An Efficient Technique of Hemoglobin Level Screening Using Machine Learning Algorithms

Conference Paper · December 2019

CITATIONS
0

READS
91

3 authors, including:

Nahiyan Bin Noor
Chittagong University of Engineering & Technology
5 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Mohammad saeid Anwar
Chittagong University of Engineering & Technology
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

An Efficient Technique of Hemoglobin Level Screening Using Machine Learning Algorithms

Nahiyan Bin Noor[1*], Md. Saeid Anwar[2*], and Mrinmoy Dey[3]

[1,2,3]Department of Electrical and Electronic Engineering, Chittagong University of Engineering and Technology, Chattagram-4349, Bangladesh.

[1]nahiyannabil@gmail.com, [2]u1402114@student.cuet.ac.bd, [3]mrinmoycuet@gmail.com

Abstract- **Hemoglobin (Hb), a very significant parameter for the human body and deficiency of it causes anemia. During pregnancy, menstruation and ICU deficiency of it can be very risky and even caused death. So, it is important to diagnose it continuously. Usually, physicians examine it by conducting a blood test to confirm it is painful, time-consuming and costly. The major concept of this study is to screen Hb levels within a short period of time. In this study, the data of clinical blood Hb levels of a total of 104 people (54 males and 50 females) are collected along with an eye conjunctiva image. The images are taken with a Smartphone camera of constant resolution and lighting. Using MATLAB, image processing method, the percentages of the red, green and blue pixels are extracted. Taking those features, the Hb level is plotted. The 104 data have been split into two sets where the first 81 data for training purposes, the remaining 23 data have been considered for testing. To train the model of 81 data, Multivariate Linear Regression (MLR), Decision Tree (Medium), Linear Support Vector Regression (SVR) are taken and the lowest percentage of error of 11.01% has been found in the Decision Tree (Medium) while testing the 23-test data.**

Index Terms— **Hemoglobin (Hb); Eye Palpebral Conjunctiva; Multivariate Linear Regression, Decision Tree; Linear Support Vector Regression (SVR).**

I. INTRODUCTION

Hemoglobin (Hb) is a metalloprotein in the red blood cells (RBC) that contain iron and transport oxygen in all vertebrates [1]. Lack of Hb is heavily linked with premature baby births, less birth weight and mortality of maternal [2]. World Health Organization (WHO) reports that [3,4] 24.8% of the total population is affected by a lack of Hb. There are many methods to calculate Hb in the blood. CBC (Complete Blood Count) is clinically recognized method where blood is taken invasively which is painful, time-consuming as well as costly. So, it is difficult to diagnose continuously specially during pregnancy and ICU patients where continuously Hb level screening is a must. It is known that 13.5 to 17.5 gm/dl is safe for men and 12 to 15.5 gm/dl is adequate for women. But clinically less than 10 gm/dl is considered as a lack of Hb on average.

In this research work, total 104 data are collected from subjects. From the total dataset 81 data are taken to train the model and 23 data kept separate for further testing the algorithm. Among 104 data, there 54 are male and 50 are female of different ages. All data are collected from Chittagong Medical College Hospital, Chittagong and Cox's Bazar Medical College Hospital. Everyone's conjunctiva picture has been taken with a smartphone camera of 12-megapixel camera. The CBC (Complete Blood Count) Report is also taken where the Hb level is mentioned and certified by doctors which is screened in a diagnostic center. We have used this CBC report for actual Hb level. This report was taken within two or three days after test. Eye conjunctiva images are captured of their and applied image processing to extract the percentage of red, green and blue. Our aim is to predict numeric hemoglobin value by regression. Different regression technique is used for training the model that measures blood Hb level such as Multivariate Linear Regression (MLR), Linear SVR, Decision Tree (Medium) Tree and compared the result by taking the most accurate method. Machine Learning Technique can produce good overall accuracy in predicting blood Hb levels, unlike the existing methods.

II. RELATED WORK

Putut Dewantoro et al. [5] formulated a non- invasive Hb measurement technique based on a smartphone. Total techniques are run by PPG signal which from patients' finger. Overall processing used linear regression method for Hb measurement. They got standard deviation of.254gm/dl and the relative standard deviation is 12% from the device. In 2016 Shahzab Bukhari et al. [6] conducted research for the screening of Hb levels using the non-invasive device. Spectrophotometry technique was used to Hb in blood flow. The device was placed in the index finger to determine the Hb. Researchers took the invasive lab results as a reference and compared the non-invasive results with them. They found 92% accuracy, 89% sensitivity, and 76% specificity.

Resit Kavasaoglu et al. [7] Hb assessment based on PPG signal where characteristics PPG signal features are used for developing a prediction. Different machine learning algorithms such as CART (classification and regression trees), GLR (generalized linear regression), LSR (least squares regression), MVLR (multivariate linear regression), GRNN (generalized regression neural network), RFS (RELIEF feature selection) and SVR (support vector regression) are used. In this research, PPG signals from 33 people were obtained and extracted 40 features from them. They received a mean square error of 0.0027 using SVR. Soumil Chugh & Jaskirat Kaur proposed a method [8] where the principle of photoplethysmography and Beer-Lambert law is used to measure the hemoglobin levels in the blood. 12 people are tested with known Hb levels and found 10% of deviation from an actual value.

Haiquan et al. illustrated a research work where they are predicted haemoglobin level using back propagation of Artificial neural network (BP-ANN).9 LEDs used to construct broadband light source along with Si photodiode array and grating spectrograph. With these they established a high-performance spectrophotometric system. 109 volunteers used for this research work. This method earned .94 correlation coefficient of network model and BP-ANN as a

978-1-7281-6040-5/19/$31.00 ©2019 IEEE

machine learning technique is used. Overall 7.41% standard error of prediction is found [14].

III. Methodology

This whole study has been done in MATLAB Software. Image Processing, color thresholding, and Machine Learning all have been done in the same software. A flow chart is presenting the whole process which is presented in fig. 1. The major part of the work is in the flow chart which will be described in detail below.

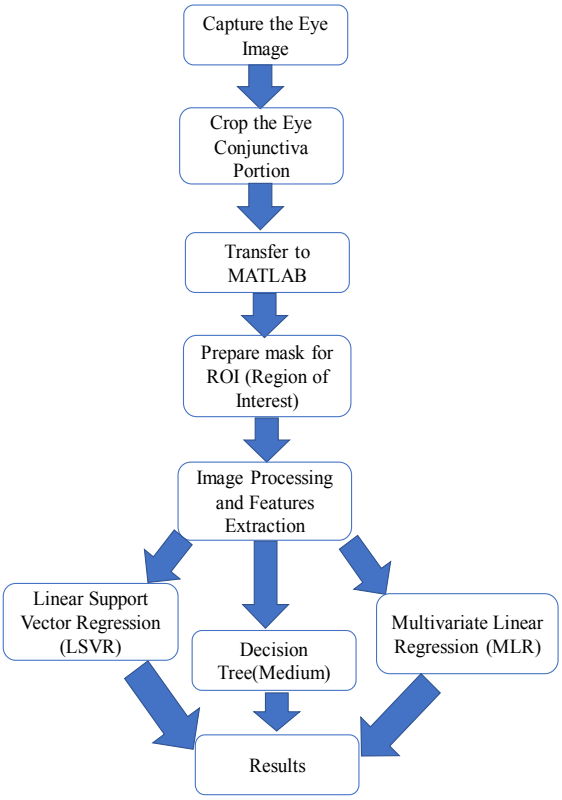

Fig. 1 The flowchart describing the overall procedure

A. *Capture the eye conjunctiva image and crop the eye conjunctiva portion*

The palpebral eye conjunctiva picture has been taken utilizing a 12-megapixel cell phone camera. It has been finished by pulling the lower eyelid and snapping the photo of the eye conjunctiva. At that point focusing front conjunctiva, the picture has been taken. The steady light is required for taking the picture so a device represented in fig. 2 is made. Utilizing that the picture is caught by keeping on the flashlight of mobile. In fig. 3 the picture appears. The device is shown in fig. 2 basically a very low-cost box that is handmade spending only $1. It creates a lightless medium where any light can to enter as a result a constant medium of light is created by this box. This box is designed in such a way where a smartphone can easily be placed and by using that smartphone palpebral eye conjunctiva image can be captured. Then the image is cropped and taken only the conjunctiva portion. After cropping the image, it is taken to MATLAB for further processing. As everyone uses a smartphone nowadays, the cost of a smartphone camera is neglected as a part of our device. It is very user-friendly as there is no complication in wearing the box shown in fig. 2 and anyone can take the picture with their smartphones.

B. *Prepare mask for ROI using MATLAB image thresholding app*

The taken image is cropped, kept the Region of Interest (ROI) and eliminated the white background color of the image by using the MATLAB thresholding app. When it turned totally black this is called masking where unwanted pixels are eliminated. After the image looks like fig. 4

Fig. 2 Device for creating constant light

Fig. 3 Palpebral Conjunctiva

Fig. 4 Masked and Photo with ROI

C. *Image processing and feature extraction in MATLAB*

Suner et al. [9] postulated that color characteristics of the conjunctiva are correlated to Hb concentration and that the representation of these characteristics in a digital format is based on relative shades of red, green and blue for each pixel. So, from that, we are used eye image and using MATLAB image processing and feature extraction. Here the percentage of RGB pixel from eye image using MATLAB.
The percentage red, green and blue pixels are extracted as follows.

1. *The Red pixel percentage extraction*

Using equation (1) the red pixel percentage is extracted. In an RGB image, there are three layers Red, Green, and Blue. The percentage of red is more effective for determining the palpebral pallor. There a correlation between green pixel percentage and clinical Hb level shown in fig. 5. The equation is below.

The percentage of red pixel=

$$\frac{\text{Total red pixel}}{\text{Total Red pixel+Total Green pixel+Total blue Pixel}} * 100 \quad (1)$$

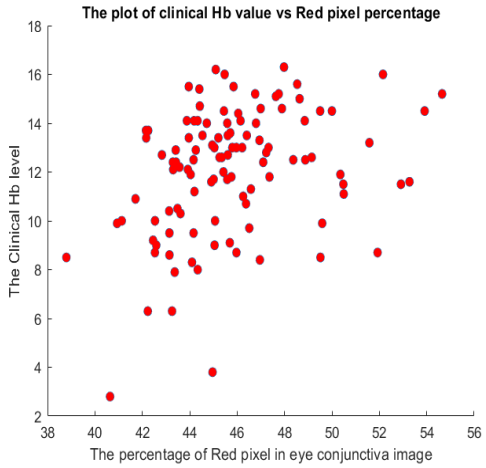

Fig. 5 The plot of clinical Hb level vs Red pixel percentage.

2. *The Green pixel percentage extraction*

The percentage of the green pixels is extracted using equation (2). There a correlation between green pixel percentage and clinical Hb level shown in fig. 6.
The percentage of green pixel=

$$\frac{Total\ Green\ pixel}{Total\ Red\ pixel+Total\ Green\ pixel+Total\ blue\ Pixel} * 100 \quad (2)$$

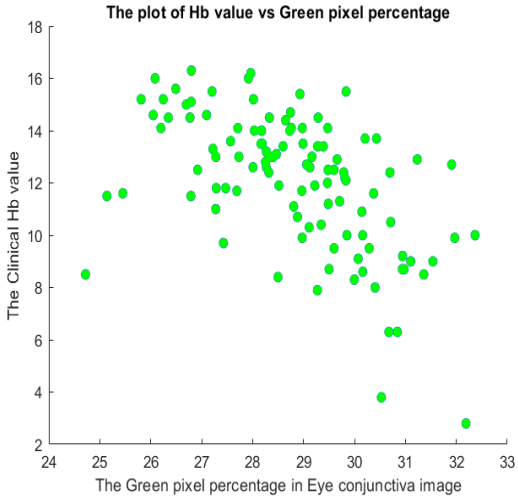

Fig. 6 The plot of clinical Hb level vs Green pixel percentage.

3. *The Blue pixel percentage extraction*

The percentage of the blue pixel is extracted using equation (2). There a correlation between green pixel percentage and clinical Hb level shown in fig. 7.

The percentage of blue pixel=

$$\frac{Total\ Blue\ pixel}{Tota\ Red\ pixel+Total\ Green\ pixel+Total\ blue\ Pixel} * 100 \quad (3)$$

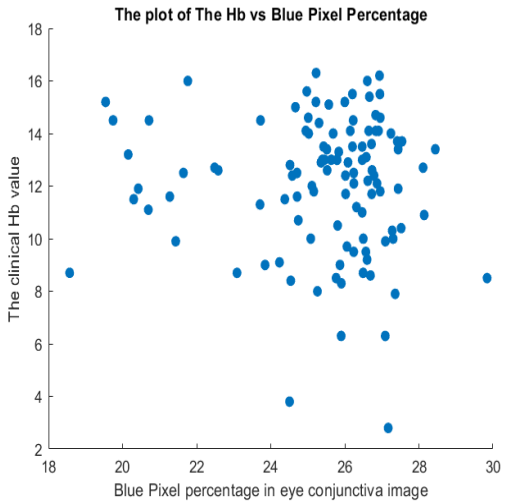

Fig. 7 The plot of clinical Hb level vs Blue pixel percentages.

Using equation (1), (2) and (3) the most significant three features Red, Green, and Blue pixels are extracted. In Table I. 6 people's data are given out of 104 people with their red pixel percentage, Green Pixel percentage, blue pixel percentage along with their Hb value.

TABLE. I: LIST FOR PEOPLE WITH THEIR PERCENTAGE OF RED, GREEN AND BLUE PIXEL ALONG WITH BLOOD HB LEVEL.

Sl	Name	Red	Green	Blue	Hb (g/dl)
1.	Subject 1	49.5037	26.7716	23.7246	14.5
2.	Subject 2	38.7968	31.3618	29.8414	8.5
3.	Subject 3	47.7549	26.2457	25.9994	15.2
4.	Subject 4	44.1609	29.5993	26.2399	9.5
5.	Subject 5	42.2324	30.6757	27.0919	6.3
6	Subject 6	54.6478	25.8109	19.5413	15.2

D. Regression

The MATLAB software is used where Regression Learner and Classification Learner automatically built-in. A Microsoft Excel file is formed containing all data with Red, Green, Blue pixel percentage along with Hb level and Anaemic condition. The excel file imported as CSV file to regression learner where Red, green and blue pixels are taken as feature and Hb is taken as a response. A total of 81 data are taken for training. After training the model is saved for further prediction. Then this model is tested with the rest 23 data and got a result that is slightly deviated from an actual value. In this study, Multiple Linear Regression, Tree and Linear SVM algorithms are used.

1. *Multivariate Linear Regression*

Linear Regression is a common predictive model that identifies the correlation among the variables [10,11]. Apart from a single variable and multiple variable data types the concept is linear. Linear regression can be either simple linear or multiple linear regression. In this study, multiple linear regression is used. Multiple linear regression is the process of

prediction with more than one independent variable which is described in Eq. (4)

$$y = a+b_1x+b_2x+b_3x+\ldots\ldots b_nx+c \quad (4)$$

2. Decision Tree (Medium)

The decision tree builds regression models in the form of a tree structure. The whole dataset is broken into smaller subsets with an increase in depth of the tree. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. The leaf node represents a classification or decision. The topmost decision node in a tree that corresponds to the best predictor is called the root node. Decision trees can handle categorical as well as numerical data [12].

3. Linear SVR (Support vector regression):

Support Vector Regression uses linear kernel functions for regression than is similar to support vector machines but SVR sets the tolerance margin (ε) to approximation, not like SVM which should be taken from the problem. Support Vector Regression with linear kernel function is described in Eq. 5.

$$y = w.x+b \quad (5)$$

IV. RESULTS & DISCUSSION

In the Regression Learner app in MATLAB relatively good Accuracy is found in the Medium tree linear regression technique has. The Actual value predicted value along with the error in terms of red percentage is shown in fig. 8 which is for 81 trained data in the Decision Tree (Medium) algorithm. In fig. 8 it can be seen that 4 or 5 predicted values have a great deviation from actual value and the other values has less deviation from an actual value. It is found that the Root Mean Square Error (RMSE) in Decision Tree, LSVR, MLR is 2.1005, 2.0161 and 2.0202 respectively.

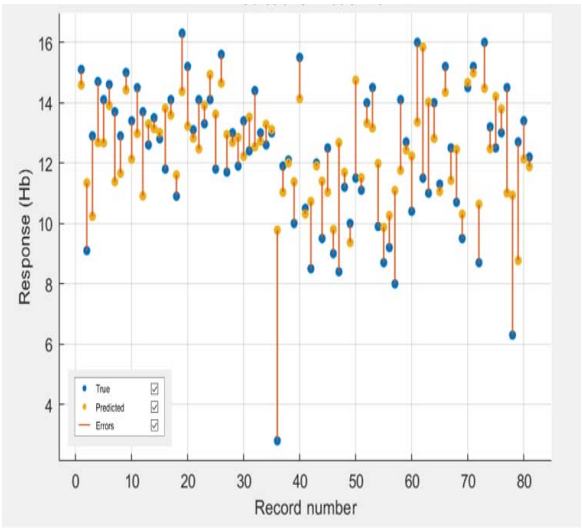

Fig. 8 The Actual value, predicted value along with the error in terms of red percentage for 81 test data

In fig. 9 the plot depicts the correlation of the predicted values and the actual values where blue dots are true response and orange dots are predicted response. In fig. 10 the residual values that are a deviation from actual values are illustrated for 81 data which are taken as training data. The training model result is then tested with the rest 23 data.

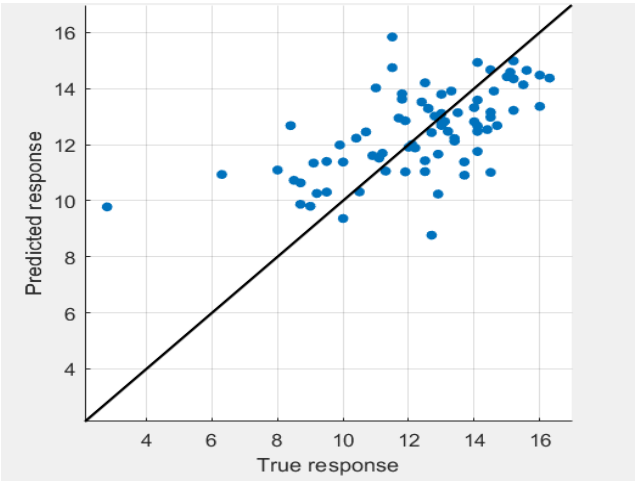

Fig. 9 The predicted value plot for 81 train data

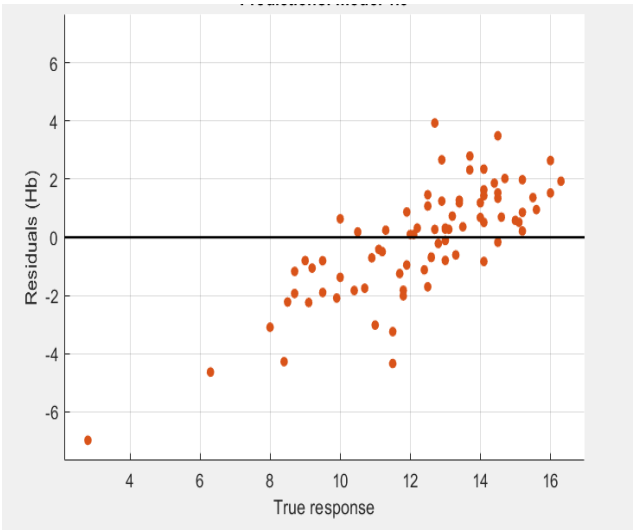

Fig. 10 The residuals value deviated from the centre for 81 train data.

As it is mentioned earlier that three regression algorithms are used to determine blood Hb level named Decision Tree, LSVR and MLR. Applying these three techniques a comparative study has done to find the best method. For this, the accuracy test is done on 23 test datasets after training the model with 81 datasets. The three methods are depicted in the table below.

In Table II, the actual value and predicted value are depicted which is determined by using the Decision Tree algorithm. The average percentage of error in this algorithm is 11.01% The percentage of error is also shown in the table for an individual record.

In Table III, the actual value and predicted value is illustrated which is determined by using LSVR (Linear Support Vector Regression) algorithm. The average percentage of error in this algorithm is 15.056% The percentage of error is also depicted in the table for an individual record.

In Table IV, the actual value and predicted value is illustrated which is determined by using MLR (Multivariate Linear Regression) algorithm. The average percentage of error in this algorithm is 18.03%. The Percentage of error is also illustrated in the table for an individual record.

TABLE II: LIST OF PEOPLE WITH ACTUAL HB LEVEL, PREDICTED THROUGH MEDIUM TREE ALGORITHM AND THE PERCENTAGE OF ERROR

Sl.	Name	Actual Hb(g/dl)	Predicted Hb (Medium Tree)	% Error
1	Subject 1	6.3	9.4211	49.54
2	Subject 2	13.5	13.6417	1.05
3	Subject 3	11.7	13.6417	16.596
4	Subject 4	13.5	13.6417	1.05
5	Subject 5	12.4	9.4211	24.02
6	Subject 6	16.2	13.125	18.98
7	Subject 7	8.6	9.4211	9.55
8	Subject 8	10.3	12.2	18.45
9	Subject 9	13	12.2	6.15
10	Subject 10	9.7	13.125	35.31
11	Subject 11	12.6	12.2	3.17
12	Subject 12	15.4	13.6417	11.41
13	Subject 13	14.1	14.5	2.83
14	Subject 14	9	9.4211	4.68
15	Subject 15	14.6	14.5	0.685
16	Subject 16	14	13.6417	2.56
17	Subject 17	10	9.4211	5.789
18	Subject 18	8.3	9.4211	13.51
19	Subject 19	13.6	13.125	3.49
20	Subject 20	9.9	9.4211	4.837
21	Subject 21	11.6	9.4211	18.78
22	Subject 22	12.2	12.2	0
23	Subject 23	12.1	12.2	0.82

TABLE III: LIST OF PEOPLE WITH ACTUAL HB LEVEL, PREDICTED THROUGH LINEAR SVR ALGORITHM AND THE PERCENTAGE OF ERROR

Sl.	Name	Actual Hb(g/dl)	Predicted Hb (Linear SVR)	% Error
1	Subject 1	6.3	9.4211	49.54
2	Subject 2	13.5	13.6417	1.05
3	Subject 3	11.7	13.6417	16.596
4	Subject 4	13.5	13.6417	1.05
5	Subject 5	12.4	9.4211	24.02
6	Subject 6	16.2	13.125	18.98
7	Subject 7	8.6	9.4211	9.55
8	Subject 8	10.3	12.2	18.45
9	Subject 9	13	12.2	6.15
10	Subject 10	9.7	13.125	35.31
11	Subject 11	12.6	12.2	3.17
12	Subject 12	15.4	13.6417	11.41
13	Subject 13	14.1	14.5	2.83
14	Subject 14	9	9.4211	4.68
15	Subject 15	14.6	14.5	0.685
16	Subject 16	14	13.6417	2.56
17	Subject 17	10	9.4211	5.789
18	Subject 18	8.3	9.4211	13.51
19	Subject 19	13.6	13.125	3.49
20	Subject 20	9.9	9.4211	4.837
21	Subject 21	11.6	9.4211	18.78
22	Subject 22	12.2	12.2	0
23	Subject 23	12.1	12.2	0.82

TABLE IV: LIST OF PEOPLE WITH ACTUAL HB LEVEL, PREDICTED THROUGH LINEAR REGRESSION ALGORITHM AND THE PERCENTAGE OF ERROR

Sl.	Name	Actual Hb(g/dl)	Predicted Hb (Linear Regression)	% Error
1	Subject 1	6.3	9.5271	51.22
2	Subject 2	13.5	12.9698	3.93
3	Subject 3	11.7	12.6926	8.48
4	Subject 4	13.5	12.1355	10.11
5	Subject 5	12.4	10.2652	17.22
6	Subject 6	16.2	13.2757	18.05
7	Subject 7	8.6	10.8976	26.72
8	Subject 8	10.3	12.0781	17.26
9	Subject 9	13	11.3141	12.96
10	Subject 10	9.7	13.7733	41.99
11	Subject 11	12.6	11.9076	5.5
12	Subject 12	15.4	12.217	20.67
13	Subject 13	14.1	14.9919	6.33
14	Subject 14	9	9.6372	7.08
15	Subject 15	14.6	15.3303	5.002
16	Subject 16	14	11.7709	15.92
17	Subject 17	10	10.3812	3.812
18	Subject 18	8.3	11.0059	32.6
19	Subject 19	13.6	13.1075	3.6
20	Subject 20	9.9	8.9972	9.11
21	Subject 21	11.6	10.4947	9.52
22	Subject 22	12.2	11.307	7.31
23	Subject 23	12.1	10.6988	11.58

Comparing three algorithms it can be found that Decision Tree is the best technique with only 11.01% of error. The 23-test data is compared with their clinical Hb value in the Decision Tree algorithm and plotted in fig. 11 where record numbers from 1 to 23 are along X-axis and actual and predicted Hb levels are along Y-axis.

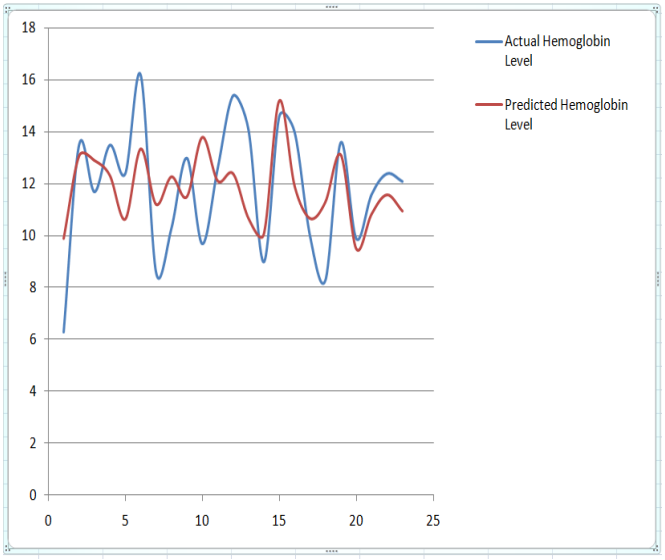

Fig. 11 Comparison between clinical value and predicted value

Table V depicts the comparison between other work and this work.

TABLE V: COMPARATIVE STUDY BETWEEN OTHER METHODS VS THIS METHOD

Sl.	Author's name	Method	Technique/Algorithm	Dataset number	Accuracy/Error
01.	Putut Dewantoro et al. [5]	A smartphone-based method that is run by PPG signal which from patients' finger.	Multiple Linear Regression	30	Accuracy of 88%
02.	Shahzab Bukhari et al. [6]	Machine learning and Spectrophotometry technique	Linear Regression	591	92% for male and 74% for female
03.	A. Resit Kavasaoglu et al. [7]	Machine learning and PPG signal	CART, LSR, SVR, RFS, GLR	33	Mean Square error of 0.0027
04.	Haiquan ding et al. [14]	Machine learning, spectrophotometric system	Back propagation ANN	109	SEP (standard error of prediction) 7.41%
05.	Edward J. Wang et al. [13]	Machine learning, PPG signal and smartphone-based	Linear Regression	32	RMSE (Root Mean Square Error) 1.93 with outliers and RMSE 1.27 without outliers
06.	This Work	Taking the picture of palpebral eye conjunctiva and image processing method.	Decision Tree, Multivariate Linear Regression, Linear SVR	104 (84 train data 23 test data)	Accuracy of 88.99%

By observing the above result, it can be said that our method is different from all invasive and non-invasive technique of measuring blood Hb level as it is less time consuming, cost-efficient along with that it has more accuracy than [5], [6]. The results obtained are comparable to data obtained in [7], [13], [14]. So, it is said to an efficient technique in a new way that uses eye palpebral pallor and red, green and blue pixel percentage of eye conjunctiva image. The novelty of this work is the collection of dataset, applying different machine learning algorithms on those data such Linear Regression, Support Vector Regression and Decision Tree and selecting the best algorithm which is Decision Tree having 88.99% of accuracy.

V. Conclusion

It is an efficient method of Hb screening along with painless, less time consuming and cost-effective. Normally it cost 400 BDT (almost $5), but this method is cost-free. It is a crucial tool in the fight against the lack of red blood cells or Hb in the lesser developed regions of the world. This can result in a significant increase in the health conditions in those areas along with sufficiently fewer deaths caused by the effects of lack of Hb. Moreover, this method can also be employed to detect the lack of blood caused by other means like blood loss during menstruation, pregnancy, and physical injury of different organs of the human body.

VI. Acknowledgement

The authors would like to thank the authority, doctors and intern doctors of Chittagong Medical College Hospital and Cox's Bazar Medical College Hospital for their support to collect data. The authors also would like to thank the Department of Electrical and Electronic Engineering of Chittagong University of Engineering and Technology for the help in this research work.

VII. References

[1] Maton, Anthea, Jean Hopkins, Charles William, Mclaughlin, Susan Johnson, Maryanna Quon Warner, David LaHart, Jill D. Wright (1993). *Human Biology and Health. Englewood Cliffs*, New Jersey, USA: Princeton Hall. ISBN 0-13-981176-1.

[2] Monika Malhotra, "Severe Anemia linked to poor outcomes for pregnant women and their babies", *International Journal Gynecology and obstetrics,* vol. 79 p p93-100, 2002

[3] World Health Organization "Nutritional Anemias. Report of a WHO scientific group," World Heal Organ—Tech Rep Ser, 405:5-37, 1968.

[4] Prevalence of Anemia, "Department of Nutrition for Health and Development, World Health Organization," Geneva, Switzerland. WHO Vitamin and Mineral Nutrition Information System, Public Health Nutr. April 2009.World Health Organization. Worldwide prevalence of anemia. WHO Rep. 51, 2005.

[5] Putut Dewantoro, Clinton Elian Gandana, Resti Oktia Rahman Hasballah Zakaria, Yoke Saadia Irawan, "Development of Smartphone-based Non-Invasive Hb Measurement".

[6] Shahzaib Bukhari et al. "Transforming Community-based Screening of Total Hb using Non-invasive Devise," *IEEE Conference on Technologies for Sustainability (SusTech),2016*

[7] A. Resit Kavasaoglu, Kemal Polat, M. Hariharan, "Non-invasive prediction of Hb level using machine learning techniques with PPG signal's characteristics features," *Elsevier Science Publishers B. V. Amsterdam*, vol. 37, Issue C, pp. 983-991, December 2015

[8] Soumil Chugh, Jaskirat Kaur, "Non-Invasive Hb Monitoring Device," *2015 International Conference on Control, Communication & Computing India (ICCC)* | 19-21 November 2015 ,Trivandrum.

[9] S. Suner, G. Crawford, J. Mcmurdy, G. Jay, "Non-invasive determination of Hb by digital photography of palpebral conjunctiva", J. Emerg,Med,33 (2)(2007): 10

[10] GAF Seber, AJ Lee, "Linear regression analysis," *Wiley Series in Probability and Statistics*, 2012.

[11] DC Montgomery, EA Peck, GG Vining, "Introduction to linear regression analysis", *Wiley Series in Probability and Statistics*, 2015

[12] https://medium.com/@rishabhjain_22692/decision-trees-it-begins-here-93ff54ef134. Access date: September, 2015

[13] Edward J. Wang, William Li, Junyi Zhu, Rajneil Rana, and Shwetak N. Patel, "Noninvasive Hb Measurement Using Unmodified Smartphone Camera and White Flash," *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017

[14] Haiquan Ding,* Qipeng Lu, Hongzhi Gao, and Zhongqi Peng, "Non-invasive prediction of hemoglobin levels by principal component and back propagation artificial neural network," *Biomedical Optics Express,* vol. 5, issue no. 4, pp. 1145-1152, March 2014