

# Non-invasive prediction of hemoglobin level using machine learning techniques with the PPG signal's characteristics features



A. Reşit Kavsaoglu<sup>a,\*</sup>, Kemal Polat<sup>b</sup>, M. Hariharan<sup>c</sup>

<sup>a</sup> Department of Biomedical Engineering, Faculty of Engineering, Karabuk University, 78050 Karabuk, Turkey

<sup>b</sup> Department of Electrical and Electronics Engineering, Faculty of Engineering and Architecture, Abant İzzet Baysal University, 14280 Bolu, Turkey

<sup>c</sup> School of Mechatronic Engineering, Universiti Malaysia Perlis (UniMAP), Campus Pauk Putra, 02600 Perlis, Malaysia

## ARTICLE INFO

### Article history:

Received 11 February 2015

Received in revised form 16 March 2015

Accepted 3 April 2015

Available online 14 April 2015

### Keywords:

Photoplethysmography (PPG)

Hemoglobin (Hb)

Feature extraction

Regression models

Feature selection

Prediction

## ABSTRACT

Hemoglobin can be measured normally after the analysis of the blood sample taken from the body and this measurement is named as invasive. Hemoglobin must continuously be measured to control the disease and its progression in people who go through hemodialysis and have diseases such as oligocythemia and anemia. This gives a perpetual feeling of pain to the people. This paper proposes a non-invasive method for the prediction of the hemoglobin using the characteristic features of the PPG signals and different machine learning algorithms. In this work, PPG signals from 33 people were included in 10 periods and 40 characteristic features were extracted from them. In addition to these features, gender information (male or female), height (as cm), weight (as kg) and age of each subjects were also considered as the features. Blood count and hemoglobin level were measured simultaneously by using the “Hemocue Hb-201TM” device. Using the different machine learning regression techniques (classification and regression trees – CART, least squares regression – LSR, generalized linear regression – GLR, multivariate linear regression – MVLRL, partial least squares regression – PLSR, generalized regression neural network – GRNN, MLP – multilayer perceptron, and support vector regression – SVR). RELIEFF feature selection (RFS) and correlation-based feature selection (CFS) were used to select the best features. Original features and selected features using RFS (10 features) and CFS (11 features) were used to predict the hemoglobin level using the different machine learning techniques. To evaluate the performance of the machine learning techniques, different performance measures such as mean absolute error – MAE, mean square error – MSE,  $R^2$  (coefficient of determination), root mean square error – RMSE, Mean Absolute Percentage Error (MAPE) and Index of Agreement – IA were used. The promising results were obtained (MSE-0.0027) using the selected features by RFS and SVR. Hence, the proposed method may clinically be used to predict the hemoglobin level of human being clinically without taking and analyzing blood samples.

© 2015 Elsevier B.V. All rights reserved.

## 1. Introduction

Hemoglobin (Hb) is the basic component of the red blood cells. The Hb concentration in human blood is an important parameter both for each blood count and the evaluation of the physiologic state of an individual. Many invasive methods, (in which blood is taken from the patient physically and then analyzed) are used to measure the Hb concentration in today's technology. Besides the ailment experienced while taking the blood samples, the disadvantage of such invasive methods is the delay between the bloodletting and blood analysis [1].

Knowing the hemoglobin content in blood helps doctors to diagnose many diseases and to know about their healthy status as well. For instance, the anemia, which is responsible for a major part of death, is easily diagnosed with the decrease in the hemoglobin content in blood. Similarly, the daily measurement of the hemoglobin content in blood is significant for the treatment of patients with dengue fever [1,2].

In this study, photoplethysmography (PPG) signal was used (non-invasively) to measure the hemoglobin concentration level in blood. Below, information is given about the PPG signal.

PPG is obtained through the light that illuminates the relevant area of the body, is reflected or transmitted. In order to obtain the PPG, a light source with a wavelength of  $\lambda$  is placed on one side of an extension, for example a finger, and the photodetector, which will catch the light transmitted right across the source, is positioned on the other side. A typical PPG signal is comprised of a big DC

\* Corresponding author. Tel.: +90 370 433 20 21; fax: +90 370 433 32 90.

E-mail addresses: [kavsaoglu@hotmail.com](mailto:kavsaoglu@hotmail.com) (A.R. Kavsaoglu), [kemal.polat2003@yahoo.com](mailto:kemal.polat2003@yahoo.com) (K. Polat), [hari@unimap.edu.my](mailto:hari@unimap.edu.my) (M. Hariharan).

component which passes through skin, muscle and bone without moving through the blood vessels, a small AC component which directly passes through blood vessels by leaving the skin, muscle and bone and the component in the heart beat frequency, which results from the light passing through the arterial blood vessels as well. The blood amount in arteries increases right after the systole, and therefore the light intensity received decreases. The blood amount in arteries is reduced during the diastole and an increase is observed in the light transmission. In a basic measurement area, 99% of the signal comes from the skin, muscle and bone, 0.9% from the vein and 0.1% from the arteries [3].

PPG is a photoelectrical method, which is used for measuring the tissue blood volume based on the change in the blood volume at every heartbeat. PPG has been prevalently used for the detection of many basic physiologic parameters such as the oxygen amount in blood [4], heart rate, respiration [5] and blood pressure [6] so far [7]. PPG can also reflect some other significant cardiovascular (related to heart and blood vessels) parameters like embolism [8]. Therefore, it is meaningful and important to deeply analyze the PPG signal for the purpose of discovering various physiologic parameters with high accuracy [9].

There are numerous studies available in the literature with regard to the non-invasive measurement of the hemoglobin level in blood. Among these, Masakazu Iwasaka et al. used optical absorption to measure the hemoglobin and red blood cells under the magnetic field in 2001 [10]. In 2007, Gregory D. Jay et al. used the fiber optic reflectance spectroscopy to measure the value of the hemoglobin level [11]. In 2009, Linda M. Head et al. developed the wireless fNIR spectroscopy for the detection of Hb (hemoglobin) and HbO<sub>2</sub> (oxy-hemoglobin concentration) [12]. In 2009, U. Timm et al. developed an optical sensor system to determine the hemoglobin value [13]. In 2011, they used the near infrared spectroscopy to detect the hemoglobin concentration of the newborns while breathing [14]. In 2013, Luangrat et al. developed an impact detector system, which can be used at home, in order to measure the hemoglobin level in blood non-invasively and used the single infrared LED for calculating the hemoglobin value [15]. In 2014, Ding et al. formed a broadband light source consisting of nine LEDs with the grating spectrography and Si (silicium) photodiode array and developed a high-performance spectrophotometric system to measure the hemoglobin value [16]. In 2014, Nirupa J. et al. suggested a system relying on the estimation with empirical formulas from the PPG signal based on two different wavelengths to measure the oxygen saturation in blood, change in the heart rate and hemoglobin value [1].

In this study, a method based on machine learning associated with characteristic features of the PPG signal has been developed to measure the hemoglobin value non-invasively. Hence, PPG signals of 33 people were included in 10 periods and 40 characteristic features were extracted from the PPG signal of each person. In addition to these features, gender information (male or female), height (as cm), weight (as kg) and age information about them were also added and the number of features was increased up to 44. Blood count and hemoglobin measurement were performed simultaneously by using the “Hemocue Hb-201TM” device during the reception of the PPG signals. Afterwards, these 44 features were entered into the machine learning algorithms and hemoglobin values of the people were estimated. Additively, the methods of RELIEFF feature selection (RFS) and correlation based feature selection (CFS) have been used to decrease the number of features. For evaluating the performance of the machine learning methods, MAE (mean absolute error), MSE (mean square error),  $R^2$ , RMSE (root mean square error), MAPE (mean absolute percentage error) and Index of Agreement (IA) were used.

The outline of this paper is organized as follows. Section 2 depicts the material and method. The results and discussion have been explained in the Section 3. Section 4 concludes the paper.

## 2. Materials and methods

### 2.1. Hemoglobin measurement setup and dataset based on PPG signal

In this study, the PPG signal was received from 33 subjects through the data obtainment card without taking any blood sample. Meanwhile, blood count and hemoglobin measurement of these subjects were also carried out simultaneously while receiving the PPG signals through the “Hemocue Hb-201TM” device. 10-PPG signals were received from 33 subjects and Table 1 gives the statistical information about those subjects. Measurement mechanism is given in Fig. 1.

The PPG signal is produced with the periodical heartbeat. Therefore, heart rate has a close relation to the beating period because it can be extracted. Another aspect is that this blood flow in vessels influences the vessel flexibility and blood viscosity (resistance against fluidity). Moreover, plenty of cardiovascular (related to heart and blood vessels) information can be selected from the PPG signal. All the physiologic parameters of this type are reflected within the feature points of the PPG signal as shown in Fig. 2 [9,37].

In a typical PPG signal, the section from S to M point is the fast beating phase and the part from M to P point is the delaying phase. M is the main peak point of the PPG signal. At this point, the blood pressure (BP) has the highest value within the whole period. Another peak is the dicrotic wave of the artiole, which is called Q and reflects its adaptation. The P point is named as dicrotic notch. Heart rate is obtained from the duration between two main peaks (tpp). The M–Q time slot is an index of the arteriosclerosis (vessel stiffness). However, it is difficult to directly obtain the feature points for the real PPG signal because of the presence of various factors. For this purpose, it is important to find useful pre-processing and feature extraction methods to determine the PPG signal accurately [9,37].

The PPG signals were obtained from their right index fingers while the people were seated in a calm position. Prior to the acquisition of data, information regarding the study was given to the volunteers, the questions they were curious about were answered, the explanations as to the privacy of the data and their personal information and keeping them within the limits of scientific studies were made, and they were given a form to fill out the information about their consent and approval of volunteering to participate in this study.

### 2.2. Acquisition of PPG signals and feature extraction process

In this work, the PPG signals have been obtained by means of an STM32F103xC ARM-based PPG data-acquisition card with a 2 kHz sampling frequency, a 32-bit microcontroller, and a DCM03 reflection sensor on it. The circuit schematic diagram of used PPG data acquisition system has been shown in Fig. 3.

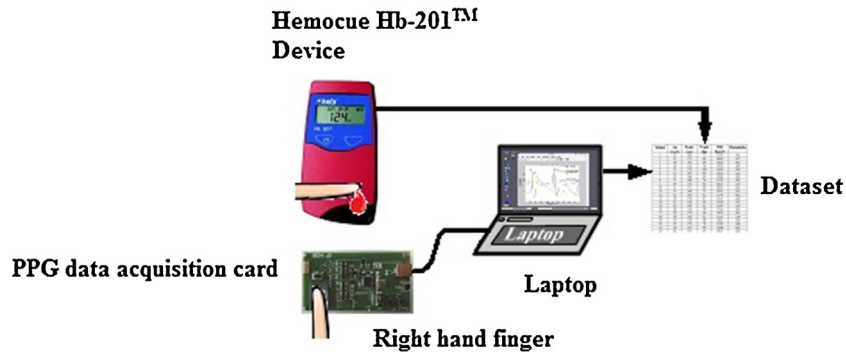
Various pre-processing steps were used to eliminate the noise in the PPG signals after their reception. In order to remove the noise in PPG signals as a pre-processing, a low-pass FIR filter with an  $N = 200$  points and 10 Hz cut-off frequency has been used.

During the feature extraction from the PPG signal, 40 time-domain features were extracted by using the status of the PPG signal after the pre-processing, the first derivative of its status after the pre-processing and the second derivative of its status after the pre-processing. Distribution of these features was obtained as follows: 21 time-domain features from the PPG signal, 8 time-domain

**Table 1**

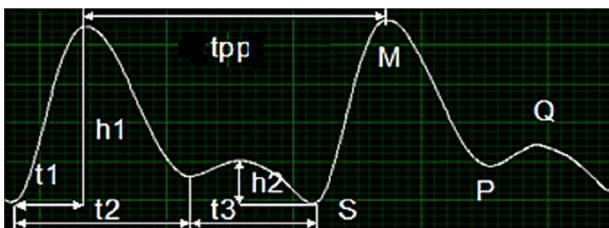
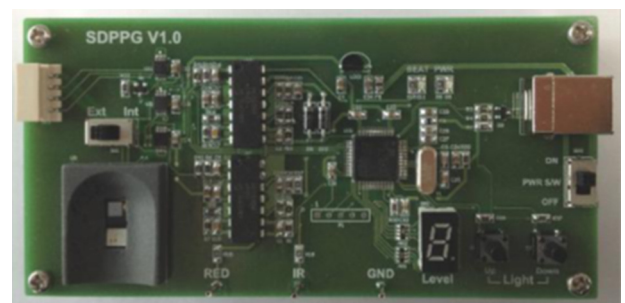
The statistical data belonging to subjects in this study.

Subject	Age (years)	Height (cm)	Weight (kg)	BMI (kg/m <sup>2</sup> )	Hemoglobin level measured by hemocue Hb-201™ device
1	35	183	72	21.50	15.1
2	22	172	90	30.42	15.7
3	18	162	69	26.29	10.1
4	18	176	79	25.50	14.5
5	21	170	100	34.60	14.0
6	21	167	65	23.31	11.3
7	20	186	69	19.94	16.5
8	21	186	74	21.39	16.5
9	34	175	93	30.37	17.4
10	21	171	60	20.52	14.2
11	31	169	72	25.21	15.2
12	22	153	51	21.79	10.2
13	19	178	67	21.15	14.4
14	31	158	55	22.03	12.3
15	23	150	51	22.67	11.4
16	28	170	79	27.34	14.5
17	19	170	58	20.07	16.1
18	18	155	51	21.23	11.8
19	19	178	58	18.31	13.0
20	20	179	70	21.85	14.8
21	28	157	49	19.88	10.5
22	22	172	57	19.27	13.4
23	20	165	61	22.41	10.5
24	19	160	53	20.70	10.3
25	18	150	57	25.33	12.0
26	20	175	71	23.18	13.9
27	17	158	57	22.83	13.3
28	20	158	54	21.63	11.4
29	21	173	70	23.39	16.0
30	20	175	65	21.22	15.2
31	41	178	98	30.93	16.2
32	19	171	65	22.23	11.4
33	18	160	60	23.44	13.2

**Fig. 1.** Hemoglobin level measurement device based on PPG signal.

features from the first derivative of the PPG signal, 7 time-domain features from the second derivative of the PPG signal and the remaining 4 features from the combination of the first and second derivatives of the PPG signal. Fig. 4 shows the feature extraction block diagram from PPG signal, and its derivatives.

In this stage, the 40 characteristic properties are obtained from the first and second derivations of PPG signals. Fig. 5 shows the relationship of the PPG signals with the first and second derivations. By using these relationships, the points of properties are defined.

**Fig. 2.** Standard PPG signal and its characteristic points [9].**Fig. 3.** PPG data acquisition card [17,35].

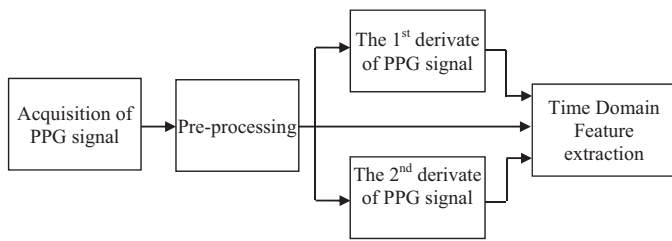


Fig. 4. Feature extraction block diagram from PPG signal and its derivatives [36].

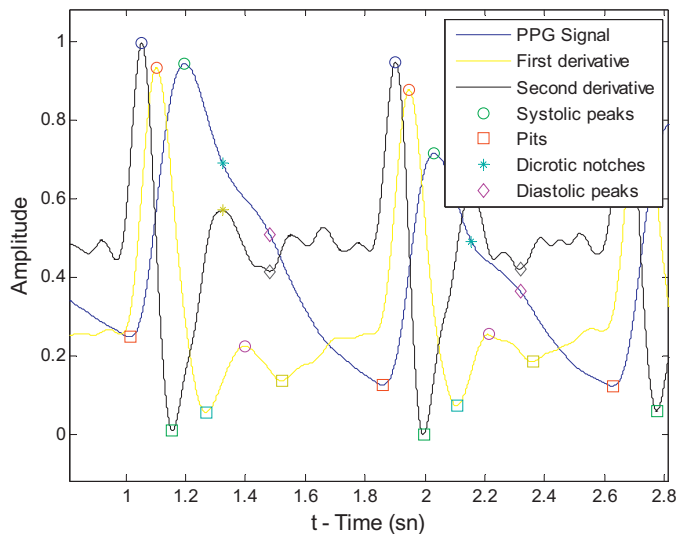


Fig. 5. The PPG signal, the first and second derivatives of the PPG signal, and the characteristic feature points [36].

For more information about these forty characteristic points from both PPG signal and its derivatives, the readers can refer to [36].

### 2.3. The stages of used methods

In this paper study, a new method has been suggested based on the PPG signal, which is a non-invasive method without taking blood sample from the human body, and machine learning methods for the non-invasive and distant prediction of the hemoglobin value. First of all, blood count and hemoglobin measurement were simultaneously performed by using the “Hemocue Hb-201TM” device. Meanwhile, 10-period PPG signals were received from the related person right after the measurement through the device. Then, 40 characteristic features were extracted from the PPG signal and the number of the features was increased up to 44 with the addition of the physical attributes that belonged to the person. Two methods were used to decrease the number of the features: RELIEF feature selection and correlation based feature selection. The total data amount was increased up to 330 with the 10-period signals of 33 people by using a dataset with 44 features. The numbers of the features were respectively reduced down to 10 and 11 after the application of the RFS and CFS methods onto the dataset with 44 features. After the feature selection process, eight different machine learning methods were implemented to estimate the hemoglobin values.

#### 2.3.1. Feature selection process

The feature selection algorithms are often used for the purpose of selecting the relevant features associated with the output label and minimizing the computational time during classification process of classification. There are mainly two objectives in feature selection: the first is to minimize the classification cost by reducing

the number of features and thus, forming a more compact classification model; and the second is to improve the classification performance [38]. Here, we have used two different feature selection methods including RELIEF feature selection and correlation based feature selection.

**2.3.1.1. RELIEF based feature selection (RFS).** RELIEF is a feature selection algorithm suggested by Kira and Rendell in 1992 and used in the binary data classification [20]. Being independent on the heuristic structure it has many strong aspects and it needs only linear time for the number of the given features and training samples. It has a structure, which can be applied to continuous or binary data. However, it cannot make a good distinction in case of the existence of unnecessary features and less training samples. Kononenko et al. made improvements in the system by applying the RELIEFF algorithm on attributes such as generalization for multiclass datasets and strength in incomplete data [18,19]. After RELIEFF based feature selection being applied to PPG dataset with hemoglobin prediction, the number of features has been decreased from 44 to 10. The selected features are as follows:

The selected features by RELIEF based feature selection:

The selected features by means of RFS method (from 44 features to 10 features)									
41	43	42	44	10	1	26	37	27	16

**2.3.1.2. Correlation-based feature selection (CFS).** Correlation is a criterion used to measure whether a feature is relevant to other features in dataset and relevant to outputs or not. A feature is good if it is relevant to the class but is not redundant to any of the other relevant features. Applied with correlation, the good side of a feature is measured according to whether it is highly correlated with the class or not highly correlated with any of the other features [21,22]. Here, we have used the correlation coefficient probability ( $p$ -value) to select the significant features from all dataset. In this work, the  $p$  values has been chosen as  $p < 0.01$ . After CFS method applied to PPG dataset with hemoglobin prediction, the number of features has been decreased from 44 to 11. The selected features are as follows:

The selected features by means of CFS method:

The selected features by means of CFS method (from 44 features to 11 features)										
14	18	23	26	27	34	38	41	42	43	44

#### 2.3.2. Machine learning methods used

In this study, 40 characteristic features were extracted from the PPG signal of each person; in addition to these features, gender information (male or female), height (as cm), weight (as kg) and age information about them were also added and the number of features was increased up to 44. Afterwards, these 44 features were entered into the machine learning algorithms and hemoglobin values of the people were estimated. The used machine learning algorithms are as follows: CART (classification and regression trees), LSR (Least Square Regression), GLR (generalized linear regression), MVLRL (multivariate linear regression), PLSR (partial least squares regression), GRNN (generalized regression neural network), MLP (multilayer perceptron), and SVR (support vector regression). The machine learning methods used are explained under titles below. The parameters used in the machine learning methods are shown in Table 2.

In this work, conventional validation was performed. 70% of the data were used for training and the remaining 30% of the data were used for testing. Same training and testing data were used for all the classifiers. Training and testing were repeated for 25 times and the results were presented in terms of mean and standard deviation.

**2.3.2.1. Classification and regression trees (CART).** Classification and regression trees are machine learning methods for



**Table 2**  
The parameters and their values used in the machine learning methods.

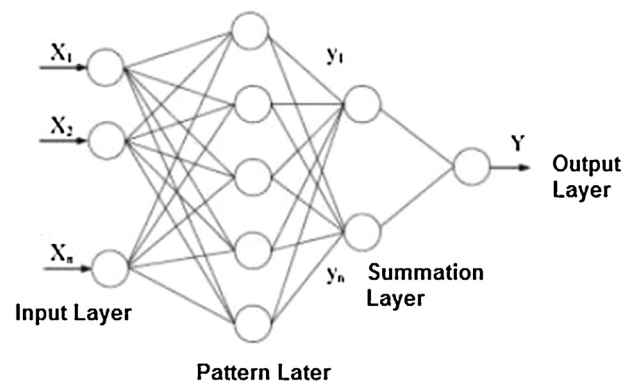
The used parameter in machine learning method	The used parameters and their values
CART	Default parameters were used
LSR	Default parameters were used. “Linear” Model type was used
GLR	Default parameters were used. “Linear” Model type was used
MVLR	Default parameters were used
PLSR	22 PLS components were selected before feature selection. Only five PLS components were selected after feature selection
GRNN	Spread factor = 0.4.
MLP	Number of input neurons = 44 (all features), 10 (using RELIFF), 11 (CFS) number of hidden neurons = 22 number of output neuron = 1. Activation function = ‘tansig’ training algorithm = ‘trainlm’. Momentum factor = 0.9, number of epoch = 2000, goal = 0.001
SVR	Gaussian’ kernel was used. $c = 500$ , $\epsilon = 0.005$

constituting prediction models from the data. Models are obtained with successive data space partitioning and the concordance of a simple prediction method within each division. As a result, splitting can be shown in a schema as a decision tree. Classification trees are designed for dependent variables, which take the foreseen error measured in terms of wrong classification costs (expenses) and the irregular values at a finite number. Regression trees are generally for variables, which take permanent or sequential disjoint values and depend on the estimation error measured with the quadratic difference between the observed and predicted values [23].

**2.3.2.2. Least squares regression (LSR).** The least squares method is a standard approach for the approximate solution of the over-determined systems; that is, they are the equation sets in which there are more equalities than unknowns. “The least squares” method is the minimization of the sum of the errors made in equality result within the general solution. The least squares method divides the problems into two categories: linear or ordinary least squares and nonlinear least squares, depending on whether each unknown value of the redundancies is linear. The linear least squares problem occurs in statistical regression analysis; there is a closed-form solution available. A closed-form solution (or a closed expression) is a formula, which can assess the standard operation at a limited number. A nonlinear problem doesn’t have a closed-form solution and it is generally solved through the iterated treatment, which is approximated to the system by a linear problem at each repetition and in which the kernel calculation is made similar in both situations [24].

**2.3.2.3. Generalized linear regression (GLR).** In statistics, the generalized linear regression is a flexible generalization of the ordinary linear regression, which provides response variables – error distribution models out of a normal distribution. The generalized linear regression generalizes the linear regression by allowing the associated linear model between a connection function and the response variable and providing the variance (deviation) size of each measurement as a function of the expected value [25].

**2.3.2.4. Multivariate linear regression (MVLR).** Multivariate regression models aim at defining the behavior of many response variables in terms of other variables. Multivariate regression models expand the main thought (idea) of only one response variable or linear regression models with many response variables. These models have broad applicability and can be used in training area as well. In this paper, easily digestible mathematical formulation



**Fig. 6.** The general structure of GRNN model [28].

of the multivariate linear regression model is given. The assembly model is described with the least squares method and the implementation of the monte-regression model is shown with a simple example [26].

**2.3.2.5. Partial least squares regression (PLSR).** Partial least squares regression (PLS regression) is a statistical method, basic components of which carry some relationships of the regression; it discovers a linear regression model by reflecting the estimation variables and observable variables on a new space instead of finding the minimum variance hyper-planes between the response and independent variables, because both  $X$  and  $Y$  data design new spaces. The PLS method family is known as bilinear factor models. The partial least squares method is a variety of the Discriminant Analysis (PLS-DA) that is used when  $Y$  is categorical. PLS is a latent variable approach for modeling covariance structures in these two spaces; that is, it is used to find the basic relationships between two matrices ( $X$  and  $Y$ ). A PLS model tries to find the multidimensional direction in  $X$  space explaining the maximum dimensional variance in  $Y$  space. The PLS regression model is especially applicable when there are more variables in feature predictions than the observations, and multiple connections exist between  $X$  values. On the contrary, standard regression is unsuccessful in these situations [27].

**2.3.2.6. Generalized regression neural network (GRNN).** GRNN is a variety of the radial basis neural networks based on kernel regression networks [28]. GRNN doesn’t require a repeated training procedure like back-propagation networks. These training data directly approximate the function estimation to an arbitrary function between plotting, input and output vectors. Additionally, the estimation error approaches zero as the size of the training set expands.

GRNN consists of four main layers; these are input, model, summation and output layers as shown in Fig. 6. The number of input units in the input layer depends on the total number of the observation parameters. The first layer is connected to the pattern layer and each neuron in this layer points at a training pattern and output. Model layer is connected to the summation layer. The summation layer is comprised of two different parts as a single division unit and summation units. Summation and output layers have an influence normalizing the output set. And the network training uses the radial basis and linear activation functions as hidden and output layers [28,39,40].

**2.3.2.7. Multilayer perceptrons (MLP).** Multilayer perceptron (MLP) is a feed-forward artificial neural network that maps sets of input data onto a set of appropriate outputs. A MLP consists of many layer structures that combine each layer, which is fully connected to the

next one, and directed nodes in a graph. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back-propagation for training the network. [30] [31]. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable [29].

**2.3.2.8. Support vector regression (SVR).** The Support Vector Machine algorithm (maximum-margin) can be used as a regression method by protecting all the basic features characterized. The support vector regression (SVR) method utilizes the same principles with the classified SVM structure with only a few small differences. First of all, it becomes quite difficult to predict the present information with endless possibilities, because the output is a real number. In case of this regression, tolerance (epsilon) is the approximated setting of the SVM model that is already claimed from the problem. Besides this reality, there is however another complicated reason, and therefore, the situation to consider is more complexes. But the main idea is always the same: customization of the maximized margin sub plane for the minimization of the error by keeping in mind that the part of the error is tolerated [32].

### 3. Results and discussions

#### 3.1. Performance evaluating criteria

In this section, the performance evaluating criteria to test the performance of used machine learning methods are the mean absolute error (MAE), mean square error (MSE),  $R^2$  (coefficient of determination), root mean square error (RMSE), mean absolute percentage error (MAPE), and index of agreement (IA) values. And also, to train and test the machine learning algorithms, the 70 training – 30% testing division method has been used.

If  $y_1, y_2, \dots, y_n$  are  $n$  observed values and  $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$  are the corresponding values predicted, then the mean absolute error (MAE), mean square error (MSE),  $R^2$ , root mean square error

(RMSE), mean absolute percentage error (MAPE), and index of agreement (IA) values are as follows [33,34]:

$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j| \quad (1)$$

$$MSE = \frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2 \quad (2)$$

$$RMSE = \sqrt{MSE} \quad (3)$$

$$IA = 1 - \frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2}{\sum_{j=1}^n \left[ \left| y_j' \right| + \left| \hat{y}_j' \right| \right]^2} \quad (4)$$

$$MAPE = \frac{1}{n} \sum_{j=1}^n \left| \frac{y_j - \hat{y}_j}{y_j} \right| \quad (5)$$

$$y_j' = y_j - y_m \quad \text{and} \quad \hat{y}_j' = \hat{y}_j - y_m \quad (6)$$

$y_m$  is the average of observed points.

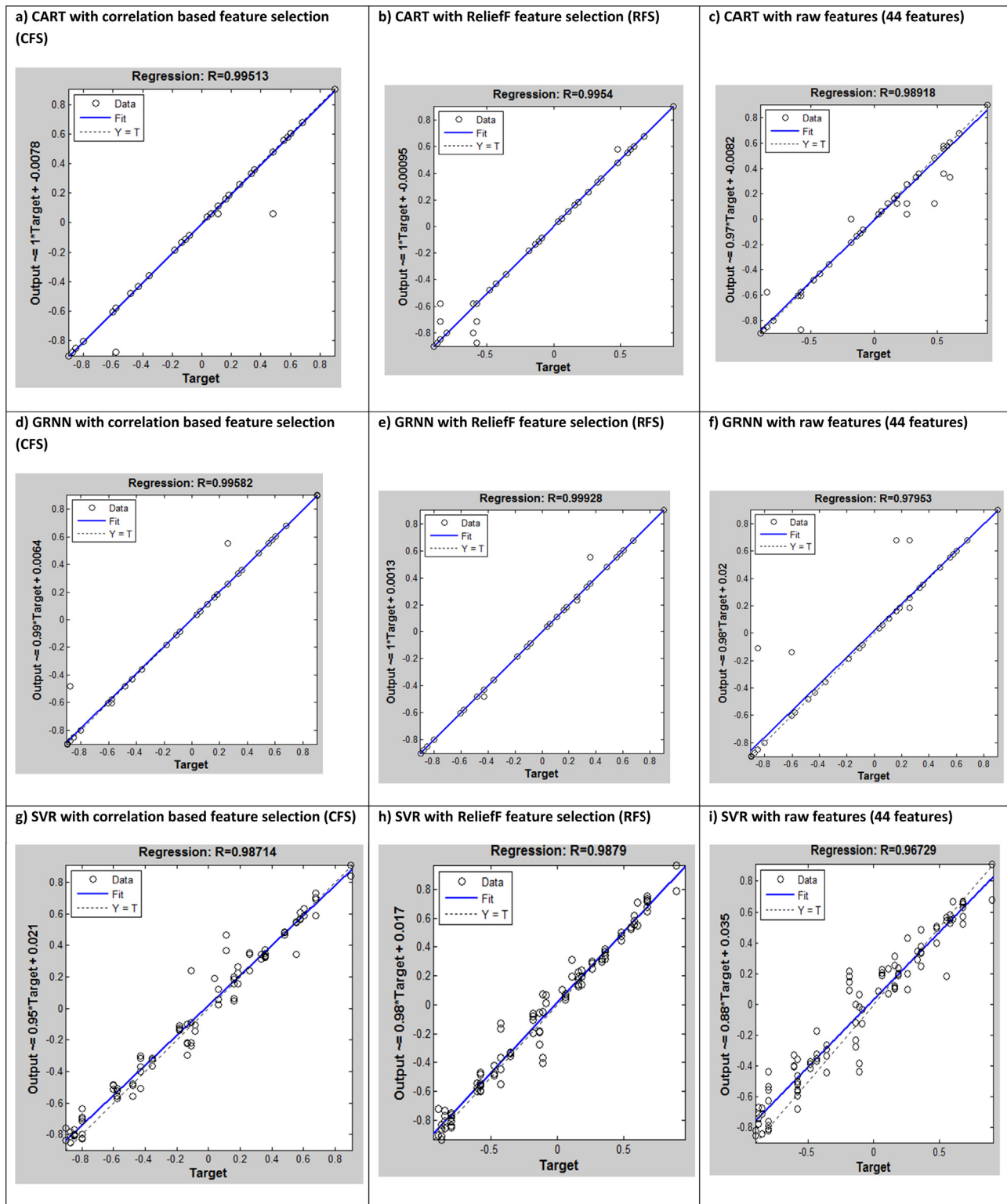
#### 3.2. Results obtained

In this work, non-invasive hemoglobin concentration level prediction system has been proposed based on combining regression methods and feature selection with time domain features extracted from PPG signal. PPG signals of 33 people have been included in 10 periods and 40 characteristic features have been extracted from the PPG signal of each person; in addition to these features, gender information (male or female), height (as cm), weight (as kg) and age information about them have been also added and the number of features has been increased up to 44. Blood count and hemoglobin measurement have been performed simultaneously by using the “Hemocue Hb-201TM” device during the reception of the PPG signals. During the feature extraction from the PPG signal, 40 time-domain features were extracted by using the status of the PPG signal after the pre-processing, the first

**Table 3**

The hemoglobin prediction results using eight different machine learning techniques with 44 time domain features obtained PPG signal.

Machine learning models	With 44 features (raw features)					
	Performance measure criteria					
	MAE	MSE	$R^2$	RMSE	MAPE	IA
CART						
Mean	0.0564	0.0215	0.9211	0.1410	0.2700	0.9796
Std. Dev.	0.0224	0.0122	0.0461	0.0408	0.1526	0.0123
LSR						
Mean	0.1987	0.0634	0.7671	0.2508	0.8900	0.9352
Std. Dev.	0.0145	0.0111	0.0488	0.0216	0.1331	0.0132
GLR						
Mean	0.1987	0.0634	0.7671	0.2508	0.8900	0.9352
Std. Dev.	0.0145	0.0111	0.0488	0.0216	0.1331	0.0132
MVLR						
Mean	0.1982	0.0629	0.7686	0.2500	0.8887	0.9356
Std. Dev.	0.0145	0.0113	0.0496	0.0218	0.1335	0.0134
PLSR						
Mean	0.1979	0.0597	0.7803	0.2438	0.8673	0.9374
Std. Dev.	0.0142	0.0084	0.0401	0.0171	0.1355	0.0111
GRNN						
Mean	0.0403	0.0221	0.9181	0.1452	0.1917	0.9797
Std. Dev.	0.0151	0.0097	0.0388	0.0324	0.0872	0.0093
MLP						
Mean	0.1245	0.0334	0.8776	0.1805	0.6542	0.9689
Std. Dev.	0.0174	0.0114	0.0419	0.0291	0.1174	0.0111
SVR						
Mean	0.1026	0.0213	0.9220	0.1444	0.4204	0.9788
Std. Dev.	0.0125	0.0065	0.0233	0.0221	0.0716	0.0067



**Fig. 7.** The relationships between target (the real measured data) and predicted hemoglobin values using CART, GRNN, and SVR regression methods and two different feature selection including RELIEF based feature selection and correlation-based feature selection

derivative of its status after the pre-processing and the second derivative of its status after the pre-processing. Distribution of these features was obtained as follows: 21 time-domain features from the PPG signal, 8 time-domain features from the first derivative of the PPG signal, 7 time-domain features from the second derivative of the PPG signal and the remaining four features from

the combination of the first and second derivatives of the PPG signal.

In this paper, two feature selection methods have been used to decrease the number of features in PPG dataset:

RELIEF feature selection (RFS) and correlation based feature selection (CFS). The total number of data has been obtained as the

**Table 4**  
The hemoglobin prediction results using eight different machine learning techniques with 10 time domain features selected by RELIEF based feature selection from PPG dataset.

Machine learning models	With the 10 selected features via RFS method					
	Performance measure criteria					
	MAE	MSE	$R^2$	RMSE	MAPE	IA
CART						
Mean	0.0356	0.0137	0.9504	0.1101	0.1711	0.9872
Std. Dev.	0.0190	0.0098	0.0349	0.0402	0.1208	0.0091
LSR						
Mean	0.2113	0.0644	0.7634	0.2531	1.0085	0.9298
Std. Dev.	0.0141	0.0087	0.0414	0.0172	0.1518	0.0107
GLR						
Mean	0.2113	0.0644	0.7634	0.2531	1.0085	0.9298
Std. Dev.	0.0141	0.0087	0.0414	0.0172	0.1518	0.0107
MVLR						
Mean	0.2156	0.0662	0.7568	0.2566	1.0230	0.9273
Std. Dev.	0.0140	0.0093	0.0429	0.0183	0.1504	0.0111
PLSR						
Mean	0.2114	0.0641	0.7644	0.2526	1.0069	0.9300
Std. Dev.	0.0137	0.0085	0.0406	0.0168	0.1518	0.0107
GRNN						
Mean	0.0158	0.0070	0.9741	0.0780	0.0791	0.9936
Std. Dev.	0.0098	0.0053	0.0201	0.0310	0.0553	0.0050
MLP						
Mean	0.0691	0.0114	0.9583	0.1045	0.3559	0.9895
Std. Dev.	0.0124	0.0046	0.0167	0.0215	0.0973	0.0041
SVR						
Mean	0.0678	0.0092	0.9664	0.0951	0.3110	0.9914
Std. Dev.	0.0079	0.0027	0.0093	0.0135	0.0528	0.0024

**Table 5**  
The hemoglobin prediction results using eight different machine learning techniques with 11 time domain features selected by CFS method from PPG dataset.

Machine learning models	With the 11 selected features via CFS method					
	Performance measure criteria					
	MAE	MSE	$R^2$	RMSE	MAPE	IA
CART						
Mean	0.0246	0.0079	0.9711	0.0833	0.1049	0.9927
Std. Dev.	0.0135	0.0060	0.0227	0.0311	0.0837	0.0058
LSR						
Mean	0.2030	0.0612	0.7753	0.2469	0.8921	0.9333
Std. Dev.	0.0109	0.0074	0.0351	0.0148	0.1412	0.0100
GLR						
Mean	0.2030	0.0612	0.7753	0.2469	0.8921	0.9333
Std. Dev.	0.0109	0.0074	0.0351	0.0148	0.1412	0.0100
MVLR						
Mean	0.2028	0.0609	0.7763	0.2463	0.8857	0.9336
Std. Dev.	0.0113	0.0077	0.0360	0.0155	0.1390	0.0103
PLSR						
Mean	0.2101	0.0651	0.7609	0.2545	0.9983	0.9282
Std. Dev.	0.0142	0.0088	0.0411	0.0174	0.1623	0.0114
GRNN						
Mean	0.0160	0.0072	0.9738	0.0807	0.0867	0.9935
Std. Dev.	0.0070	0.0042	0.0155	0.0261	0.0526	0.0038
MLP						
Mean	0.0667	0.0139	0.9499	0.1120	0.3411	0.9873
Std. Dev.	0.0132	0.0104	0.0352	0.0370	0.0849	0.0093
SVR						
Mean	0.0711	0.0126	0.9539	0.1099	0.3636	0.9882
Std. Dev.	0.0083	0.0056	0.0200	0.0232	0.0975	0.0052

data of 330 with 10 periods belonging to 33 people using dataset with 44 features. After RBS and CFS methods applied to dataset with 44 features, the number of reduced features was 10 and 11, respectively. Afterwards, to predict the hemoglobin level based on PPG signals, the eight different machine learning methods were used. Table 3 gives the hemoglobin prediction results using eight different machine learning techniques with 44 time domain features obtained from PPG signal. Table 4 denotes the hemoglobin prediction results using eight different machine learning techniques with 10 time domain features selected by RELIEF based feature selection

from PPG dataset. Table 5 gives the hemoglobin prediction results using eight different machine learning techniques with 11 time domain features selected by CFS method from PPG dataset (Fig. 7).

In addition to these above results, we have given the relationships between target (the real measured data) and predicted hemoglobin values using data fit on MATLAB™. Several regression methods like CART, GRNN and SVR machine learning methods have been applied. Many feature selection methods like RFS and CFS have been used. In the given figure, we have used the  $R^2$  performance criteria to compare methods and to select the



best one on the prediction of hemoglobin level based on the time domain features extracted from PPG signal. According to the obtained results, support vector regression has been inferred as the best machine learning method using the mean square error criteria in the prediction of hemoglobin level. As the hybrid model, the combination of RFS and SVR has obtained very good results in the prediction of hemoglobin level. The obtained results have shown that the combination of RFS and SVR could be confidently used in the prediction of hemoglobin level as the non-invasive method based on PPG signal.

#### 4. Conclusions

The measurement of the hemoglobin content in blood does not only give information about the health of a patient, but also helps the doctor to decide on a treatment for many diseases. Hemoglobin in blood is normally measured after taking blood sample from the body. This causes inconvenience for the patients whose hemoglobin is continuously measured. A new non-invasive hemoglobin measurement method, which is based on PPG signal and machine learning, has been proposed in this study. In this method, the hemoglobin values of the people were estimated by receiving PPG signals from them, extracting time-domain features from those signals and entering them into machine learning algorithms. 8 different regression methods were used for the prediction of the hemoglobin value, including the CART (classification and regression trees), LSR (least squares regression), GLR (generalized linear regression), MVLR (multivariate linear regression), PLSR (partial least squares regression), GRNN (generalized regression neural network), MLP (multilayer perceptron) and SVR (support vector regression) methods. Besides, the methods of RELIEF feature selection (RFS) and correlation based feature selection (CFS) methods were used to decrease the number of features. The number of the features was considerably decreased down to 10 and 11 (RFS and CFS) with the application of these feature selection algorithms onto the dataset with 44 features. According to the  $R^2$  performance criteria, the support vector regression has been found to be the best machine learning method in the prediction of hemoglobin level. As the hybrid model, the combination of RFS and SVR has obtained very good results in the prediction of hemoglobin level. The obtained results have shown that the combination of RFS and SVR could be confidently used in the prediction of hemoglobin level as the non-invasive method based on PPG signal.

#### References

- [1] J.L.A. Nirupa, V.J. Kumar, Non-invasive measurement of hemoglobin content in blood, in: 2014 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 2014, pp. 1–5.
- [2] J. Kraitl, U. Timm, H. Ewald, E. Lewis, Non-invasive measurement of blood components, in: 2011 Fifth International Conference on Sensing Technology (ICST), 2011, pp. 253–257.
- [3] J. Bailey, M. Fecteau, N.L. Pendleton, Wireless Pulse Oximeter, Worcester Polytechnic Institute, 2008, April 24, Bachelor Degree Thesis.
- [4] P.A. Kyriacou, S. Powell, R.M. Langford, D.P. Jones, Investigation of oesophageal photoplethysmographic signals and blood oxygen saturation measurements in cardiothoracic surgery patients, *Physiol. Meas.* 23 (2002) 533–545, August.
- [5] A. Johansson, Neural network for photoplethysmographic respiratory rate monitoring, *Med. Biol. Eng. Comput.* 41 (2003) 242–248, May.
- [6] S.H. Binns, D.D. Sisson, D.A. Buoscio, D.J. Schaeffer, Doppler ultrasonographic, oscillometric sphygmomanometric, and photoplethysmographic techniques for noninvasive blood-pressure measurement in anesthetized cats, *J. Vet. Intern. Med.* 9 (1995) 405–414, November–December.
- [7] J. Allen, Photoplethysmography and its application in clinical physiological measurement, *Physiol. Meas.* 28 (2007) R1–R39, March.
- [8] S.B. Usman, M.A.B. Ali, M.M.B. Reaz, K. Chellapan, Second derivative of photoplethysmogram in estimating vascular aging among diabetic patients, in: 2009 International Conference for Technical Postgraduates (Techpos 2009), vol. 357, 2009, pp. 5–7.
- [9] W. Chen, S. Lei, L. Guo, Y. Chen, M. Pan, Study on conditioning and feature extraction algorithm of photoplethysmography signal for physiological parameters detection, in: 2011 4th International Congress on Image and Signal Processing (CISP), 15–17 October 2011, vol. 4, 2011, pp. 2194–2197.
- [10] M. Iwasaka, J. Miyakoshi, S. Ueno, Optical absorbance of hemoglobin and red blood cell suspensions under magnetic fields, *IEEE Trans. Magn.* 37 (4) (2001) 2906–2908, July.
- [11] G.D. Jay, J. Racht, J. McMurphy, Z. Mathews, A. Hughes, S. Suner, G. Crawford, Point-of-care noninvasive hemoglobin determination using fiber optic reflectance spectroscopy, in: Proceedings of the 29th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society, August 2007, 2007, pp. 2932–2935.
- [12] L.M. Head, R.M. Pierson, Functional near infrared spectroscopy for Hb and HbO<sub>2</sub> detection using remote sensing, in: Sensors Applications Symposium (SAS), 23–25 February, 2010, 2010, pp. 166–169.
- [13] U. Timm, D. McGrath, E. Lewis, J. Kraitl, H. Ewald, Sensor system for non-invasive optical hemoglobin determination, *IEEE Sens.* 2009 (2009) 1975–1978, October.
- [14] T. Yichao, Y. Li, X. Hou, Non-invasive detecting cerebral hemoglobin concentration of newborns during oxygen inhalation using near-infrared spectroscopy, in: IEEE/ICME International Conference on Complex Medical Engineering (CME), May 2011, 2011, pp. 121–126.
- [15] K. Luangrat, P. Suriyaphol, Y. Wongsawat, Hemoglobin prediction system from pulse signal, 2013 Asia-Pacific Signal and Information, in: Processing Association Annual Summit and Conference (APSIPA), 2013, pp. 1–4.
- [16] 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, <http://dx.doi.org/10.1364/BOE.5.001145>
- [17] A.R. Kavsaoglu, K. Polat, M.R. Bozkurt, Temassız Elde Edilen Fotoplethysmografi Sinyalleri İle Gerçek Zamanlı Kalp Hızı Tespiti, 196–199, 22, in: IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı 23–25, Nisan, Karadeniz Teknik Üniversitesi, Trabzon, 2014.
- [18] [http://en.wikipedia.org/wiki/Relief.%28feature\\_selection%29](http://en.wikipedia.org/wiki/Relief.%28feature_selection%29) (accessed February, 2015).
- [19] I. Kononenko, et al., Overcoming the myopia of inductive learning algorithms with RELIEFF, *Appl. Intell.* 7 (1) (1997) p39–p55.
- [20] K. Kira, L. Rendell, The feature selection problem: traditional methods and a new algorithm, in: AAAI-92 Proceedings, 1992.
- [21] S.S. Durduran, A decision making system to automatic recognize of traffic accidents on the basis of a GIS platform, *Expert Syst. Appl.* 37 (12) (2010) 7729–7736.
- [22] K. Polat, V. Kirmaci, Determining of gas type in counter flow vortex tube using pairwise fisher score attribute reduction method, *Int. J. Refrig.* 34 (6) (2011) 1372–1386.
- [23] [www.stat.wisc.edu/~loh/treeprogs/guide/wires11.pdf](http://www.stat.wisc.edu/~loh/treeprogs/guide/wires11.pdf) (accessed February, 2015).
- [24] [http://en.wikipedia.org/wiki/Least\\_squares](http://en.wikipedia.org/wiki/Least_squares), (accessed February, 2015).
- [25] [http://en.wikipedia.org/wiki/Generalized\\_linear\\_model](http://en.wikipedia.org/wiki/Generalized_linear_model) (accessed February, 2015).
- [26] S. Ganesh, Multivariate linear regression, *Int. Encycl. Educ.* (3rd ed.) (2010) 324–331.
- [27] [http://en.wikipedia.org/wiki/Partial\\_least\\_squares\\_regression](http://en.wikipedia.org/wiki/Partial_least_squares_regression) (accessed February, 2015).
- [28] <http://www.ijcaonline.org/volume7/number13/pxc3871799.pdf> (accessed February, 2015).
- [29] [http://en.wikipedia.org/wiki/Multilayer\\_perceptron](http://en.wikipedia.org/wiki/Multilayer_perceptron) (accessed February, 2015).
- [30] F. Rosenblatt, Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, Spartan Books, Washington DC, 1961.
- [31] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, in: D.E. Rumelhart, J.L. McClelland, PDP Research Group (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations, MIT Press, 1986.
- [32] [http://www.saedsayad.com/support\\_vector\\_machine\\_reg.htm](http://www.saedsayad.com/support_vector_machine_reg.htm) (accessed February, 2015).
- [33] Y. Yildirim, M. Bayramoglu, Adaptive neuro-fuzzy based modelling for prediction of air pollution daily levels in city of Zonguldak, *Chemosphere* 63 (2006) 1575–1582.
- [34] K. Polat, V. Kirmaci, Application of the output dependent feature scaling in modeling and prediction of performance of counter flow vortex tube having various nozzles numbers at different inlet pressures of air, oxygen, nitrogen and argon, *Int. J. Refrig.* 34 (6) (2011) 1387–1397.
- [35] Advanced Prime Materials Korea (2015). <http://www.apmkr.com/> (February, 2015).
- [36] A.R. Kavsaoglu, K. Polat, M.R. Bozkurt, A novel feature ranking algorithm for biometric recognition with PPG signals, *Comput. Biol. Med.* 49 (2014) 1–14.
- [37] A.R. Kavsaoglu, K. Polat, M.R. Bozkurt, An innovative peak detection algorithm for photoplethysmography signals: an adaptive segmentation method, *Turk J Elec Eng Comp Sci* (2014), <http://dx.doi.org/10.3906/elk-1310-177>
- [38] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, *JMLR* (2003) 3.
- [39] R. Das, A. Sengur, Evaluation of ensemble methods for diagnosing of valvular heart disease, *Expert Syst Appl* 37 (7) (2010) 5110–5115.
- [40] R. Das, I. Turkoglu, A. Sengur, Diagnosis of valvular heart disease through neural networks ensembles, *Comput Methods Progr Biomed* 93 (2) (2008) 185–191.