Non-invasive Estimation of Clinical Severity of Anemia Using Hierarchical Ensemble Classifiers

Sushovan Chakraborty1 · Krity Kansara1 · R Dinesh Kumar1 · Dhivya Swaminathan1 · Kiran Aatre1 · Soumyadipta Acharya2

Received: 3 October 2021 / Accepted: 26 August 2022 / Published online: 19 October 2022
© Taiwanese Society of Biomedical Engineering 2022

Abstract
Purpose Current techniques of anemia classification are either invasive, expensive or inaccurate, making them ill-suited for community health-worker based screening programs. In this study, we propose an Artificial Intelligence (AI) based anemia classification method using a multi-wavelength non-invasive photometry device.

Methods A finger mounted photo-plethysmogram (PPG) device was designed to acquire PPG signals at four wavelengths (590, 660, 810, and 940 nm). A set of 13 attenuation and ratio-of-ratio features, derived using the peak and trough information extracted from the PPG signals, were used to develop a three-way hierarchical ensemble classification scheme using a machine-learning algorithm. PPG data from the device and true hemoglobin data from laboratory-based cell counters was collected for 1583 women of childbearing age and subjects were classified into either healthy (Hemoglobin, Hb > 11 g/dL), anemic (Hb: 7–11 g/dL) or severely anemic (Hb < 7 g/dL) categories.

Results We report a classification sensitivity of 92% (p < 0.05) and specificity of 84% (p < 0.05) in differentiating anemic and non-anemic women. We also report a sensitivity of 76% (p < 0.05), and specificity of 74% (p < 0.05) in identifying severe anemia.

Conclusion We believe that the proposed anemia classification algorithm, along with the associated sensor has the potential to be productized as a low-cost non-invasive anemia-screening device to rapidly determine next steps in clinical decision making in widespread community interventions.

Keywords Anemia · Ensemble classifiers · Photoplethysmograph · Hemoglobin

1 Introduction
Anemia is a condition characterized by reduced levels of hemoglobin (Hb) in the blood, impacting the amount of oxygen in circulation [1], affecting billions of people globally. Women and children are especially susceptible to the condition in both urban and rural settings, with very high prevalence and mortality in Low- and Middle-income countries (LMICs) [2]. Additionally, frequent monitoring of Hemoglobin is required in several other conditions, including infectious diseases such as dengue fever [3, 4].

The gold standard for Hb estimation involves an intravenous blood draw followed by testing the sample in a laboratory, making this method unsuitable for large scale communities and/or rural areas. As an inexpensive and appropriate alternative, the World Health Organization (WHO) released the Hemoglobin Color Scale (HCS) in 2001, a method to assess Hb concentrations outside the laboratory using frontline community health-care workers (CHWs) at the last mile. [5]. Although this color scale based semi-qualitative method serves as a useful tool in primary healthcare settings, its sensitivity and specificity in field conditions has been largely inconsistent [6] due to the subjectivity in interpretation, adherence to instructions, and ambient lighting conditions. The invasive nature of the method also poses a barrier to adoption due to hesitation from subjects, and the heightened risk and awareness of blood borne infections [7, 8].
Other conventional methods of estimating Hb concentrations include photometric techniques measuring absorbance of Hb in blood samples (Hemocue®, Hematocrit, among others) [9, 10]. These methods have acceptable clinical accuracies and can be used for on-the-spot analysis but are invasive, and expensive [11]. The issue of invasive techniques has been addressed by the advent of non-invasive methods to estimate Hb. Devices like Masimo’s Pronto and Orsense’s NBM 200 have recently introduced non-invasive hemoglobinometers. However, several studies have reported variations in sensitivity and specificity of these devices, as well as low levels of precision and large limits of agreement [12, 13]. Additionally, these devices are cost prohibitive for large scale use in public health programs in LMICs.

A study by Swaminathan et al. provides a quick, non-invasive, machine learning based method to estimate Hb value [14]. Although advantageous, for population-based screening of anemia especially by CHWs, clinical severity of anemia (normal, mild, moderate, or severe as defined by WHO) determines the next steps in clinical decision making, and decimal level accuracy is seldom needed. For instance, severe gestational anemia requires urgent referral to a hospital for possible blood transfusion or parenteral iron infusion, whereas mild anemia during early stages of pregnancy can be managed usually by increased iron supplementation and dietary modifications. Hence, instead of estimating/predicting the Hb value, interpreting the anemia level based on age and gender and then deciding the treatment plan for an individual, a CHW could be equipped with an objective easy-to-use classification tool to rapidly determine the clinical severity of anemia thereby improving the efficiency of the health care facilities provided in most LMICs.

Additionally, a report on the recommendations for anemia treatment by the National Health Mission, Govt. of India [15] clearly details that in a public healthcare setting, the therapeutic regimen for the categories 7–9 g/dL and 9–11 g/dL do not differ significantly. In order to enhance the efficacy of CHWs, we intended to develop a rapid three way classification technique for anemia screening, which exhibits better classification accuracy in comparison to the conventional classification scheme.

In this study, we leverage recent advancements in machine learning based approaches [16] to propose a non-invasive method of anemia classification, combining a multi-wavelength photoplethysmographic sensor and a novel classification algorithm based on hierarchical ensemble classifiers. We trained and tested this method against laboratory-based gold standard cell counters and achieved improved accuracies and reliability over the WHO color scale [6].

### 2 Materials and Methods

#### 2.1 Hardware

An in-house PPG acquisition device with a mechanical lever operated, fiber-optics based sensor developed at Robert Bosch Engineering and Business Solutions Pvt. Ltd. (RBEI), India in collaboration with the Center of Bioengineering Innovation & Design (CBID) at the Johns Hopkins University, USA was used in the study (Fig. 1). The device architecture consists of a single-point light source, a finger slot, a photodetector, an analog front-end board and a single board computer to orchestrate the proposed solution. Light from four LED sources of different wavelengths (590 nm, 660 nm, 810 nm and 940 nm) were made to pass through the subject’s finger from the dorsal end and the attenuated intensity of light was detected at the ventral end using a

---

**Fig. 1** Schematic representation of PPG collection device, automated peak and trough detection, and the classification scheme implemented in the machine learning algorithm.
photo detector. The four LEDs were switched ON at pre-programmed intervals of 10 ms. Only one LED was ON at any point of time. The intensity of light absorbed by the finger was de-multiplexed to correspond to the four LEDs, resulting in four independent PPG signals at the respective wavelengths. Actual values of total blood hemoglobin were calculated using the gold-standard automated cell counters. Automated Beckman Coulter LH780, Artocell-200 [Haematology Analyzer], Nihon Kohden Celltac and Mek-6420P were used during the process of data collection. Further details of the architecture and hardware and the system design is provided as part of another study [17].

2.2 Data Collection Protocol

With the assistance of the Prasanna School of Public Health, Manipal Academy of Higher Education (MAHE), India and in compliance with the Manipal University Ethics Board (MUEC/012/2016-17), data collection was conducted at public sector hospitals of Gangavathi and Koppal and a private sector hospital of Karkala. A total of 1583 women (17–52 years), with 825 of them being pregnant, provided the requisite informed consent to participate in the study, in accordance with the procedure laid out by the ethics committee. A set of four PPG signals of 60 s duration each was recorded for each subject. Each participant also provided an intravenous blood sample to determine the true hemoglobin concentration using automated cell counters (Beckman Coulter LH780, Artocell-200 [Haematology Analyzer], Nihon Kohden Celltac and Mek-6420P). The demographic distributions of the participants of the study is described in Table 1.

2.3 Data Analysis Protocol

2.3.1 Preprocessing

2.3.1.1 Peak and Trough Detection All PPG signals were bandpass filtered with a second order Butterworth filter (passband = 0.25–10 Hz). This band was chosen to eliminate signal artefacts due to breathing (baseline drifts), motions artefacts and high frequency noise [18]. Since the information of the PPG signal is in the frequency range of 0.5–4.0 Hz, signal characteristics remained unaffected by the application of the filter. Peaks and troughs in the PPG signal were detected using the methods outlined in Wang et al. [19]. Figure 1 illustrates an example of the automated peak and trough detection on a representative segment of acquired PPG.

2.3.1.2 Visual Inspection for Motion Artifact Removal The peak and trough labeled PPG signals were visually checked to remove signals that showed no PPG like waveforms and/or displayed significant motion artefacts, thus eliminating 67 of the 1583 total subjects from the dataset used for this analysis.

2.3.2 Feature Extraction

Estimation of Hb from PPG signals relies on the application of the Beer-Lambert’s law which states the relationship between the incident intensity (from the LEDs), attenuated intensity (captured by the photodetector) and the concentration of an analyte (blood molecules):

\[ I' = I_0 e^{-\varepsilon c d} \]

where \( I \) is the LED intensity, \( I' \) is the attenuated intensity, \( c \) is concentration, \( \varepsilon \) is the extinction coefficient, and \( d \) is the path length traversed by LED light within the medium.

The absorption of incident intensity by the steady-state components of the blood (tissues, plasma, etc.) form the DC component and the absorption due to the pulsatile component of the blood corresponds to the AC component of the PPG signal [16]. The four LED wavelengths (590 nm, 660 nm, 810 nm and 940 nm) were chosen to maximize the absorption of the four Hb components (Oxygenated Hb, Deoxygenated/Reduced Hb, Carboxy Hb, and Methemoglobin) based on their absorption spectra [14]. A set of features representing the amount of light absorbed (Attenuation features) and the pairwise ratio-of-ratios parameters to capture the influence of the blood hemoglobin concentration on the attenuated signal were chosen for the development of this classification algorithm.

\[ R_i = \frac{r_i^{\text{peak amplitude}}}{r_i^{\text{trough amplitude}}} \]

where \( i \) is the index of the peak or trough, \( n \) is the total number of peaks/troughs detected in one PPG signal at wavelength \( \lambda \).

### Table 1 Demographic details of data collection

| n = 1583 | Age (years) | Weight (kg) | Hb (g/dL) |
|---------|------------|-------------|-----------|
| Missing data | 1 | 539 | 0 |
| Range | 17–52 | 24–102 | 1.6–14.8 |
| Mean | 26.12 | 50.06 | 9.51 |
| Standard deviation | 6.08 | 9.46 | 2.3 |
The calculated $R_i$ values were plotted against actual hemoglobin values for all subjects at all wavelengths. The set of $R_i$ values showing higher correlation with actual hemoglobin values (used for training the machine learning algorithm, MLA) on a percentile scale were subsequently utilized for building the feature set. As a representative value of $R_i$, $p$th percentiles of $R_i$ where $p \in \{70, 80, 90, 100\}$ were chosen for feature derivation for training the MLA.

Pairwise ratio of ratios for the four wavelengths at four percentiles were calculated according to Eq. 4.

$$R_{\lambda_1/\lambda_2}|_p = \left|\frac{R_{\lambda_1}}{R_{\lambda_2}}\right|_p = \frac{\ln([r_1, r_2, \ldots, r_n]_{\lambda_1})}{\ln([r_1, r_2, \ldots, r_n]_{\lambda_2})}_p \quad (4)$$

where $R_{\lambda_1/\lambda_2}$ is the ratio of ratios between any two wavelengths where $\lambda_1$ and $\lambda_2 \in \{590, 660, 810, 940\}$ calculated by considering $p$th percentile values of $R_{\lambda_1}$ and $R_{\lambda_2}$ where $p \in \{70, 80, 90, 100\}$. The pairwise combinations used are shown in Table 2.

### Table 2: Set of features chosen to train the MLA

| Ratio of ratios                          | MLA              |
|----------------------------------------|------------------|
| R660/590_80                            | A990             |
| R660/590_90                            | A660             |
| R660/810_80                            | A810             |
| R660/810_90                            | A940             |
| R810/590_70                            |                  |
| R940/590_90                            |                  |
| R940/660_100                           |                  |

### 2.3.2.2 Attenuation Values
A value defined as the logarithm of the difference between the peak amplitude and the subsequent trough amplitude were also considered as one of the features to train MLA (Eq. 5).

$$A = \text{median} [\log (I_n)] \quad (5)$$

$$I_n = \text{peak}_i – \text{trough}_i \quad (6)$$

where $A$ is attenuation value, $I_n$ is the set of all peak to peak amplitude in the PPG signal, $i$ represents the index of the peak or trough.

For each of the four different LEDs, four distinct values ($A_{590}, A_{660}, A_{810}, A_{940}$) were created for each subject and used as features to train the MLA.

The final feature set used to train the MLA consisted of nine ratios of ratios and four attenuation values, listed in Table 2.

### 2.3.3 Machine Learning Algorithm (MLA)

#### 2.3.3.1 Classification Structure
A hierarchical classification scheme, with two classifiers, was adopted to develop the proposed MLA to classify subjects as healthy, mild to moderately anemic or severely anemic (Table 3).

Since the combined performance of a committee of models is expected to outperform a single model [19], both classifiers: Stage-1 and Stage-2 were modelled using an ensemble of eleven MLAs comprising of four types of classification models each. Eight Support Vector Machine (SVM) models with different parameter tunings, one Logistic regression model, one Random forest model and one Boosted trees model were used. Each model within an ensemble was a binary classifier. The Stage-1 models classified a subject into healthy or anemic category, while The Stage-2 models classified the anemic subjects into moderately or severely anemic category. The final classification, described in Fig. 1, of a subject to a category at each stage was determined by the majority of votes across the models in that ensemble.

#### 2.3.3.2 Model Training and Testing
The data was randomly split into three mutually exclusive and exhaustive partitions of 70% for training, 15% each for validation testing. For the Stage-1 classifier, all subjects from the training set were considered for training the ensemble models, while for the Stage-2 classifier, subjects from the training set with true Hb values less than 11 g/dL were considered as depicted in Fig. 2. In order to train the models independently, each model within the ensemble was trained using randomly chosen subset of the training data.

To estimate a distribution of the model performance, the procedure was repeated for multiple iterations, and test sets generated at every iteration were kept aside for evaluation of the model performance.

Each model generated a probability estimate, indicating the chance of a subject belonging to one of the two categories (at each stage). Based on this value being above or below a threshold, the subject was classified into either category.

For a model with probability estimate $p$ and threshold $x$, an individual was categorized as per the following rule:

$$\text{Category of the individual} = \begin{cases} \text{Category 1 if } p \geq x \\ \text{Category 2 if } p < x \end{cases} \quad (7)$$

To optimize the performance of each classifier, an optimal value of threshold was derived from a range

---

Table 3: Anemia classification limits

| Hb value intervals (g/dL) | Anemia status          |
|---------------------------|------------------------|
| <7                        | Severely anemic        |
| 7–11                      | Mild to moderately anemic |
| ≥11                       | Healthy (non-anemic)   |

---

For a model with probability estimate $p$ and threshold $x$, an individual was categorized as per the following rule:

$$\text{Category of the individual} = \begin{cases} \text{Category 1 if } p \geq x \\ \text{Category 2 if } p < x \end{cases} \quad (7)$$

To optimize the performance of each classifier, an optimal value of threshold was derived from a range
of threshold values using the validation set. The set of threshold values considered for this purpose consisted of 100 equally spaced values from 0 to 1. The predictions of classifiers at these thresholds for the validation set were calculated and an array of sensitivity and specificity values were generated. This process was repeated 30 times using random splits resulting in a distribution of sensitivity and specificity values at each threshold for every model. A receiver operating characteristic (ROC) curve was generated from the median values of the distribution of sensitivity and specificity values. ROC curves for the eleven models arranged row-wise (Eight Support Vector Machine (SVM) models with different parameter tunings, one Logistic regression model, one Random forest model and one Boosted trees model) in Stage-1 and Stage-2 classifier ensembles are shown in Fig. 3. The threshold value corresponding to the best trade-off between sensitivity and specificity was then finalized for a model by calculating the Youden index \( y_x \) [20] for the threshold values from the ROC. The final threshold \( x_{opt} \) for a model was chosen according to the equation:

\[
x_{opt} = \max \{ y_x = \text{sensitivity}_x -(1 - \text{specificity})_x \}
\]  

(8)

Once the two classifiers were optimized and successfully modelled, the proposed MLA for anemia classification was assembled and the tested.

3 Results

3.1 Relationship Between Ratio of Ratios and True Hb

3.1.1 Feature Extraction

A plot of the \( R \) values against actual hemoglobin values for all subjects (in the training set) at all wavelengths depicted better correlation at higher values of \( R \), as shown in Fig. 4. (The ratios for \( R_{590} \) differed from the other wavelengths because the gain settings for the 590 nm LED was kept higher than the other three wavelengths to compensate for the higher absorption of light by the finger at lower wavelengths which resulted in a very low amplitude signal.) The higher percentile values (namely 70th, 80th, 90th and 100th) of \( R \) were chosen for feature development. Using these \( R \) as input, a further set of pairwise ratio of ratios for the four wavelengths at four percentiles were calculated as features. The top two pairwise ratios with maximum correlation with true Hb values in each combination of wavelength were then selected as input features to the MLA (see Table 4).
3.1.2 Model Selection

To select the optimal number of models constituting the ensemble, we observed the effect of number of models on the ensemble performance. The sensitivity and specificity values did not change significantly for Stage-1 Classifier (Fig. 5A, C) but indicated improvement in sensitivity values with increasing number of models for Stage-2.
Fig. 4 Distribution of $R$ values versus actual hemoglobin values for subjects belonging to the training set at A 660 nm LED, B 590 nm LED, C 940 nm LED, and D 810 nm LED

Table 4 Correlation between ratio of ratios and actual Hb values at four percentiles

| Ratio of ratios | Correlation | Ratio of ratios | Correlation |
|----------------|-------------|----------------|-------------|
| $R_{810/590}_{70}$ | 0.51 | $R_{940/660}_{70}$ | 0.092 |
| $R_{810/590}_{80}$ | 0.51 | $R_{940/660}_{80}$ | 0.096 |
| $R_{810/590}_{90}$ | 0.504 | $R_{940/660}_{90}$ | 0.105 |
| $R_{810/590}_{100}$ | 0.461 | $R_{940/660}_{100}$ | 0.114 |
| $R_{940/590}_{70}$ | 0.459 | $R_{940/810}_{70}$ | -0.158 |
| $R_{940/590}_{80}$ | 0.466 | $R_{940/810}_{80}$ | -0.152 |
| $R_{940/590}_{90}$ | 0.465 | $R_{940/810}_{90}$ | -0.16 |
| $R_{940/590}_{100}$ | 0.431 | $R_{940/810}_{100}$ | -0.164 |
| $R_{660/590}_{70}$ | 0.33 | $R_{660/810}_{70}$ | -0.405 |
| $R_{660/590}_{80}$ | 0.344 | $R_{660/810}_{80}$ | -0.41 |
| $R_{660/590}_{90}$ | 0.339 | $R_{660/810}_{90}$ | -0.416 |

*These ratios were selected as features to be input into the MLA

3.1.3 Model Performance

The trained models within the ensembles were evaluated in terms of their sensitivity and specificity values at each stage [21, 22]. The models reported a median sensitivity of 0.92 (p < 0.05) for Stage-1 classifier and a median specificity of 0.84 (p < 0.05) for Stage-2 classifier. The median value of sensitivity and specificity were noted as 0.76 (p < 0.05) and 0.74 (p < 0.05) respectively. Stage-1 sensitivity reflects the ability of the trained classifier to detect the presence of anemia and specificity reflects the ability to detect its absence. Stage-2 sensitivity reflects the ability of the trained classifier to detect the presence of severe anemia and specificity reflects the ability to detect the absence of severe anemia. Stage-1 metrics indicate that the trained classifier was highly sensitive and precise in detecting anemia cases (> 11 g/dL) whereas Stage-2 classifier, although significantly sensitive, lacks precision in detecting severely anemic cases. The Positive Predictive Value (PPV) and Negative Predictive Value (NPV) indicate the tendency of the trained classifier to exhibit bias towards a particular class thereby inflating the sensitivity or recall value of a particular class.
The median PPV values for Stage-1 and Stage-2 classifiers were 0.92 and 0.43, whereas the median NPV values were 0.8 and 0.9 respectively. The low PPV value for Stage-2 classifier could be attributed to the imbalance in distribution of hemoglobin values across the dataset. The performance of the MLA, as calculated across multiple test iterations, is illustrated in Fig. 6.

Table 5 summarizes the sensitivity and specificity values of the individual models of the ensemble along with the performance of the voting scheme. The voting classifier
performed better than the individual models in Stage-2 and a moderate performance improvement is observed in Stage-1.

### 3.1.4 Confusion Matrix

The confusion matrix, shown in Fig. 7a, reflects the accuracy of classification (healthy, mild to moderately anemic and anemic). The matrix was calculated taking into consideration the average performance of the MLA across all test sets. While the rows of the confusion matrix represent the actual class of anemia, the columns represent the class assigned by the MLA to the subjects.

As evidenced by the results, the risk of extreme misclassification (i.e., an error of more than one class shift) was minimal. None of the subjects in the maximum risk category i.e., severe anemia were misclassified as healthy and among all the healthy subjects across all test sets around 2% were wrongly classified as severely anemic.

#### 3.1.5 Four Way Anemia Classification

An attempt was also made to develop a 4-way anemia classifier to better differentiate the ‘mild to moderate anemia’ category into ‘mild anemia’ and ‘moderate anemia’. The middle bucket of Anemic individuals (7–11 g/dL) in the three way classifier was subject to bifurcation into two sub-classes, as ‘moderate anemia’ (Hb 7–9 g/dL) and mild anemia (Hb 9–11 g/dL).

From the training sample, a regression model was developed for subjects with Hb values between in the 7–11 g/dL. The performance of the trained regression model was tested on the data set of observations that were previously assigned to the anemic category of anemia by the three way classifier. The performance of the newly trained four way classifier was assessed in similar manner as discussed in the previous sections, and a four way confusion matrix was computed by averaging the performance on all the test sets (Fig. 7b). The bucket for Hb level (9–11 g/dL) reported a
p-value (0.01). The p-value confirms the hypothesis that the developed classifier indeed works better than a random guess. For the bucket (7–9 g/dL) the observed sensitivity is around 0.38 and the p-value reported is 0.01.

4 Discussion

In this study we demonstrate the capability of a non-invasive opto-electronic device coupled with machine learning techniques for the non-invasive classification of Anemia in community health settings.

Anemia is an easily treatable but rampant condition across emerging economies. Technology utilized in the field today includes venous blood-draw based laboratory techniques, finger prick based point of care methods, and sophisticated non-invasive devices. Each have their challenges related to accessibility and/or accuracy in field settings.

By using a rapid point of care, handheld, battery operated automated Anemia classification device, doctors, nurses and/or community health workers can easily diagnose and treat a patient in accordance with the WHO therapeutic guidelines.

The non-invasive nature of the device facilitates increased acceptance by patients and reduces the problems of poor follow up, safety, biohazard disposal and supply chain requirements. Photoplethysmographs from an individual’s finger provided enough information to a machine learning algorithm to reliably predict Hb ranges. The study described in this paper is a step in this direction.

Features derived from the PPG signals, proposed in this paper, were sufficient to develop a supervised machine learning algorithm with a three way classification of anemia-severity with accuracies that are comparable to current community based (invasive) anemia screening tools. The algorithm developed in this paper is unique in terms of the data volume and heterogeneity. It has been trained on a significantly large cohort size (n = 1516) spread over a wide range of hemoglobin values from 1.6 to 14.8 g/dL. Most studies in the past have been implemented on smaller and less diverse data sets. Kavsaoglu et al. [16] proposed machine learning based approach for estimating hemoglobin non-invasively using data from 33 subjects with a considerably narrow range of hemoglobin values: 10.1–17.4 g/dL. Ding et al. [23] proposed an MLA for non-invasive hemoglobin monitoring based on principal components analysis and artificial neural networks which included 109 subjects with hemoglobin values ranging from 11 to 18 g/dL.

The area under the ROC curve of the individual models quantify the overall ability of the model to discriminate between healthy and anemic individuals by Stage-1 classifier, and moderate and severe anemia by the Stage-2 classifier. The proposed machine learning based classifier could differentiate between a healthy and anemic individual with a sensitivity of 92%, and a moderate to severely anemic individual with a sensitivity of 76%. Another study [18] proposing a smart phone based non-invasive monitoring set up, reports sensitivity and specificity values for a two way classification (presence or absence of anemia) as 85.7 and 70.6 respectively calculated from a set of 31 individuals.

Although the reported median PPV for severely anemic cases is low (43%), we believe that it is significantly influenced by the class prevalence [24], which was around 14% in the data set. The performance of the proposed model may be further enhanced by including more data from the higher end (> 14 g/dL) and lower end (< 7 g/dL) of the Hb range.

Kavsaoglu et al. [16] has proposed machine learning based approach for estimating hemoglobin non-invasively from features extracted from PPG signals on 33 subjects but the range of hemoglobin values in the data set was considerably narrow (10.1–17.4 g/dL). Another study [18] proposing a smart phone based non-invasive monitoring set up, reports sensitivity and specificity values for a two way classification (presence or absence of anemia) as 85.7 and 70.6 respectively calculated from a set of 31 individuals. Ding et al. [23] proposed a MLA for non-invasive hemoglobin monitoring based on principal components analysis and artificial neural networks, which included 109 subjects with hemoglobin values ranging from 11 to 18 g/dL.

For the data used in this study, the performance of the four way classification was observed to be low for the mild and moderate anemia classes. Since the clinical handling of the intermediate buckets is not significantly different [25], we propose that, a three way classification scheme might provide a more suitable and sufficient method for community health interventions.

We believe that the results from the trained machine-learning algorithm presented in this study show significant potential to be implemented into a non-invasive classification device useful for rapid and widespread hemoglobin screening for diagnosis of anemia. A rapid point-of-care, automated anemia classification device would eliminate problems related to subjectivity in interpretation due to ambient lighting conditions as posed on the HCS. It could enable doctors, nurses and CHWs to prescribe on-spot treatment plans and aid in rapid clinical decision making thereby strengthening healthcare facilities. Non-invasive nature of such a device would further facilitate increased acceptance by patients, reduce the problems of poor follow up, safety, biohazard disposal, and supply chain requirements.

The data in the current study was focused on the female population. The results of this study presume that the PPG signal acquired from the device architecture is of good quality, finger motion artefacts are removable and the finger nails are not polished/cracked. Further improvisations to the study could be aimed at evaluating the effects of other confounding factors like age and gender. Also, the variations in light
absorption characteristics due to differences in skin pigmentation and the impact of pregnancy status should be investigated in order to make the machine learning algorithms robust to the regional and ethnicity variations. Other applications in a similar direction have explored the possibility of evaluating critical health conditions non-invasively, like non-invasive estimation of blood glucose and bilirubin.

Acknowledgements We would like to acknowledge the valuable inputs and efforts of Dr. Sanjay Pattanshetty, Dr. Navya Vyas and Dr. Shilpa Drakshi from Prasanna School of Public Health, MAHE for facilitating the study. We would also like to show our gratitude towards Dr. Keerthi Ballal, Chief Medical Officer from TMA Pai Rotary Hospital, Karkala; and Health officials and staff of Gangavathi Taluk Hospital and Koppal District Hospital for their assistance in data collection.

References
1. World Health Organization. (2011). Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity. In: Vitamin and Mineral Nutrition Information System (VMNIS). Retrieved from https://www.who.int/vmnis/indicators/haemoglobin/en
2. World Health Organization. (2008). WHO Global Database on Anaemia. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/43894/9789241596657_eng.pdf
3. Barker, S. J., Shander, A., & Ramsay, M. A. (2016). Continuous noninvasive hemoglobin monitoring: A measured response to a critical review. Anesthesia & Analgesia, 122(2), 565–572. https://doi.org/10.1213/ane.0000000000000605
4. Nirup, J. L. A., & Kumar, V. J. (2014). Non-invasive measurement of hemoglobin content in blood. In: Medical Measurements and Applications (MeMeA). IEEE International Symposium. https://doi.org/10.1109/MeMeA.2014.6860140
5. Marn, H., & Alison Critchley, J. (2016). Accuracy of the WHO Haemoglobin Colour Scale for the diagnosis of anaemia in primary health care settings in low-income countries a systematic review and meta-analysis. The Lancet Global Health, 4(4), 251–265. https://doi.org/10.1016/s2214-109x(16)00005-x
6. Darshana, L. G. T., & Uluwaduge, D. I. (2014). Validation of the WHO hemoglobin color scale method. Anemia, 2014(1), 1–4. https://doi.org/10.1155/2014/531670
7. Srivastava, T., Negandhi, H., Neogi, S. B., Sharma, J., & Saxena, R. (2014). Methods for hemoglobin estimation: A review of ‘What Works.’ Journal Hematofus transfusion, 2(3), 1028.
8. Lindner, G., & Exadaktylos, A. K. (2013). How noninvasive hemoglobin measurement with pulse CO-oximetry can change your practice: An expert review. Emergency Medicine International, 2013, 1–4. https://doi.org/10.1155/2013/701529
9. Akhtar, K., Sherwani, R. K., Rahman, K., Hasan, J., & Shahid, M. (2008). HemoCue photometer: A better alternative of hemoglobin estimation in blood donors? Indian Journal Hematology Blood Transfusion, 24(4), 170–172. https://doi.org/10.1007/s12288-008-0041-8
10. Brown, J., Theis, L., Kerr, L., Zakhidova, N., O’Connor, K.,uthman, M., Oden, Z. M., & Richards-Kortum, R. (2011). A hand-powered, portable, low-cost centrifuge for diagnosing anemia in low-resource settings. The American Journal of Tropical Medicine Hygiene, 85(2), 327–332. https://doi.org/10.4269/ajtmh.2011.10-0399
11. Bond, M. M., & Richards-Kortum, R. R. (2015). Drop-to-drop variation in the cellular components of fingerprick blood implications for point-of-care diagnostic development. American Journal of Clinical Pathology, 144(6), 885–894. https://doi.org/10.1309/ajcip117dkmppchp
12. Gayat, E., Aulagnier, J., Matthieu, E., Boisson, M., & Fischler, M. (2012). Non-invasive measurement of hemoglobin: Assessment of two different point-of-care technologies. PLoS ONE, 7(1), e30065. https://doi.org/10.1371/journal.pone.0030065
13. Morey, T. E., Gravenstein, N., & Rice, M. J. (2011). Let’s think clinically instead of mathematically about device accuracy. Anesthesia & Analgesia, 113(1), 89–91. https://doi.org/10.1213/ane.0b013e318219e290
14. Acharya, S., Swaminathan, D., Das, S., Kansara, K., Chakraborty, S., Kumar, R. D., Francis, T., & Aatre, K. R. (2019). Non-invasive estimation of hemoglobin using a multi-model stacking regressor. IEEE Journal Biomedical Health Informatics, 24(6), 1717–1726. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6876720/
15. Kavsaoglu, A. R., Polat, K., & Harirhan, M. (2015). Non-invasive prediction of hemoglobin level using machine learning techniques with the PPG signal’s characteristics features. Applied Soft Computing, 37, 983–991. https://doi.org/10.1016/j.asoc.2015.04.008
16. Dinesh Kumar, R., Guruprasad, S., Kritty Kansara, K. N., Rao, R., Mohan, M., Reddy, M. R., Prabhu, U. H., Prakash, P., Chakraborty, S., Sreeatam Das, K. N., & Madhusoodanan, S. (2021). A novel noninvasive hemoglobin sensing device for anemia screening. IEEE Sensors Journal, 21(13), 15318–15329.
17. Wang, E. J., Li, W., Hawkins, D., Gernsheimer, T., Norby-Slycord, C., & Patel, S. N. (2012). HemaApp: Noninvasive blood screening of hemoglobin using smartphone cameras. UbiComp’16. https://doi.org/10.1145/2971648.2971653
18. Zhou, Z. H. (2012). Ensemble methods—Foundations and algorithms. Chapman & Hall/CRC.
19. Ruopp, M. D., Perkins, N. J., Whitcomb, B. W., & Schisterman, E. F. (2008). Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. Biometrical Journal, 50(3), 419–430. https://doi.org/10.1002/bimj.200710415
20. Goh, Y. C., Cai, X. Q., Theseira, W., Ko, G., & Khor, K. A. (2020). Evaluating human versus machine learning performance in classifying research abstracts. Scientometrics, 125, 1197–1212.
21. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of Ensemble methods—Foundations and algorithms. Journal of Machine Learning Research, 50, 119–145. https://doi.org/10.1016/j.jmlr.2009.03.002
22. Ding, H., Lu, Q., Gao, H., & Peng, Z. (2014). Non-invasive prediction of hemoglobin levels by principal component and back propagation artificial neural network. Biomedical Optics Express, 5(4), 1145. https://doi.org/10.1364/BOE.5.001145
23. Guyatt, G. H., Tugwell, P. X., Feeny, D. H., Brian Haynes, R., & Drummond, M. (1986). A framework for clinical evaluation of diagnostic technologies. Canadian Medical Association Journal, 134, 587–594.
24. Tandon, R., Jain, A., & Malhotra, P. (2018). Management of iron deficiency anaemia in pregnancy in India. Indian Journal of Hematology Blood Transfusion, 34(2), 204–215.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.