spectrum is significantly dependent on the water content, making this method very sensitive to the degree of hydration. Thus spectroscopy based techniques are of interest to estimate and validate the correlation between the blood glucose and parameters obtained by optical principles. With respect to spectroscopy, Near Infra-red (NIR) overcomes all above limitations so special emphasis is given to non-invasive blood glucose measurement (NIBGM) based on IR spectroscopy [6].

Thus, use of an NIR based optical sensor in determination of blood glucose content to provide a low cost, non-invasive solution is our primary focus. Photoplethysmography (PPG) is a simple, optical based non-invasive technique used in the development of advanced health care [7, 8]. The PPG signal is recorded with a sensor similar to pulse oximeter which works only in near infrared spectrum. A sensor is built with a pair of photo transmitter and photo receiver operating at near infrared region (920 nm). A PPG signal is obtained by illuminating the skin and measuring changes in the light absorption at this wavelength. The absorptions depend on the heart beat because the blood vessels in the finger expand and contract with each heart-beat. The reflected signal is reproduced as pulse wave near the detector. 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, 8–11]. Ramashyamam *et al* [12] performed non-invasive blood glucose estimation using NIR spectroscopy using transmission photoplethysmography (PPG). The NIR spectroscopy has been performed based on NIR LED and photo detector constituting an optode pair, the analog front end system has been implemented to get the PPG signal at

the near infra-red wavelengths of 1070 nm, 950 nm, 935 nm. The PPG signal that has been obtained is processed and double regression analysis has been carried out with the artificial neural network using FPGA for estimating the glucose levels. Monte-Moreno [1] presented a system for a simultaneous non-invasive estimate of the blood glucose level (BGL) using a Photoplethysmograph (PPG) and machine learning techniques. The signal is sensed using a photoplethysmography technique, an activity detection module selects a window of 1 min duration (4500 samples) containing a clean signal. The module eliminates the artifacts and prevents loss of signal caused by movement of finger. The signal processing module extracts the main features and forms the feature vector. The output of this module is a fixed length vector. Finally, a machine learning module infers the function that relates the output of signal processing module to BGL. The system is trained and tested with cross validation in a population of 410 individuals. Various machine algorithms such as Linear Regression, Neural network, Support vector machines and Random forest are implemented on novel data and coefficient of determination is computed. Random forest regression technique shows less variability on the cross validation amongst the other three measures with an $R^2 = 0.88$ and overall system performance with $R^2 = 0.9$.

The feature vector computed in [1] is at window and frame level. The difficulty that lies here is that if the frame length is changed, the pulse features also changes. The start and end point detection of the pulse waveform in the frames may change according to the frame duration which would also affect the computed statistical feature vectors. Here, we propose an algorithm which detects each pulse in a frame with a fixed start and end point. Here we consider the periodicity of pulse and local minima for separating the pulses within a given window. This technique is called as SPA (single pulse analysis). Thus pulse variations are more effectively observed in this technique and the features corresponding to these variations are computed. The feature vectors act as an input to neural network and BGL values are estimated using nonlinear regression. The BGL estimation is improved by a significant factor for nonlinear regression implemented using neural network. This signifies that the SPA contributes effectively in BGL estimation.

In this paper, we present comparative results of the technique implemented in [1], along with our proposed algorithm on a dataset of 611 individuals. Initially, we present the comparative results obtained with the features referred from [1], along with some additional features used to train the neural network. Features like age, weight, saturation oxygen level, and body mass index have been eliminated from the referred feature vector as we feel that these features would add as person-specific dependent parameters for estimation. To reduce the dependency of these features in the BGL estimation we eliminated them from feature vector. It is observed that addition of few new features in time domain improves the correlation factor. We considered this as the

basis to investigate single pulse features in time domain. The final comparative results are presented by implementing single pulse analysis technique with neural network. Using neural network, nonlinear regression shows improvement in the BGL estimation performance. A part of this work was funded by Board of College and University Development (BCUD) Savitribai Phule Pune University. The data collection, preliminary experimentation work of finalizing feature sets and optimizing neural network topology is carried out under this work [13]. With the help of the framework developed in this study, further experimentations on increased dataset, additional feature sets and fine-tuned neural network topology are carried out and the results are presented in this work.

2. Proposed work

The proposed work is divided into four parts: (1) System Architecture, (2) Database, (3) Methodology, and (4) BGL estimation experimentation, as explained in the following sub-sections.

2.1 System architecture

The proposed system architecture is as shown in figure 1. The system consists of a Pulsed Data Acquisition module which records the PPG signal of the subject. Simultaneously the BGL value measurement of the subject is made using self-monitoring blood glucometer Accu-Check® machine [14], as an actual BGL reference value. The database consisted of 611 individuals whose PPG signals are recorded for 3 minute duration along with their BGL values. A pre-processing module eliminates the various artifacts from the recorded signal and a clean 1 minute signal is extracted. Further, various features relating to alterations in PPG signals are computed. These feature vectors along with target BGL values act as an input to the neural network which is trained with nonlinear regression.

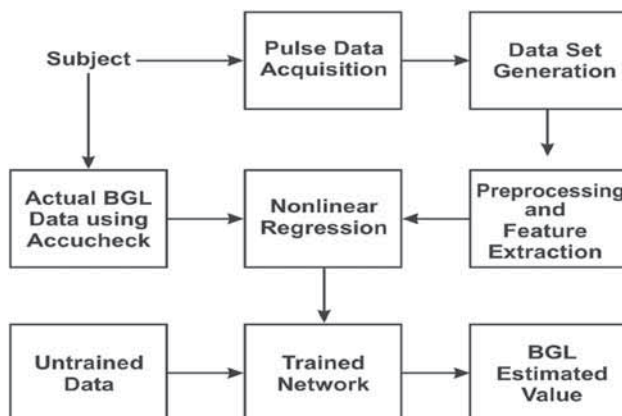


Figure 1. Block schematic of the proposed system architecture.

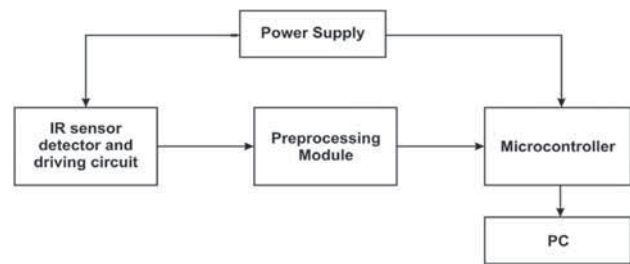


Figure 2. Pulsed data acquisition system.

The trained network is used for testing purpose and the performance of BGL value estimation is analyzed using R^2 and Clarke Error Grid.

The Pulsed data acquisition system consists of PPG sensor, analog signal conditioning circuit and processor to interface digitized pulse signal for recording and storing the data as shown in figure 2.

2.2 Database

The PPG data of 611 individuals is recorded over a 3 minute duration each. For diabetic subjects the data is specifically recorded at Jahangir Medical and Research Centre, India and at Freedom from Diabetes Organization India., whereas for normal subjects date were recorded at Vishwakarma Institute of Information Technology, India. Initially, the PPG input measurement is done by the data acquisition system which is built and tested in house. The actual data acquisition system is built as per our design specifications by Biokit, India. This company is in the development of biomedical equipments required for carrying out experimentation in biomedical engineering field. Finally, using this system the PPG signal of the individuals are recorded along with their BGL values on the Accu-Check® machine simultaneously. The range of BGL values varied from 70 to 450 mg/dl. Table 1 shows the database details.

2.3 Methodology

2.3a Acquisition of PPG signal: The recorded PPG signal is a digitized signal of 3 minute duration with a sampling rate of 100 samples per second giving a total of 18000 samples. A clean data of 1 min duration is extracted from 3 min recorded signal resulting in 6000 samples. This is done to eliminate the artifacts and get a clean signal. The gain of the system is set to 25.

Table 1. Database details.

| | Male | Female | Age group in years |
|----------------|------|--------|--------------------|
| Diabetic (233) | 134 | 99 | 4 to 70 |
| Normal (378) | 210 | 168 | 18 to 70 |
| Total | 611 | | |

$$\text{ADC reference} = \frac{5}{1024} \text{ Volts} \tag{1}$$

$$\text{Voltage Conversion} = \text{Reading} \times \text{ADC Reference}/\text{Gain} \tag{2}$$

Figure 3 shows the 1 minute extracted signal. The signal is represented in physical unit as shown in figure 4.

2.3b *Pre-processing of the input signal:* The PPG signal quality depends on the subject’s skin tone at measurement, blood oxygen saturation, blood flow rate and temperature of the skin. The signal strength will be affected by power line interference, motion artifacts, muscle artifacts, high frequency artifacts and low amplitude. Thus, it is necessary to apply pre-processing techniques to remove the noise introduced in the signal which may affect feature extraction and overall BGL measurement. Thus, the acquired signal is now given as an input to the pre-processing module. This input signal is represented as

$$\text{Input signal} = x(n) = \{x_1, x_2, x_3, \dots, x_n | 0 < n < S_{\text{window}}\} \tag{3}$$

The filters used for pre-processing are moving average filter and Savitzky Golay filter. The base line wandering is removed by using moving average filter. MA filter is used to remove base line wandering. The output of filter is represented by Eq. (4).

$$Y_s(i) = \frac{1}{2N + 1} [y(i + N) + y(i + N - 1) + \dots + y(i - n)] \tag{4}$$

Figure 5 shows the input signal and red line shows its base line variation. Figure 6 shows the zoomed input signal and figure 7 shows the baseline removed signal. The noise and other artifacts are removed using Savitzky-Golay filter. Savitzky-Golay is also called as digital smoothening polynomial filters or least squares smoothening filters [15]. It preserves the shape of the pulse and high frequency components of data while smoothening it. It gives us a clean and smooth PPG signal as shown in figure 8.

2.3c *Feature extraction:* In this section, we summarize the features computed from the PPG signal. A functional relationship exists between the PPG signal and blood

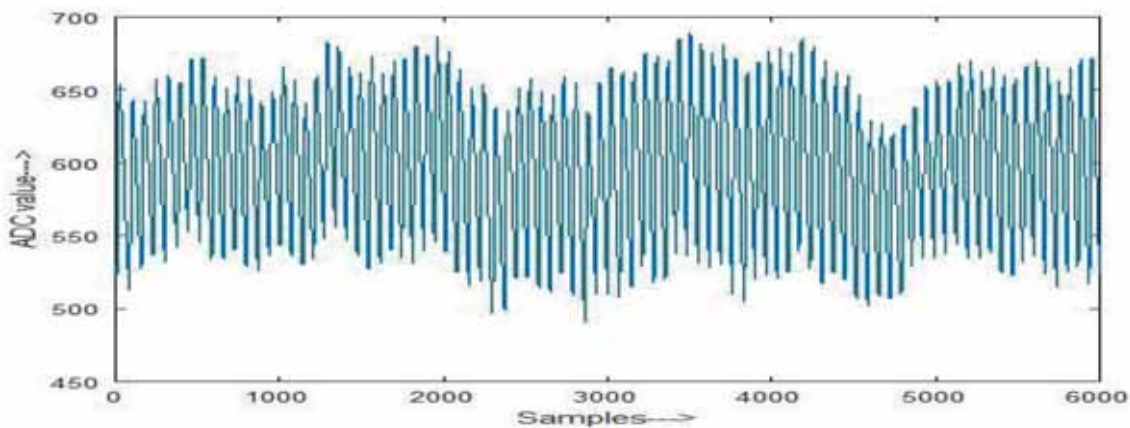


Figure 3. One minute input signal.

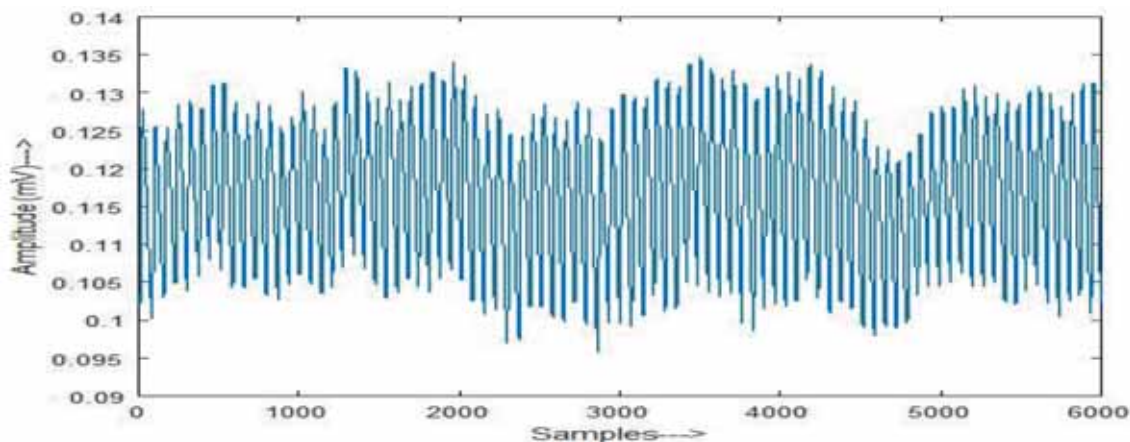


Figure 4. Input signal in physical unit.

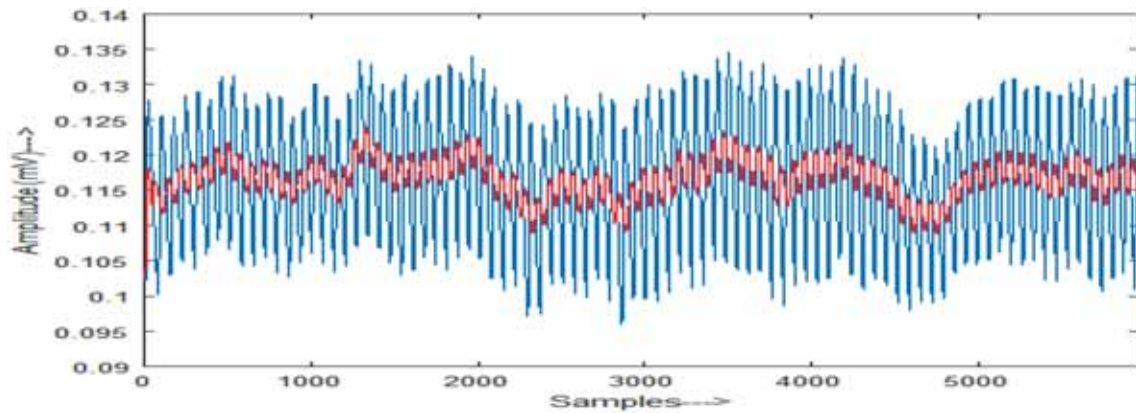


Figure 5. Input signal with base line wandering.

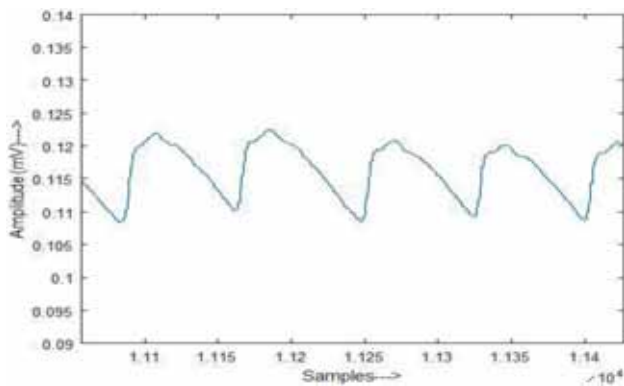


Figure 6. Input signal (zoomed).

glucose levels [1]. Thus, the physiological alterations are measured from the shape of the PPG waveform. These physiological properties are used as a basis to select the

features. The output of this module is a Feature vector (FV) that consists of all features relating to statistical variations in the PPG of the subject. These features reflect the changes in subject’s hemodynamic conditions [1].

The feature vector consists of global features, which is computed at window level and local features which are computed at frame level as, (a) Global features computed from S_{window} and (b) Local features computed from L_{frame} . The length of window S_{window} is 6000 samples and length of the frame L_{frame} is 500 samples. For each frame, the time length of 5 s is selected so as to get atleast four to five pulses in a frame [1]. Figure 9 represents the signal window and frames.

Following section explains the feature extraction details of PPG signal:

I) *Auto Regressive (AR) Model of PPG waveform*: AR coefficients model the spectral envelope of PPG signal. These coefficients capture the change in the shape of the

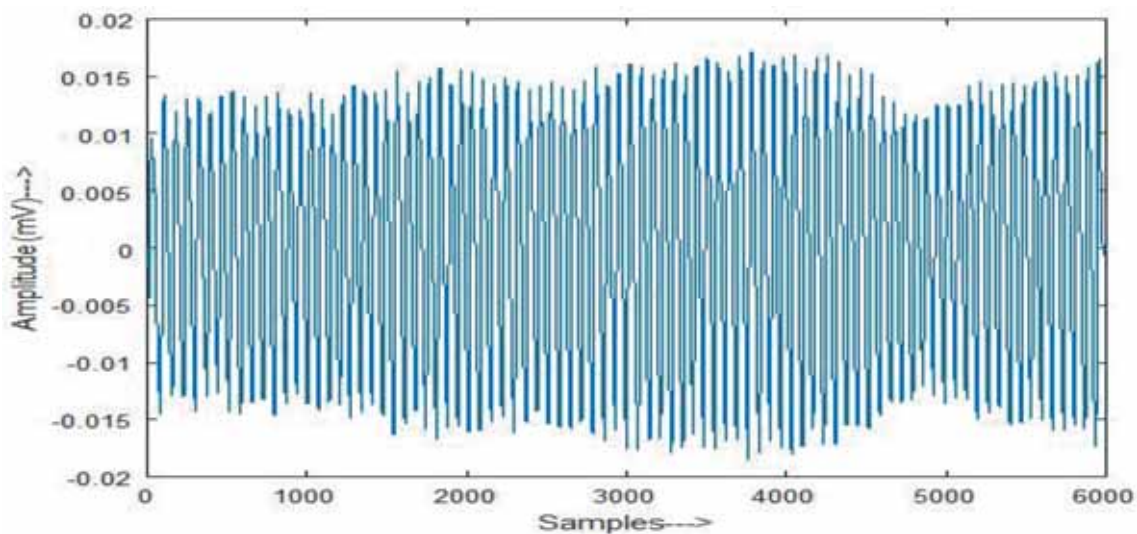


Figure 7. Input signal with removed baseline wandering.

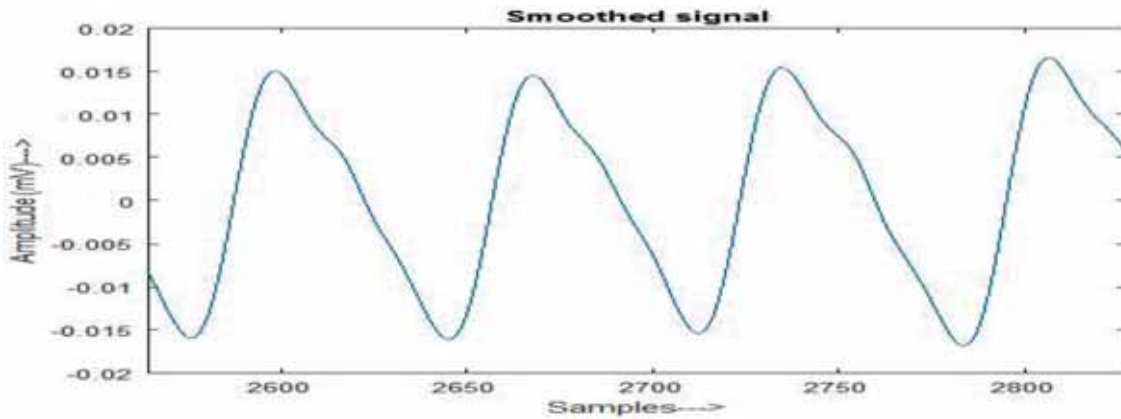


Figure 8. Input signal smoothed using Savitzky-Golay smoothing Filter.

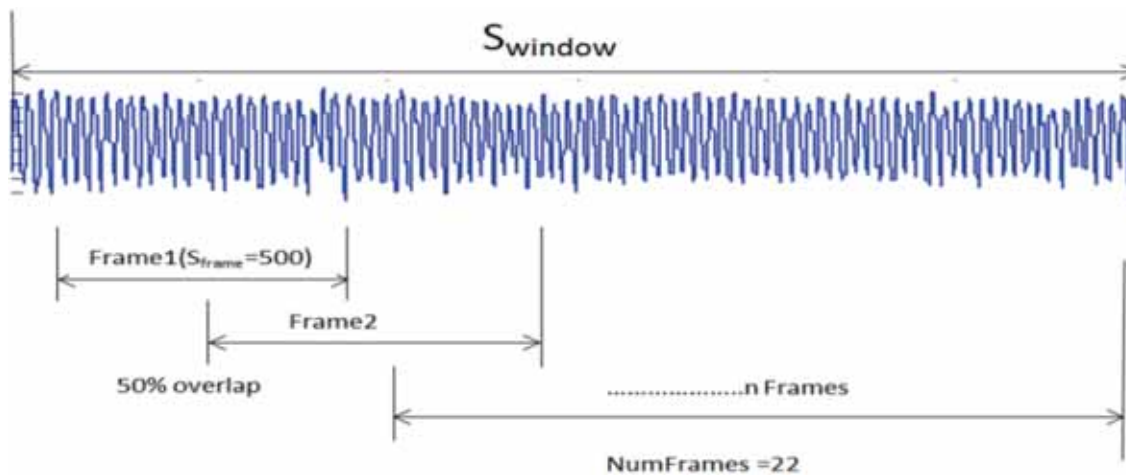


Figure 9. Signal represented in Window and Frames.

pulse occurring due to change in blood flow, flowing through the different arteries, veins and capillaries [16]. AR coefficients also model the shape of basic pulse. It is required that the shape of the AR power spectrum of a given pulse follow correctly the periodogram of PPG signal.

$$\text{Power Spectrum} = S(f) = \frac{\Delta t}{N} \left| \sum_{n=0}^{N-1} x_n e^{-j2\pi fn} \right|^2 \quad (5)$$

We have selected 5th order AR model. AR coefficients are computed using Yule Walker equations [16], for the given periodogram which were derived from sample covariances as mentioned in Eq. (6) for $l > 0$

$$\sum_{k=1}^N a_k \gamma_{xx}[l - k] = -\gamma_{xx}[l] \quad (6)$$

Five AR_{ppg} coefficients are extracted from Eq. (6) and are represented as C₁, C₂, C₃, C₄, and C₅ to form the feature vector FVAR_{ppg} s follows

$$FVAR_{ppg} = \left\{ \begin{array}{l} AR_{ppg}C_1, AR_{ppg}C_2, AR_{ppg}C_2, AR_{ppg}C_3, \\ AR_{ppg}C_4, AR_{ppg}C_5 \end{array} \right\}$$

II) *Kaiser Teager Energy (KTE)*: KTE of the PPG signal is a well-known method used for finding energy profiles of signals with periodic signal components [17, 18]. The property of KTE indicates whether the signal is a clean signal or a noisy signal. For a periodic waveform if the mean value of KTE is high then it is a clean signal and if low it indicates the presence of noise, transients or artifacts in the signal. The energy profiles are computed at window level and also at the frame level. At frame levels the energy is computed using Eq. (7). The KTE operator for real valued signal is given by

$$KTE(t) = x(t)^2 - x(t + 1)x(t - 1) \quad (7)$$

The Teager energy at window level is computed using Eq. (8)

$$KTE(t) = S_{window}(t)^2 - S_{window}(t+1)S_{window}(t-1) \quad (8)$$

From this sequence of KTE (t), we estimated the AR model coefficients of order 5 using matlab function *aryule* and thus KTE_{AR} vector is generated. At frame level the Teager energy for n^{th} frame is further computed using Eq. (9)

$$KTE_n(t) = S_{frame}^2(t, n) - S_{frame}^2(t+1, n)S_{frame}^2(t-1, n) \quad (9)$$

For each frame the statistical parameters like mean KTE, variance KTE, interquartile range KTE and skewness KTE are calculated. The statistical averages of these parameters for n frames are computed as follows:

$$KTE_{frame} = \{KTE_n^\mu, KTE_n^\sigma, KTE_n^{iqr}, KTE_n^{skew}\}$$

Thus, Feature vector FV of KTE consists of the following

$$FVKTE = \{KTE_{AR}, KTE_n^\mu, KTE_n^\sigma, KTE_n^{iqr}, KTE_n^{skew}\}$$

III) *Heart rate statistics (HR)*: There exists a functional relationship between altered glucose levels due to diabetes and heart rate variability (HRV) [18]. Hence we used HR variability as a feature, which is computed at window and frame level. This feature is calculated by finding peaks from the signal window. The statistical parameters of heart rate such as, mean, variance, interquartile range and skewness are also computed from the entire window. Thus, the feature vector FVHR is represented as follows:

$$FVHR = \{Hr_n^\mu, Hr_n^\sigma, Hr_n^{iqr}, Hr_n^{skew}\}$$

IV) *Spectral Entropy Statistics*: The spectral entropy is a scalar obtained by computing the entropy function from normalized power spectrum [19, 20]. This feature measures the damping of pulses, the spectral shape harmonic components and presence of noise. Power spectrum of the n^{th} frame is computed by taking short time Fourier transform and normalizing by the squared magnitude of each bin by total power of the frame and is represented by Eqs. (10) and (11).

$$X^n = FFT(S_{frame}(\tau, n), L_{FFT}) \quad (10)$$

$$P_x^n[k] = \frac{|X^n[k]|^2}{\sum_{j=1}^{L_{FFT}} |X^n[j]|^2} \quad k = 1 \dots L_{FFT} \quad (11)$$

Finally, the spectral entropy is computed using power spectral density as represented by Eq. (12).

$$H_n^s = \sum_{k=1}^{L_{FFT}} P_x^n[k] \log(P_x^n[k]) \quad (12)$$

Similarly, to the KTE_{frame} measures, statistical parameters such as mean, variance, interquartile range and

skewness of spectral entropy are computed for each of the n frames. The feature vector FVHs represents the average spectral entropy parameters as follows:

$$FVHs = \{Hs_n^\mu, Hs_n^\sigma, Hs_n^{iqr}, Hs_n^{skew}\}$$

V) *Energy profile of the PPG signal*: Past research work by Leonard *et al* [21] have shown that rate of respiration affects blood pressure and in diabetics they are closely related. To notify these changes, the log energy of the PPG signal is calculated at frame level in order to estimate respiratory rate.

$$LogE_n = \left(\sum_{\tau=1}^{L_{frame}} S_{frame}^2(\tau, n) \right) \quad (13)$$

Using $LogE_n$ we compute the AR coefficients of order 5 using matlab function *aryule*. $LogE_n$ is a model of the fluctuations of energy of the PPG relative to the respiratory rate. Also the average statistical parameters like variance and interquartile range () are also computed. Hence FVLogE feature vector is obtained as follows:

$$FVLogE = \left\{ \begin{array}{l} LogE_n^{AR} C_1, LogE_n^{AR} C_2, LogE_n^{AR} C_3, \\ LogE_n^{AR} C_4, LogE_n^{AR} C_5, LogE_n^\sigma, LogE_n^{iqr} \end{array} \right\}$$

Although the morphology of the PPG signal looks similar to the arterial pressure pulse, but the wave contour does not remain the same. The relationship between the PPG signal and the pressure pulse has been quantified by Millasseau *et al* [22]. Hence we hypothesized that the time domain features should contribute significant information in BGL estimation (which take into account the alteration in pulse wave shape over the time). The features computed based on the pulse alteration nature are shown in figure 10 and the details about the computation is explained in the following section.

VI) *Pulse transit Time (PTT)*: It is the time difference between the onsets of two consecutive pulses. PTT is computed as:

$$PTT = TP_2 - TP_1 \quad (14)$$

Where TP_1 is the starting point of the pulse and TP_2 is the ending point of the pulse. PTT is calculated at frame level and the average statistical parameters of PTT for n frames are also computed and represented by FVPTT feature vector as:

$$FVPTT = \{PTT_n^\mu, PTT_n^\sigma, PTT_n^{iqr}, PTT_n^{skew}\}$$

VII) *Peak to peak Interval (PPint)*: It is the time difference between the two peaks of the consecutive pulses. The R-R interval in the ECG signal correlates closely with the *interval* as both represent a completed heart cycle. The *Peak to Peak interval* has been used to detect the condition of heart in PPG signals [23–26]. The peak to peak interval is calculated at frame level and average statistical

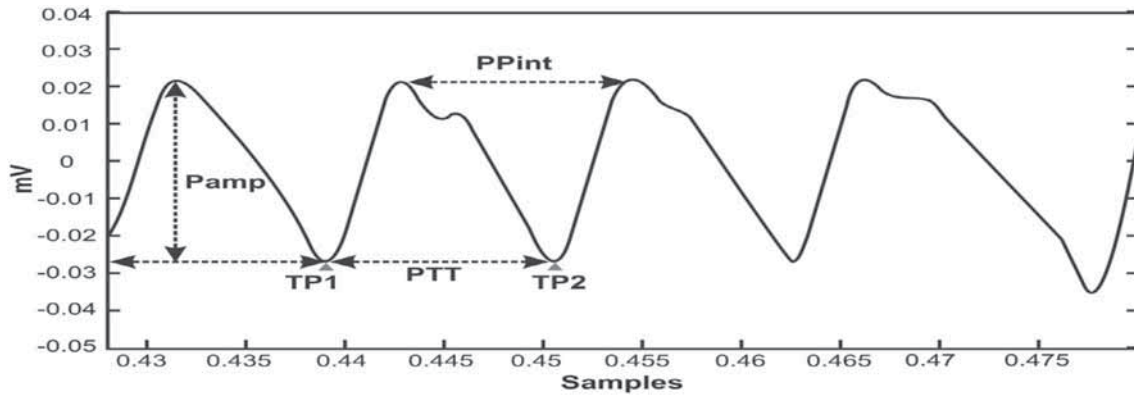


Figure 10. Pulse features.

parameters for n frames are computed and the feature vector is represented as

$$FVPPint = \{PPint_n^\mu, PPint_n^\sigma, PPint_n^{igr}, PPint_n^{skew}\}$$

In 2008, Lu *et al* [27] compared the HRV in PPG signals with the HRV using R-R intervals in ECG signals. Their results demonstrated that HRV in PPG and ECG signals are highly correlated. They strongly suggested that PPG signals could be used as an alternative measurement of HRV. As mentioned in [18], there exists a functional relationship between heart rate variability and glucose hence the above features PTT and PPint are computed as a measure of heart rate variability in PPG signal.

VIII) *Pulse Amplitude*: It is the peak amplitude of the pulse also known as systolic amplitude. The systolic amplitude is an indicator of the pulsatile changes in blood volume caused by arterial blood flow around the measurement site [28, 29]. Systolic amplitude has been related to stroke volume [30]. Thus, the pulse amplitude is calculated at frame level and average statistical parameters over n frames are also computed and FVPamp feature vector is represented as:

$$FVPamp = \{Pamp_n^\mu, Pamp_n^{igr}, Pamp_n^{skew}\}$$

Thus, the proposed feature vector consists of time and frequency domain features and is represented as 35 dimensional feature vector (excluding 5 features from above mentioned 40 features. Detail description is given in the section 2.4c(II)) as:

$$FV = \{FVAR_{PPG}, FVKTE, FVHr, FVHs, FVLogE, FVPTT, FVPint, FVPamp\}$$

So far the features are computed at frame or at window level. If the frame length is changed the feature vector computed also changes. If the start and end point detection of the frame changes, this effectively changes the nature of signal in the specified frame length. Thus, due to this limitation there is a need to analyze feature of each single pulse acquired in the signal. If the variation in each single

pulse is considered, then the effect of variation of the pulses within a window can be captured in more depth. Hence a new feature vector FV_{new} is proposed which is computed for each single pulse.

2.3d *Single Pulse Analysis (FV_{new})*: We considered periodicity of a pulse and local minima for separating the pulses within a given window. After separating each pulse we computed the time and frequency domain features. The separated pulses of a given window are plotted in figure 11. It is observed that the pulses show time variant nature and this information are captured by computing following features. Figure 12 shows the sample representation of a single pulse extracted from the given window. The features such as energy of pulse, pulse transit time, peak amplitude1, end amplitude, peak amplitude 2, peak on set time, peak end set time and pulse interval are extracted for each of the pulse and averaged over n pulses for a given window. The energy of each pulse is calculated by taking FFT of each pulse in the dataset. The Pulse transit time (PTT) is the total time period of the pulse. From figure 11 it is observed that each pulse has different pulse duration. Here, Peakamp 1 represents the total peak amplitude, End Amplitude corresponds to amplitude of the pulse arising after zero crossing of pulse i.e., negative peak amplitude and Peakamp 2 represents the difference between Peakamp 1 and End

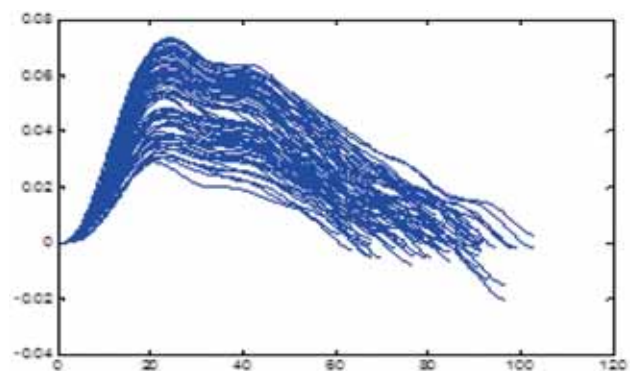


Figure 11. Pulse separation observed of a given window.

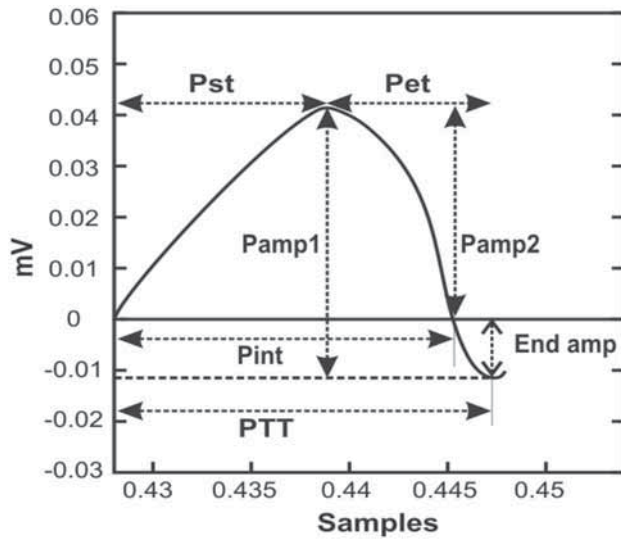


Figure 12. Single Pulse Features.

amplitude. Pulse onset time (Pst) is the time difference from start to peak of pulse duration. Pulse end set time (Pet) is the time difference from peak of pulse to end of pulse duration. Pulse interval (Pint) is the time period of the pulse starting from zero point upto the next zero crossing point. Thus, we computed seven distinct features for each of the pulse in a given window and a feature vector is given by:

$$PFV = \{Energy, Peakamp_1, Peakamp_2, Endamp, PTT, P_{int}, P_{st}, P_{et}\}$$

Further, we computed their statistical parameters such as mean, variance, interquartile range and skewness for each of the seven features for a given window. This forms a 28 dimensional feature vector FV_{new} (7 features * 4 statistical parameters):

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

Overall, we extracted two different types of feature sets, (1) time and frequency domain at frame level and (2) single pulse analysis. Time and Frequency domain feature vector (FV) consists of 35 dimension and single pulse analysis results into 28 dimension feature vector. Further experiments are carried out using these two types of feature sets.

2.4 BGL estimation

In this study we implement neural network based BGL estimation for two feature sets, i) Time and Frequency domain at frame level and ii) Single Pulse Analysis. We also perform baseline experiments on our dataset. Following sections describes in detail about datasets, neural network topology and detail results.

2.4a Training and testing dataset: Pulse plethysmograph signal (PPG) is recorded using Pulse Data acquisition

system as shown in figure 2. Section 2.1 describes in detail about the data acquisition system. We recorded three minute of PPG signal from each of the 611 individuals along with their BGL values. Each data signal was observed thoroughly for its continuity over the entire three minute duration along with the amount of power line interference, and baseline drift. Signal samples highly corrupted with noise, extreme baseline drift and any discontinuity are eliminated from training data set. To reduce the computational complexity, we extracted a one minute window signal from three minute duration PPG signal. For the reliable performance of neural network we require large number of data samples for training purpose. Apart from this, the distribution of the samples from normal range to extreme diabetic condition should be evenly distributed to generate reliable neural network models. For this purpose, we extracted multiple instances of one minute window out of three minute recorded PPG signal of 611 individuals with suitable time overlap. We extracted 3 to 4 one minute window from each individual PPG data. With this process we obtained around 1900 window segments of 1 min duration. By randomly observing the window segments we selected 1500 window segments which are noise, power line interference free and useful for experimentation.

Out of 1500 window segments we distributed these signals into 3 sets of 500 segments each namely Train1, Train2 and Test datasets. Train1 and Train2 datasets are used exclusively for training purpose and test dataset for testing. 500 distinct segments in each of the training dataset, Train1 and Train2 are selected with approximately equal distribution of BGL values (range 70 to 450 mg/dl) from total of 1500 window segments. Remaining 500 segments are used as test dataset. Even though the BGL value distribution of segments in the test data is not uniform but it covers all the range of BGL values from 70 to 450 mg/dl. The distribution of BGL values in each dataset is represented by the histogram shown in figure 13.

2.4b Neural network configuration details: We used three hidden layer neural network topology. Preliminary experimentation on 29 dimension feature vector (baseline) is carried out with varying neurons in each layer to estimate the best neural network topology. We started with 2 hidden layers and (12, 8) neurons in layer1 and 2, respectively. Further, we increased one more layer with 10, 8, 4 neurons in layer 1, 2, 3, respectively. Further experimentation with varying neurons is carried out and we found that 20, 15, 10 neuron configuration performs best with maximum R^2 value of 0.73. The results are shown in the section 2.4c. We finalize 3 hidden layers with (20, 15, 10) neuron configuration for baseline and time and frequency domain feature vector and (30, 20, 10) for Single Pulse Analysis feature vector.

2.4.c Results: We performed experiments with three different feature sets, (i) baseline system (29 feature), (ii) time and frequency domain features (with 35 dimensions) and

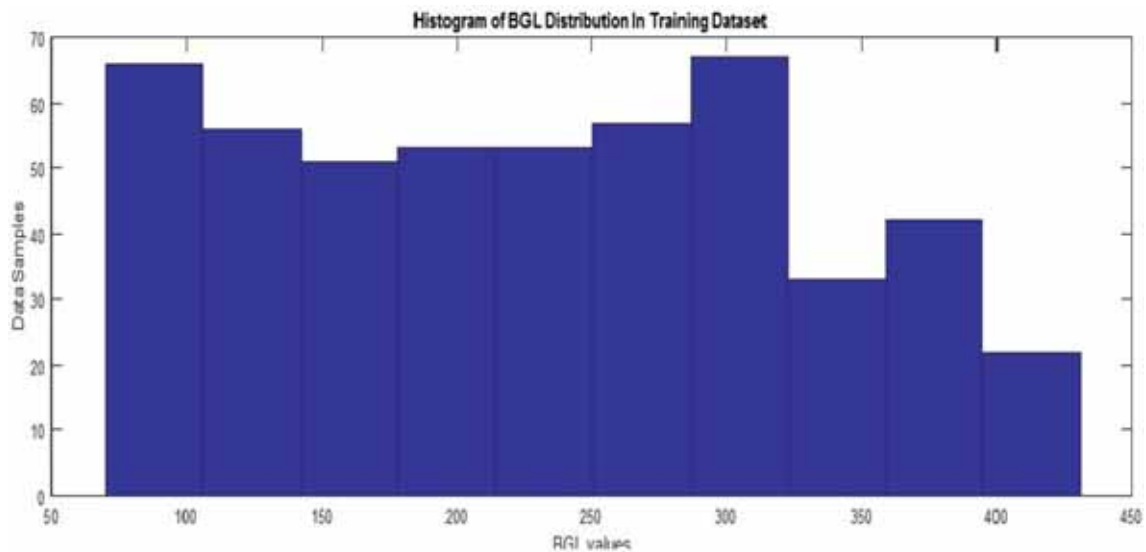


Figure 13. Histogram of BGL distribution of training dataset.

(iii) Single Pulse Analysis (28 features). We used coefficient of determination (i.e., R^2) and Clarke Error Grid Analysis [31], as a performance measure. We also computed Spearman coefficient of correlation and Pearson coefficient of correlation. Clarke Error Grid Analysis is clinically accepted for validation of BGL estimation. The grid is divided into 5 regions [31]. Clarke grid scatter plot is plotted using Matlab Function [32]. Region A represents prediction with 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 baseline system, time and frequency domain features and Single Pulse. We trained neural network separately with (i) Train set 1, (ii) Train set 2 and (iii) Train set 3 (Train set 1 and Train set 2 combined) dataset.

We performed the testing in the following manner:

- i. Training network :Train set 1, Testing : Train set 2 and Test dataset
- ii. Training network: Train set 2, Testing: Train set 1 and Test dataset.
- iii. Training network : Train set 3, Testing : Test dataset

I) *Baseline system*: We considered twenty nine dimensional features vector for our baseline system. The feature vector is similar to [1], where author used 33 dimensional feature set. We excluded age, weight, body mass index, and oxygen saturation features and considered 29 dimensional features. One of the reasons to exclude these features is to

Table 2. Baseline results and results obtained by [1].

| | Reference [1] | Baseline System Train set 1 and testing on Test dataset | Baseline System Train set 2 and testing on Test dataset | Baseline System Train set 3 and testing on Test dataset |
|------------------|---------------|---|---|---|
| Features | 33 | 29 | 29 | 29 |
| Testing Data Set | 410 | 500 | 500 | 500 |
| R^2 | 0.54 | 0.73 | 0.81 | 0.71 |

eliminate the person-specific dependency. Table 2 represents baseline results on our test data set and results obtained by Monte-Moreno [1].

The results obtained on our dataset using our baseline feature vector (29 dimensional) are comparable with results obtained by Monte-Moreno [1]. Considering R^2 value, both the training dataset show comparable results on test data. Train2 dataset shows slightly more R^2 value which indicates more balanced training data for a given test conditions. We use these results as a reference for our further experimentation and the results validate our methodology, neural network configuration settings and feature extraction process.

II) *Time and Frequency Domain Features*: Here, we considered thirty five dimensional feature vector. These features are computed based on autoregressive coefficients, Kaiser Teager energy, heart rate, spectral entropy, log energy, etc. The details about these feature extraction is explained in section 2.3c. We call this feature vector FV and are given by

Table 3. Performance of Time and Frequency domain features.

| Class | Train set 1 Training Network | | Train set 1 Training Network | | Train set 2 Training Network | | Train set 2 Training Network | | Train set 3 Training Network | | Train set 3 Training Network | |
|----------------------|------------------------------|------------|------------------------------|------------|------------------------------|------------|------------------------------|------------|------------------------------|------------|------------------------------|------------|
| | Test: Test Dataset | | Test: Train set 2 Data | | Test: Test Dataset | | Test: Train set 1 Data | | Test: Test Dataset | | Test: Test Dataset | |
| | No of samples predicted | Percentage | No of samples predicted | Percentage | No of samples predicted | Percentage | No of samples predicted | Percentage | No of samples predicted | Percentage | No of samples predicted | Percentage |
| A | 376 | 75.2 | 386 | 77.2 | 349 | 69 | 369 | 73.8 | 403 | 80.6 | | |
| B | 108 | 21.6 | 91 | 18.2 | 129 | 25.8 | 112 | 22.4 | 87 | 17.4 | | |
| C | 8 | 1.6 | 15 | 3 | 17 | 3.4 | 14 | 2.8 | 5 | 1 | | |
| D | 8 | 1.6 | 8 | 1.6 | 5 | 1 | 5 | 1 | 5 | 1 | | |
| E | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | |
| R ² | 0.78 | | 0.74 | | 0.77 | | 0.73 | | 0.84 | | | |
| Spearman Coefficient | 0.88 | | 0.86 | | 0.87 | | 0.86 | | 0.92 | | | |
| Pearson Coefficient | 0.88 | | 0.88 | | 0.88 | | 0.85 | | 0.92 | | | |

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

In addition to base line features (excluding KTE mean, KTE variance, KTE interquartile range, Heart rate variance, Heart rate interquartile range), we added time domain features like pulse transit time PTT, peak to peak interval PPint, and Pulse amplitude.

Table 3 represents performance of Time and Frequency domain features for BGL estimation using R² and Clarke Error Grid Analysis. Figure 14 represents Clarke Error Grid plot for the train set 3. From table 3, considering 3 different training networks, the performance of network Train set 3 shows highest R² values. Also the R² values across three different training networks are consistent i.e., variations observed in R² value is small (0.73 to 0.84).

Comparing with the baseline results given in table 2, R² value is significantly improved from 0.71 to 0.84 for train set 3 network. It is also observed that the correlation coefficient also shows consistency across the training network and high correlation (max 0.92).

As shown in figure 14, region A represents 80.6% prediction within 20% of actual BGL value. Region B represents 17.4%, region C, D and E are negligible.

III) *Single Pulse Analysis*: Here, we considered twenty eight dimensional feature vector. The features are computed based on Pulse features like energy, pulse transit time, pulse interval, pulse amplitude, etc. The details about the feature extraction are explained in section 2.3d. We call this feature vector as FV_{new} and is given by

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

Table 4 represents performance of Single Pulse for BGL estimation using R² and Clarke Error Grid Analysis. Figure 15 represents Clarke Error Grid plot for the Train set 3. Table 4 shows highest R² value and correlation coefficient across all 3 training networks and test data (R² = 0.91)

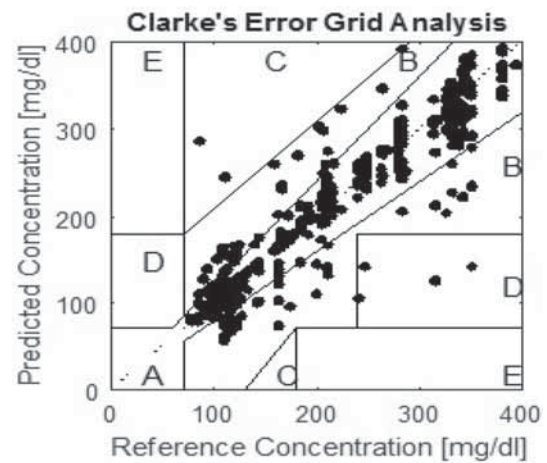


Figure 14. Clarke Error Grid Analysis of Time and Frequency domain features, training on Train set 3 and testing on test data.

Table 4. Performance of Single Pulse Analysis.

| Class | Train set 1 Training Network | | Train set 1 Training Network | | Train set 2 Training Network | | Train set 2 Training Network | | Train set 3 Training Network | |
|----------------------|------------------------------|------------|------------------------------|------------|------------------------------|------------|------------------------------|------------|------------------------------|------------|
| | Test: Test Dataset | | Test: Train set 2 Data | | Test: Test Dataset | | Test: Train set 1 Data | | Test: Test Dataset | |
| | No of samples predicted | Percentage | No of samples predicted | Percentage | No of samples predicted | Percentage | No of samples predicted | Percentage | No of samples predicted | Percentage |
| A | 394 | 78.8 | 423 | 84.6 | 397 | 79.4 | 386 | 77.2 | 415 | 83 |
| B | 104 | 20.8 | 75 | 15 | 96 | 19.2 | 93 | 18.6 | 87 | 17 |
| C | 2 | 0.4 | 1 | 0.2 | 7 | 1.4 | 14 | 2.8 | 0 | 0 |
| D | 0 | 0 | 1 | 0.2 | 0 | 0 | 7 | 1.4 | 0 | 0 |
| E | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| R ² | 0.83 | | 0.87 | | 0.91 | | 0.77 | | 0.91 | |
| Spearman Coefficient | 0.92 | | 0.94 | | 0.94 | | 0.88 | | 0.94 | |
| Pearson Coefficient | 0.91 | | 0.93 | | 0.95 | | 0.88 | | 0.95 | |

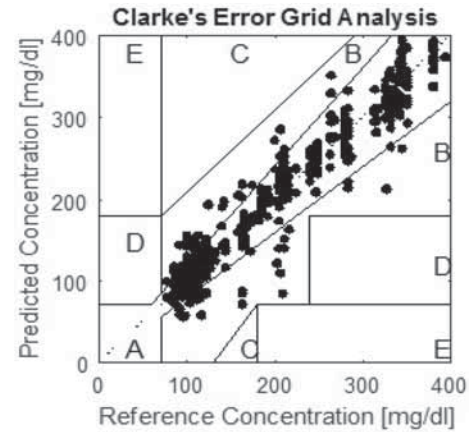


Figure 15. Clarke Error Grid Analysis of Single Pulse Analysis features training on Train set 3 and testing on test data.

(Spearman coefficient = 0.94 and Pearson coefficient = 0.95). Also the R² value, Spearman coefficient and Pearson coefficient across all 3 training networks and all test data sets show more consistency as compared to results obtained by time and frequency domain (table 3). Combined Training dataset i.e., Train set 3 shows more balanced distribution s of BGL samples ranging from 70 to 450 mg/dl and reflected in highest R² value and highest prediction accuracy. Class A and Class B BGL determination are clinically accepted and hence almost 100% prediction accuracy is achieved by this technique.

3. Discussion and conclusions

In this work, we considered two sets of feature vector; (1) Time and frequency domain and (2) Single Pulse Analysis. Initially, we performed experiments with feature vector proposed by Monte-Moreno [1]. We obtained marginal improvement in the results. We used this as baseline results and to validate our methodology. Using Time and Frequency domain features (35 dimensional feature set) we observed significant improvement in R² (0.84) value as compared to our baseline results (i.e., R² = 0.71) for train set 3. This feature vector includes time domain features which were not included in the baseline system. This shows that including the time domain features helps in estimating BGL value better. Further, experimentation based on Single Pulse Analysis which takes into account time varying nature the pulses showed further improvement in BGL value estimation (i.e., R² = 0.91) as compared to Time and Frequency domain (i.e., R² = 0.84) for train set 3.

To summarize the work, 3 different datasets are used for training of neural networks. Train set 1 and Train set 2 each consists of 33% of entire data and 33% data is used exclusively for test set. Both the feature sets i.e., Time and frequency domain and Single Pulse analysis show comparable performance across 3 different training data sets. This

shows that 33% data of training set which consists of balanced BGL values (i.e., from 70 to 450 mg/dl), can estimate BGL values with high prediction accuracy. This indicates that the features that were used for training the networks, contribute significantly in prediction of BGL value. The performance of train set 3 which is almost 67% of entire data (balanced) shows the highest R^2 value with both the feature sets ($R^2 = 0.84$ for Time and frequency domain, $R^2 = 0.91$ for Single Pulse Analysis) and correlation coefficient (0.92 Time and frequency domain and 0.95 for Single Pulse Analysis). Among the Time and Frequency domain features and Single Pulse Analysis, results obtained using Single Pulse analysis technique shows highest $R^2 = 0.91$ and prediction accuracy. Single Pulse analysis technique effectively captures time varying nature of the pulses present in the acquired PPG signal. We captured these variations by separating the pulses and extracted single pulse features. From table 4 it is observed that this technique shows significant contribution in BGL value estimation.

Currently, we used only time domain features extracted for Single Pulse analysis method but further, frequency domain features could be explored to investigate its significance in BGL value estimation. The system implemented in this study excluded subject-dependent features and hence can be directly adopted to implement a low cost modular non-invasive glucometer for commercial use.

Acknowledgement

A part of this work is funded by Board of College and University Development (BCUD) Savitribai Phule Pune University. We would like to thank BCUD for providing funding for this work. Diabetic subjects data is recorded at Jahangir Medical and Research Centre, India and at Freedom from Diabetes Organization India. We would like to thank Dr. Anuradha Khadilkar and Dr. Pramod Tripathi for allowing us to record the data at their research centre.

References

- [1] Monte-Moreno E 2011 Non-invasive estimate of blood glucose and blood pressure from a photoplethysmograph by means of machine learning techniques. *Artificial Intelligence in Medicine* 53(2): 127–138
- [2] Losoya-Leal A, Camacho-León S, Dieck-Assad G and Martínez-Chapa S 2012 State of the art and new perspectives in non-invasive glucose sensors. *Revista Mexicana De Ingeniería Biomédica* 33(1): 41–52
- [3] March W 1974 Noninvasive automatic glucose sensor system US3958560A
- [4] So C, Choi K, Wong T and Chung J 2012 Recent advances in noninvasive glucose monitoring. *Medical Devices (Auckland, NZ)*, 5, p.45
- [5] Ciudin A, Hernández C and Simó R 2012 Non-invasive methods of glucose measurement: current status and future perspectives. *Current Diabetes Reviews* 8(1): 48–54
- [6] Habbu S, Ghongade R, Aher M 2013 Noninvasive techniques for Blood Glucose Measurement In: *Proceedings of International Conference on Electrical, Electronics and Computer Engineering (ICEECS)CHE-016*
- [7] Fortino G and Giampà V 2010 PPG-based methods for non-invasive and continuous blood pressure measurement: an overview and development issues in body sensor networks. In: *Proceedings of IEEE International Workshop on Medical Measurements and Applications* pp.10–13
- [8] Poon Y and Zhang T 2006 Cuff-less and noninvasive measurements of arterial blood pressure by pulse transit time. In: *Proceedings of 27th annual conference of IEEE engineering in medicine and biology* pp. 5877–5880
- [9] Spigulis J 2005 Optical noninvasive monitoring of skin blood pulsations. *Applied Optics* 44(10): 1850–1857
- [10] Allen J 2007 Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement* 28(3): p.R1
- [11] Anderson R and Parrish A 1981 The optics of human skin. *Journal of investigative Dermatology* 77(1): 13–19
- [12] Ramasahayam S, Haindavi K and Chowdhury S 2015 *Non-invasive estimation of Blood glucose concentration using near infrared optodes* Springer, Smart Sensors Meas. Instrum. 12, pp. 67–82
- [13] Habbu S 2017 *Noninvasive technique for blood glucose measurement using optical methods and soft computing*, unpublished project report, BCUD, SPPU Pune
- [14] Accu-check aviva. Last accessed May 2, 2011. URL <https://www.accucheck.com/us/glucose-meters/aviva.html>
- [15] Ronald S 2011 *Notes in IEEE Signal Processing Magazine*, Issue 4 vol. 28 pp. 111–117 <https://doi.org/10.1109/msp.2011.941097>
- [16] Rabiner L, Juang B and Rutledge J 1993 *Fundamentals of speech recognition* Englewood Cliffs: PTR Prentice Hall Vol. 14
- [17] Dunn R, Quatieri T and Kaiser J 1993 Detection of Transient signals using the energy operator. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal processing (ICASSP93)*. p.145–8
- [18] Ducher M, Cerutti C, Gustin P, Abou-Amara S, Thivolet C, Laville M, Paultre Z and Fauvel P 1999 Noninvasive exploration of cardiac autonomic neuropathy. Four reliable methods for diabetes? *Diabetes Care* 22(3): 388–393
- [19] Renevey P and Drygajlo A 2001 Entropy based voice activity detection in very noisy conditions. In: *Proceedings of Seventh European Conference on Speech Communication and Technology*
- [20] Shen J, Hung W and Lee L 1998 Robust entropy-based endpoint detection for speech recognition in noisy environments In: *Proceedings of International Conference on Spoken Language Processing Sydney, Australia*, p.1015–8
- [21] Leonard P, Douglas J, Grubb R, Clifton D, Addison P and Watson J 2006 A fully automated algorithm for the determination of respiratory rate from the photoplethysmogram. *Journal of Clinical Monitoring and Computing*, 20(1): 33–36
- [22] Millasseau S, Guigui F, Kelly R, Prasad K, Cockcroft J, Ritter J and Chowiecnyk P 2000 Noninvasive assessment of the digital volume pulse: comparison with the peripheral pressure pulse. *Hypertension* 36(6): 952–956

- [23] Jubadi W and Sahak S 2009 Heartbeat monitoring alert via SMS. In: *Proceedings of IEEE Symposium on Industrial Electronics & Applications* Vol. 1. pp.1–5
- [24] Fu T, Liu S and Tang K 2008 Heart rate extraction from photoplethysmogram waveform using wavelet multi-resolution analysis. *Journal of Medical and Biological Engineering* 28(4): 229–232
- [25] Linder S, Wendelken S, Wei E and McGrath S 2006 Using the morphology of photoplethysmogram peaks to detect changes in posture. *Journal of Clinical Monitoring and Computing* 20(3): 151–158
- [26] Gil E, Orini M, Bailón R, Vergara J, Mainardi L and Laguna P 2010 Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions. *Physiological Measurement* 31(9): 1271
- [27] Lu S, Zhao H, Ju K, Shin K, Lee M, Shelley K and Chon K 2008 Can photoplethysmography variability serve as an alternative approach to obtain heart rate variability information? *Journal of Clinical Monitoring and Computing* 22(1): 23–29
- [28] Asada H, Shaltis P, Reisner A, Rhee S and Hutchinson R 2003 Mobile monitoring with wearable photoplethysmographic biosensors. *IEEE Engineering in Medicine and Biology Magazine* 22(3): 28–40
- [29] Chua C and Heneghan C 2006 Continuous blood pressure monitoring using ECG and finger photoplethysmogram. In: *Proceedings of International Conference of IEEE Engineering in Medicine and Biology Society* pp. 5117–5120
- [30] Murray W and Foster P 1996 The peripheral pulse wave: information overlooked. *Journal of Clinical Monitoring* 12(5): 365–377
- [31] Clarke W, Cox D, Gonder-Frederick L, Carter W and Pohl S 1987 Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care* 10(5): 622–628
- [32] Codina E G 2008 Clarke error grid analysis; Matlab central file exchange. Last accessed May 2, 2011. URL <http://www.mathworks.com/matlabcentral/fileexchange/20545-clarke-error-grid-analysis>