Biomedical Signal Processing and Control 67 (2021) 102478

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Hemoglobin and glucose level estimation from PPG characteristics features of fingertip video using MGGP-based model

Md. Asaf-uddowla Golap *, S. M. Taslim Uddin Raju, Md. Rezwanul Haque, M.M.A Hashem

Department of Computer Science and Engineering, Khulna University of Engineering & Technology, Khulna 9203, Bangladesh

ARTICLE INFO

Keywords:
Multigene genetic programming (MGGP)
Hemoglobin (Hb)
Glucose (Gl)
Photoplethysmogram (PPG)
Feature selection
Feature extraction

ABSTRACT

Hemoglobin and the glucose level can be measured after taking a blood sample using a needle from the human body and analyzing the sample, the result can be observed. This type of invasive measurement is very painful and uncomfortable for the patient who is required to measure hemoglobin or glucose regularly. However, the non-invasive method only needed a bio-signal (image or spectra) to estimate blood components with the advantages of being painless, cheap, and user-friendliness. In this work, a non-invasive hemoglobin and glucose level estimation model have been developed based on multigene genetic programming (MGGP) using photoplethysmogram (PPG) characteristic features extracted from fingertip video captured by a smartphone. The videos are processed to generate the PPG signal. Analyzing the PPG signal, its first and second derivative, and applying Fourier analysis total of 46 features have been extracted. Additionally, age and gender are also included in the feature set. Then, a correlation-based feature selection method using a genetic algorithm is applied to select the best features. Finally, an MGGP based symbolic regression model has been developed to estimate hemoglobin and glucose level. To compare the performance of the MGGP model, several classical regression models are also developed using the same input condition as the MGGP model. A comparison between MGGP based model and classical regression models have been done by estimating different error measurement indexes. Among these regression models, the best results (±0.304 for hemoglobin and ±0.324 for glucose) are found using selected features and symbolic regression based on MGGP.

1. Introduction

Blood is an important component of the human body, and hemoglobin (Hb) and glucose (Gl) are two key components of human blood. A number of biological activity is done with blood. Hemoglobin carries oxygen around the body from the lungs. Both the deficiency of hemoglobin and excessive hemoglobin cause disease. The body produces glucose from foods that supply energy to all the cells in the body. But, if too much glucose remains in the blood, it causes problems. Diabetes is one of the most common diseases in the world [1]. It occurs when the body system cannot control the sugar level in the blood. Long term diabetes is very risky because it can increase the risk of stroke and other heart diseases, damage kidneys and nerves, and lead to blindness [2]. Continuous monitoring of glucose levels is crucial for diabetes patients. Likewise, regular measurement of hemoglobin level is very important for dengue fever patient [3], anemia patient [4] and premature babies [5].

Various invasive methods are being used for blood component measurement. Most of these methods measure blood components drawing blood via venipuncture from the body. These methods are painful and have a risk of infection as blood is drawn using a needle, and take much time to produce result analyzing the blood sample. On the other hand, the non-invasive methods are more convenient to the patients as only bio-signal (optical or spectral) is enough to measure the blood components instantly. Although invasive techniques are more reliable, it is often costly and required a well-equipped diagnostic center with adequately trained personnel. For the same task, non-invasive methods are fairly new and can generate relatively accurate results using the current technology. In 2020, COVID-19 spreads all over the world and unfortunately it is not easy to control because of its speed and reach of spread depends on many social and environmental issues [6]. During the COVID-19 pandemic, the importance of non-invasive measurement has been realized as everyone has to maintain a distance. Moreover, the doctors and nurses are incapable of serving the COVID-19

* Corresponding author.
E-mail addresses: asaf.golap@iict.kuet.ac.bd (Md.A.-u. Golap), raju.taslim@cse.kuet.ac.bd (S.M.T.U. Raju), haque1407028@stud.kuet.ac.bd (Md.R. Haque), hashem@cse.kuet.ac.bd (M.M.A Hashem).

https://doi.org/10.1016/j.bspc.2021.102478
Received 26 October 2020; Received in revised form 10 January 2021; Accepted 30 January 2021
Available online 4 March 2021
1746-8094/

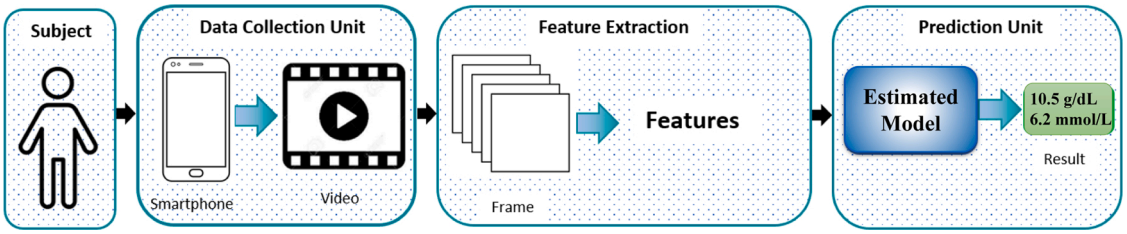


Fig. 1. Phase involved in non-invasive prediction system.

affected patient as it is a highly infectious disease [7,8]. As a contactless method, the non-invasive procedure is highly recommended in this situation.

Noninvasive systems have mainly three functional units shown in Fig. 1: (1) a data collection unit which gathers raw optical or spectral data from the subject; (2) a feature engineering unit that extracts features reprocessing the row biosignal; (3) a prediction unit that estimates and validates results using different prediction models (e.g. machine learning model, MGGP model, etc.). Photoplethysmography (PPG) is an optically obtained plethysmogram often acquired by pulse oximeter to discover the variation in blood volume in microvascular tissue bed. There is much research based on the PPG signal for healthcare monitoring over the past 10 years. For example, hemoglobin level [3,9,10], heart rate monitoring [11], sleep monitoring [12], real-time monitoring of hemodynamic parameters [13], blood pressure estimation [14]. PPG based near-infrared spectroscopy is the most admired technique by the researchers because of its get-at-able and cheap setup [15]. In recent days, there are several smartphones which have builtin sensor system to instantly measure heart rate, oxygen saturation based on photoplethysmography (PPG). This type of non-invasive methods are becoming much more attractive to the people by its evident advantages such as user-friendly, pain-free, no risk of infection, and can generate result instantly [16]. However, for the advancements of the technology smartphone camera has the ability to being used as a sensor. For example, in the year 2015, Wu et al. [17] used the image captured by iPhone-4 and Devadhasan et al. [18] used Samsung to predict Glucose level. Zaman et al. [19] in 2012 used video captured by iPhone-4/5/6 to calculate heart rate. In the same year Anggraeni et al. [20] developed a system using Asus Zenfone 2 Laser to estimate hemoglobin level from the image.

In this paper, we have proposed a non-invasive blood component (hemoglobin and glucose) measurement method using symbolic regression of multigene genetic programming (MGGP) and fingertip video. Fingertip video data is collected using a near-infrared light-emitting diode (NIR-LED) by a smartphone camera. Processing the video, PPG signal is generated by applying several filleting methods. Then, features are extracted from the PPG signal, its first and second derivatives, and using Fourier analysis on the signal. After feature extraction, a correlation-based feature selection method using a genetic algorithm (GA) has been applied to select the best features and discard redundant and irrelevant features for hemoglobin or glucose measurement. Finally, we have developed two independent models to estimate hemoglobin and glucose level using MGGP. The major contributions of this paper are:

- Generation of PPG signal from fingertip video.
- Extraction of PPG characteristic features from PPG signal.
- Reduction of features using a correlation-based feature selection approach using a genetic algorithm (GA).
- Development of MGGP based mathematical models for hemoglobin and glucose measurement. The models can be used for smartphone app-based noninvasive estimation.

Rest of the paper is organized as follows: an overview of the existing works related to our work is briefly discussed in Section 2. A brief overview of MGGP is given in Section 3. The proposed methodology is explained elaborately in Section 4. Performance analysis of our work along with the comparison of results with previous works are demonstrated in Section 5. Finally, the paper ends with conclusions and future directions of this work in Section 6.

2. Related work

In this section, previous experiments and literature surveys are reviewed. There are numerous noninvasive hemoglobin or glucose measurement method available related to our work. Among these, Wang et al. [3,21] developed HemaApp, a smartphone application to monitor Hb concentration level using a camera and some light sources. When collecting data, they point light source towards the finger and carry out chromatic analysis to estimate hemoglobin level. The dataset they used, focus on a range of Hb concentration level (8–16 g/dL), which covers a certain level of anemia and normal people. However, this is not sufficient for the people whose hemoglobin level is <8 (heavily anemic). They achieved Pearson correlation between 0.69 and 0.82 and RMSE rate 1.26–1.56 g/dL using visible IR lights and obtained Pearson correlation and RMSE 0.62 and 1.27 g/dL respectively for white LED.

A PPG signal based non-invasive Hb concentration level prediction method analyzing PPG characteristics features using several machine learning techniques was introduced by Kavsaoğlu et al. [9]. Datasets (PPG signals) were collected from 33 people illuminating light at the finger in 10 periods and after analyzing, 40 characteristic features were extracted. Before the construction of eight different regression models, RELIEFF feature selection (RFS) and correlation-based feature selection (CFS) was employed to select the best features. The support vector-based regression model was performed better than the other model according to prediction results.

Hasan et al. [22] developed SmartHeLP, a smartphone-based hemoglobin estimation system with an artificial neural network (ANN) and fingertip video. The authors collected a 10-s video form 75 participants of 20–56 years of age and hemoglobin level from 7.6 to 13.5 g/dL. Red, green, blue pixels intensities were separated for feature extraction and ANN is used to develop a model using these features to estimate hemoglobin level. They obtained $R^2 = 0.93$ and also identified a specific region of interest in the image frame to decrease the necessary feature space.

In [23] Yuan et al. present a NIR spectroscopy-based non-invasive Hb estimation method using indium gallium arsenide (InGaAs) detector array and plane grating spectrometer. 91 volunteers' fingertip blood spectra data were collected and divided into three categories. Two prediction tests were performed to improve the accuracy. In each test, PLS, MSC coupled with PLS, DOSC coupled with PLS methods were analyzed respectively. For PLS prediction accuracy in terms of relative RMSEP was 6.31% and 7.61% respectively for two tests. For MSC coupled with PLS method, relative RMSEP was 7.0% and 8.09% and relative RMSEP was 6.16% and 6.08% for DOSC coupled with PLS.

In the article [24], they proposed a spectrophotometric non-invasive Hb concentration level prediction system combining a broadband light source which consists of nine LED, a grating spectrograph, and a Si

photodiode array. Fingertip spectra of 109 people were collected and principal component analysis (PCA) was applied on spectra to minimize the dimension of the data. The correlation coefficient of the developed model was 0.94 with 11.29 g/L standard error of calibration (SECal).

Robert et al. [25] developed a smartphone app-based technology to detect anemia (caused by the deficiency of hemoglobin) from fingernail photos. User capture photo of fingernail and upload to their system via the app. Analyzing the color of the image pixels and metadata of the fingernail bed result is instantly displayed on the screen with an accuracy of ±2.4 g/dL. They used multi-linear regression with a bisquare weighting algorithm to estimate hemoglobin value.

Al-Baradie and Bose [26] designed a portable non-invasive hemoglobin measurement sensor using several hardware modules consists of light source and receiver (phototransistor), ARM processor LPC 2148, LCD. 670 nm NIR-LED was used for data collection and the Hemo Cue instrument was used to calibrate the sensor system. Analyzing the result, a linear relationship was tracked out between the Hb concentration level and Hb coefficients measured by the sensor.

Glovanni et al. [27] developed a non-invasive device to estimate anemia from conjunctiva image. Conjunctiva image was collected using smartphones. They conducted tests on 113 persons both healthy and anemic and obtain a correlation coefficient of 0.65 between real hemoglobin value and the observed value using blood sample. K-nearest neighbor classification with 10-fold cross validation is used to determine the risk of anemia.

Soni et al. [28] developed a glucose estimation system using saliva and a smartphone. A sensor was fabricated by immobilization of glucose oxidase enzyme along with a pH-responsive dye on a filter paper-based strip so that it changes color with the variation of glucose present in saliva. Changes in color were detected by a smartphone camera through RGB analysis and the result displayed on the screen. The effect of the smartphone model on the sensor is also studied as the specification of the camera is changed with the model. They found that within the dark box, changes in pixel intensity with respect to smartphone models were not significant. However, in ambient light condition variations in pixel intensity with the smartphone model were noticeable. They obtained a correlation of 0.44, 0.64, and 0.94 for healthy, prediabetic, and diabetic people respectively.

Ramasahayam et al. [29] reported the NIR spectroscopy technique based on NIR LED and a photodetector constituting an optode pair, using transmission photoplethsymography (PPG). The spectroscopy has been performed at the second overtone of glucose which falls in the NIR region to estimate blood glucose concentration. After obtaining the PPG signal from 1070, 950, and 935 nm near-infrared signal, it processes PPG signal and double regression has been applied with an artificial neural network to estimate blood glucose concentration. RMS error of the estimation was 5.84 mg/dL.

Malin et al. [30] demonstrate a non-invasive glucose prediction method using NIR diffuse reflectance over the 1050–2450 nm. To predict glucose concentration of blood two approaches were used. In one approach, random NIR spectra of 7 diabetic patients were examined over 35 days. In another approach, 3 non-diabetic people were tested using oral glucose tolerance tests over several days. Total 20 NIR spectra were taken over the 3.5 h test for analyte analysis. The mean standard error for two approaches were 1.41 mmol/L (25 mg/dL) and 1.1 mmol/L (20 mg/dL) respectively. The mean standard error for the independent tests was equivalent to 1.03 mmol/L (19 mg/dL). These results indicate that NIR diffuse reflectance spectroscopy can be used to estimate blood glucose concentrations noninvasively.

Pai et al. [31] developed a cloud computing-based glucose monitoring technique using near-infrared photoacoustic spectroscopy. A portable embedded system collects photoacoustic signals from tissue using a FPGA and in the back-end, the signal is denoised at a very high speed. Multiple features of the photoacoustic signal were applied to a kernel-based regression algorithm to predict the glucose concentration of blood. The system is connected to the cloud using a mobile phone and

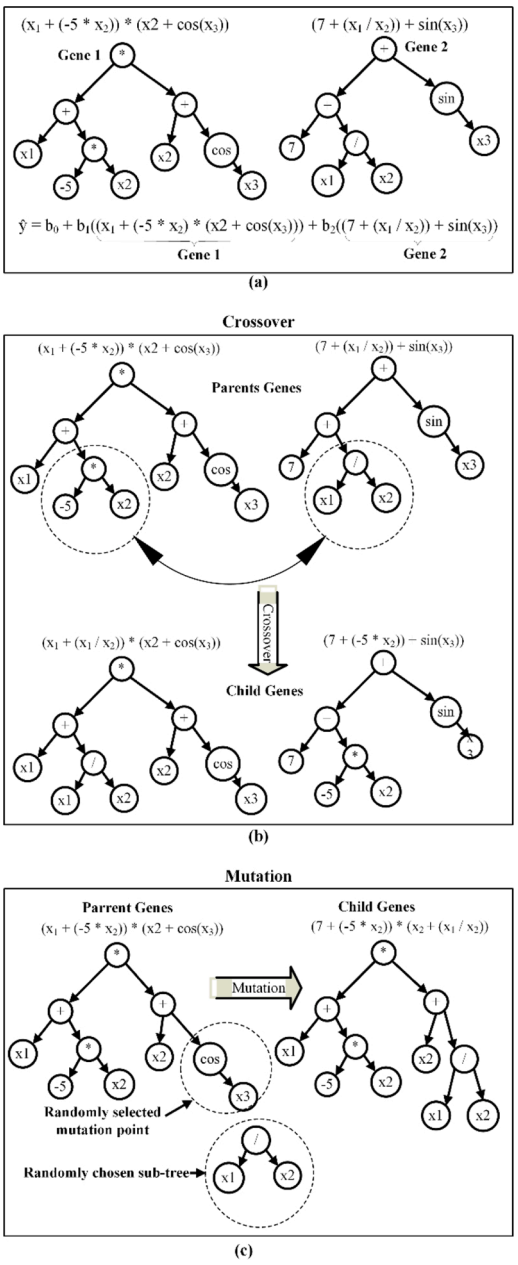


Fig. 2. (a) Tree structure of symbolic regression model, (b) crossover, (c) mutation.

its perform calibration tasks, store data, and analysis measurement data for monitoring and treatment. The mean absolute relative difference of the calibration algorithm was 8.84%.

With this study, we observed that video or digital images of different parts of the body (eye, finger, conjunctiva, etc.) are capable to measure the blood components such as hemoglobin or glucose. Several existing techniques used special sensors to capture signals from the subjects. Most of the techniques used machine learning algorithms and classical regression methods to estimate the results. However, we used a symbolic regression to generate a mathematical model to estimate the results. As a mathematical model, it can be run on any device without requiring special hardware requirements. A user can easily grasp the terms of the predictive equation generated by MGGP which help to trust the model [32]. Human can easily insights into the MGGP model as it is visible to everyone, unlike typical back box predictive models. It is difficult to overdraw the importance of trust and understanding in predictive models. Unlike most of the soft-computing techniques such as artificial

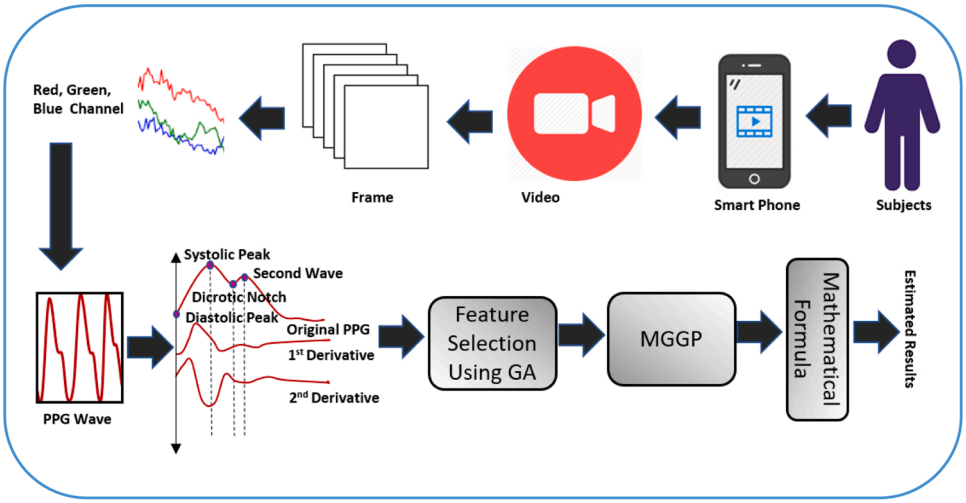


Fig. 3. Main steps of proposed system.

neural network (ANN), deep learning (DNN) or support vector machine (SVM), no specialized software environment setting is needed to deploy the trained MGGP model. The model equation can easily be used in any environment outside MATLAB for structural transparency of the model. A user can easily convert the model equation to any modern computing language without prior knowledge of MGGP.

3. Multigene genetic programming

Genetic programming is a search-based optimization technique derived from the biological phenomenon of natural selection [33]. It works on population, a pool of possible solutions. Initial population is generated randomly and mimicking the Darwinian evolution, it continues until the satisfactory fitness is found or touches the iteration limit. Typically symbolic regression is performed using GP where each individual represents a single tree structure. On the other hand, in multigene genetic programming (MGGP), each individual is composed of one or more gene and each gene form a traditional GP tree. The output of a symbolic regression model is a weighted linear summation of genes. Typically, a symbolic regression model encodes a mathematical expression or function to predict an output using a corresponding inputs variables (1).

$$\widehat{y} = f(F_1, F_2, \ldots, F_M) \tag{1}$$

where

- $\widehat{y}$: output response
- f: symbolic non-linear function or a collection of non-linear functions
- $F_1, F_2, \ldots, F_M$: input variables (features)
- M: number of input variables (features)

In the beginning, a symbolic expression is randomly generated using simple tree building algorithm that randomly picks node from the supplied function set (e.g. $+$, $-$, *, /), input variables, and randomly generated constants. Nodes are then randomly assembled maintaining tree depth or size limitation and form tree structure of symbolic expressions. Evaluating the population for a number of generation by genetic operations, the tree expression with the best fitness is selected as the solution. A tree structure of symbolic regression model with three input variable and two genes is illustrated in Fig. 2(a), where $\widehat{y}$ is output, x_1, x_2 and x_3 are input variables, b_0 is bias and b_1 and b_2 are weight variables. Two fundamental genetic recombination operations are crossover and mutation. Crossover is performed on two parent trees, one sub-tree from each tree is selected randomly and exchanged between them to generate two new child trees (Fig. 2(b)). As presented in the figure sub-tree (-5*x2) and (x1/x2) from parents was interchanged to formulate new child genes. On the other hand, the mutation is performed on a single parent tree. A sub-tree (cos*x3) is deleted randomly from the parent and generate a sub-tree (x1/x2) randomly using the same tree building algorithm as the initial state to insert in position from where the sub-tree is deleted (Fig. 2(c)).

Table 1
Statistical descriptions of dataset.

Participants demographics ($N = 111$)	
Gender	Male 65, female 46
Age (years)	0–79 ($\mu = 32.87, \sigma = 16.91$)
Gold standard value	
Hemoglobin (g/dL)	7.6–21.49 ($\mu = 12.77, \sigma = 2.20$)
Glucose (mmol/L)	2.67–21.11 ($\mu = 6.47, \sigma = 2.76$)

4. Methodology

In this section, our proposed methodology is explained elaborately. Data collection, PPG signal generation, feature extraction, feature selection and model construction process are explained here. Main steps of our proposed system are depicted in Fig. 3.

4.1. Data collection

In this work, fingertip videos (.mp4) of 111 various type of patients including male and female as well as infant are collected. Moreover, in our dataset both diabetic (Type-1 and Typ-2) and non-diabetic patient are included. Premature baby as well as pregnant patient are also included in our dataset. At the same time, the true value (gold standard) for hemoglobin (complete blood count (CBC)) is measured with Sysmex XS-800i Haematology Analyzer, and glucose (random blood sugar test-RBS) is measured with Thermo Scientific Konelab 60i, Chemistry Analyzer respectively in the clinical laboratory. Data description is presented in Table 1. Before the data collection collector was trained properly on how to collect fingertip video. Data collection is a very vital step as data can be corrupted within a moment for a simple mistake and affect all the next processes and even all the effort can be destroyed.

The previous works we have studied used 600–1400 nm NIR-LED light for the data collection. For example, Ramasahayam et al. [29] used 935, 950, and 1070 nm NIR to estimate glucose level. In HemaApp by Wang et al. [3,21] used 500, 700–1300 nm NIR to measure

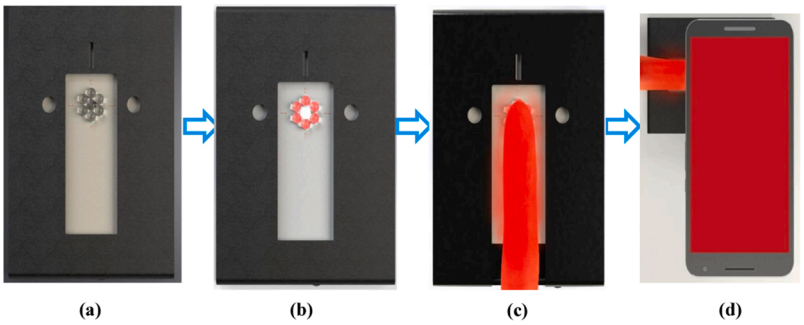


Fig. 4. Data collection: (a) NIR-LED box, (b) power on of the box, (c) finger placed on the box, (d) capturing video.

hemoglobin concentration. However, in this paper, 850 nm NIR-LED board has been used for illuminating the finger while recording the video using smartphone camera. It is consisting of six infra-red (NIR) LEDs and one flash LED. Traditionally, PPG signal acquired using optical techniques like sensor-based device. However, we used smartphone camera as a sensor to capture variation of light intensity reflected from a finger caused by the change of blood volume in systolic and diastolic cycles. The device needs to wear only during examination not all the time. From the experience of data collection, we have found out some challenges and recommendations:

1. Index finger was clean and dry before collecting the fingertip video otherwise we cannot obtain the PPG feature correctly. Nail polish was strictly prohibited.
2. The NIR-LED board was constructed user friendly so that participant can easily put the finger on the board. The index finger was recommended for capturing the fingertip video.
3. While capturing the video, the finger was placed with normal pressure and without any movement (wait some time for camera stabilization) otherwise, noise may be inserted. A slightest movement may disrupt the data and if extra pressure is produced on the finger then, the arterial wall may be deformed and leads to wrong reading [34].
4. Another problem in data collection is, if a camera is on for a long time then heat can be produced by the camera that leads to wrong reading.
5. Same condition (room temperature and light, camera device) should be maintained during the acquisition of data from any participant so that all data remain in same scale.

After the installation process following these recommendations, the finger is placed on the NIR-LED board and illuminating the finger with the light. Finally, a 15 s video is captured placing a smartphone (Nexus-6p) camera on the finger according to Fig. 4. From each video data first 3 s and last 2 s are discarded to avoid unstable frame.

4.2. PPG signal generation and feature extraction

The PPG signal reflects the blood movement in the vessel, which goes from the heart to the fingertips and toes through the blood vessels in a wave-like motion. According to heart blood circulation patterns, arteries carry more blood in the systolic period than the diastolic period. As a result, the absorb (blood) of light in the tissue with arteries varies with these two blood circulation period. During the systolic period, the diameter of arteries is higher than the diastolic period and light passes through a longer path. In contrast, light passes a shorter path during the diastolic period. For these reasons, the light intensity is changed with time and the pattern is called PPG-wave.

Region of interest (ROI) has a significant effect on PPG signal generation [44,45]. For PPG generation from an image frame, the choice of channel (RED, GREEN, BLUE) is a vital issue. It is seen that if pixel

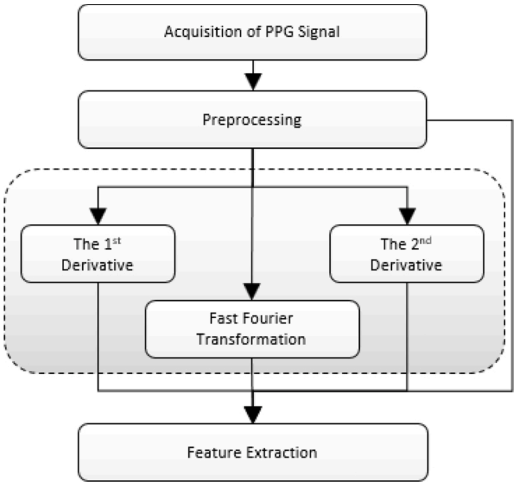


Fig. 5. Block diagram of feature extraction.

intensity in an image is <200 then, it is very difficult to obtain all the characteristic features from the generated PPG signal. RED channel intensity was between 225 and 240. On the contrary, GREEN and BLUE Channel intensity were 0–3 and 15–25 respectively. HemaApp [3] used the central region of an image frame and averaged the intensity for each channel, while Scully et al. [46] took 50 × 50 pixels frame on the green channel and Jonathan et al. [47] used 10 × 10 pixels block of mean intensity from the central region. SmartHeLP partition the image into 10 × 10 pixels block to locate the best position for strong PPG signal. However, we have considered only the highest intensity channel (RED) with the highest positive peck of 500 × 500 pixels from right to left.

The following steps are followed to obtain PPG Wave from collected videos:

1. Video is formed by a sequence of frames of a scene. From each video 300 image frames $(10\,s \times 30\,\text{fps})$ are extracted.
2. Each image frame has three different channels (RED, GREEN, BLUE) and channels are separated using chromatic analysis. Then, the channel with the highest intensity is identified and used in the next step.
3. Bandpass Butterworth Filter [34] is applied to the output of the previous step. Butterworth filter flattens the frequency response as much as possible.
4. Peak detection algorithm [48,49] is applied to find the peaks on the signal. Best PPG wave (the wave which has highest positive peak) is identified from PPG signal and stored for feature extraction.

At first, various prepossessing action had been applied to reduce the noise as raw PPG signals are prone to the noise and motion artifacts. A finite impulse response (FIR) filter was applied to eliminate noise. After

Table 2
Extracted features from PPG.

Feature	Descriptions	Feature	Descriptions
F_1: Age	Age of patient in years	F_{25}: t_{e1}	Interval time from point f_1 to point e_1
F_2: Gender	Male or Female	F_{26}: t_{f1}	Interval time from point f_1 to next f_1 point
F_3: x	Systolic peak [35]	F_{27}: b_2/a_2	Ratio of b_2 and a_2 [36, 37]
F_4: y	Diastolic peak [35]	F_{28}: e_2/a_2	Ratio of e_2 and a_2 [36, 38]
F_5: z	Dicrotic notch	F_{29}: $(b_2+e_2)/a_2$	Ratio of (b_2+e_2) and a_2 [38]
F_6: t_{pi}	Pulse interval	F_{30}: t_{a2}	Time interval from point f_2 to next point a_2
F_7: y/x	Augmentation index [36]	F_{31}: t_{b2}	Time interval from point f_2 to next point b_2
F_8: (x-y)/x	Alternative augmentation index [39]	F_{32}: t_{a1}/t_{pi}	Ratio between time interval of $a_1(t_{a1})$ and pulse interval
F_9: z/x	Ratio of dicrotic notch and systolic peak [40]	F_{33}: t_{b1}/t_{pi}	Ratio between time interval of $b_1(t_{b1})$ and pulse interval
F_{10}: (y-x)/x	Negative relative augmentation index	F_{34}: t_{e1}/t_{pi}	Ratio between time interval of $f_1(t_{pi})$ and pulse interval
F_{11}: t_1	Systolic peak time	F_{35}: t_{f1}/t_{pi}	Ratio between time interval of $f_1(t_{pi})$ and pulse interval
F_{12}: t_2	Dicrotic notch time	F_{36}: t_{a2}/t_{pi}	Ratio between time interval of $t_{a2}(t_{pi})$ and pulse interval
F_{13}: t_3	Diastolic peak time	F_{37}: t_{b2}/t_{pi}	Ratio between time interval of $b_2(t_{pi})$ and pulse interval
F_{14}: ΔT	Time between systolic and diastolic peaks	F_{38}: $(t_{a1}+t_{a2})/t_{pi}$	Ratio of $(t_{a1}+t_{a2})$ pulse interval
F_{15}: $t_1/2$	Time of half systolic peak point	F_{39}: $(t_{b1}+t_{b2})/t_{pi}$	Ratio of $(t_{b1}+t_{b2})/t_{pi}$ and pulse interval
F_{16}: A_2/A_1	Inflection point area ratio-IPA [41,42]	F_{40}: $(t_{e1}+t_2)/t_{pi}$	Ratio of $(t_{e1}+t_2)/t_{pi}$ and pulse interval
F_{17}: t_1/x	Systolic peak rising curve	F_{41}: $(t_{f1}+t_3)/t_{pi}$	Ratio of $(t_{f1}+t_3)/t_{pi}$ and pulse interval
F_{18}: $y/(t_{pi}-t_3)$	Diastolic peak downward curve	F_{42}: f_{base}	Fundamental component frequency acquired from FTT
F_{19}: t_1/t_{pi}	Ratio of t_1 and t_{pi}	F_{43}: $\lvert s_{base}\rvert$	Fundamental component magnitude acquired from FTT
F_{20}: t_2/t_{pi}	Ratio of t_2 and t_{pi}	F_{44}: f_{2nd}	2nd harmonic frequency acquired from FTT
F_{21}: t_3/t_{pi}	Ratio of t_3 and t_{pi}	F_{45}: $\lvert s_{2nd}\rvert$	2nd harmonic magnitude acquired from FTT
F_{22}: $\Delta T/t_{pi}$	Ratio of Time between systolic and diastolic peaks and pulse time interval	F_{46}: f_{3rd}	3rd harmonic frequency acquired from FTT
F_{23}: t_{a1}	Interval time from point f_1 to point a_1	F_{47}: $\lvert s_{3nd}\rvert$	3rd harmonic magnitude acquired from FTT
F_{24}: t_{b1}	Interval time from point f_1 to point b_1	F_{48}: sVRI $= V_2/V_1$	Stress-induced vascular response index [43]

prepossessing, 46 characteristic features of PPG are extracted analyzing the PPG wave (F_2 to F_{22} and F_{48}), its first and second derivative (F_{23} to F_{41}) [50] and using fourier analysis (Fast Fourier Transformation (FTT)) (F_{42} to F_{47}). Additionally, age (as year) and gender (as male or female) are added to the feature set. Block diagram of the feature extraction process is depicted in Fig. 5. All the features are listed in Table 2 and features with characteristics point of PPG are illustrated in Fig. 6.

4.3. Feature selection

Feature selection is the most important step before model construction as the prediction power of a model depends on the features. Redundant, irrelevant, or partially relevant features can negatively affect model performance. There are several benefits of performing feature selection before developing the model. Firstly, reduces overfitting opportunities by discarding redundant features. Secondly, this process discards irrelevant features which reduce misleading opportunities and improves model accuracy. Lastly, it reduces the number of features, hence reduces the complexity of the algorithm and model trains faster. There are a number of feature selection methods for feature selection. In this work, a correlation-based feature selection (CFS) using a genetic algorithm is applied [51]. It works based on the genetic algorithm (GA). Initially, it selects a subset of features randomly as initial population (a pool of possible solutions) and computes fitness by an objective function (6) that measures the suitability of the features. According to GA, a number of genetic operations are performed on the population to generate a new population. Fitness is calculated in the same way as the initial population and repeat the process. Features (individuals) in each generation are selected by the search technique based on fitness value. The individual with a larger fitness value has a higher probability of selection. This process ends when it reaches the maximum generation limit or touches the satisfactory fitness level. To select the best features using GA, we used following control parameter: population limit = 75, generation limit = 100, probability of crossover = 0.10, probability of crossover = 0.20, crossover criteria = 2-point, selection criteria = roulette wheel. Feature selection process is illustrated in Fig. 7. The probability of selecting kth individual from the population is indicated in (2):

$$p_i = \frac{\mathrm{fit}_k}{\sum_{j=1}^{n}\mathrm{fit}_j} \tag{2}$$

where

- p_i: probability of selection of ith individual.
- fit_k: fitness of kth individual.

A suitable objective function is required to test fitness. Here, a correlation-based fitness function is used. Let, $o_1, o_2, \ldots, o_N$ be the output values and $F_1, F_2, \ldots, F_N$ be the corresponding n-dimensional feature vectors. Distance between any two output values, D_o can be expressed as (3) and the distance between corresponding feature vectors, D_F can be express as (4) and (5):

$$D_o = o_i - o_j \tag{3}$$

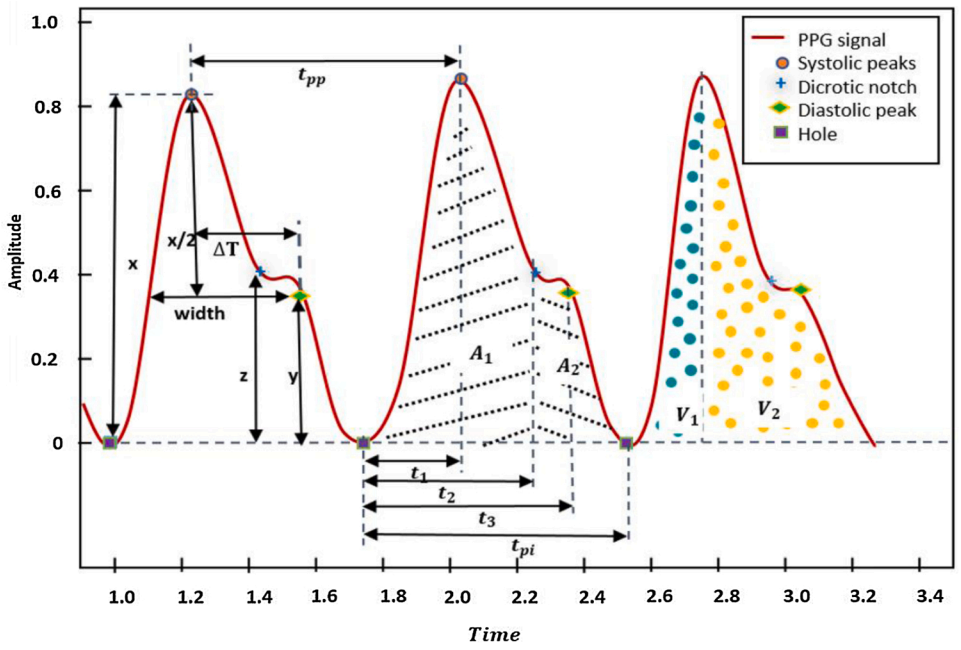
- If $D_o \geq 0$

$$D_F = \sqrt{\frac{\sum_{k=1}^{n}(F_{i,k}-F_{j,k})^2}{n}} \tag{4}$$

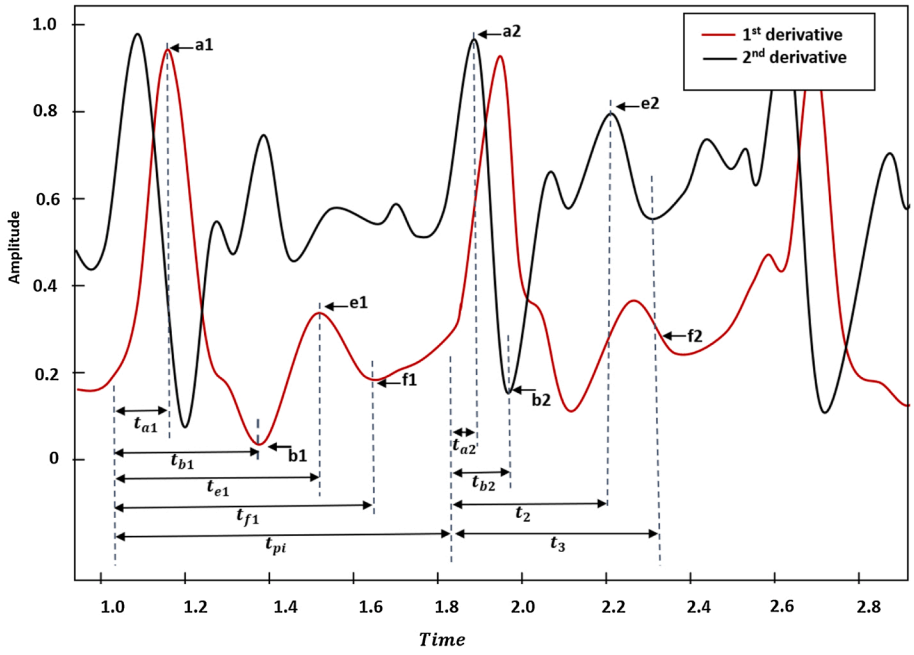
- If $D_o < 0$

$$D_F = -\sqrt{\frac{\sum_{k=1}^{n}(F_{i,k}-F_{j,k})^2}{n}} \tag{5}$$

Finally, correlation, r (as objective function) between D_F and D_o can be calculated as (6):



(a)



(b)

Fig. 6. The characteristic features: (a) acquired from PPG signal, (b) acquired from the first and second derivatives.

$$r = \frac{\frac{\sum_i (D_{F_i} - \bar{D}_F)(D_{o_i} - \bar{D}_o)}{n-1}}{\sqrt{\frac{\sum_i (D_{F_i} - \bar{D}_F)^2}{n-1} \frac{\sum_i (D_{o_i} - \bar{D}_o)^2}{n-1}}} \quad (6)$$

As the goodness of fit increases, r becomes larger. The goal is to maximize the value of r. Finally, we obtained 22 features for hemoglobin and 31 features for glucose shown in Table 3, and their corresponding fitness values are 0.66 and 0.74, respectively (Fig. 8).

4.4. Model construction and validation

Two independent models using MGGP were constructed on the basis of the features selected in the feature selection stage by GA. In this study, GPTIPS toolbox [52] written in MATLAB was used to generate two mathematical model for estimating both hemoglobin and glucose. To develop the MGGP models, a single computer (Asus Intel® Core™ i7, 32 GB RAM, NVIDIA GeForce GTX 980M GPU) with Windows – 10 operating system was used. These two models were trained with GPTIPS separately with the selected feature set. Parallel computing paradigm is

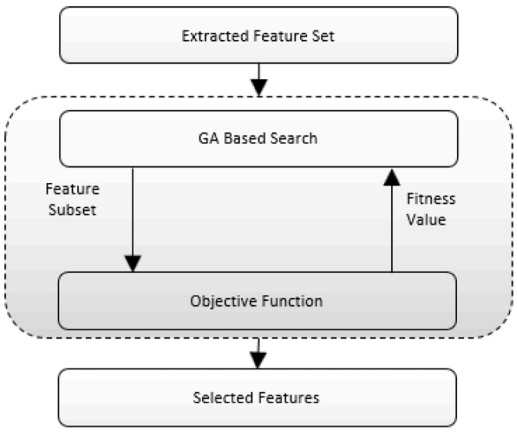


Fig. 7. Block diagram of feature selection.

Table 4
Control parameter for MGGP.

Parameter	Value
Population size	2000
Maximum generation	1000
Selection method	Tournament
Type of crossover	2-point
Maximum tree depth	9
Number of gene	8
Probability of crossover	0.85
Probability of mutation	0.10
Elite fraction	0.15
Function set	+, −, x, /, sqrt, sin, cos, tan, tanh, exp, log, power, abs, ifthe, negexp,gth,lth
Fitness function	GPTIPS [52] default (root mean square error)

available in GPTIPS and we activated the feature with six workers that reduced the training time significantly. Before the model construction, setting control parameter is a very important issue as it is intrinsic to the model design and effect the estimation capability of the model. Model complexity directly depends on proper number of populations, maximum tree depth and the number of generations. It can be concluded that model speed is decreased with the increases in the complexity of the derived equation. So, these facts need to be kept in mind when setting control parameters. Control parameters that were applied for both the model listed in Table 4. These parameters were set after several preparatory runs and on the basis of previous knowledge of predictive modeling with other dataset. So they may not be optimal. Random permutation was done before model construction to avoid overfitting. In this paper, 5-fold cross-validation was employed to validate models [53]. Input dataset is randomly partitioned into five equal size subsets and each round a single subset is used as validation data for validating the model, and the remaining four subsets are used to train the model. The cross-validation process is then repeated 5 times and each subset is used exactly once as a validation set. For 5-cross fold validation, this process ends with $K = 5$ iterations which was described algorithmically in Algorithm 1. To investigate the competency of the new developed model, the dataset is also trained with classical regression methods such as linear regression (LR), support vector regression (SVR), and random forest regression (RFR) using the same input condition as MGGP based models. Best MGGP models are selected based on fitness value and model complexity.

Table 3
Selected features.

Dataset	Selected Features														Count
Hemoglobin	F_1	F_2	F_4	F_5	F_6	F_7	F_9	F_{10}	F_{16}	F_{17}	F_{18}	F_{19}	F_{28}	F_{31}	22
	F_{32}	F_{33}	F_{34}	F_{36}	F_{39}	F_{40}	F_{43}	F_{45}							
Glucose	F_1	F_2	F_3	F_4	F_5	F_{11}	F_{12}	F_{13}	F_{14}	F_{19}	F_{21}	F_{22}	F_{23}	F_{24}	31
	F_{25}	F_{26}	F_{32}	F_{34}	F_{35}	F_{36}	F_{38}	F_{39}	F_{40}	F_{41}	F_{42}	F_{43}	F_{44}	F_{45}	
	F_{46}	F_{47}	F_{48}												

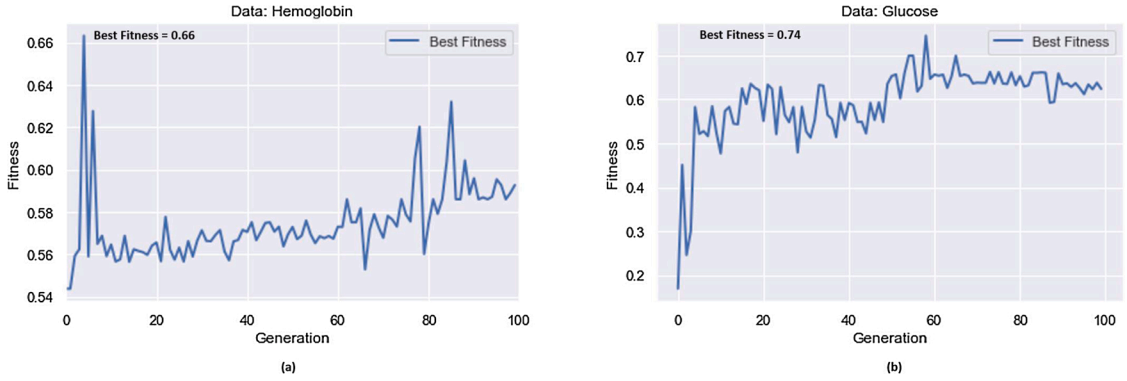


Fig. 8. Feature selection results: (a) hemoglobin, (b) glucose.

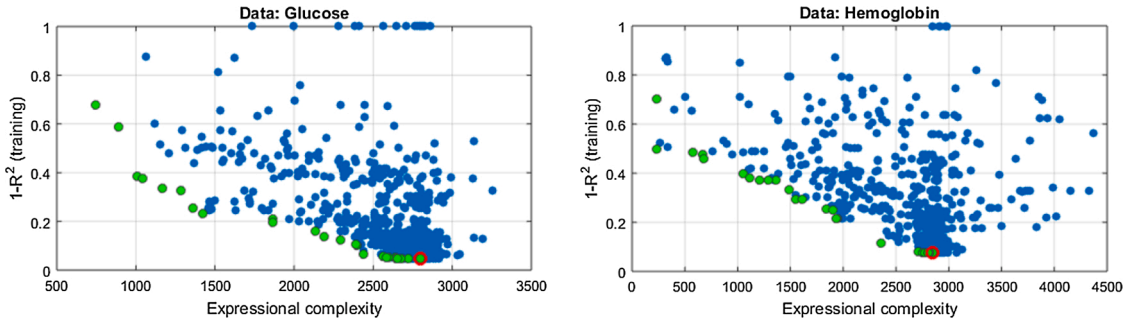


Fig. 9. All models developed using MGGP (solid blue circles), Pareto front results obtained by non-dominated sorting algorithm (solid green circle), and the selected MGGP model (solid green circle with red border).

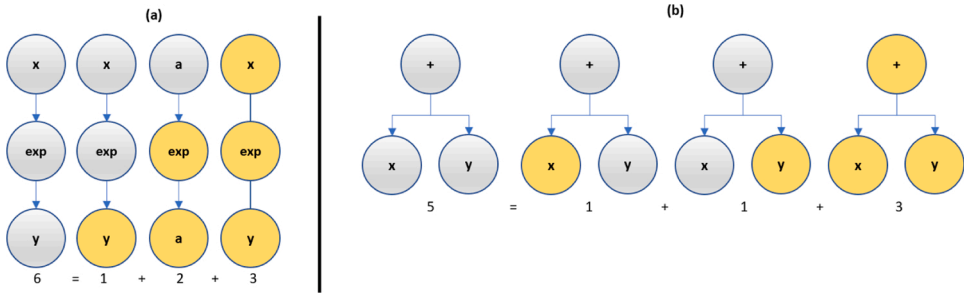


Fig. 10. Complexity calculation of two 3-node simple model.

Algorithm 1. K-fold cross-validation for each model to estimate blood hemoglobin level and glucose level

Input: input dataset $\mathcal{D} = \{\mathcal{X}_i, y_i\}_{i=1}^{m}$ where, $(\mathcal{X} \in \mathbb{R}^n,\ y \in \mathcal{Y})$, $\mathcal{P}_{set}$ is the set of parameter for each model, $\mathcal{M}$.

Output: Final performance evaluation indices.

for $j \leftarrow 1$ **to** $\mathcal{K}$ **do**
  Divide $\mathcal{D}$ into $\mathcal{D}_j^{train}$ and $\mathcal{D}_j^{test}$ for j^{th} split ;
  Train the model $\mathcal{M}$ with $\mathcal{D}_j^{train}$ using $\mathcal{P}_{set}$;
  Evaluate the performance $\mathcal{P}_j$ for $\mathcal{M}$ with $\mathcal{D}_j^{test}$;
Calculate mean of the performance for each regression model $\overline{\mathcal{P}}_{\mathcal{M}} = \frac{1}{\mathcal{K}} \sum_{n=1}^{\mathcal{K}} \mathcal{P}_n$;
return $\overline{\mathcal{P}}_{\mathcal{M}}$;

4.5. Experimentation

Two independent MGGP symbolic regression model with 1000 number of generations, 2000 size of the population, 48 input variables, 17 functions set for estimation hemoglobin and glucose have been developed using the GPTIPS toolbox. Fig. 9 shows all models developed using MGGP in terms of its predictive performance and model complexity characteristics. Each model is presented as a solid circle where the y axis represents the performance $(1 - R^2)$, and the x axis represents the complexity of the model. The complexity of a model is the sum of the expressional complexities of its constituent genes (trees). For a single tree, complexity can be calculated by summing together all its nodes and nodes of all possible full sub-trees (a leaf node is also considered as a full sub-tree). Hence, two trees with equal number of node, the balanced tree has lower expressional complexity than the deeper tree. For example, two simple 3-nodes model is presented in Fig. 10, total number of nodes of the model (a) (Fig. 10(a)) and its all possible full sub-trees is 6. On the other hand, the total node counts for the model (b) and its all possible full sub-trees are 5. Pareto front in the population is obtained by a non-dominated sorting indicated by the green dot. The red circled green dot represents the best model in the whole population in terms of performance. One can easily trade-off between the performance and the complexity of the model from the Pareto front members and choose the best model. For each model, linear weights (coefficients) are computed using the ordinary least squares method from the training data. Thus, MGGP combines the power of ordinary linear regression with the ability to capture non-linear behavior without needing to pre-specify the structure of the non-linear model.

5. Results and discussion

In this section experimental results and performance of MGGP model will be described briefly.

GPTIPS toolbox automatically generated two individual compact mathematical equation to estimate hemoglobin and glucose. A user can easily grasp the terms of the predictive equation which help to trust the model [32]. Human can easily insights into the MGGP model as it is visible to everyone, unlike typical back box predictive models. It is difficult to overdraw the importance of trust and understanding in predictive models. Unlike most of the soft-computing techniques such as artificial neural network (ANN), deep learning (DNN) or support vector machine (SVM), no specialized software environment setting is needed to deploy the trained MGGP model. The model equation can easily be used in any environment outside MATLAB for structural transparency of the model. A user can easily convert the model equation to any modern computing language without prior knowledge of MGGP. Mathematical formula for hemoglobin and glucose estimation are provided in Appendices A and B respectively.

Performance of the models has been evaluated using sing five metrics: mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), coefficient of determination (R^2). Mathematical equations for these metrics can be expressed as follows:

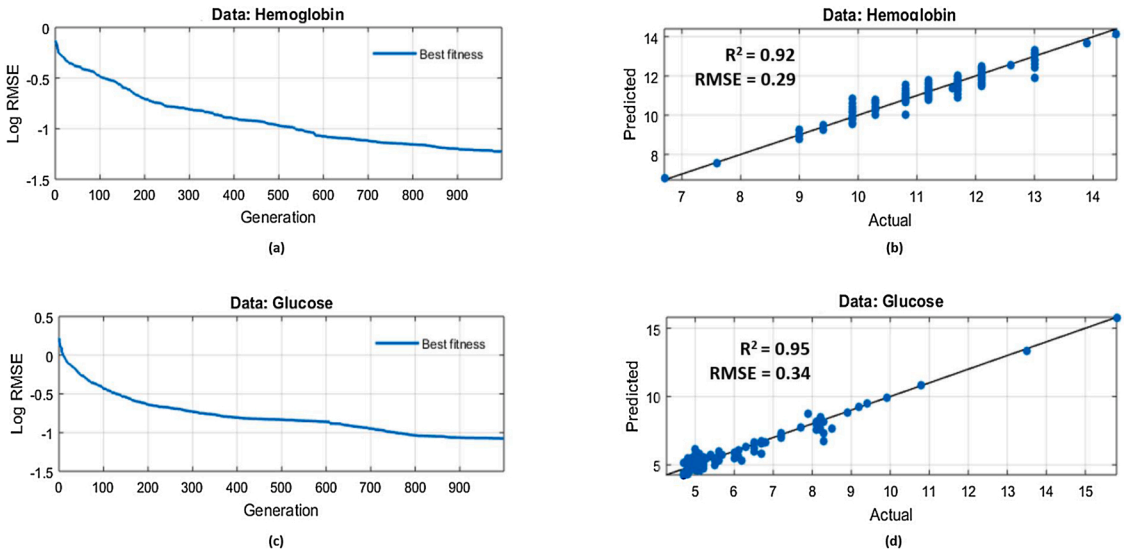


Fig. 11. Relationship and convergence curve at training for MGGP based model: (a) convergence curve (hemoglobin), (b) relationship (hemoglobin), (c) convergence curve (glucose), (d) relationship (glucose).

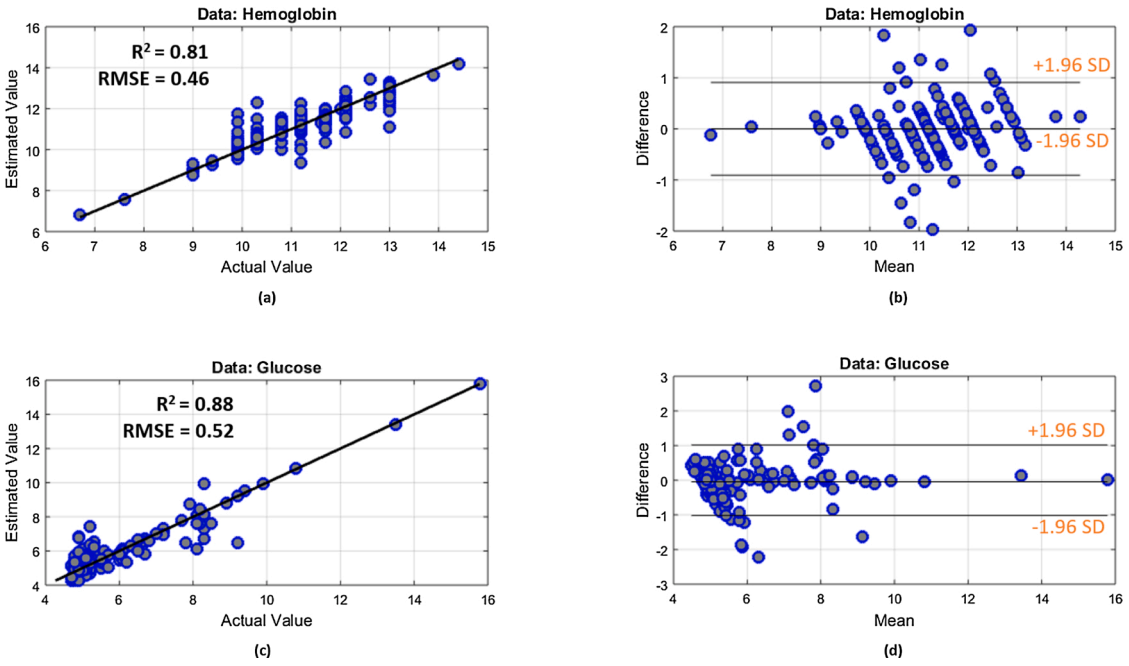


Fig. 12. Relationship and agreement between predicted value and actual Value at testing for MGGP based model: (a) relationship (hemoglobin), (b) agreement (hemoglobin), (c) relationship (glucose), (d) agreement (glucose).

$$\text{MAE} = \frac{\sum_1^n |y_i - \hat{y}_i|}{n} \tag{7}$$

$$\text{MSE} = \frac{\sum_1^n (\hat{y}_i - y_i)^2}{n} \tag{8}$$

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{9}$$

$$R^2 = 1 - \frac{\sum_0^{n-1} (y_i - \hat{y}_i)^2}{\sum_0^{n-1} (y_i - \bar{y}_i)^2} \tag{10}$$

where $\bar{y} = \frac{1}{n} \sum_0^{n-1} y_i$

where $y_1, y_2, y_3, \ldots, y_n$ are true values and $\hat{y}_1, \hat{y}_2, \hat{y}_3, \ldots, \hat{y}_n$ are corresponding estimated values.

Convergence shows, further refinement in the model will lead it towards better results or not. Lack of convergence indicates that the data do not fit the model sufficiently. Convergence characteristics of MGGP models are shown in Fig. 11. It indicates that the fitness curve becomes flat after 700 generations for hemoglobin and 500 generations for glucose and changes in fitness function are not significant. It signifies that increasing generation value not likely to get a better result. However, the best solution is obtained at 997, 999 generations for hemoglobin and glucose respectively.

It is noteworthy that the proximity of R^2 to 1 indicates the strength of the relationship between model output and actual results. The RMSE, MSE show relative error, and MAE shows absolute error. R^2 and RMSE

Table 5
Performance comparison of different models.

Model	Hemoglobin				Glucose			
	R^2	MAE	MSE	RMSE	R^2	MAE	MSE	RMSE
LR	0.247	0.695	0.815	0.902	0.320	0.954	2.136	1.461
SVR	0.188	0.727	0.878	0.937	0.201	0.843	2.696	1.642
RFR	0.181	0.748	0.886	0.941	0.698	0.610	0.946	0.972
MGGP	**0.807**	**0.304**	**0.214**	**0.462**	**0.881**	**0.324**	**0.270**	**0.520**

Table 6
Comparison of our proposed MGGP based model with several exiting non-invasive methods.

Reference	Purpose	Device	Participant	Data	Signal	Algorithm(s)	Performance
HemaApp [3,21]	🩸	Nexus-6p	31	Video	PPG	LR, SVR	R = 0.62–0.82
Kavsaoğlu et al. [9]	🩸	Hemocue Hb-201TM	33	–	PPG	CART, LSR, GLR, MVLR, PLSR, GRNN, MLP, SVR	R^2 = 0.757–0.974
SmartHeLP [22]	🩸	Nexus-4p	75	Video	PPG	ANN	R^2 = 0.93
Ding et al. [24]	🩸	–	109	–	Spectra	ANN	R = 0.94
Yuan et al. [23]	🩸	Spectrometer	91	–	Spectra	PLS	R = 0.60–0.81
Soni et al. [28]	🩸	Samsung Galaxy SIII	167	Image	–	DNS	R = 0.44–0.94
Malin et al. [30]	🩸	–	10	–	Spectra	PLS	*MSE* = 1.03
Glovanni et al. [27]	🩸	Iphone 4s, Huawei p7	113	Image	–	KNN	R = 0.52–0.65
Robert et al. [25]	🩸	–	337	–	Spectra	MLR, BWA	R = 0.82
Anggraeni et al. [20]	🩸	Asus ZenFone 2 Laser	20	Image	–	LR	R^2 = 0.81
Pai et al. [31]	🩸	–	24	–	Photo acoustic	KBR	*RMSEP* = 9.64
Ramasahayam et al. [29]	🩸	–	–	–	PPG	ANN	*RMSE* = 5.84
In our study	**Nexus-6p**	**111**	**Video**	**PPG**	**MGGP**	R^2 = 0.81	
In our study	**Nexus-6p**	**111**	**Video**	**PPG**	**MGGP**	R^2 = 0.88	

🩸 = hemoglobin, 🩸 = glucose, LR = linear regression, SVR = support vector regression, CART = classification and regression trees, LSR = least square regression, GLR = generalized linear regression, MVLR = multivariate linear regression, PLSR = partial least squares regression, GRNN = generalized regression neural network, MLP = multilayer perceptron, ANN = artificial neural network, PLS = partial least squares, DNS = dinitrosalicylic acid spectroscopic, KNN = K-Nearest neighbor, MLR = multi-linear regression, BWA = bisquare weighting algorithm, KBR = kernel-based regression, RMSEP = root mean squared error of prediction.

for both hemoglobin and glucose on training datasets are illustrated in Fig. 11. From the illustrations, we can observe that higher R^2 values exceeding 0.92 are obtained for both hemoglobin and glucose which indicates that the models are fitted well with the input dataset.

In the next stage, MGGP models generalization is assessed on the test datasets. Fig. 12(a) and (c) represent the measure of how well the feature samples are likely to be predicted by the model. The coefficient of determination is a measurement used to explain how much variability of one factor can be caused by its relationship to another related factor. We achieved RMSE, 0.462 and R^2, 0.807 for hemoglobin and 0.520 and 0.881 for glucose estimation.

The Bland–Altman plot [54] is a graphical method to compare two measurements techniques. In this graphical method the differences between the two techniques are plotted against the averages of the two techniques. Horizontal lines are drawn at the mean difference, and at the limits of agreement, which are defined as the mean difference plus and minus ±1.96 times the standard deviation of the differences. From Fig. 12(b) and (d), we observed that only 6% value for hemoglobin and 4% value are out of the limit of agreement (±1.96 SD) for the testing dataset which is an evidence of a good agreement between the actual value and the estimated value.

The estimation accuracy for classical regression methods and MGGP based symbolic regression method are illustrated in Table 5. As presented in the illustration, the R^2, MAE, MSE and RMSE value for hemoglobin model are 0.803, 0.304, 0.214 and 0.462 respectively whereas other models have $R^2 \leq 0.247$, MAE ≥ 0.695, MSE ≥ 0.815 and RMSE ≥ 0.902. Similar favorable result is also found for glucose with R^2 = 0.881, MAE = 0.324, MSE = 0.270 and RMSE = 0.520. Overall, it is clear that, in all the error measurement indexes MGGP models achieved the best performance over classical regression models.

5.1. Comparison of results

A comparison study is drawn in Table 6 for estimating hemoglobin and glucose levels to validate our contributions with respect to exiting works.

In [3,21], authors predicted hemoglobin level non-invasive way; they used a smartphone (Nexus-6p), captured video data, used the LR and SVM models, and gained correlation coefficient $R = 0.82$ as performance. In [9], they used a device (Hemocue Hb-201TM) to collect the PPG signal for each subject, developed eight regression models and gained coefficient of determination R^2 = 0.757–0.974 as performance. In [22], the authors developed a mobile application named SmartHeLP to predict the blood hemoglobin level. Using a smartphone (Nexus-4p), 75 video data were collected and applied in an ANN model. The estimated accuracy of their proposed system was R^2 = 0.93. Similarly, in [20,24,27,23,25], the authors have worked for the same purpose, but the way of data collection and their proposed methodology varies from each other.

On the other, Soni et al. [28], authors estimated glucose level non-invasive way; they used the smartphone device (Samsung Galaxy SIII), captured image data, used the DNS model, and obtained a correlation coefficient (R) range 0.44 to 0.94 as performance. Similarly, Ramasahayam et al. [29], developed an ANN model for the same purpose, and RMSE was 5.84. Most of the existing methods used classical regression methods to estimate the values and different authors used

Table 7
Mathematical expression of MGGP model for hemoglobin.

$\hat{y} = 0.5316 \sin(\text{pdiv}(\cos(x_1), \tanh(\tanh(\tanh(x_{14}))))) + 5.934 \tanh(\tanh(x_2 + \exp(x_{17})) + \tanh(x_{15} + \text{gth}(\text{lth}(\tan(x_2), \exp(x_{10})), \sin(\tan(x_{21})^2)))) + 0.3441 \exp(\cos(\text{iflte}(\sin(\text{pdiv}(\cos(x_1), x_{14})), \text{iflte}(x_5, \tan(x_2), \sin(\text{plog}(\tanh(x_{13}))), x_9), \text{iflte}(x_{14}, \tan(x_{21}), x_{13}, \text{iflte}(x_{21}, x_3, \text{pdiv}(\cos(x_1), x_{14}), x_4)), 2.547))) + 25588.0 \tanh(x_{15} - 4.892) + 0.8041 \text{gth}(-4.909\, x_{10}\, x_{15}\, x_{21}\, \text{abs}(\text{plog}(x_4))\, \text{iflte}(\text{gth}(\text{gth}(-3.407, \text{plog}(x_{13})), x_6), \tanh(x_3), x_{22} x_1 + 4.892, x_1^2), \text{plog}(\text{iflte}(\tanh(\tanh(x_{14})), x_{10}, x_{13}, \text{abs}(x_3 + x_8 + \text{abs}(x_3 + x_8 + x_{14})) + \text{gth}(x_5, \text{pdiv}(x_{13}, x_{17}))))) - 0.5961 \text{lth}(\sin(\text{iflte}(x_{13}, \sin(\text{pdiv}(7.269, x_{17})) + \text{lth}(\text{iflte}(\text{abs}(\text{plog}(x_4)), x_2, -4.892, x_{20}), \tan(x_5)), \text{iflte}(x_{13}, \text{gth}(\cos(x_4), x_4), -4.892, x_{20}) (x_{12}^3 (9.143\, x_3 + 9.143\, \text{plog}(x_7))), \text{iflte}(x_{14}, x_{18}, \text{iflte}(x_{14}, \text{abs}(\cos(x_1)), \text{iflte}(x_{13}, x_8, x_6, \text{iflte}(x_{10}, \text{gth}(x_{16}, x_6), x_{13}, \text{pdiv}(x_{13}, x_{17}))), \text{plog}(-1.0)), \sin(x_7)))), \text{lth}(\sin(5.498\, x_{11}), x_{17})) + 2.125 \text{gth}(x_{19} x_{21}^2 - 2.082, \text{plog}(\text{iflte}(x_{14}, x_{10}, x_{13}, \text{iflte}(x_{13}, x_{21}, -0.7615, x_{13})))) + 0.3441 x_{15} x_{22} (\tan(\text{gth}(x_{14}, \exp(-\tan(x_8))) (x_8 - 1.0 \tanh(x_{15})) \text{iflte}(\text{gth}(\tan(x_2), \text{pdiv}(x_{13}, x_{17})), x_{21}, -4.898, \text{pdiv}(\cos(x_1), x_{14}))) + x_6^2) - 0.01898 \text{iflte}(x_{16}, x_{10}, \text{iflte}(x_{13}, \text{gth}(-0.7615, \text{pdiv}(x_{13}, x_{17})) (x_8 - \text{lth}(\sin(5.743\, x_7), \text{plog}(x_{22}^{1/2}) + \tan(x_{21}))) \text{iflte}(\text{lth}(\text{abs}(x_2 + x_{14} + x_{21}), \exp(\sin(x_1))), \text{plog}(x_3), x_{22}, \text{pdiv}(\cos(x_1), x_{14})), -7.498, \text{iflte}(x_{14}, x_6, \text{iflte}(\cos(\text{plog}(x_{13})), \text{abs}(x_2 + x_{14} + x_{21}), \text{iflte}(x_{13}, \tanh(x_4), \tan(x_1) - x_1 + 1.498\, x_1\, x_{13}, \text{iflte}(x_{14}, x_{17}, x_{11} + x_{13} + 23.83, x_1)), \sin(x_1 x_{21})), x_1 + \text{pdiv}(\cos(x_1), x_{16}) + \text{iflte}(\text{abs}(x_2 + x_{14} + x_{21}), \text{lth}(x_{10}, \tan(x_{21})), -4.892, \text{pdiv}(x_{13}, x_{17})))), x_1 + \text{iflte}(\text{iflte}(x_9, \text{gth}(x_{16}, x_6), x_{13}, \tanh(\tanh(x_{14}))), \text{gth}(\text{gth}(-3.558, \text{plog}(x_{20})), x_7), x_5^2, \text{gth}(\sin(\text{gth}(x_6, x_{15})^2), \sin(x_7^2)))^2 + \text{iflte}(\cos(x_1), \text{gth}(\text{gth}(-3.558, \text{plog}(x_{13})), \tan(x_9)^2), x_5^2, \exp(\sin(x_1)))^2) + 0.3441 \text{iflte}(\text{iflte}(\tanh(\text{lth}(\tan(\exp(-\tan(x_8)))), x_{10})), \text{gth}(\tanh(\sin(\tan(x_{21})^2)), \tanh(x_{19})), x_5, \tan(x_2)), \tan(\text{gth}(\text{lth}(\tan(x_2), \text{iflte}(x_{13}, \tanh(x_{16}), x_5, x_9)^2), \sin(\tan(x_3 + x_8 + x_{19})^2))), \sin(\text{plog}(\tanh(\text{iflte}(x_9, \text{abs}(\text{plog}(x_4)), x_{13}, \tanh(\tanh(x_{14}))))))), \tanh(\text{iflte}(\text{plog}(x_4) x_1 + 4.892, x_2 - x_1 + \exp(-\tan(x_8)) + \tan(x_{13})^2, \text{iflte}(x_{13}, x_{21}, x_1, x_{13}), x_{10})) \tan(x_5) \tan(x_{20})) + 25588.0$

Where $\hat{y}$: estimated value and x_i: denotes corresponding feature value F_i.

Table 8
Mathematical expression of MGGP model for Glucose

$\hat{y} = 0.02087\, x_1 + 2.124 \tanh(\tan(\tanh(\text{plog}(\text{plog}(x_{23}) - x_3\, x_{23} - \exp(-x_4) \exp(-x_{23}) \sin(x_4)))) \tanh(\tanh(\tanh(x_3\, x_{21}) (4.0\, x_2 + x_3 - 5.0\, x_{15} + x_{28} + \sin(x_4) + 2.0 \text{pdiv}(x_{23}, x_4))))) - 0.02087 \tan(\exp(-1.0\, x_4) \exp(-1.0\, x_{23}) \sin(x_4)) + 0.02087 \text{plog}(\text{pdiv}(x_{25} - x_5 + \text{pdiv}(x_5, x_4), \tan(x_4) \tan(x_{29}) (x_7 - x_{25}))) + 0.02087 \text{plog}(x_{22} (x_4 - x_{25})) - 0.2112 (\text{abs}(x_9 + \text{abs}(x_9) + \text{iflte}(x_9, x_{23}, \text{lth}(x_{23}, x_3), x_{15})^2) + \tanh(x_{21} (x_2 - 1.0\, x_{26} + \tanh(\text{pdiv}(x_{13}, x_3)) - 1.0 \text{abs}(x_3) + \text{pdiv}(x_{13}, x_3))) + x_9^2)^4 + 0.0002381 (x_3 + x_4 - 1.0\, x_5 + \tan(\exp(x_{25})) + \exp(x_{25}) - 1.0 \exp(x_{25}) \tan(x_{25}) - 1.0 \text{pdiv}(x_{13}, x_{20}) + \text{pdiv}(x_5, x_8 - x_{23}) + \text{pdiv}(x_1, \tanh(x_2)) + \text{pdiv}(x_9^{1/2}, x_8 - x_5))^2 + 0.02087 \text{pdiv}(x_{28}, x_8 - x_5) + 0.9536 (\exp(-1.0 \exp(-1.0 \text{pdiv}(2\, x_{26} - 2\, x_{25}, \text{abs}(\text{plog}(x_{23})))) \text{abs}(\sin(0.3211 \text{plog}(x_{22})))) + x_{22} \text{plog}(x_8 - x_{25} - \tan(x_{25}) + \text{pdiv}(x_5, x_4)))^2 + 0.1531 \text{abs}(\text{plog}(\text{plog}(\text{pdiv}(x_5 + \text{pdiv}(x_{21}, x_{20}), x_{14} x_{24} \text{pdiv}(x_{14}, x_9)))) + \exp(5.954\, x_8 x_{14} x_{29})) \text{abs}(\sin(\tan(x_{20}) (\text{plog}(x_{23}) - 1.0\, x_{25} \exp(-1.0\, x_{14}) \sin(x_4)))) + 0.02087 \text{pdiv}(\tan(x_{25}) + ((x_5 - x_{14})^2) \sin(x_4), \text{plog}(x_{23}) - x_{25} \sin(\sin(x_4)) \exp(-x_{28})) + 0.3072 \text{iflte}((-4.627) x_{12}, \text{gth}(\tanh(x_2\, x_8 \sin(x_4) (x_{22} - x_4 + \exp(x_{25}) + \tan(x_{15}))), \text{gth}(x_5, \tanh(\text{plog}(x_{23}) - x_{25} - x_2\, x_{21} \exp(-x_{23}))^2)), \text{abs}(\text{gth}(\exp(x_{25}), \text{abs}(x_{29})) \text{pdiv}(\exp(-\exp(-x_9)), \text{iflte}(\text{abs}(\text{gth}(x_{24}, \tanh(x_{23}))), x_{25}, (x_{21} + \exp(-x_4))^2, x_{23})) \text{pdiv}(x_5, \text{plog}(x_{23}) - \text{pdiv}(x_{14}, x_9))), \text{iflte}(x_{23}, x_{14}, x_{29} \tanh(\tanh(x_{18} x_{27})) \exp(-x_3 (x_{21} + \sin(x_5))^2) \sin(x_4), x_3 + x_{12})) - 0.02087\, x_4 \tan(\tan(x_{25})) \tan(\exp(-1.0\, x_4)) \tan(x_{25}) - 0.0005346\, x_5 \exp(x_{25} - 1.0\, x_{26}) (\tan(x_5) + \text{pdiv}(x_5, x_8 - x_{23}) + \text{pdiv}(x_1, \tanh(x_2)) - 1.0\, x_5 \exp(-1.0\, x_4) \exp(-1.0\, x_{24}))^2 \text{iflte}(x_4 + x_8 + x_{15}, \tan(\text{pdiv}(x_{21}, x_4)) (3\, x_2 + x_3 - x_8 - x_9 + \tanh(x_8 - x_{21}) + \exp(x_{25}) + \sin(x_4)), \text{pdiv}(x_8, x_{16}^2 x_{29}^2), x_{23} \sin(\tan(\tan(x_6)) \text{plog}(x_{22}))) + 3.79$

Where $\hat{y}$: estimated value and x_i: denotes corresponding feature value F_i.

their own device (sensor) to collect data from the subjects. To measure the goodness of their developed models, they used different indexes. However, used symbolic regression method to estimate hemoglobin and glucose level. Our primary aim is to developed a model that can used in any device without requiring any special hardware. Thus, we have used MGGP based symbolic regression and developed a mathematical model which can be easily implemented in real life applications without prior knowledge of MGGP.

6. Conclusions

In this study, an attempt was drawn up to estimate hemoglobin and glucose level based on multigene genetic programming (MGGP). MGGP was used to formulate two different mathematical equations for calculating hemoglobin and glucose level from PPG features extracted from fingertip video. We have also discussed the challenges and recommendations for noise-free data collection. In this work, we wanted to compare classical regression models with a symbolic regression model for medical data. To evaluate the performance of the MGGP based approach, results are compared with LR, SVR, and RFR model. From the MSE, RMSE and MAE value, it has concluded that MGGP based symbolic regression method had better estimation accuracy (± 0.304 with R^2, 0.807 for Hb and ± 0.324 with R^2, 0.881 for Gl) over conventional regression methods.

Our study shows that the PPG signal has sufficient information regarding blood hemoglobin and glucose level and any variations in blood hemoglobin or glucose level affect the PPG signal.

In the future, we will upgrade estimation accuracy with the smallest possible number of features. Firstly, the dataset will be extended from a heterogeneous people (e.g. different geographical region, patient of different disease and disorder, etc.) to make our model more universal and robust against unseen data. Then, we will provide a smartphone application to measure hemoglobin and glucose using the MGGP model. The application will capture data from the user and send it to a cloud server. Processing the data system will return the result to the user's display with some interpretation of the value with visualization. We also try to reduce features using a more precise analysis of the feature characteristics and reduce testing time.

In this work, we collected true value of hemoglobin and glucose for a wide range of value but unfortunately, we did not consider arrhythmia, hypoglycemia, hypertension when collecting data. In future, we hope to extend our work considering these issues.

Acknowledgment

The authors would like to thanks all the voluntary participants of this study and Dr. Moniruzzaman, Managing Director, physicians, nurses, and supporting members from Medical Centre Hospital, Chittagong, Bangladesh, for their constant support during the study. This work was financially supported in part by the Khulna University of Engineering & Technology (Khulna-9203, Bangladesh). The authors also thank anonymous reviewers for suggestions on improving this manuscript.

Conflict of interest

All authors have no conflict of interest to disclose this research.

Appendix A

Table 7

Appendix B

Table 8

References

[1] J. Li, C. Fernando, Smartphone-based personalized blood glucose prediction, ICT Express 2 (4) (2016) 150–154, https://doi.org/10.1016/j.icte.2016.10.001.
[2] D. Naumann, R. Meyers, Encyclopedia of Analytical Chemistry, John Wiley & Sons, Chichester, UK, 2000, pp. 102–131.
[3] E.J. Wang, W. Li, D. Hawkins, T. Gernsheimer, C. Norby-Slycord, S.N. Patel, Hemaapp: noninvasive blood screening of hemoglobin using smartphone cameras, in: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2016, pp. 593–604.
[4] S.S. Morris, M.T. Ruel, R.J. Cohen, K.G. Dewey, B. de la Brière, M.N. Hassan, Precision, accuracy, and reliability of hemoglobin assessment with use of capillary blood, Am. J. Clin. Nutr. 69 (6) (1999) 1243–1248.
[5] M.K. Hasan, N. Sakib, R.R. Love, S.I. Ahamed, Analyzing the existing noninvasive hemoglobin measurement techniques, in: 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), IEEE, 2017, pp. 442–448.
[6] M. Wieczorek, J. Siłka, M. Woźniak, Neural network powered covid-19 spread forecasting model, Chaos Solitons Fract. 140 (2020) 110203.
[7] X. Yang, Y. Yu, J. Xu, H. Shu, H. Liu, Y. Wu, L. Zhang, Z. Yu, M. Fang, T. Yu, et al., Clinical course and outcomes of critically ill patients with sars-cov-2 pneumonia in

Wuhan, China: a single-centered, retrospective, observational study, Lancet Respir. Med. (2020).
[8] J.S.M. Peiris, C.-M. Chu, V.C.-C. Cheng, K. Chan, I. Hung, L.L. Poon, K.-I. Law, B. Tang, T. Hon, C. Chan, et al., Clinical progression and viral load in a community outbreak of coronavirus-associated sars pneumonia: a prospective study, Lancet 361 (9371) (2003) 1767–1772.
[9] A.R. Kavsaoğlu, K. Polat, M. Hariharan, Non-invasive prediction of hemoglobin level using machine learning techniques with the ppg signal's characteristics features, Appl. Soft Comput. 37 (2015) 983–991.
[10] M.A.-u. Golap, M. Hashem, Non-invasive hemoglobin concentration measurement using mggp-based model, in: 2019 5th International Conference on Advances in Electrical Engineering (ICAEE), IEEE, 2019, pp. 1–6.
[11] S. Kwon, H. Kim, K.S. Park, Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone, in: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2012, pp. 2174–2177.
[12] R. Nandakumar, S. Gollakota, N. Watson, Contactless sleep apnea detection on smartphones, in: Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services, ACM, 2015, pp. 45–57.
[13] R. Girčys, E. Kazanavičius, R. Maskeliūnas, R. Damaševičius, M. Woźniak, Wearable system for real-time monitoring of hemodynamic parameters: implementation and evaluation, Biomed. Signal Process. Control 59 (2020) 101873.
[14] F. Miao, N. Fu, Y.-T. Zhang, X.-R. Ding, X. Hong, Q. He, Y. Li, A novel continuous blood pressure estimation approach based on data mining techniques, IEEE J. Biomed. Health Informatics 21 (6) (2017) 1730–1740.
[15] J. Allen, Photoplethysmography and its application in clinical physiological measurement, Physiol. Meas. 28 (3) (2007) R1.
[16] J. Kraitl, U. Timm, H. Ewald, E. Lewis, Non-invasive measurement of blood components, in: 2011 Fifth International Conference on Sensing Technology, IEEE, 2011, pp. 253–257.
[17] Y. Wu, A. Boonloed, N. Sleszynski, M. Koesdjojo, C. Armstrong, S. Bracha, V. T. Remcho, Clinical chemistry measurements with commercially available test slides on a smartphone platform: colorimetric determination of glucose and urea, Clin. Chim. Acta 448 (2015) 133–138.
[18] J.P. Devadhasan, H. Oh, C.S. Choi, S. Kim, Whole blood glucose analysis based on smartphone camera module, J. Biomed. Opt. 20 (11) (2015) 117001.
[19] R. Zaman, C. Cho, K. Hartmann-Vaccarezza, T. Phan, G. Yoon, J. Chong, Novel fingertip image-based heart rate detection methods for a smartphone, Sensors 17 (2) (2017) 358.
[20] M. Anggraeni, A. Fatoni, Non-invasive self-care anemia detection during pregnancy using a smartphone camera, in: IOP Conference Series: Materials Science and Engineering, vol. 172, IOP Publishing, 2017, 012030.
[21] E.J. Wang, W. Li, J. Zhu, R. Rana, S.N. Patel, Noninvasive hemoglobin measurement using unmodified smartphone camera and white flash, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2017, pp. 2333–2336.
[22] M.K. Hasan, M.M. Haque, R. Adib, J.F. Tumpa, A. Begum, R.R. Love, Y.L. Kim, I. A. Sheikh, Smarthelp: Smartphone-based hemoglobin level prediction using an artificial neural network, in: AMIA Annual Symposium Proceedings, vol. 2018, American Medical Informatics Association, 2018, 535.
[23] J. Yuan, H. Ding, H. Gao, Q. Lu, Research on improving the accuracy of near infrared non-invasive hemoglobin detection, Infrared Phys. Technol. 72 (2015) 117–121.
[24] H. Ding, Q. Lu, H. Gao, Z. Peng, Non-invasive prediction of hemoglobin levels by principal component and back propagation artificial neural network, Biomed. Opt. Express 5 (4) (2014) 1145–1152.
[25] R.G. Mannino, D.R. Myers, E.A. Tyburski, C. Caruso, J. Boudreaux, T. Leong, G. Clifford, W.A. Lam, Smartphone app for non-invasive detection of anemia using only patient-sourced photos, Nat. Commun. 9 (2018).
[26] R.S. Al-Baradie, A.S.C. Bose, Portable smart non-invasive hemoglobin measurement system, in: 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13), IEEE, 2013, pp. 1–4.
[27] G. Dimauro, D. Caivano, F. Girardi, A new method and a non-invasive device to estimate anemia based on digital images of the conjunctiva, IEEE Access 6 (2018) 46968–46975.
[28] A. Soni, S.K. Jha, Smartphone based non-invasive salivary glucose biosensor, Anal. Chim. Acta 996 (2017) 54–63.
[29] S. Ramasahayam, K.S. Haindavi, S.R. Chowdhury, Noninvasive estimation of blood glucose concentration using near infrared optodes, in: Sensing Technology: Current Status and Future Trends IV, Springer, 2015, pp. 67–82.
[30] S.F. Malin, T.L. Ruchti, T.B. Blank, S.N. Thennadil, S.L. Monfre, Noninvasive prediction of glucose by near-infrared diffuse reflectance spectroscopy, Clin. Chem. 45 (9) (1999) 1651–1658.
[31] P.P. Pai, P.K. Sanki, S.K. Sahoo, A. De, S. Bhattacharya, S. Banerjee, Cloud computing-based non-invasive glucose monitoring for diabetic care, IEEE Trans. Circuits Syst. I: Regular Papers 65 (2) (2017) 663–676.
[32] G.F. Smits, M. Kotanchek, Pareto-front exploitation in symbolic regression, in: Genetic Programming Theory and Practice II, Springer, 2005, pp. 283–299.
[33] J. Koza, Genetic programming as a means for programming computers by natural selection, Stat. Comput. 4 (2) (1994), https://doi.org/10.1007/bf00175355.
[34] A. Chatterjee, A. Prinz, Image analysis on fingertip video to obtain ppg, Biomed. Pharmacol. J. 11 (4) (2018) 1811–1827.
[35] D. McDuff, S. Gontarek, R.W. Picard, Remote detection of photoplethysmographic systolic and diastolic peaks using a digital camera, IEEE Trans. Biomed. Eng. 61 (12) (2014) 2948–2954.
[36] K. Takazawa, N. Tanaka, M. Fujita, O. Matsuoka, T. Saiki, M. Aikawa, S. Tamura, C. Ibukiyama, Assessment of vasoactive agents and vascular aging by the second derivative of photoplethysmogram waveform, Hypertension 32 (2) (1998) 365–370.
[37] I. Imanaga, H. Hara, S. Koyanagi, K. Tanaka, Correlation between wave components of the second derivative of plethysmogram and arterial distensibility, Jpn. Heart J. 39 (6) (1998) 775–784.
[38] H.J. Baek, J.S. Kim, Y.S. Kim, H.B. Lee, K.S. Park, Second derivative of photoplethysmography for estimating vascular aging, in: 2007 6th International Special Topic Conference on Information Technology Applications in Biomedicine, IEEE, 2007, pp. 70–72.
[39] U. Rubins, A. Grabovskis, J. Grube, I. Kukulis, Photoplethysmography analysis of artery properties in patients with cardiovascular diseases, in: 14th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics, Springer, 2008, pp. 319–322.
[40] S.A. Esper, M.R. Pinsky, Arterial waveform analysis, Best Pract. Res. Clin. Anaesthesiol. 28 (4) (2014) 363–380.
[41] E. Seitsonen, I. Korhonen, M. Van Gils, M. Huiku, J. Lötjönen, K. Korttila, A. Yli-Hankala, Eeg spectral entropy, heart rate, photoplethysmography and motor responses to skin incision during sevoflurane anaesthesia, Acta Anaesthesiol. Scand. 49 (3) (2005) 284–292.
[42] L. Wang, E. Pickwell-MacPherson, Y. Liang, Y. Zhang, Noninvasive cardiac output estimation using a novel photoplethysmogram index, in: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2009, pp. 1746–1749.
[43] X. Zhang, Y. Lyu, X. Hu, Z. Hu, Y. Shi, H. Yin, Evaluating photoplethysmogram as a real-time cognitive load assessment during game playing, Int. J. Human-Comput. Interact. 34 (8) (2018) 695–706.
[44] T.-H. Fu, S.-H. Liu, K.-T. Tang, Heart rate extraction from photoplethysmogram waveform using wavelet multi-resolution analysis, J. Med. Biol. Eng. 28 (4) (2008) 229–232.
[45] L. Chen, A.T. Reisner, J. Reifman, Automated beat onset and peak detection algorithm for field-collected photoplethysmograms, in: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2009, pp. 5689–5692.
[46] C.G. Scully, J. Lee, J. Meyer, A.M. Gorbach, D. Granquist-Fraser, Y. Mendelson, K. H. Chon, Physiological parameter monitoring from optical recordings with a mobile phone, IEEE Trans. Biomed. Eng. 59 (2) (2011) 303–306.
[47] E. Jonathan, M. Leahy, Investigating a smartphone imaging unit for photoplethysmography, Physiol. Meas. 31 (11) (2010) N79.
[48] A.R. KAVSAOĞLU, K. Polat, M.R. Bozkurt, An innovative peak detection algorithm for photoplethysmography signals: an adaptive segmentation method, Turk. J. Electr. Eng. Comput. Sci. 24 (3) (2016) 1782–1796.
[49] H.S. Shin, C. Lee, M. Lee, Adaptive threshold method for the peak detection of photoplethysmographic waveform, Comput. Biol. Med. 39 (12) (2009) 1145–1152.
[50] A.R. Kavsaoğlu, K. Polat, M.R. Bozkurt, A novel feature ranking algorithm for biometric recognition with ppg signals, Comput. Biol. Med. 49 (2014) 1–14.
[51] R. Tiwari, M.P. Singh, Correlation-based attribute selection using genetic algorithm, Int. J. Comput. Appl. 4 (8) (2010) 28–34.
[52] D.P. Searson, D.E. Leahy, M.J. Willis, Gptips: an open source genetic programming toolbox for multigene symbolic regression, in: Proceedings of the International Multiconference of Engineers and Computer Scientists, vol. 1, IMECS Hong Kong, 2010, pp. 77–80.
[53] J.D. Rodriguez, A. Perez, J.A. Lozano, Sensitivity analysis of k-fold cross validation in prediction error estimation, IEEE Trans. Pattern Anal. Mach. Intell. 32 (3) (2009) 569–575.
[54] J.M. Bland, D. Altman, Statistical methods for assessing agreement between two methods of clinical measurement, Lancet 327 (8476) (1986) 307–310.