IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 July 2023, accepted 25 July 2023, date of publication 31 July 2023, date of current version 3 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3300293

APPLIED RESEARCH

Noninvasive Hemoglobin Measurements With Photoplethysmography in Wrist

VLADISLAV V. LYCHAGOV[1], VLADIMIR M. SEMENOV[1], ELENA K. VOLKOVA[1], DMITRII I. CHERNAKOV[1], JOONGWOO AHN[2], AND JUSTIN YOUNGHYUN KIM[3]

[1]Samsung Research and Development Institute Russia, 127018 Moscow, Russia
[2]Samsung Electronics, Suwon 16677, South Korea
[3]Mobile eXperience Business Division, Samsung Electronics, Suwon 16677, South Korea

Corresponding author: Vladislav V. Lychagov (lychagov.vladislav@gmail.com)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Commission of Saratov State Medical University, Russia, under Protocol No. 12 of June 30, 2022.

ABSTRACT This paper describes the application of multiwavelength photoplethysmography (MW-PPG) in reflectance mode for noninvasive measurements of total hemoglobin concentration. Consumer wrist-wearable devices (smartwatches, wristbands) typically contain a photoplethysmography sensor operating in reflectance mode, as opposed to clinical pulse-oximetry sensors operating in transmittance mode. We assume that the generally accepted approach to the analysis of the transmittance-mode MW-PPG signal, based on the direct implementation of the Beer-Lambert law, is not applicable to the analysis of reflectance-mode MW-PPG signals. We propose that the shape of the MW-PPG signal carries information regarding the distribution of optical paths at different wavelengths within the tissue, which is crucial for the correct estimation of the absorption of light in the probe volume. We have developed and tested several machine-learning algorithms to analyze the reflectance-mode MW-PPG signal and to estimate the total hemoglobin concentration based on this analysis. To train and validate the algorithms, we collected a dataset of 840 MW-PPG signals measured from 170 volunteers by using a wrist-wearable PPG sensor. Three reference devices were used to label the data and test the performance of the developed algorithms: an invasive laboratory blood test, minimally invasive HemoCue Hb 201+, and noninvasive Masimo Radical-7 transmittance mode pulse-oximeter. The best performance achieved with the developed algorithm, mean absolute error MAE $\approx$ 12.6$\pm$1.7 g/L and correlation coefficient R $\approx$ 0.66$\pm$0.09, is comparable with the performance of the clinical noninvasive device.

INDEX TERMS Machine learning, hemoglobin, noninvasive, photoplethysmography, smartwatch.

I. INTRODUCTION

Wearable devices are capable of measuring several valuable physiological parameters: including heart rate (HR), HR variability (HRV), oxygen saturation ratio (SpO_2), and bio-impedance (BIA) [1], [2]. These markers are important for managing stress levels, amount of physical activity, sleep quality, and body mass index, [3], [4], [5]. Despite the importance of all the mentioned parameters, information is being provided to users with little or no recommendations on possible improvements in the health state or mitigation of existing problems. This is mostly because of the insufficiency and uncertainty of the parameters currently available in consumer devices. For example, oxygen saturation indicates the ability of the blood to carry oxygen bound to hemoglobin [6]. However, this relative parameter still provides no information about the actual amount of oxygen in the blood stream because of a lack of knowledge of hemoglobin concentration and hematocrit. Meanwhile, the absolute amount of oxygen determines the ability of the body to release accumulated energy and deliver it to organs. The efficiency of the energy supply mechanism governs cognitive and muscle functions, to a large extent, and understanding the energy budget plays a crucial role in planning any daily life or fitness activity.

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

In this paper, we propose a method for noninvasive measurements of hemoglobin concentration using wearable multiwavelength photoplethysmography (PPG) sensor and an advanced algorithm for the time-spectral analysis of PPG data. Wrist wearable sensors working in reflectance mode impose additional requirements on signal analysis because of the more complicated processes of light propagation and light-tissue interaction compared to the approach that is widely accepted in conventional transmittance pulse-oximetry [7], [8], [9].

II. PPG SIGNAL IN REFLECTANCE MODE

The commonly accepted approach for describing the operation principle of a photoplethysmography sensor is based on the Beer-Lambert law. The derivation of the Beer-Lambert law relies on counting the incremental changes in the intensity of light while it passes sequentially through a number of infinitely thin layers composing a sample of absorbing substance, characterized by the total thickness and attenuation coefficient. By measuring the ratio of the intensities of the incident light to the light passed through the sample, one can calculate the absorbance of the sample and, if the thickness is known, the attenuation coefficient of the sample. It is assumed that different portions of the light travel approximately the same path inside the sample. This does not hold, in general, because of the scattering of light in turbid media such as biological tissue.

The processing of a photoplethysmography signal usually involves decomposing the signal into two parts: a fluctuating AC component and a slowly varying DC component, both of which comply with Beer-Lambert law. The AC component corresponds to the pulsatile part of the blood volume and the DC component corresponds to the constant part of the blood volume and other tissues. Dividing the AC component by the DC component eliminates the optical path length term in the Beer-Lambert law, thus yielding the absorbance of the pulsatile blood regardless of the sample geometry [10]. By measuring the absorbance of pulsatile blood at multiple wavelengths, one can recover the spectral properties of the blood and calculate some parameters, such as the oxygen saturation ratio.

Skin tissue consists of multiple layers with different optical properties and is occupied by different types of blood vessels. The depth arrangement of different blood vessels inside the tissue is shown schematically in Fig. 1. It is often stated that changes in blood volume occur due to the pulsatile motion of blood in the arteries and, correspondingly, changes in the amplitude of the AC component of the photoplethysmography signal are due to changes in the absorbance of pulsatile arterial blood, that is, the AC component of the photoplethysmography signal corresponds to blood solely. Few researchers have suggested that the pulsatile motion of arterial blood can induce mechanical stress in the surrounding and upper layers of tissue, including but not limited to blood vessels (capillaries) [11], [12]. Pulsations in mechanical and, consequently, optical density of tissues cause changes in

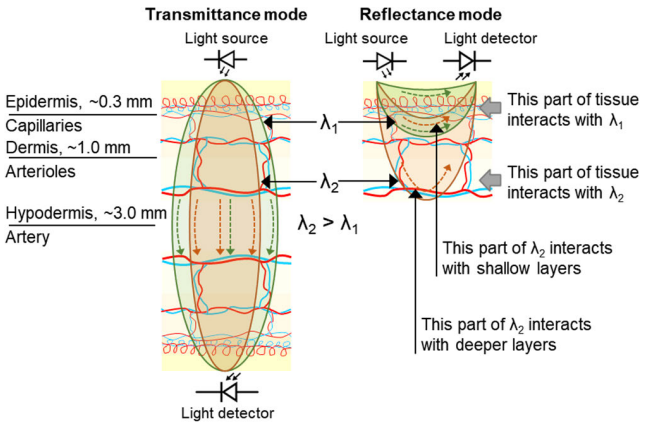

FIGURE 1. Difference between photoplethysmography measurements in reflectance and transmittance modes.

absorbance, synchronously with the heart rate. This means that the fluctuation of the AC component is no longer proportional to the pulsatile arterial blood, but includes a signal from other tissues as well. Furthermore, despite being triggered by the propagation of a pulse wave in the arteries, deformations in the other layers may have different intensities and time delays. The superposition of individual signals from different layers of tissue can produce a total plethysmography signal of different shapes. The shape of the total pulse wave depends on the ratio of signals originating from the shallow to deep layers of the tissue [11], [12].

In transmittance mode, the mutual alignment of the light source and the light detector on the opposite sites of a sampling volume ensures the the light (of different wavelengths) travels approximately the same path through all layers of tissue and all types of blood vessels (Fig. 1). The directivity of the light source, receiving aperture of the detector, and scattering of light largely determine the actual distribution of the optical paths (the so-called ''banana-shape'') in the probe volume. In Fig. 1, for example, the optical paths for λ_1 and λ_2 are slightly different. In practice, the difference in optical paths is often omitted and this approximation still provides correct results in most cases of processing photoplethysmography signals in the transmittance mode. From Fig. 1, we can conclude that the relative contributions of different layers of tissue and blood vessels to the AC and DC components of plethysmography signals at wavelengths λ_1 and λ_2 are similar. At the same time, the contribution of pulsatile arterial blood to the AC component prevails over all other pulsating components of the tissue. Therefore, the difference in the pulse wave shape caused by the superposition of individual signals from shallow and deep layers of tissue can also be ignored because of the fixed probe volume and significantly larger absorbance in the arteries.

In reflectance mode, the impact of the abovementioned effects is crucial and unavoidable. First, light of different wavelengths penetrates at significantly different depths, and hence interacts with different types of vessels and tissues.

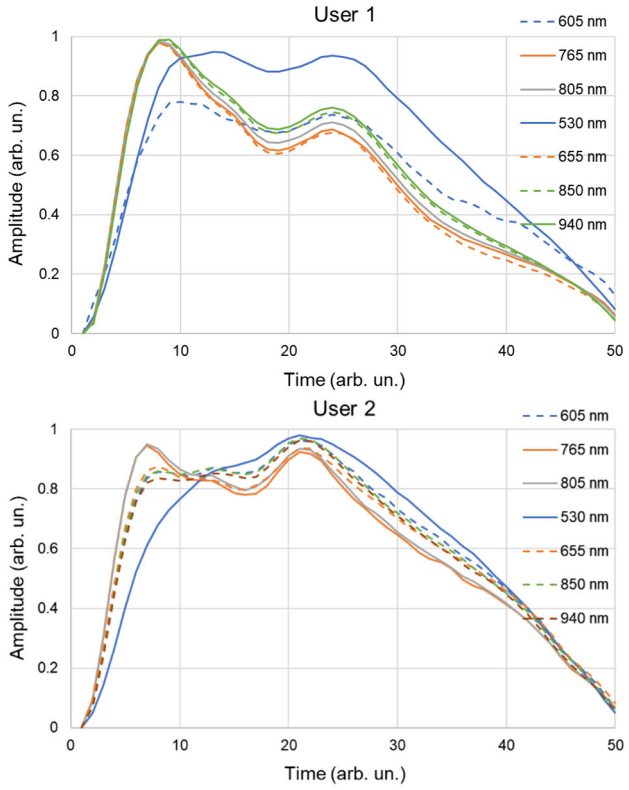

FIGURE 2. Dependence of the pulse wave shape on the wavelength of light and on the user in reflectance mode PPG.

A shorter wavelength λ_1 interacts with the upper layers of tissue and small vessels and capillaries, with little or no interaction with large vessels, arterioles and arteries. A longer wavelength λ_2 interacts mostly with large vessels, arteries, and arterioles, and to a lesser extent with capillaries. The superposition of individual signals from these vessels results in considerably different shapes of the total plethysmography signal. Fig. 2 shows the plethysmography signals recorded in reflectance mode with light of different wavelengths. It should be noted that the dependence of the pulse wave shape on the wavelength is also user specific because of the differences in the tissue properties of the two users. Furthermore, light of any particular wavelength propagates through the probe volume, which is distributed non-uniformly over several layers of tissue. Some parts of the light travel a shorter optical path, penetrate slightly into the tissue, and interact with smaller vessels, while some parts of the light travel a longer optical path, penetrate deeper, and interact with larger vessels. Decomposition of the total plethysmography signal, which is a superposition of all these parts, into AC and DC components is no longer valid. Hence, direct calculation of the absorbance of the pulsatile blood in the reflectance-mode PPG through the AC/DC ratio is not correct. We must characterize the exact shape of the pulse wave at each wavelength. A set of features describing the shape of the pulse wave at multiple wavelengths can be associated with the absorption properties of the pulsatile blood. However, this empirical statement may not have a straightforward analytical solution. Therefore, we employed a numerical approach based on supervised machine learning.

III. MATERIALS AND METHODS

A. REFERENCE DEVICES

Labeling the data before fitting the algorithm requires the use of a reference device that provides reliable and accurate measurements of the target parameter. However, the reference device should be easy to use and allow high-throughput screening of the subjects under study during the collection of a large-scale dataset. Laboratory venous or arterial blood test provides the most accurate estimation of hemoglobin concentration. However, such an analysis is not suitable for frequent and repetitive in-situ measurements. To simplify the data collection protocol, we used portable device for point-of-care testing HemoCue Hb 201+. According to the information available in the user manual, the operating principle of HemoCue Hb 201+ is based on the azide methhemoglobin method, similar to that often used in laboratory blood tests. A rigorous analysis of the accuracy and precision of HemoCue Hb 201+ has been reported recently [13].

Possibility of noninvasive measurements of hemoglobin is available in several commercial devices manufactured by different companies, including the most reputable ones: Masimo and Cercacor. Masimo and Cercacor offer various solutions for both professional and consumer markets. A distinctive feature of these devices is that they operate in transmittance mode only, even those that are available in a wearable design intended for personal use. According to the available information [14], Masimo products operate in the short-wavelength near-infrared (SWIR) spectral range, as opposed to the VIS-NIR range commonly used in pulse oximetry. Masimo claimed an accuracy of ± 10 g/L within the working range of 80-170 g/L, which was largely confirmed by independent studies [13]. However, it should be noted that this accuracy was calculated within ± 1 standard deviation ($\sim$68% of data points) and more than 30% of the data points were outside the ± 10 g/L error range. Since Masimo provides state-of-the-art performance for noninvasive measurements of hemoglobin concentration in transmittance mode, it is of particular interest to compare measurements in reflectance mode with this baseline.

B. REFLECTANCE MODE PPG PROTOTYPE

The MW-PPG signals were recorded using a custom-designed wrist wearable prototype. The prototype consists of eight light-emitting diodes arranged in two custom SMD modules and two SFH 2706 photodiodes. The LED modules and photodiodes were provided by OSRAM. The mutual arrangement of LEDs and PDs in the prototype is shown in Fig. 3. To prevent crosstalk caused by specular reflections, a 3 mm high light-tight shield separates the LEDs and PDs in the 3D-printed housing of the prototype. The emission spectra of the LEDs are shown in Fig. 4. We did not use any

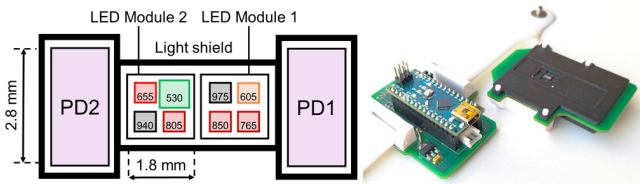

FIGURE 3. Schematic and overall view of the wrist wearable reflectance mode PPG sensor prototype.

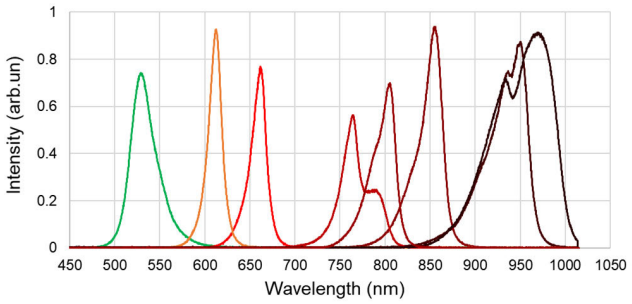

FIGURE 4. Emission spectra of light emitting diodes in the PPG sensor.

additional diffuser or focusing lens for beam shaping. The emission half-angle of all LEDs was approximately 60°.

An analog front-end AFE4500 from Texas Instruments was used for driving of LEDs, signal acquisition, amplification, and sampling. The LEDs were pumped with a 125 mA pulse with a duration of 54.7 μs, during which the signals from both PDs were simultaneously sampled. The gain of the differential transimpedance amplifier was set to 5k for the nearest PD and to 10k for the furthest PD. The MW-PPG measurements consisted of sequential recording of signals of eight wavelengths and recording of the ambient phase (LEDs are off) with a repetition rate of 40 Hz. Eighteen signals (8 LED $\times$ 2 PD + 2 PD ambient phases) in total, each approximately 5 min long, were recorded in a single measurement.

C. DATASET

The clinical dataset consisted of 840 MW-PPG signal samples measured in 170 volunteers (81 males and 89 females). The data collection protocol was reviewed and approved by the Commission of Saratov State Medical University, Russia (protocol no. 12 of June 30, 2022). Before the trials, all the subjects received a detailed explanation of the clinical tests and signed an informed consent form. The data were anonymized and used for the intended research purposes only.

The reference data consisted of hemoglobin values measured invasively with the HemoCue Hb 201+ analyzer and noninvasively with the Masimo Radical-7 pulse-oximeter. Additionally, each subject underwent a laboratory blood test once before the clinical trials. The results of the laboratory blood test were intended for the prescreening of volunteers and were not used for data labeling during the algorithm development. Fig. 5 shows the distribution of volunteers in the dataset and the consistency between the HemoCue

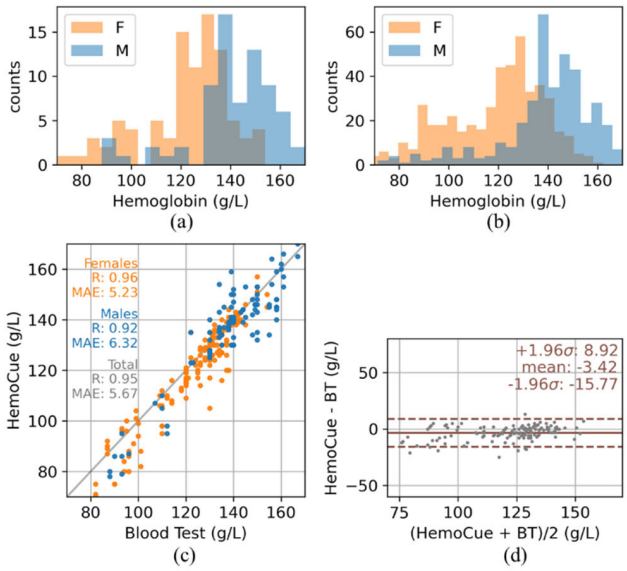

FIGURE 5. Contents and quality of the dataset: (a) distribution of volunteers according to blood test results; (b) distribution of volunteers according to HemoCue measurements; (c) and (d) agreement between blood test results and HemoCue measurements.

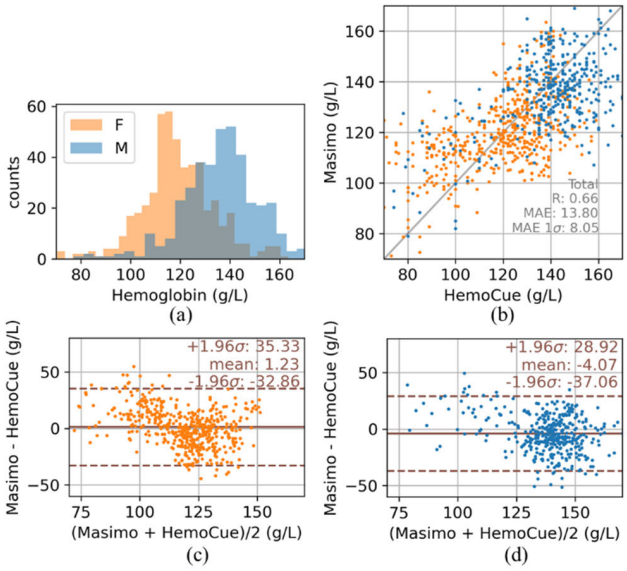

FIGURE 6. Performance of Masimo measurements for users in the dataset: (a) distribution of users according to Masimo measurements; (b) correlation between Masimo measurements and HemoCue; (c) and (d) agreement between Masimo measurements and HemoCue for females and males, correspondingly.

measurements and laboratory blood test results. From these plots, we can conclude that HemoCue provides sufficiently reliable measurements of hemoglobin concentrations that are suitable for labeling MW-PPG data within the entire measurement range. Fig. 6 shows the distribution of volunteers in the dataset according to the Masimo measurements and the analysis of agreement between Masimo and HemoCue. It should be noted that the actual error of Masimo is much higher than 10 g/L, although MAE within $\pm 1\sigma$ is in good agreement

with the performance claimed in the specifications. The most important conclusion drawn from these plots is the much lower performance of Masimo at low hemoglobin concentrations. In the Bland-Altman plots for males and females, there are two clearly distinguishable elbow points located at ∼130 g/L and ∼120 g/L, respectively. Masimo consistently overestimated hemoglobin concentrations below these points.

IV. ALGORITHMS FOR ESTIMATION OF HEMOGLOBIN CONCENTRATION

A. TRAIN-TEST SPLIT AND VALIDATION

The performance of the algorithm was evaluated using the following commonly accepted metrics: the mean absolute error (MAE), standard deviation of the error σ, mean absolute error within $\pm\,\sigma$ interval (MAE 1σ) and Pearson correlation coefficient R. The performance of the random guess model can be considered a baseline for these metrics. Given the dataset, random sampling from a normal distribution with the same mean and standard deviation yielded MAE $\approx$ 26 g/L, MAE $1\sigma \approx 9.4$ g/L, R ≈ 0. It should be noted that the MAE 1σ for random sampling almost matches the claimed accuracy of Masimo (MAE $1\sigma = 10$ g/L). Therefore, we expect the Pearson correlation coefficient R to be a more valuable metric than MAE because of its potentially higher dynamic range.

The training and validation procedures were performed on separate subsets of the dataset. Because different MW-PPG signals of one user are typically correlated, the dataset was split by users, so that all signals of one user were placed in either the train or the test subset. However, the dataset has only 170 unique people entries, which is not sufficient to provide accurate validation on a single test subset. Therefore, we employed two common approaches for cross-validation: live-one-out and live-p-out.

Leave-one-out is the most straightforward and simple approach: we keep one user out of the N users available in the dataset, train the algorithm on the remaining N-1 users, and predict hemoglobin for the user. We repeated this procedure iteratively for all users in the dataset. We then accumulated the predictions and calculated the average scores for all available predictions and models. However, this approach has two main disadvantages:

1) Instead of a single model, we have N models. Each of these N models can have different performance and properties: bias and variance of predicted hemoglobin. We cannot evaluate the performance of each model independently because of the small number of samples in the test portion of the data. Scores for accumulated predictions can be misleading and falsely low because of the significantly different properties of individual models.

2) Because of the significantly heterogeneous dataset, and especially the sparse data in the low-hemoglobin region, the test user may have a hemoglobin concentration that is very different from the hemoglobin values in the train portion of the data. Therefore, the model knows nothing about the PPG signals corresponding to this hemoglobin concentration and fails to predict it correctly.

Leave-p-out cross-validation (p out of N total users in the dataset are randomly selected for testing and the remaining N-p users are used for training) allows us to avoid the problems described above. We used the number of folds k = 200 with p = 15 users in each test. Thus, we can evaluate the scores of each model. However, for the practical implementation of this approach, the significantly heterogeneous distribution of samples by hemoglobin concentration and sex in the dataset should be taken into account. The dataset must be stratified during splitting so that the distribution of males/females and hemoglobin values in the test portion of the data matches the distribution in the train portion and in the entire dataset. The main advantages of this approach are as following:

1) We can evaluate the performance of each model independently and calculate the statistics for the multiple models.

2) Because we are using a large number of folds, each user participates in multiple tests (test subsets partially overlap), and hence, we can compare the predictions of a particular user made by different models, or average the predictions made by different models.

B. MACHINE LEARNING APPROACH #1

MW-PPG signal preprocessing begins with the calculation of the power spectrum of the raw PPG signal at each wavelength. The power spectrum of the raw PPG signal at each wavelength consists of several spectral bands (harmonics) centered at frequencies $f_1, f_2, \ldots f_n$, as shown in Fig. 7(a). Frequency f_1 corresponds to the heart rate, while the other frequencies are multiples of f_1: $f_2 = 2f_1$, $f_3 = 3f_1, \ldots$ We then calculate the set of 0-order raw features: $t^0_\lambda = \{p_0, p_1, p_2, \ldots m_1, m_2, \ldots\}$. Features $p_1, p_2, \ldots p_n$ are calculated as the total power of the corresponding spectral band within 0.3 Hz half-bandwidth. Features $m_1, m_2, \ldots m_n$ are calculated as the magnitude of the spectrum in the corresponding spectral band. The ratio between the total power p_n and the magnitude m_n is a measure of the width of the n-th harmonic. The DC-term p_0 is calculated as the average power of the signal after a low-pass filter with 0.65 Hz cut-off frequency. The set of 0-order raw features t^0_λ unambiguously describes the shape of the PPG waveform at each wavelength λ.

The development of a machine learning algorithm requires understanding of the features that can be extracted from the PPG signal and the relationship between these features and the target parameter, that is, the concentration of hemoglobin. The mutual arrangement of light sources and detectors in the prototype assumes that there are two source-detector separation distances (short and long) for each wavelength. At longer source-detector distances, the light penetrates deeper into the tissue and interacts with the blood vessels, so that the backscattered light contains more information about the blood (Fig. 7(b)). At shorter source-detector distances, the light travels at shallow depths and interacts with the upper layers of tissues with fewer or no blood vessels inside, so that

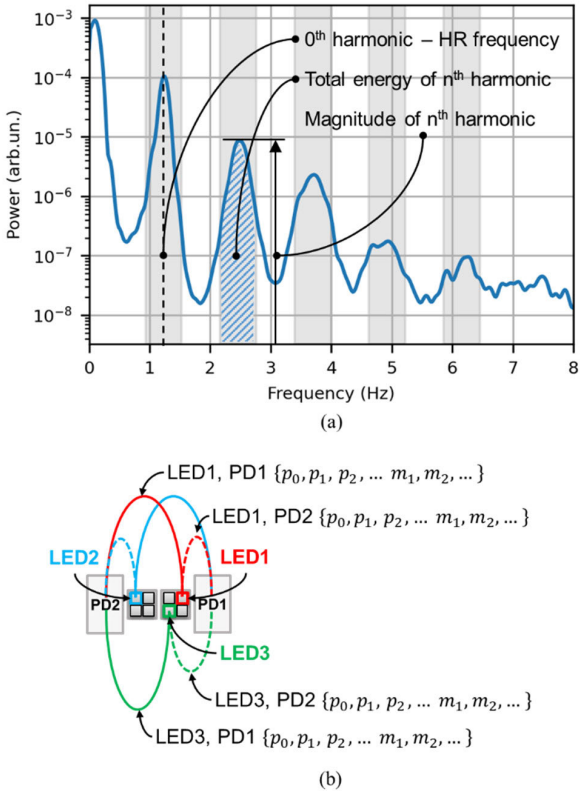

FIGURE 7. Extraction of the raw features from the power spectrum of the PPG signal: (a) typical power spectrum of the PPG signal and its relationship with the raw features; (b) the mutual arrangement of light emitting diodes and photodiodes in the sensor and corresponding distributions of light paths and penetration depths for different source-detector separation distances.

the backscattered light contains less information about the blood and more information about the spectral properties of the upper layers of the skin.

The set of 0-order raw features t^0_λ is calculated for each LED-PD pair, that is, each source-detector separation distance at each wavelength. An exhaustive approach to feature engineering relies on the following general assumptions:

1) Measurement of the relative changes in absorption at multiple wavelengths, that is, the ratio between signals measured at two different wavelengths and at the same source-detector separation distance.

2) Accounting for different optical paths and the influence of the spectral properties of the skin on the PPG signal, that is, the ratio between signals measured at the same wavelength and at different source-detector separation distances.

3) The combination of 1 and 2 provides information about the mutual changes in the absorption at two different wavelengths relative to the spectral properties of the skin at these wavelengths.

Thus, we can define the following equations for the 1st and 2nd order cross-features:

$$t^1 = \frac{t^0_{LED_i-PD_{far}}}{t^0_{LED_j-PD_{far}}} \text{ and } t^1 = \frac{t^0_{LED_i-PD_{close}}}{t^0_{LED_j-PD_{close}}}, \quad (1)$$

$$t^2 = \frac{t^0_{LED_i-PD_{far}}}{t^0_{LED_j-PD_{far}}} \Big/ \frac{t^0_{LED_i-PD_{close}}}{t^0_{LED_j-PD_{close}}} \quad (2)$$

where PD_{far} and PD_{close} represent the farthest and for the closest photodiodes relative to the corresponding light source.

C. MACHINE LEARNING APPROACH #2

Unlike oxygen saturation, the assessment of hemoglobin concentration requires measurement of the absolute values of absorbance. To calculate oxygen saturation, one can measure the transmittance or absorbance of a tissue at two wavelengths located on opposite sides of the spectrum relative to the isosbestic point, as shown in Fig. 8(a). At wavelengths shorter than the isosbestic point, absorption decreases with saturation. At wavelengths longer than the isosbestic point, absorption increases with saturation. We can measure the relative changes in the signal at these two wavelengths without having to know the absolute values of the absorbance at either of those wavelengths.

When the concentration of hemoglobin changes, the absorption changes in the same way at all wavelengths, and the difference in absorption at the two wavelengths is not directly proportional to the concentration of hemoglobin, as shown Fig. 8(b).

The main idea of an alternative approach to feature engineering for algorithm ML #2 is to make hemoglobin concentration measurements relative rather than absolute by introducing artificial "isosbestic" points. If we normalize the absorption spectrum of hemoglobin by absorption at a certain wavelength (e.g., 850 nm), then absorption at this wavelength no longer depends on the concentration of hemoglobin. Furthermore, absorption at shorter wavelengths (<850 nm) decreases with hemoglobin concentration, whereas absorption at longer wavelengths (>850 nm) increases with hemoglobin concentration (Fig. 8(c)), similar to the dependence of the absorption spectrum of the whole blood on oxygen saturation. Therefore, the difference in absorption between shorter and longer wavelengths is proportional to the concentration of hemoglobin. Normalization of shorter and longer wavelengths by exactly the same wavelength becomes meaningless when the ratio between absorptions at these wavelengths is calculated. However, several additional "isosbestic" points located at approximately 620 and 990 nm appear in the spectrum upon normalization to 850 nm, as shown in Fig. 8(c). Thus, we can split the spectrum into two parts and normalize each part using two different wavelengths. For example, shorter wavelengths can be normalized to 990 nm, longer wavelengths can be normalized to 850 nm, and vice versa. This normalization should be equivalent but not canceled out when calculating the difference in absorption between shorter and longer wavelengths. Furthermore, the absorption of light by other tissue components can be detected by the deviation of the normalized absorption at one of the "isosbestic" points from unity.

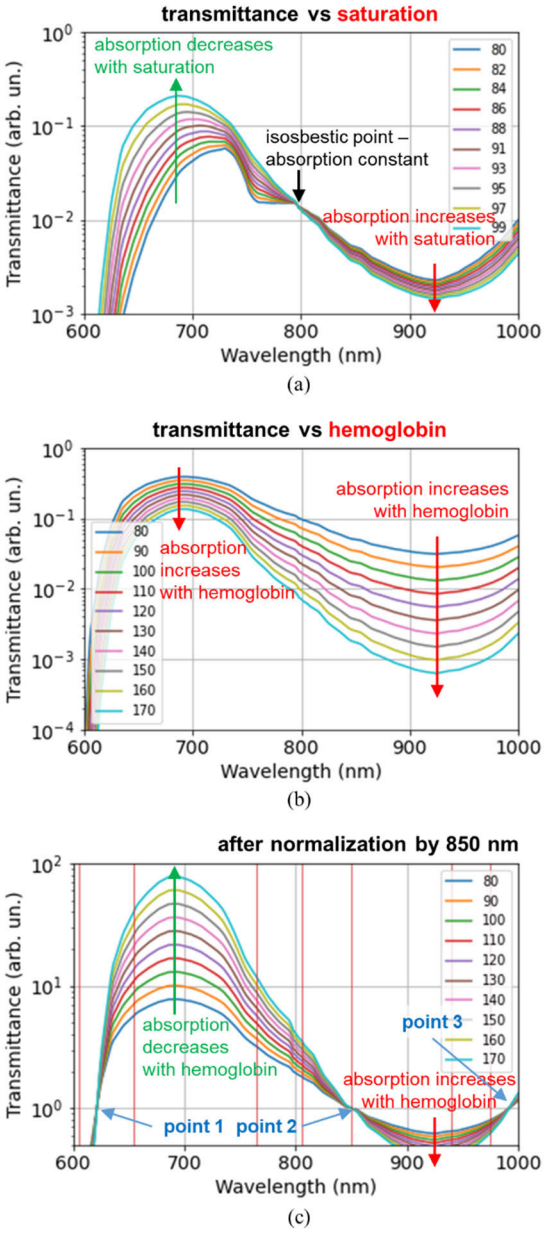

FIGURE 8. Numerically simulated spectra of the whole blood: (a) transmittance of the whole blood sample as a function of oxygen saturation ratio; (b) transmittance of the whole blood sample as a function of hemoglobin concentration; (c) transmittance of the whole blood sample after normalization by transmittance at 850 nm.

These considerations allowed us to simplify the signal preprocessing and feature extraction procedure for the ML #2 algorithm. In Fig. 8(c), the red vertical lines indicate the central wavelengths of the LEDs installed in the PPG prototype, overlapping with the absorption spectrum of hemoglobin normalized to 850 nm. Because there is no 990 nm light source in the prototype, we chose the nearest LED emitting at 975 nm and the LED at 850 nm as the reference points for normalization. We discovered experimentally that using PPG signals at 530 nm and 605 nm significantly reduced the performance of the algorithm because of the much stronger absorption at these wavelengths. Therefore, we have simplified the measurement scheme as follows:

1) 6 LEDs were split into two groups of:
   a) Group 1: 655, 765, and 805 nm normalized to 975 nm.
   b) Group 2: 940 nm normalized to 850 nm.
2) Longer LED-PD separation distances were used because of the greater penetration depth.

We then calculate the absolute difference between the signals in Groups 1 and 2. This allowed us to reduce significantly the number of features and, accordingly, the time required for training. Moreover, such a simplified scheme is compatible with the typical LED-PD arrangement in smart watches, making the implementation of the algorithm in a real product much easier.

Both the ML #1 and ML #2 algorithms are based on the XGBoost regression model with the main tunable parameters: number of estimators, learning rate, and maximum depth of a tree.

D. NEURAL NETWORK APPROACH

The application of neural networks (NNs) to PPG waveform analysis has been studied in blood pressure measurements [15], motion artifact mitigation [16], and stress monitoring [17]. In this regard, the performance of different NN architectures based on convolutional NN (CNN) [18], long-short-term memory (LSTM) NN [19], [20] and U-net [21] were evaluated.

In this work, we propose an end-to-end NN model capable of predicting hemoglobin concentration based on raw MW-PPG waveforms. The proposed NN model acts as an automatic feature-extraction procedure followed by a linear regression on the last layer. First, 16 MW-PPG signals, ~5 min long each, were split into 20-second chunks (800 points each) and arranged in an 800 × 16 array. The array is then fed to the NN input.

We evaluated the performance of several convolutional NNs (CNN's), each based on a series of 1D convolutional layers (Fig. 9). The intuition underlying the proposed architectures follows the results of the analysis of the PPG signal in reflectance mode (Section II). The main idea was to divide the CNN layers into two groups. The first group performs a series of 1D convolutions followed by average pooling (3 × 1) in the time domain (columns of the input matrix), which makes it possible to extract common time features in all spectral channels. The second group of layers starts with 1D convolution by rows (spectral domain corresponding to signals from different LED-PD pairs), which is equivalent to extracting spectral (different LEDs) or spatial (different PDs – different optical paths) information. The output of this layer is followed by three fully connected (dense) layers, which allows the extraction of more complex relationships between spectral and spatial features. The predicted hemoglobin concentration was the result of the output of a

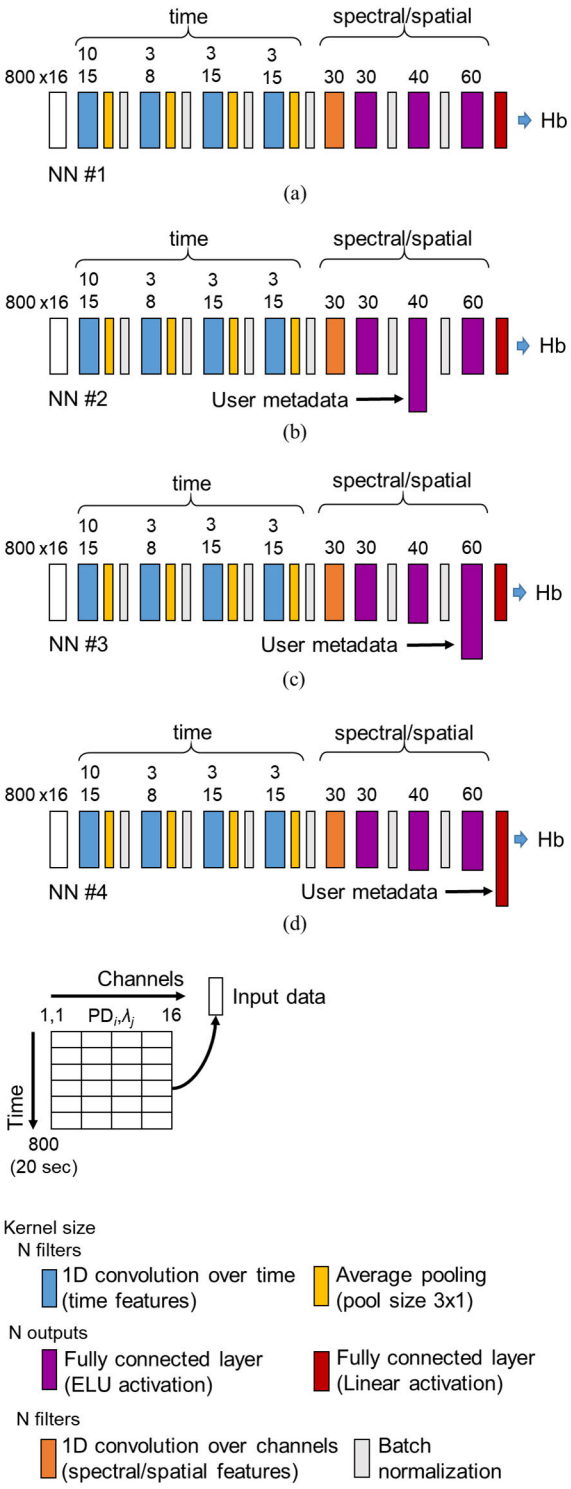

FIGURE 9. Architectures of tested convolutional neural networks based on: (a) MW-PPG signal only; (b), (c), and (d) different approaches to integration of user's metadata.

fully connected layer with linear activation. The batch normalization of the layer outputs (as shown in Fig. 9) was used to improve the learning behavior of the model. ELU activation was used for all non-linear layers. L1 regularization was applied to the last fully connected layer for feature selection, and L2 was used for the other dense layers.

In addition to PPG waveforms, user metadata (sex) can also be included in the NN model to improve hemoglobin concentration predictions. For this purpose, three optional NN models were designed, as shown in Fig. 9(b)−(d). User metadata can be included in the last dense layer, as shown in Fig. 9(d), which performs a linear regression on the features extracted from the PPG waveforms. Alternatively, including user metadata in other dense layers, as shown in Fig. 9(b)−(c), allows semantically more complex features to be extracted from PPG waveforms (i.e. different spectral features for males and females). Sex was included in the models as one-hot encoding.

We do not expect complex semantics in PPG signals (unlike image processing); hence, the CNN model does not need to be very deep, and the number of fitting parameters should be limited. This prevents overfitting and makes the model more computationally efficient. The kernel size and number of filters (N filters) for convolutional layers, as well as the number of outputs (N outputs) for dense layers, are shown in Fig. 9 for each NN architecture.

The training procedure consisted of 50 optimization epochs with a batch size of 128 and a batch shuffle. The Adam optimization algorithm [22] with the MSE loss function was used. As discussed above (Section IV), 15 people in the test subset are not sufficient to provide high-quality validation of a single model, and it is not possible to use the scores in the test subset for model selection or learning rate (LR) control, as is usually the case [23]. Therefore, we started training with LR=5×10.4 and reduced it after the 15^{th} epoch by a factor of 0.9. As the best model, we chose the result of the epoch that showed the maximum R-squared value on the train subset (usually epoch #45-50). This training procedure allowed us to balance over- and under-fitting.

V. RESULTS

Fig. 10 shows the exemplary results of hemoglobin concentration prediction achieved with both ML #1 and ML #2 algorithms in the leave-p-out test when user metadata (sex) were included in a feature vector. These plots contain predictions accumulated over all the available folds. Because of the large number of folds, test subsets in different folds may overlap, such that one true (measured) hemoglobin concentration corresponds to a number of concentrations predicted by different models. This appears as vertical rows on the scatter plots and inclined rows on the Bland-Altman plots. In general, ML #1 provides slightly better performance than ML #2. However, it should be noted that the ML #1 algorithm is slightly overfitted to the user's sex. Both the scatter plot and the Bland-Altman plot revealed clustering of the data points into two clouds (both tilted about the horizontal axis) corresponding to males and females. Even though ML #2 provides a slightly lower performance, it seems to be more resistant to overfitting to the user's sex. The Bland-Altman plots also show that the data points are more tightly grouped in the

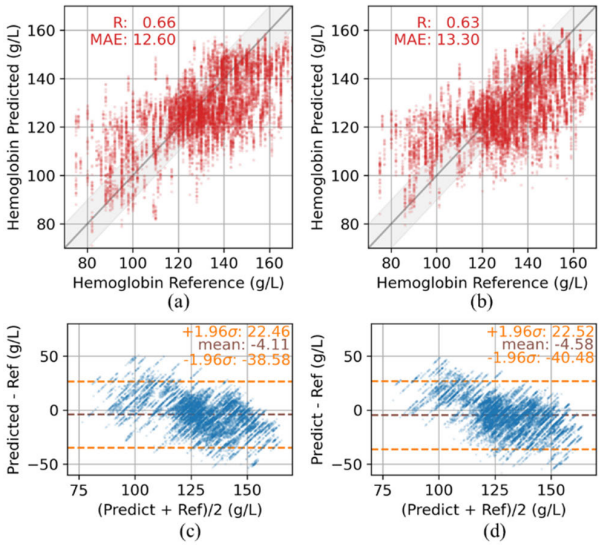

FIGURE 10. Performance of machine learning algorithms: (a), (b) correlation between hemoglobin concentration estimated using ML #1 and ML #2 algorithms, correspondingly, and reference HemoCue measurements; (c), (d) agreement between ML #1 and ML #2 algorithms, correspondingly, and HemoCue measurements.

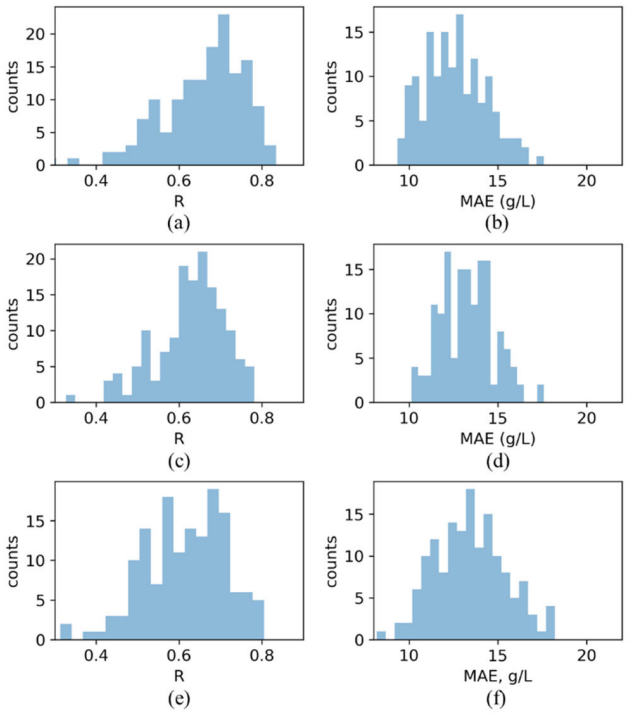

FIGURE 11. Distribution of scores in leave-p-out test for: (a) , (b) ML #1 algorithm; (c), (d) ML #2 algorithm; (e), (f) Masimo measurements validated on the same test subsets.

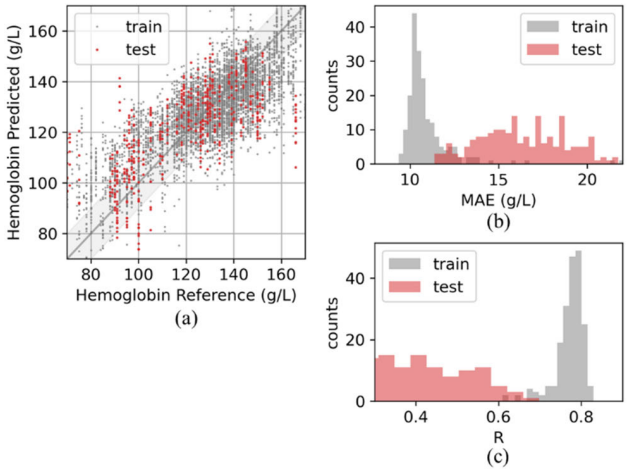

FIGURE 12. Performance of NN #3 model: (a) example of performance in a single fold (155 users in train, 15 users in test); (b), (c) distribution of scores over 200 folds.

case of ML #2, indicating less dispersion in the predicted values. The Bland-Altman plot also shows better alignment around the mean within the normal range of hemoglobin concentrations (>120 g/L) in the case of the ML #2 algorithm. Nevertheless, both models showed poor performance at lower concentrations of hemoglobin (<120 g/L), which is very similar to the results of Masimo (Fig. 6). It is worth mentioning how the performance of the individual models is distributed. Fig. 11 shows that the individual performance of a large number of models is well above the average performance shown in Fig. 10. The lower performance of some models can be explained by the quality of their corresponding test subsets. Despite the stratification, some test subsets may consist mainly of lower (and rare) hemoglobin concentrations, resulting in lower scores.

The performance of NN (for example, NN #3) on a single train and test subsets is shown in Fig. 12(a). Vertical rows in the scatter plot correspond to the predictions of the model for different 20-second chunks of the 5-minute raw signal. Predictions for individual chunks of the signal were averaged before calculating the overall score of the model. The distributions of R and MAE across the 200 models trained on different train-test splits (Fig. 12(b)−(c)) are widespread, indicating a relatively high variance in the model. However, reducing the number of parameters (N layers, filters, kernel size, etc.) did not improve the performance of the model on the test subsets.

The simplest model, NN #1, provided R ≈ 0.56 and MAE ≈ 14.5 g/L on average. The most straightforward approach to include user metadata (sex) in the linear regression layer (NN #4) did not improve performance.

The best scores (MAE ≈ 13 g/L, R ≈ 0.68) were achieved with model NN #2, in which user metadata were included in the second dense layer. However, a comparison of the scatter plots for the average test subset predictions for models NN #1 (Fig. 13(a)) and NN #2 (Fig. 13(b)) revealed cloud-like clustering around the average concentrations of hemoglobin in males and females. This effect was overcome in model NN #3 (Fig. 13(c)), which significantly reduced the error (in comparison with the NN #1 model) without data clustering.

To compare the performance of our algorithms with the standard well-developed NN architectures, we also tested MobileNet [24] using the collected dataset. Because

TABLE 1. Performance of ML and NN algorithm summary.

	NN #1	NN #2	NN #3	NN #4	ML #1	ML #1	ML #2	ML #2	MobileNet
User metadata	no	sex	sex	sex	no	sex	no	sex	no
Cross validation train score									
200 folds, 155 users in train									
MAE, g/L	11.1 ± 1.1	10.3 ± 0.8	10.4 ± 1.5	11.0 ± 1.1	--	--	--	--	7.6 ± 1.38
R	0.76 ± 0.05	0.80 ± 0.03	0.78 ± 0.05	0.76 ± 0.05	--	--	--	--	0.89 ± 0.04
Cross validation test score									
200 folds, 15 users in test									
MAE, g/L	16.9 ± 2.6	15.0 ± 2.5	16.5 ± 2.7	17.0 ± 2.7	14.3 ± 1.7	12.6 ± 1.7	15.4 ± 1.6	13.3 ± 1.5	18.8 ± 2.6
R	0.57 ± 0.14	0.67 ± 0.13	0.60 ± 0.13	0.59 ± 0.14	0.51 ± 0.16	0.66 ± 0.09	0.42 ± 0.15	0.63 ± 0.08	0.44 ± 0.16
Averaged prediction score									
200 folds, 15 users in test									
MAE, g/L	14.4	13.2	13.6	14.3	--	--	--	--	15.5
R	0.57	0.64	0.62	0.57	--	--	--	--	0.47
Live-one-out score									
MAE, g/L	15.1	13.6	14.1	15.1	15.5	13.9	16.5	14.8	--
R	0.53	0.64	0.58	0.50	0.50	0.61	0.42	0.56	--

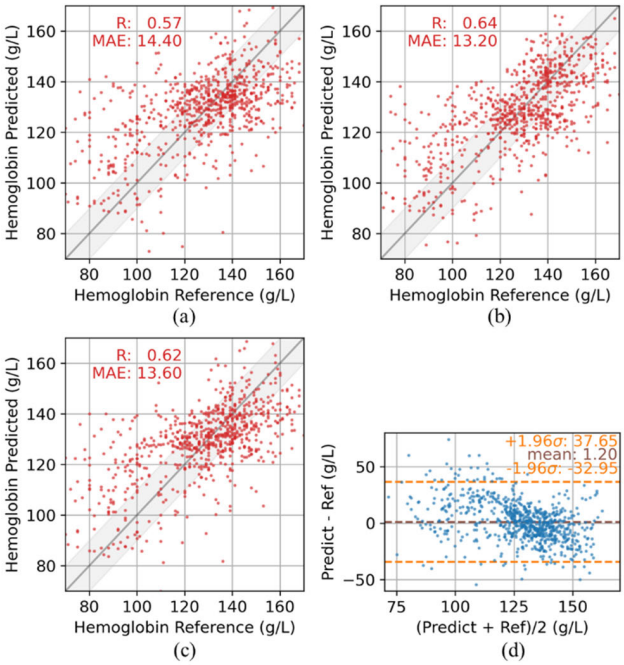

FIGURE 13. Performance of NN models in leave-p-out test with the prediction averaged over 200 folds: (a) NN #1 model; (b) NN #2 model; (c) NN #3 model; (d) agreement between NN #3 and reference HemoCue measurements.

MobileNet accepts as input a 3D tensor of minimum size (width/height) equal to 32, chunks of the raw signal were reshaped to $800 \times 32 \times 3$ with two additional dimensions filled with zeros and spectral/spatial dimensions doubled with the same PPG signal. With the large number of tuned parameters, Mobilenet tends to be overfitted: train MAE dropped down to ∼1 g/L while test MAE reached 20 g/L during ∼3-5 epochs. Therefore, the training parameters were reduced: batch size – 32, LR=5 × 10.6 with a reduction by factor of 0.9 after 10 epochs and total number of epochs 25. The best scores achieved with MobileNet (MAE ≈ 15.5 g/L, R ≈ 0.47) were significantly lower than those of model NN #1, which can be explained by the large number of tuning parameters (4.3M) and the architecture optimized for image processing.

Table 1 summarizes the overall performance of ML #1, ML #2, and NN models under various testing conditions.

VI. CONCLUSION

This paper presents the results of measuring the total hemoglobin concentration using wrist wearable PPG sensor operating in the reflectance mode. The performance of the developed prototype and algorithms is comparable to that of a high-grade noninvasive clinical device operating in the transmittance mode. Despite the fact that the transmittance mode is much more favorable for measuring the optical properties of a tissue, the proposed approach to the analysis of the time-spectral features of the reflectance-mode PPG waveform successfully competes with conventional PPG measurements in the transmittance mode.

The physically based approach to feature extraction in ML #2 makes it possible to reduce hardware requirements and to develop a simple and computationally efficient algorithm for the prediction of hemoglobin concentration. The ML #2 algorithm operates with only 120 features (compared to over 3000 features in the algorithm ML #1 and the complex architecture of CNN in Fig. 9), involves simple pre-processing for feature extraction, and therefore requires much less time to train the algorithm and make predictions.

According to our data, the performance of Masimo Radical-7 is similar to that of the prototype and algorithms proposed in this paper, compared to the HemoCue reference measurements (Fig. 6). Masimo Radical-7 also failed to predict correctly hemoglobin concentrations below ∼120 g/L.

It should be noted that, despite the slightly better average performance, the distributions of MAE and R in the Masimo Radical-7 measurements calculated for users in the test subsets in the leave-p-out split (Fig. 11(e)−(f)) are wider than the distributions of MAE and R in our measurements, as shown in Fig. 11(a)−(d). This could be attributed to the significantly unbalanced distribution of hemoglobin concentrations in the dataset (Fig. 5), with a small number of samples sparsely distributed over a wide range of low concentrations (∼60 – 120 g/L). However, we believe that the algorithm in Masimo Radical-7 has been thoroughly tuned and tested on a much larger dataset, including corner cases with extremely low and high hemoglobin concentrations. Nevertheless, Masimo Radical-7 showed a similar drop in performance in this range of hemoglobin concentrations. This suggests that the drop in performance at lower hemoglobin concentrations is caused by physiological and/or physical reasons, for example, much stronger interference with other tissue components at weaker absorption of light by hemoglobin.

The correct and timely detection of abnormal changes in hemoglobin concentration is an important step in preventive healthcare. Possible use cases include, but are not limited to, anemia detection, women's health, mental health, and physical activity planning. The presented results confirm the feasibility of noninvasive measurements of hemoglobin concentration by means of the reflectance-mode multiwavelength PPG sensor, making this approach compatible with consumer wearable devices.

ACKNOWLEDGMENT

The authors would like to thank Oxana Semyachkina-Glushkovskaya, Sergey Kapralov, Andrey Danilov, Daria Elovenko, and Victoria Adushkina for organizing and conducting clinical trials.

REFERENCES

[1] G. Prieto-Avalos, N. A. Cruz-Ramos, G. Alor-Hernández, J. L. Sánchez-Cervantes, L. Rodríguez-Mazahua, and L. R. Guarneros-Nolasco, "Wearable devices for physical monitoring of heart: A review," *Biosensors*, vol. 12, no. 5, p. 292, May 2022.

[2] J. Dunn, L. Kidzinski, R. Runge, D. Witt, J. L. Hicks, S. M. S.-F. Rose, X. Li, A. Bahmani, S. L. Delp, T. Hastie, and M. P. Snyder, "Wearable sensors enable personalized predictions of clinical laboratory measurements," *Nature Med.*, vol. 27, no. 6, pp. 1105–1112, May 2021.

[3] R. S. Vulcan, S. André, and M. Bruyneel, "Photoplethysmography in normal and pathological sleep," *Sensors*, vol. 21, no. 9, p. 2928, Apr. 2021.

[4] S. S. Coughlin and J. Stewart, "Use of consumer wearable devices to promote physical activity: A review of health intervention studies," *J. Environ. Health Sci.*, vol. 2, no. 6, pp. 1–6, 2016.

[5] T. Pereira, N. Tran, K. Gadhoumi, M. M. Pelter, D. H. Do, R. J. Lee, R. Colorado, K. Meisel, and X. Hu, "Photoplethysmography based atrial fibrillation detection: A review," *NPJ Digit. Med.*, vol. 3, no. 1, p. 3, Jan. 2020.

[6] R. N. Pittman, *Regulation of Tissue Oxygenation*. San Rafael, CA, USA: Morgan & Claypool Life Sciences, Jun. 2011.

[7] M. W. Causey, S. Miller, A. Foster, A. Beekley, D. Zenger, and M. Martin, "Validation of noninvasive hemoglobin measurements using the masimo radical-7 SpHb station," *Amer. J. Surgery*, vol. 201, no. 5, pp. 592–598, May 2011.

[8] J. Krait, H. Ewald, and H. Gehring, "An optical device to measure blood components by a photoplethysmographic method," *J. Opt. A, Pure Appl. Opt.*, vol. 7, no. 6, p. S318, May 2005.

[9] T. K. Aldrich, M. Moosikasuwan, S. D. Shah, and K. S. Deshpande, "Length-normalized pulse photoplethysmography: A noninvasive method to measure blood hemoglobin," *Ann. Biomed. Eng.*, vol. 30, pp. 1291–1298, Oct. 2002.

[10] T. Tamura, "Current progress of photoplethysmography and SPO2 for health monitoring," *Biomed. Eng. Lett.*, vol. 9, no. 1, pp. 21–36, Feb. 2019.

[11] J. Spigulis, L. Gailite, A. Lihachev, and R. Erts, "Simultaneous recording of skin blood pulsations at different vascular depths by multiwavelength photoplethysmography," *Appl. Opt.*, vol. 46, no. 10, pp. 1754–1759, Apr. 2007.

[12] J. Liu, B. P. Yan, Y.-T. Zhang, X.-R. Ding, P. Su, and N. Zhao, "Multi-wavelength photoplethysmography enabling continuous blood pressure measurement with compact wearable electronics," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 6, pp. 1514–1525, Jun. 2019.

[13] R. Hiscock, D. Kumar, and S. W. Simmons, "Systematic review and meta-analysis of method comparison studies of Masimo pulse co-oximeters (radical-7® or Pronto-7®) and HemoCue® absorption spectrometers (B-hemoglobin or 201+) with laboratory haemoglobin estimation," *Anaesth. Intensive Care*, vol. 43, no. 3, pp. 50–341, May 2015.

[14] M. K. Diab, "Pulse and active pulse spectraphotometry," U.S. Patent 696 159 8B2, Oct. 1, 2005.

[15] P. Su, X.-R. Ding, Y.-T. Zhang, J. Liu, F. Miao, and N. Zhao, "Long-term blood pressure prediction with deep recurrent neural networks," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Las Vegas, NV, USA, Mar. 2018, pp. 323–328.

[16] M. T. Islam, "SPECMAR: Fast heart rate estimation from PPG signal using a modified spectral subtraction scheme with composite motion artifacts reference generation," *Med. Biol. Eng. Comput.*, vol. 57, pp. 689–702, Oct. 2018.

[17] N. Mukherjee, S. Mukhopadhyay, and R. Gupta, "Real-time mental stress detection technique using neural networks towards a wearable health monitor," *Meas. Sci. Technol.*, vol. 33, no. 4, Jan. 2022, Art. no. 044003.

[18] G. Slapničar, N. Mlakar, and M. Luštrek, "Blood pressure estimation from photoplethysmogram using a spectro–temporal deep neural network," *Sensors*, vol. 19, no. 15, p. 3420, Aug. 2019.

[19] Y.-H. Li, L. N. Harfiya, K. Purwandari, and Y.-D. Lin, "Real-time cuffless continuous blood pressure estimation using deep learning model," *Sensors*, vol. 20, no. 19, p. 5606, Sep. 2020.

[20] L. N. Harfiya, C.-C. Chang, and Y.-H. Li, "Continuous blood pressure estimation using exclusively photoplethysmography by LSTM-based signal-to-signal translation," *Sensors*, vol. 21, no. 9, p. 2951, Apr. 2021.

[21] S. Mahmud, N. Ibtehaz, A. Khandakar, A. M. Tahir, T. Rahman, K. R. Islam, M. S. Hossain, M. S. Rahman, F. Musharavati, M. A. Ayari, M. T. Islam, and M. E. H. Chowdhury, "A shallow U-Net architecture for reliably predicting blood pressure (BP) from photoplethysmogram (PPG) and electrocardiogram (ECG) signals," *Sensors*, vol. 22, no. 3, p. 919, Jan. 2022.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[23] G. S. Na, "Efficient learning rate adaptation based on hierarchical optimization approach," *Neural Netw.*, vol. 150, pp. 326–335, Jun. 2022.

[24] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

VLADISLAV V. LYCHAGOV received the M.Sc. degree in biophysics and the Ph.D. degree in optics from Saratov State University (SSU), Russia, in 2003 and 2007, respectively.

From 2007 to 2014, he was appointed as a Research Associate and then an Associate Professor with SSU. From 2015 to 2016, he was a Visiting Researcher with the University of Colorado Boulder, Boulder, CO, USA, and the Albert Einstein College of Medicine, Yeshiva University, New York City, NY, USA. He has been with the Samsung Research and Development Institute Russia, Moscow, since 2017, and leading optic sensor design and algorithm group. He has designed and developed a custom microfluidic flow cytometry system for screening and sorting new types of fluorescent proteins. His research interests include coherent-domain optical measurements, statistical optics, Fourier optics, and practical implementation of these methods in industrial and biomedical applications.

VLADIMIR M. SEMENOV received the M.S. degree in physics and technology of semiconductor devices from the Moscow Power Engineering Institute (MPEI), Moscow, Russia, in 2011, and the Ph.D. degree in electrical engineering from the Tunable Diode Laser Spectroscopy (TDLS), A. M. Prokhorov General Physics Institute of Russian Academy of Science (GPI RAS), and the National Research Center "Kurchatov Institute," in 2014.

Since 2011, he has been with TDLS, A. M. Prokhorov GPI RAS, and the National Research Center "Kurchatov Institute." From 2014 to 2018, he was a leading Researcher with the Moscow Institute of Physics and Technology (MIPT) for laser spectroscopy applications for environmental care and planetary science. In 2018, he joined the Samsung Research and Development Institute Russia, Moscow, as a Senior Engineer in optical sensor design. His research interests include high-precision laser spectroscopy, coherent detection techniques, noise in laser systems, optical signal processing, physics of semiconductor lasers, laser applications for gas sensing, aerosol detection and environmental care, LIDARs, and 3-D vision.

ELENA K. VOLKOVA received the M.Sc. degree in optics and the Ph.D. degree in biophysics from Saratov State University (SSU), Russia, in 2010 and 2013, respectively.

Since 2013, she was appointed as a Research Fellow with the Laboratory of Optoelectronics and Measurement Techniques, University of Oulu, Finland, as an Engineer with the Institute of Nanostructures and Biosystems, SSU, and as a Research Associate with the Laboratory of Biomedical Optics, Institute of Optics and Biophotonics, SSU. In 2018, she joined the Samsung Research and Development Institute Russia, Moscow, as a Senior Engineer. She is the author or coauthor of 26 articles and holds three patents. Her research interests include wearable sensors, mobile health, optics and biophotonics, absorption, and luminescence spectroscopy of nanoparticles and crystals.

DMITRII I. CHERNAKOV received the B.Sc. degree in photonics from ITMO University, Saint-Petersburg, Russia, in 2016, and the M.Sc. degree in photonics from the University of Eastern Finland, Joensuu, Finland, in 2018.

He joined the Samsung Research and Development Institute Russia, Moscow, in 2019, with a focus on the development of optical sensors and algorithms for sensor data processing. His research interests include optical materials, optical sensing, and sensor signal processing.

JOONGWOO AHN received the B.S. degree in electrical engineering and biomedical engineering from Kyung Hee University, South Korea, in 2012, and the M.S. and Ph.D. degrees in bioengineering from Seoul National University, South Korea, in 2018.

From 2018 to 2019, he was a Research Professor with the Biomedical Research Institute, Seoul National University Hospital. Since 2020, he has been a Staff Engineer with the Digital Health Team, Samsung Electronics. His research interests include non-invasive health sensors using optics and electrodes in wearables.

JUSTIN YOUNGHYUN KIM received the Ph.D. degree in electrical engineering from the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA, USA, in 2012.

Since 2012, he has been with the Mobile eXperience Business Division, Samsung Electronics. He is the author or coauthor of ten articles and holds 69 patents. His main research interests include wearable health, mobile health, and accessory health sensors.