A Deep Learning–based PPG Quality Assessment Approach for Heart Rate and Heart Rate Variability

EMAD KASAEYAN NAEINI, University of California Irvine, USA
FATEMEH SARHADDI, IMAN AZIMI, and PASI LILJEBERG, University of Turku, Finland
NIKIL DUTT and AMIR M. RAHMANI, University of California Irvine, USA

Photoplethysmography (PPG) is a non-invasive optical method to acquire various vital signs, including heart rate (HR) and heart rate variability (HRV). The PPG method is highly susceptible to motion artifacts and environmental noise. Unfortunately, such artifacts are inevitable in ubiquitous health monitoring, as the users are involved in various activities in their daily routines. Such low-quality PPG signals negatively impact the accuracy of the extracted health parameters, leading to inaccurate decision-making. PPG-based health monitoring necessitates a quality assessment approach to determine the signal quality according to the accuracy of the health parameters. Different studies have thus far introduced PPG signal quality assessment methods, exploiting various indicators and machine learning algorithms. These methods differentiate reliable and unreliable signals, considering morphological features of the PPG signal and focusing on the cardiac cycles. Therefore, they can be utilized for HR detection applications. However, they do not apply to HRV, as only having an acceptable shape is insufficient, and other signal factors may also affect the accuracy. In this article, we propose a deep learning–based PPG quality assessment method for HR and various HRV parameters. We employ one customized one-dimensional (1D) and three 2D Convolutional Neural Networks (CNN) to train models for each parameter. Reliability of each of these parameters will be evaluated against the corresponding electrocardiogram signal, using 210 hours of data collected from a home-based health monitoring application. Our results show that the proposed 1D CNN method outperforms the other 2D CNN approaches. Our 1D CNN model obtains the accuracy of 95.63%, 96.71%, 91.42%, 94.01%, and 94.81% for the HR, average of normal to normal interbeat (NN) intervals, root mean square of successive NN interval differences, standard deviation of NN intervals, and ratio of absolute power in low frequency to absolute power in high frequency ratios, respectively. Moreover, we compare the performance of our proposed method with state-of-the-art algorithms. We compare our best models for HR-HRV health parameters with six different state-of-the-art PPG signal quality assessment methods. Our results indicate that the proposed method performs better than the other methods. We also provide the open source model implemented in Python for the community to be integrated into their solutions.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

Additional Key Words and Phrases: Deep learning, convolutional neural networks, signal quality assessment, health monitoring, Internet of Things, photoplethysmogram, heart rate variability, wearable electronics

This work was partially supported by the US National Science Foundation (NSF) grant S&CC CNS-1831918, and Academy of Finland through the SLIM Project under Grant 316810 and Grant 316811.
Authors' addresses: E. Kasaeyan Naeini, University of California Irvine, California, USA; e-mail: ekasaeya@uci.edu; F. Sarhaddi, I. Azimi, and P. Liljeberg, University of Turku, Finland; e-mails: {fatemeh.sarhaddi, iman.azimi, pakrli}@utu.fi; N. Dutt and A. M. Rahmani, University of California Irvine, USA; e-mails: {dutt, a.rahmani}@uci.edu.

2637-8051/2023/11-ART24
https://doi.org/10.1145/3616019

ACM Reference format:
Emad Kasaeyan Naeini, Fatemeh Sarhaddi, Iman Azimi, Pasi Liljeberg, Nikil Dutt, and Amir M. Rahmani. 2023. A Deep Learning–based PPG Quality Assessment Approach for Heart Rate and Heart Rate Variability. *ACM Trans. Comput. Healthcare* 4, 4, Article 24 (November 2023), 22 pages.
https://doi.org/10.1145/3616019

1 INTRODUCTION

Internet of Things technology is fundamentally changing the delivery of healthcare, enabling health monitoring applications anywhere and anytime [1–5]. Wearables, as intelligent electronic devices, play a key role in such applications; health data are collected,analyzed, and shared over a network. Such devices are becoming widely used around the globe, as they are even more miniaturized, smarter, and easier to use. Wearables can include a variety of sensing resources to continuously monitor body functions. *Photoplethysmography (PPG)* is an inexpensive and convenient method that is being employed in a variety of wearables as well as smartphones to collect various vital signs such as *heart rate (HR)*, *heart rate variability (HRV)*, SpO2, and respiration rate [6].

PPG is a simple optical method to measure plethysmogram, showing the variations of blood volume in body organs. The obtained signal can be tailored to track cardiorespiratory parameters. The PPG method mainly consists of two components placed on the skin. First, a light source is utilized to reflect light to the skin surface. The red, infrared, or green light can be selected according to the application. Second, a photodetector collects the light reflection [7]. The collected signal includes a pulsatile (AC) and a slowly fluctuating (DC) component, allowing the monitoring of cardiorespiratory parameters non-invasively and continuously. The PPG method is tailored in many clinically approved devices and commercial wearables (e.g., smartwatches and rings), as it is easy-to-implement and energy efficient [8–13].

However, input PPG signals might be distorted due to noises caused by motion artifacts and other environmental sources [14], which are ubiquitous and unavoidable in everyday life settings. For instance, the light sensors might be exposed by environment light sources, by which the collected PPG signal is distorted and the information is concealed within the signal. Specifically, we observe noise affects the collected signal differently as the users participate in various physical activities while using the PPG-based wearables. Such movements could negatively impact the signal quality [15, 16]. For example, if the engaged subject is running, then the noise power is much higher compared to that affecting the same signal acquired when the participant is sleeping [12]. Low-quality PPG signal (i.e., low signal-to-noise ratio) affects the reliability of the health parameters extracted, e.g., HR and HRV parameters. Such unreliable measurements can lead to false alarms or life-threatening decisions in healthcare applications [17].

In the literature, the PPG signal quality was investigated by proposing assessment methods to discriminate reliable and unreliable parts of the signal. Various studies introduced signal quality indicators [18–20] or utilized template matching approaches to distinguish the reliable segments of the PPG signal [21–23]. Moreover, traditional machine learning methods such as *support vector machine (SVM)*, decision tree, and K-nearest neighbors were presented to carry out PPG quality assessment using features extracted from the shape of the signal [24–30]. Recently, deep learning methods also exploited PPG quality assessment, enabling automatic PPG feature extraction [31–35].

These studies have mostly assessed the shape of the signal, focusing on the HR. In other words, the signal is classified as reliable if the cardiac cycle can be detected. We believe that such a quality assessment method is insufficient for PPG signals, from which many other parameters can also be extracted. PPG signals can be leveraged to remotely collect HRV parameters whose accuracy might be diminished due to different factors in the signal. A PPG signal might be reliable for HR detection but, for example, unreliable for *standard deviation of NN intervals (SDNN)*. Therefore, a PPG quality assessment method is essential in health monitoring

applications, determining the reliability of the PPG signal according to the health parameters. Such a method needs to provide a confidence value for each health parameter. Consequently, unreliable parameters can be removed, preventing incorrect health decision-making.

Moreover, there are available public datasets, such as WESAD, including physiological and motion data recorded using wrist- and chest-worn devices in lab settings [36]. PPG signals are often distorted due to motion artifacts and environmental noise. Therefore, data collected in lab settings are insufficient for PPG signal quality assessment studies. The signal quality assessment methods should be evaluated using the PPG data collected in free-living conditions, where the users engage in their daily routines.

In this article, we propose a quality assessment method to distinguish reliable PPG signals according to the HR and HRV values extracted from the signal. *Convolutional Neural Network (CNN)* methods are tailored in this regard to train models for each health parameter. We design (i) a customized 1D CNN method and (ii) three 2D CNN methods enabled by three deep neural networks. The proposed methods are investigated, and the best architecture is selected. Then, the performance of the selected method is evaluated, in comparison to existing rule-based, machine learning, and deep learning PPG quality assessment methods. To this end, we perform a home-based *electrocardiogram (ECG)* and PPG collection, in which the signals are acquired simultaneously and remotely for 24 hours. The evaluation includes more than 210 hours of data. For each health parameter, the PPG quality is defined in comparison to the ECG using an automatic annotation process. The main contributions of this article are as follows:

- Proposing a CNN-based PPG quality assessment method based on HR and HRV
- Customizing the method with different CNN architectures to investigate the performance
- Presenting an automatic annotation method, where the PPG signals are labeled as "reliable" or "unreliable" against the ECG as a baseline
- Evaluating the proposed methods in comparison to state-of-the-art PPG quality assessment methods via a home-based monitoring where more than 210 hours of PPG and ECG signals are collected
- Providing a portable and open source model implemented in Python[1] for the community to be integrated in their solutions

The rest of the article is organized as follows. Section 2 outlines the motivation behind this research and the related works. We describe the background of this research in Section 3. In Section 4, we present the case study, including recruitment, data collection, and data annotation. Section 5 introduces the PPG quality assessment approach. In Section 6, the results are presented. Finally, Section 7 concludes the article.

2 RELATED WORK AND MOTIVATION

Several signal quality assessment techniques have been introduced in the literature to determine whether collected PPG signals are reliable. To this end, the unreliable part of the signal is detected and removed, preventing misinterpretation of data and invalid decision-making. Such techniques are mostly focused on the morphology of the bio-signals. Different signal quality indicators—such as skewness, kurtosis, and baseline wandering of the signal—were exploited to estimate the PPG signal quality [18–20]. Rule-based methods have also been introduced to distinguish low-quality bio-signals, leveraging decision rules, e.g., signal saturation detection and beat-to-beat-interval evaluation [37–39]. Moreover, template matching techniques have been proposed to distinguish reliable and unreliable signals. For example, Sun et al. [21] proposed a template matching method based on dynamic time warping for signal quality assessment. In other similar studies, the quality assessment was performed by investigating the morphological similarity between the original signal and the template signals, which were the expected waveform or surrounding pulses [22, 23]. The PPG morphological waveform can be affected by motion artifacts, environmental noise, and cardiovascular issues. Therefore, these methods are

[1] https://github.com/HealthSciTech/pyPPQa

inaccurate due to variations in the PPG morphological features. Moreover, these methods usually need predefined thresholds that should set manually based on the data.

A number of machine learning methods have been developed for PPG signal quality assessment. Various studies in the literature proposed to train machine learning methods, utilizing morphological and time/frequency domain features of PPG signals. These methods utilized supervised and lightweight unsupervised approaches. The supervised methods include hierarchical decision rule [40], decision tree [41], random forest [24], SVM [25, 26], and neural network [29]. With this intention, Sabeti et al. [27] proposed a threshold optimization learning method and compared the method with a SVM and a decision tree. Furthermore, in Reference [28], the authors investigated SVM, K-nearest neighbors, and decision tree for the signal quality assessment in the case of atrial fibrillation. In addition, Mahmoudzadeh et al. [30] proposed a lightweight unsupervised method providing low-computational real-time PPG quality assessment. The authors of Reference [42] also proposed unsupervised PPG SQA methods based on the self-organizing map. Machine learning–based methods outperform rule-based methods in terms of accuracy. However, similarly to the rule-based method, results are affected by morphological variation of PPG signals. In addition, machine learning–based methods utilized manual feature extraction. Therefore, the accuracy and generalization of these methods are restricted due to a limited set of selected features.

Recently, deep learning methods were also introduced for the PPG quality assessment. Kasaeyan Naeini et al. [33] introduced a real-time PPG assessment approach to classify the signals according to the HR values. Perieira et al. [31] used 1-dimensional (1D) and 2D deep neural networks for signal quality assessment in case of atrial fibrillation. They considered the PPG signal as time series (1D) and also as images (2D). Authors in Reference [34] also proposed a 1D CNN model for determining reliable segment of PPG signals. In another work, Soto et al. [32] proposed a multi-task CNN called DeepBeat for signal quality assessment and atrial fibrillation detection. They improved their results by pre-training the model using *convolutional denoising auto-encoders (CDAE)*. In addition, Roh et al. [35] utilized a 2D CNN for classifying each segment of beat waveform. These methods benefited from automatic feature extraction. However, they need manual/expert labeling based on HR values or morphological waveform of PPG signals. They are inappropriate for HRV parameters analysis. A comparison of the PPG quality assessment methods is indicated in Table 1.

The PPG quality assessment methods, presented in the literature, were mostly designed to analyze the shape of a signal and assessing the reliability of the PPG signal itself focusing solely on HR. These methods tailor rules, templates, and hypothesis functions to classify a PPG as reliable if the desired signal (i.e., cardiac cycle) is appropriately retrieved. Such methods can be utilized for HR detection applications where the pulse is visible in the reliable signals. However, they are not applicable to HRV analysis, as only having an acceptable shape is insufficient, and other factors in the signal may affect the accuracy, too. Moreover, such factors could impact differently on the accuracy of the HRV parameters. For instance, a 5-minute high-quality PPG signal with two or three noisy peaks could result in an accurate/acceptable HR and SDNN but can lead to an invalid *Root Mean Square of Successive Differences (RMSSD)* value.

To emphasize the importance of developing quality assessment models for each HRV parameter, and for clarification, let us show five real motivational examples of PPG signals with different artifacts in Figures 1 and 2. These figures show that a noisy PPG signal, which results in some unreliable HRV parameters, can still be used to extract other HRV parameters accurately. Therefore, the quality of PPG signals should be evaluated according to the desired HRV parameters. In these examples, we extract the HR, *average of normal to normal interbeat (NN) intervals (AVNN)*, RMSSD, SDNN, and *low frequency/high frequency (LF/HF)* ratio of the PPG signals. Then, we specify their errors by calculating the distance between these parameters and the corresponding parameters obtained from the baseline ECG signals. In other words, the error is $ECG\ value\ -\ PPG\ value$.

(a) Figure 1(a) illustrates an example, in which less than 5 seconds of the PPG signal (highlighted in red) is distorted due to hand movements. As indicated, a few peaks are detected incorrectly in the noisy part. The PPG signal provides HR, AVNN, and SDNN values with low errors compared with the ECG. However, the

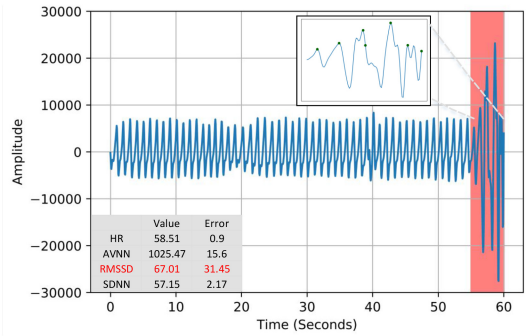
Table 1. Comparing Current PPG Quality Assessment Methods with Proposed Method for Their Features, Algorithms, Annotation Procedures, and Source-code Availability

	Reference	Automatic Features extraction	Features	Method	Automatic Labeling	Annotation	Source code Availability
	Proposed method	✓	—	1D CNN/ 2D CNN	✓	Based on HR and HRV parameters	✓
Rule based Methods	Orphanidou et al. [37]	×	Extracted HR and morphological features	Threshold based rules	×	Based on the signal shape	×
	Reddy et al. [38]	×	Predictor Coefficient	Hierarchical decision rules	×	"	×
	Tyapochkin et al. [39]	×	Statistical parameters of IBIs	Predefined rules	×	"	×
	Vadrevu et al. [40]	×	Amplitude and time-series features	Hierarchical decision rules	×	"	×
Supervised methods	Alam et al. [24]	×	Morphological features and HR value	Random forest	×	Based on the signal shape and HR values	×
	Zhang et al. [25]	×	Frequency domain and time series characteristics	SVM	×	Based on the signal shape	×
	Preira et al. [26]	×	Frequency-domain, time-domain, and non-linear features	SVM	×	"	×
	Preira et al. [28]	×	Frequency-domain, and time-domain features	SVM	×	"	×
Unsupervised methods	Mahmoudzadeh et al. [30]	×	Statistical features	Elliptical envelope	×	"	✓
	Roy et al. [42]	×	Entropy and signal complexity features	Self-organizing map	×	"	×
Deep learning methods	Preira et al. [31]	✓	—	ResNet18	×	"	×
	Soto et al. [32]	✓	—	Multi-task CNN (pre-training with CDAE)	×	"	×
	Goh et al. [34]	✓	—	1D CNN	×	"	×
	Roh et al. [35]	✓	—	2D CNN	×	"	×
	Naeini et al. [43]	✓	—	1D CNN	✓	Based on HR	×

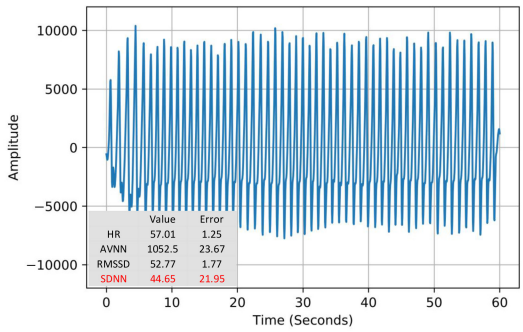
error for the RMSSD is high. The reason is that RMSSD is correlated to the short-term variation in the PPG signal, and a small corrupted part in the signal could affect its accuracy.

(b) Figure 1(b) shows a PPG signal with no distorted peaks. However, the variation of the intervals is not similar to the ECG. This signal provides reliable HR, AVNN, and RMSSD but unreliable SDNN. The reason is that SDNN is associated with long-term variation in *normal-to-normal interbeat intervals (NNIs)*, and the noises that affect the variation of intervals will impact SDNN accuracy.

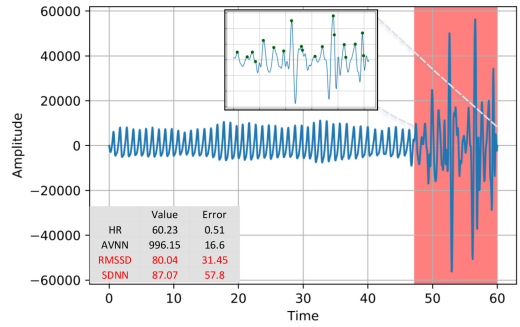
(c) Figure 1(c) shows a 1-minute window where the PPG signal is partially corrupted (i.e., 20%). This ratio of noise in the signals affects both short-term and long-term variations in the interval distribution. Therefore, both RMSSD and SDNN are unreliable in this signal, while the extracted HR and AVNN are acceptable.



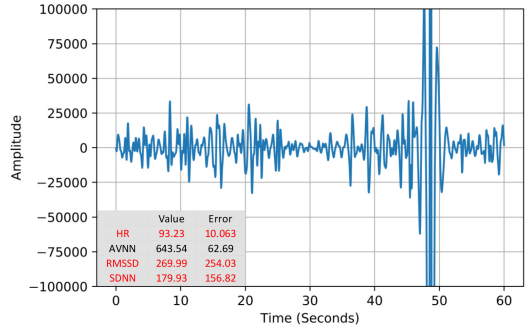
(a) PPG with accurate HR, AVNN, and SDNN, and inaccurate RMSSD



(b) PPG with accurate HR, AVNN, and RMSSD, and inaccurate SDNN



(c) PPG with accurate HR and AVNN, and inaccurate RMSSD and SDNN



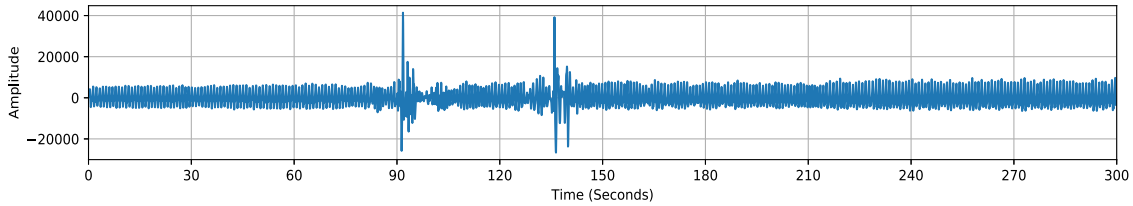
(d) PPG with (relatively) accurate AVNN, and inaccurate HR, RMSSD, and SDNN

Fig. 1. One-minute windows of filtered PPG signals, from which HRV parameters with different accuracy are extracted.

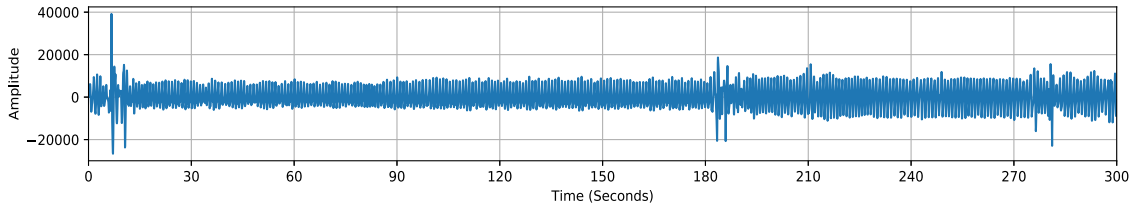
(d) Figure 1(d) depicts a noisy PPG signal. The noise corrupted a considerable part of the signal and affected detected peaks and long-term and short-term variation of NNIs. Therefore, extracted HR, RMSSD, and SDNN are inaccurate. However, the error of the AVNN is relatively low as AVNN represents the mean of the intervals and is more resistant to outliers and noises. This example shows that we can extract some HRV parameters with acceptable accuracy despite the high ratio of noise.

(e) Figure 2(a) and (b) show two 5-minute samples of PPG signal with accurate HR, AVNN, SDNN, and RMSSD. However, the signal indicated in Figure 2(a) is reliable for extracting LF/HF, while the signal in Figure 2(b) is unreliable for the LF/HF. The reason is that noises with same frequency of low- and high-frequency bands of NNIs affect the accuracy of LF/HF. Figure 2(c) and (d) show *power spectral density (PSD)* of NNIs for these two samples in Figure 2(a) and (b), respectively. As indicated, although the two PPG signals (in the time domain) are similar, the power of the NNIs signals in the low-frequency and high-frequency bands are different. As indicated in these figures, noise affects the frequency bands of the signals and, consequently, the accuracy of LF/HF.

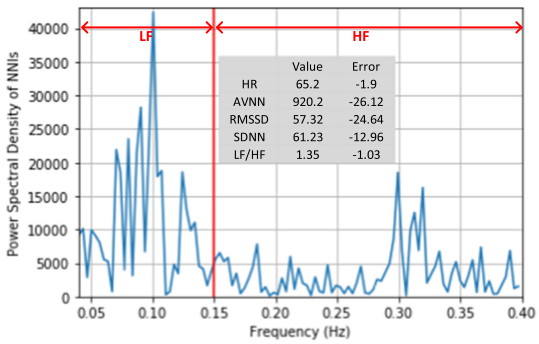
These examples show that HRV parameters are of essence for PPG quality assessment and those PPG quality assessment methods that are solely based on HR or the morphology of the PPG itself are not efficient and can result in many adverse consequences, some of which may be life-threatening due to not fully correct



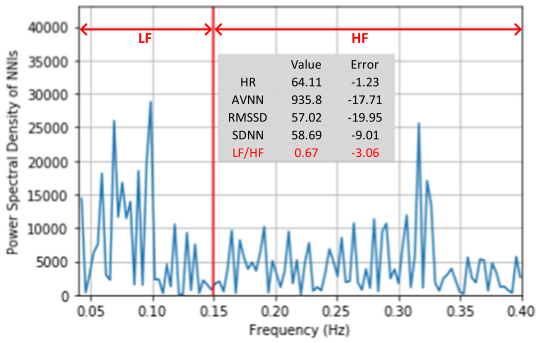
(a) PPG with accurate HR, AVNN, SDNN and RMSSD, and LF/HF



(b) PPG with accurate HR, AVNN, SDNN and RMSSD, and inaccurate LF/HF



(c) PSD of NNIs of the PPG signal in Figure 2a



(d) PSD of NNIs of the PPG signal in Figure 2b

Fig. 2. Two five-minute samples of filtered PPG signals with the corresponding PSD of NNIs.

decisions made by doctors. Moreover, state-of-the-art PPG quality assessment methods were mostly evaluated using the simulated data or data collected in controlled lab settings with limited motion artifacts. We believe that the confidence models need to be trained using the data collected in everyday settings where the subjects engage in several physical activities in various environments. This way, the model can learn about the validity of the signal in different conditions with different artifacts.

3 BACKGROUND

Recent advances in information and communication technology provide an opportunity to enable remote health monitoring using wearable electronics. Such wearables can measure biomedical signals, allowing continuous monitoring of the individual's health condition. ECG is a non-invasive method that can be used to remotely track cardiorespiratory parameters using portable monitors [8, 44]. The method includes limb and chest electrodes, which collect electrical signals generated from the action potentials of the heart cells. The ECG is the gold standard in HR detection and diagnosis of cardiovascular diseases. However, it cannot be performed for long-term monitoring due to its complicated data collection. Alternatively, PPG is a more convenient method to monitor cardiorespiratory variables. The PPG acquires the rate of blood flow in the tissue (e.g., wrist) as controlled by the

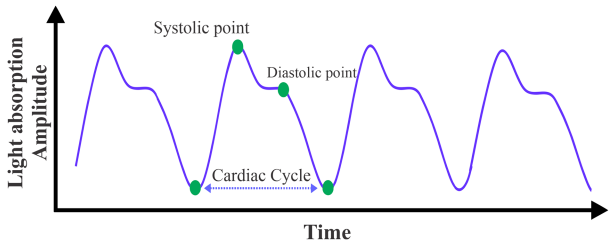


Fig. 3. A window of a filtered PPG waveform.

heart's pumping action. The method leverages an optical sensor in conjunction with a light source to collect the signals. In the following, we outline background on the PPG, HRV, and CNNs as a method we used for the PPG analysis.

3.1 Photoplethysmography

PPG is an optical measurement method that records the variation of blood flow by emitting a light onto the surface of the skin and measuring the light absorption. The PPG signal consists of a pulsatile (AC) component and a non pulsatile (DC) component [6]. The AC component reflects the pulsations in the interrogated blood volume with each heartbeat, whereas the DC component contains the low frequency fluctuations, including absorption from the tissue and bones as well as static blood absorption [45]. The AC component—oscillated with the contraction and relaxation of the heart—enables the measurements of cardiac cycles by detecting the peaks (i.e., maximum values) in this signal. The PPG signal calculated with this procedure can provide the real-time measurements of HR. Moreover, variation in time intervals between the successive pulse peaks, called HRV, allow us to obtain more information about the *Autonomic Nervous System (ANS)*. Figure 3 shows an individual PPG signal.

The PPG method is convenient, economic, and easy to set up [46] using an optical sensor in conjunction with a light source such as a green LED. The method is already used in various commercial and clinical devices. The PPG with green light is utilized in the optical sensors of most consumer wearable devices such as smartwatches for HR calculation. In addition, the PPG with red and infrared LEDs are employed in pulse oximeters to monitor peripheral capillary oxygen saturation (SpO_2).

3.2 HRV Analysis

HRV consists of the fluctuations in the time periods between successive heartbeats [47]. In the literature, HRV analysis has been introduced to examine the ANS correlated with pain intensity, stress level, sleep quality, to name but a few [43, 48–50]. Conventionally, HRV values are calculated from the ECG signal by extracting the cardiac cycles, i.e., the RR-intervals in the signal. Alternatively, the HRV can be also obtained using the PPG signals, where the peak-to-peak intervals—also called NNIs—indicate the cardiac cycles. Studies show that there is a high correlation between the HRV obtained from the ECG and PPG [51–53]. The HRV values are extracted over a period of the ECG/PPG signal, which is in long term over 24 hours, in short term over 5 minutes, or in ultra-short term over 1 minute [49, 54, 55].

The HRV can be obtained from the signal both in time domain and frequency domain. The time domain features of HRV are statistical features that quantify the amount of variability in measurements of the *inter-beat-interval (IBI)*, which is a time interval between adjacent heartbeats. In contrast, the frequency domain features of HRV are estimations of distribution of absolute or relative power into four frequency bands mainly based on PSD. Heart rate oscillations is divided into ultra-low-frequency, *very low-frequency (VLF)*, LF, and HF bands force [56]. Some common HRV features in both time domain and frequency domain are described in Table 2.

Table 2. Time Domain and Frequency Domain HRV Features and Their Descriptions

Feature	Units	Description
AVNN	ms	Mean of NN intervals
SDNN	ms	Standard deviation of NN intervals
RMSSD	ms	Root mean square of successive NN interval differences
SDSD	ms	Standard deviation of successive NN interval differences
nnXX	ms	Number of NN interval differences greater than the specified threshold
pnnXX	%	Percentage of successive NN intervals that differ by more than x ms
VLF power	s^2	Absolute power in very low frequency band (≤0.04)
LF power	s^2	Absolute power in low frequency band (0.04–0.15)
HF power	s^2	Absolute power in high frequency band (0.15–0.4)
LF peak	Hz	Peak frequency in low frequency band (0.04–0.15)
HF peak	Hz	Peak frequency in high frequency band (0.15–0.4)
Total Power	s^2	Total power over all frequency bands
LF/HF	%	Ratio of LF-to-HF power

The HRV features obtained from the PPG show different characteristics within different sampling frequencies. The lower the sampling frequency is, the more variation in the peak locations, meaning the more errors in the HRV analysis [57]. Considering this fact and limitation of our PPG data collection (the sampling rate was 20 Hz), we only focus on the HR and the following four HRV features. These features, which show insignificant error rate at $f_s \geq 20$, are AVNN, RMSSD, SDNN from the time domain and LF/HF (ratio of LF power and HF power) from the frequency domain [57]. We select these HRV in this study, as they are important for various health application such as stress monitoring [49].

3.3 Convolutional Neural Networks

CNN are a class of deep neural networks, also known as the state-of-the-art models for image recognition problems [58]. CNNs are capable of automatically learning spatial hierarchies of features from low- to high-level patterns. CNN is a hierarchical model consists of convolutional, subsampling, and fully connected layers. The first two layers—convolution and subsampling—carry out the feature engineering part and the third layer (a fully connected layer) performs the classification.

The state-of-the-art deep architectures such as VGG, ResNet, and MobileNet implement different approaches for the classification tasks. The VGG was proposed to increase the depth of the convolutional structure of the model for achieving better performance [59]. VGG obtained a top-5 error rate of 7.32%. However, only increasing the depth of the network saturates the accuracy and then degrades it rapidly. Therefore, the ResNet was introduced to address this problem, exploiting the shortcuts or parallel blocks of convolutional filters while building deeper models [60]. ResNet, however, achieved a top-5 error rate of 3.57%. In contrast, MobileNet was proposed as a lightweight model to run deep neural networks on personal mobile devices. MobileNet obtained a top-5 error of 7.5%, almost the same as VGG Network. We leverage the significant performance of these deep architectures for the PPG quality assessment. In this regard, we convert the PPG signals to PPG snapshots to feed the images to the CNNs.

Furthermore, CNNs can be harnessed on one-dimensional time-series data sharing the same characteristics and the same approach as in image-based CNNs [61]. The difference is the structure of the input data and how the filter, i.e., convolution kernel or feature detector, slides across the data. The model learns to extract features using a technique called sliding window with a one-dimensional filter over the time series, followed by a non-linear function to learn non-linear decision boundaries. The model can learn an internal representation of the time-series data automatically.

4 METHODS AND MATERIALS

The data used in this work are part of a multipurpose study on remote health monitoring. In the following, we briefly describe the participants, recruitment, data collection, and data annotation in this study.

4.1 Participants and Recruitment

The study was conducted in southern Finland during July–August 2019 by inviting healthy individuals who were between 18 and 55 years old. The exclusion criteria were if the possible candidates had a diagnosed cardiovascular disease, symptoms of illness during the recruitment, and restrictions on the physical activity or using the devices in the daily routines. The recruitment started by personally contacting students and staff members of the University of Turku. Then snowball sampling was used to reach a convenient number of participants. We aimed for an equal number of female and male participants as gender affects HRV parameters. In face-to-face meetings, the selected candidates were informed about the purpose of the study and the instructions to use the devices, i.e., a Shimmer device [62] and a Samsung Gear Sport smartwatch [63]. Written informed consent forms were also provided to the participants. Forty-six individuals, who agreed to participate in this study, were asked to wear the devices for 24 hours continuously. The data can be shared by signing a data use agreement. This can be obtained from the corresponding author upon request.

4.2 Ethics

The study was conducted according to the ethical principles based on the Declaration of Helsinki and the Finnish Medical Research Act (No. 488/1999). The study protocol received a favorable statement from the ethics committee (Unversity of Turku, Ethics committee for Human Sciences, Statement No. 44/2019). The participants were informed about the study both orally and in writing, before their consent was obtained. Participation was voluntary and all the participants had the right to withdraw from the study at any time and without giving any reason. To compensate the time used for the study, each participant got a gift card to grocery store (20 euro) at the end of the monitoring period when returning the devices.

4.3 Data Collection

We performed home-based ECG and PPG collection, in which the signals were acquired simultaneously and remotely for 24 hours. In this study, the ECG signal was collected, employing the Shimmer3 ECG [62]. The ECG test included four limb leads placed on the left arm, right arm, left leg, and right leg. This device provides raw data using medical grade sensors. It also provides accelerometer, gyroscope, and magnetometer data. Participants were instructed to place the ECG unit on their chest and to attach the four skin electrodes. The 512-Hz sampling rate was used in this study, as suggested for clinical trials.

In addition, the Samsung Gear Sport smartwatch was selected for collecting the PPG signal, considering availability of the raw PPG signals and configurability of the data recording. The watch also has a built-in inertial measurement unit by which daily physical activity data are extracted. Participants were asked to wear the watch on the non-dominant hand continuously and tightly enough. Considering constraints of the battery in the Gear sport watch, we programmed the watch to collect 16 minutes PPG signal in every 30 minutes. In this settings, there is no need to charge the smartwatch during the one-day experiment. Each PPG record contains 1 minute of unreliable data (due to sensor calibration) and 15 minutes PPG signals. The PPG signals were collected with a 20-Hz sampling frequency that is suitable to extract HR and HRV parameters.

4.4 Automatic PPG Annotation

The PPG signals should be annotated to develop a quality assessment method. Traditionally, the signals are manually annotated, where experts label windows of the signals into “reliable” or “unreliable” according to the shape/structure of the signals. As described in Section 2, such a method is inaccurate when multiple parameters

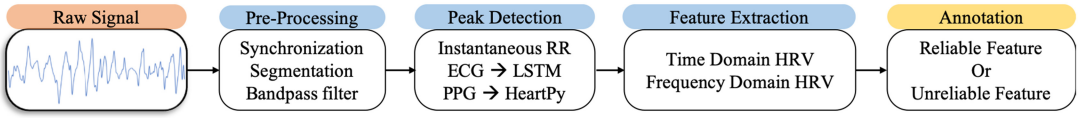


Fig. 4. Automatic PPG annotation pipeline.

are obtained using the PPG. The PPG signals should be labeled according to the accuracy of the desired applications/health parameters. Therefore, different labels should be allocated to a window of the signal, e.g., the window is reliable for HR and SDNN while it is unreliable for RMSSD (see Figures 1 and 2).

To address this issue, we develop an automatic PPG annotation method, where the PPG signals are labeled according to the health parameters. In this regard, the obtained parameters are compared with the parameters extracted from the ECG as the baseline method. The signal is labeled as "reliable" for a parameter if the error is insignificant. Otherwise, it is labeled as "unreliable." A schematic representation of the automatic PPG annotation method is shown in Figure 4.

4.4.1 Pre-Processing. The ECG and PPG signals were collected by two different wearables for 24 hours. Therefore, there might be a time shift between the two signals (e.g., seconds). To address this issue, we first synchronize the two signals. We used a cross-correlation method to synchronize the data provided by the ECG device and the smartwatch. The ECG device and PPG-based smartwatch collected acceleration signals with the same frequency of ECG and PPG signals, respectively. We extracted the cross-correlation of the signal vector magnitudes of the acceleration signals. The output indicated the possible time shift between the two signals. Then, the ECG signals was shifted with respect to the PPG signals if needed. We then segment the 24-hour PPG data into 5-minute windows using sliding technique with an stride of 10 seconds. Considering HR between 30 to 220 beats per minute, a Butterworth filter [64] was set to only pass heartbeat signals (i.e., 0.5–3.7 Hz).

4.4.2 Peak Detection. The first step of the HRV analysis is to calculate the RR intervals. In the ECG analysis, the RR intervals are calculated by detecting the QRS complex—with the highest peak and slope in the signal—and extracting the distance between two adjacent R peaks. For the QRS detection, we used the method proposed by Laitala et al. [65], as it shows more robust and accurate QRS detection in comparison to the traditional QRS detection methods such as Pan-Tompkins [66], Christov [67], Hamilton [68], and Engzee [69, 70]. This method uses a Long Short Term Memory network to obtain the probabilities and locations of the peaks, followed by extra processes to remove outliers based on anomalous peak–peak distances and obtain the valid peaks. The last step is to remove noise from the RR intervals—the time intervals between two successive R-peaks—obtained by subtracting the time of two successive peaks. We used a quotient filter [71] to remove outliers from the RR intervals that are used for the feature extraction.

In the PPG analysis, HRV parameters are obtained from the the subtle change of pulse periods, i.e., IBI, generated as a result of the heart activity. To extract the IBI from the PPG signals, we use a peak detection method proposed by Van Gent et al. [72]. They showed the method obtains an acceptable accuracy to extract HRV values in comparison to reference ECG signals using Pan-Tompkins QRS algorithm [66] and an open source algorithm called HRVAS ECGViewer [73]. This method uses an adaptive threshold to accommodate morphology variation in the PPG waveform, followed by an outlier detection/rejection to extract valid peaks in the signal.

4.4.3 Feature Extraction. The HRV parameters can be calculated using the extracted peaks. In this study, we only extract four HRV parameters, as the PPG sampling frequency is 20 Hz [57] (see Section 3.2). Within each time window of the both ECG and PPG, HR and the four HRV features—three time domain and one frequency domain—are extracted. The time domain metrics are obtained using the NN intervals: AVNN is the Average Value of NN intervals, RMSSD is the Root Mean Square of Successive Differences between normal heartbeats, and SDNN is the Standard Deviation of NN intervals. LF/HF is also the ratio of the low-frequency to

high-frequency power. In our setup, these features are extracted from the ECG and PPG signals using the HeartPy Python package [72].

4.4.4 Labeling. As previously mentioned, the PPG signals are divided into 5-minute windows, from which five features—HR, AVNN, RMSSD, SDNN, and LF/HF—are extracted. The features are tailored to label the PPG windows as "reliable" or "unreliable." For each window, the extracted features are compared with the values obtained from the corresponding ECG signal using an Euclidean distance function. The window is labeled as "reliable" for a feature if the distance is less than a threshold value obtained according to the range of the feature [74]. Otherwise, the window is labeled as "unreliable." Noted that the threshold can be selected according to the desired accuracy of the feature. This process is performed for the five features. Therefore, five binary labels are allocated to each PPG window.

5 PPG QUALITY ASSESSMENT APPROACH

In this section, we present two *deep learning– (DL)* based methods using CNNs for PPG Quality Assessment. CNN architectures are capable of handling the challenging feature engineering of PPG signals automatically. This is an advantage over the traditional PPG Quality Assessment methods that were mostly designed to extract features based on the morphology of the PPG itself. The traditional methods use extracted features to generally classify the PPG signal as "reliable" or "unreliable" solely based on HR; however, we create a specific model for each of the HR and HRV parameters separately. The classification is performed using the labels generated from our automatic ECG-based annotation method. In addition, our data include more scenarios and corner cases with different artifacts compared to a lab setting based data collection, as our data were collected in the course of everyday events.

CNN can perform feature extraction and classification without having any knowledge about the data collection. Moreover, CNN architectures naturally can handle an input with any dimensionality; two-dimensional and one-dimensional inputs are the most common ones. PPG signals as a time series have potential to be converted to an image if a 2D model is desired. With this privilege, PPG signals can be fed into the CNN in a usual 1D time-series signal or in an encoded 2D image. We will leverage three state-of-the-art 2D CNN architectures (VGG16, ResNet50, and MobileNetV2) that are pre-trained on a huge dataset and transfer the extensive knowledge gained from other image classification tasks by repurposing the models and fine-tuning them for our problem.

We train a separate model for each feature extracted from the PPG window individually. The input to each model is a 5-minute PPG signal, and the output is a label obtained via the automatic PPG annotation method. An overview of the CNN-based approaches is shown in Figure 5. This shows that the 5-minute PPG time series and one of the HR-HRV features are fed to 1D CNN in the training phase. Moreover, the pre-trained 2D CNN models (VGG16, ResNet50, and MobileNetV2) are fed with the 5-minute PPG images—encoded using *Gramian Angular Field (GAF)*—along with one of the HR-HRV features. In the following, we describe the CNN-based approaches with different architectures leveraged to perform PPG signal quality assessment.

5.1 1D CNN

We design a customized CNN method trained with the PPG time-series segments as having two convolution layers with one-dimensional filters trailed by a non-linear ReLU activation unit. Our convolutional layers are followed by a batch normalization layer [75], by which the changes in the hidden unit values are reduced. Moreover, the batch normalization reduces overfitting, since it has a slight regularization effect. The output of the convolution block is then fed into a maxpooling layer to reduce the dimensionality of the data. The learned features obtained through the convolutional block are flattened to one long vector and pass through a fully connected neural network before the output layer used to make a prediction. The fully connected layer ideally provides a buffer between the learned features and the output with the intent of interpreting the learned features before making a prediction. For our customized model, we will use a standard configuration of kernel size of 3 and

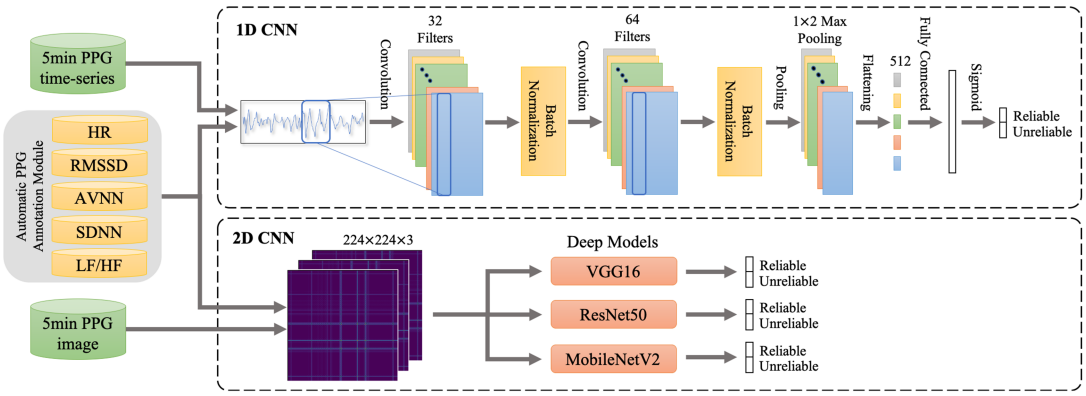


Fig. 5. Overview of the CNN-based architectures for the PPG quality assessment.

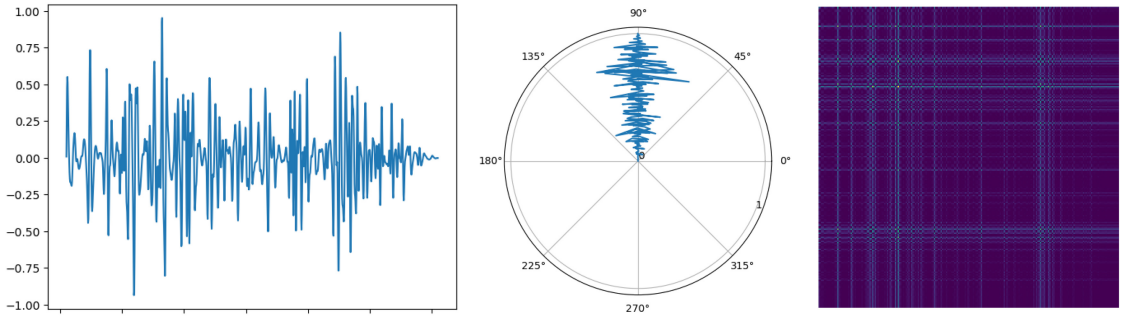
32 and 64 parallel feature maps for the first and second convolutional layers, respectively. The feature maps are the number of times the input is processed or interpreted, whereas the kernel size is the number of input timesteps considered as the input sequence is read or processed onto the feature maps.

A grid search is also carried out to optimize the hyperparameters of the model. The efficient Adam version of stochastic gradient descent with a learning rate of 0.00001 is used to optimize the network, and the binary cross entropy loss function is used given that we are learning a binary-class classification problem. We analyze the performance of the customized 1D CNN model trained separately on every single feature of HR, AVNN, RMSSD, SDNN, and LF/HF to describe the reliability and unreliability of the signal with respect to that feature. For the training phase, 5-minute PPG signals along with the label for each specific feature are fed to the 1D CNN.

5.2 2D CNN

We utilize three powerful pre-trained CNN networks (VGG, ResNet, and MobileNet) and repurpose them and fine-tune them with the PPG images. To create suitable inputs for the CNN, the PPG time series need to be converted to PPG images, while preserving the temporal dependency of the time series. Therefore, we encode the PPG signals as images using the GAF method [76]. The GAF also contains temporal correlations—similarly to an image—making the image proper for the CNN. The GAF is created based on the Gram Matrix defined by the dot product of every couple of vectors. The dot product shows the similarity of the set of vectors. Since the Gram Matrix produces a matrix with the size of time series squared, we need to reduce the dimensionality of the input to decrease the amount of computation. In this regard, we use *Piece-wise Aggregate Approximation (PAA)* to reduce the dimensionality of the input time series by dividing them into equal-sized, non-overlapping windows and extract the average in each segment [77]. To build GAFs, PPG as a time-series signal is scaled into [−1,1] with a Min-Max scaler. Next, PPG length is decreased using the PAA algorithm to a 224×224 image. Then, PPG is converted into the polar coordinates rather than keeping it in the typical Cartesian coordinates. The polar encoding is then followed by a Gram Matrix like operation on the resulting angles. Three stages of the PPG encoding is shown in Figure 6.

To leverage the significant performance of CNNs, we test our approach with three architectures: VGG, ResNet, and MobileNet. VGG stacks the convolutional layers with an increasing number of filters but with the same size of 3×3, since two 3×3 filters almost cover a 5×5 filter and are also more lightweight in multiplications [59]. Among VGG architectures, VGG16 and VGG19 are the most popular, since both followed the same strategy. VGG19 is deeper than VGG16, but it is observed that the accuracy is not improved and saturated. The other architecture that we investigate is ResNet [60]. The success recipe of ResNet for training a deep network is the residual connections, where each layer is connected not only to the previous layer but also the layer behind the



(a) Scaled PPG time-series

(b) PPG time-series in polar coordinates

(c) PPG image converted from polar coordinates using Gramian Angular Field

Fig. 6. Three stages of encoding PPG time series to PPG image with GAF.

previous layer. With this intention, each layer has more information. The last architecture considered in this study is MobileNet [78]. MobileNet brought a novel idea of replacing a standard convolution with a depthwise convolution, followed by a pointwise convolution. This way of convolving performs a solitary convolution on every color channel as opposed to joining every one of the three and smoothing it. This way of convolving helps to build a smaller model and smaller complexity, which makes it suitable to run on mobile devices. In this article, we implement the three 2D CNN models (VGG16, ResNet50, and MobileNetV2) and train them for each HR-HRV feature to describe the reliability and unreliability of the signal with respect to that feature. We pass the encoded PPG images as 2D inputs to these models along with the corresponding label obtained via the automatic PPG annotation method of each feature.

6 EXPERIMENTAL RESULTS

In this section, we investigate the performance of the four approaches (i.e., one 1D CNN and three 2D CNN) described in Section 5. The approaches are utilized to classify the PPG signals according to the reliability of five PPG-based parameters, i.e., HR, AVNN, RMSSD, SDNN, and LF/HF ratio. In this regard, four models are trained for each health parameter, resulting in a total of 20 models.

Moreover, we evaluate the performance of our method in comparison to existing methods. In this regard, our best models for the health parameters are compared with six different PPG signal quality assessment methods. In the following, we first outline our setup and the data used for the training and testing the models. Then, we evaluate and compare the models in terms of their performance.

6.1 Setup

In our setup, we use Keras Sequential API from TensorFlow to train and evaluate the CNN models [79]. We need to pass a couple of parameters to the Keras API including an optimizer, a loss function, and metrics that will be used in the training and validation phase of the model. Adam optimizer [80], binary cross-entropy loss as a loss function, and accuracy, f1-score, and *area under curve (AUC)* metrics are used to compile our CNN models. Our dataset has skew over one label, much more samples are from Reliable class for most of the classification models, which makes our classification problem to be an imbalance binary classification. Therefore, during the training and validation phase, we monitor the training and validation AUC and loss in each epoch [81]. The model with the highest AUC for the validation set during the optimization process is selected as the best model. That model will be the checkpoint for next epochs.

Table 3. Distribution of Train and Test Dataset on Each Annotation Label

	Train (23) + Validation (6)		Test (7)	
Total Segments (# of Subjects)	58,588		14,628	
	Reliable	Unreliable	Reliable	Unreliable
HR	46,175	12,413	11,544	3,104
AVNN	50,925	7,663	12,732	1,916
RMSSD	16,327	42,261	4,082	10,566
SDNN	25,822	32,766	6,456	8,192
LF/HF ratio	42,388	16,200	10,597	4,050

Table 4. Validation Set Accuracy (ACC), f1-score (F1), and AUC Performance Results of CNN Models for Various HR-HRV Features

Label	HR			AVNN			RMSSD			SDNN			LF/HF		
Metric	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
Model															
1D_CNN	**95.63**	**96.11**	**96.21**	**96.71**	**97.68**	**97.71**	**91.42**	**91.48**	**91.69**	**96.01**	**96.97**	**97.09**	**97.71**	**97.71**	**97.71**
MobileNet	94.58	93.93	94.63	95.68	95.95	96.93	89.68	89.44	89.46	92.66	93.91	93.66	91.92	91.95	92.93
ResNet50	92.49	92.44	93.52	90.43	90.22	91.32	86.47	85.81	87.17	92.52	92.12	93.66	90.22	90.22	91.31
VGG16	92.38	94.42	94.87	93.36	94.29	94.85	85.72	85.09	85.79	93.42	93.82	93.99	94.36	94.29	93.99

In this table, any bold value represents higher value in each column.

6.2 Training and Test Data Distribution

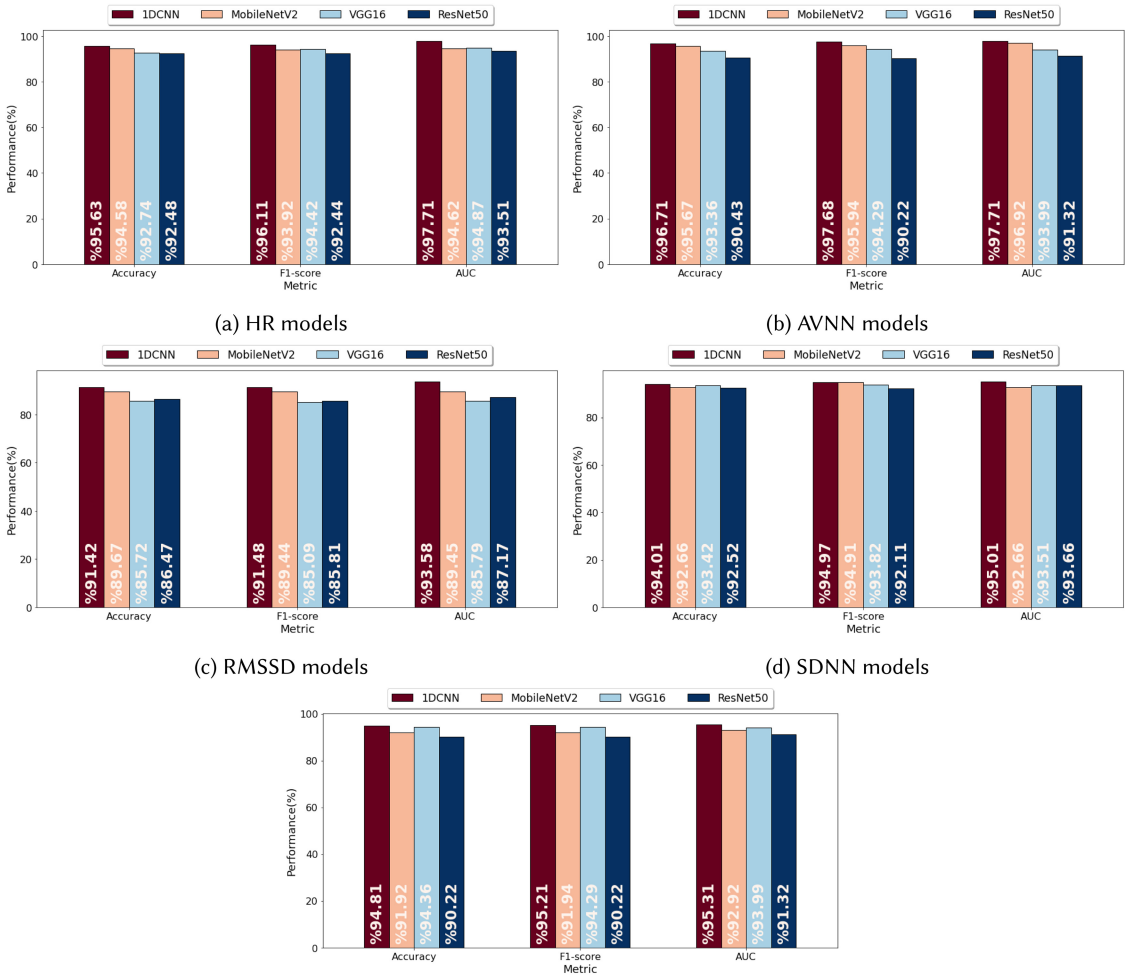
A total of 36 subjects are recruited for this study. We split the dataset into independent train set, validation set, and test set. Twenty percent of the whole dataset (7 subject) is used for test set, and the rest is split into training and validation set, 80% (23 subject) and 20% (6 subject), respectively. For the sake of fair comparison, the models are trained and validated on the same train and validation dataset. To evaluate the performance of the DL models, we use an independent test set. The detail of the distribution of the train, validation, and test set is shown in Table 3.

6.3 Proposed Method Evaluation

As described in Section 5, four CNN models are trained for each PPG parameter, i.e., HR, AVNN, RMSSD, SDNN, and LF/HF. For each parameter, the model with the maximum AUC score is chosen. In the following, we evaluate the models on the validation set and test set. Finally, the best CNN models on the test set are selected to be used for the comparison with the state-of-the-art methods.

6.3.1 HR Models. The overall performance of our proposed CNN models created using the automatic annotated label for the HR feature is shown in Table 4. Classification accuracy, f1-score, and *Receiver Operating Characteristic (ROC)* area of all DL models for the test set is shown in Figure 7(a). It can be seen that for the HR feature the *1D CNN* model works the best among all the DL models with an accuracy of 95.63%, f1-score of 96.10%, and AUC of 96.21%. As a result, the assessment of the reliability of PPG signal w.r.t. HR can be performed with high confidence using a one-dimensional CNN classifier that is less complex compared to the two-dimensional models.

6.3.2 AVNN Models. Table 4 summarizes the overall performance of our proposed CNN models created using the automatic annotated label for the AVNN feature, and Figure 7(b) shows the classification accuracy, f1-score, and ROC area of the DL models for the test set. It can be seen that for the AVNN feature, the DL models show promising performances in all the metrics. However, the *1D CNN* model outperforms the other DL models with



(a) HR models

(b) AVNN models

(c) RMSSD models

(d) SDNN models

(e) LF/HF ratio models

Fig. 7. Comparison of test set accuracy, f1-score, and AUC performance metrics for different DL models.

an accuracy of 96.71%, f1-score of 97.71%, and AUC of 97.71%. Following these results, quality assessment of PPG signals w.r.t. the AVNN can be done using a one-dimensional CNN classifier with a marginal error.

6.3.3 RMSSD Models. The overall performance of the proposed CNN-based models for RMSSD is shown in Table 4. The classification accuracy, f1-score, and ROC area of the DL models for the test set are shown in Figure 7(c). The DL models, built for the RMSSD feature, show a decent performance classifying the PPG signals as Reliable or Unreliable. For this feature, the 2D models show a mediocre performance comparing to the 1D CNN model. The performance of the RMSSD feature is worse than HR and AVNN due to the distribution of Reliable and Unreliable labels, which can be found in Table 3. Unlike HR and AVNN models with a very high Reliable to Unreliable class ratio, RMSSD models have the lowest Reliable to Unreliable ratio. This makes it difficult for the optimization process to find the best model. The *1D CNN* model outperforms the other DL models with an accuracy of 91.42%, f1-score of 91.48%, and AUC of 91.68%. Consequently, PPG quality classification based on the RMSSD feature can still be performed using a one-dimensional CNN classifier.

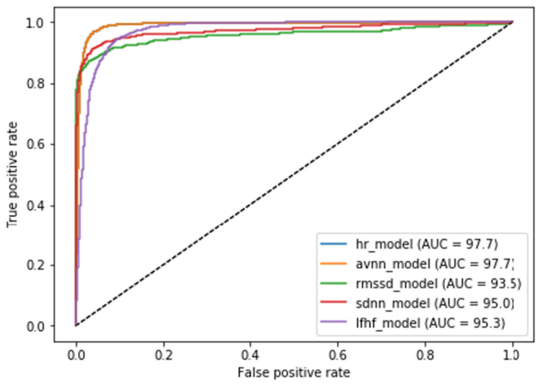


Fig. 8. ROC curve of the reliable class for 1D CNN models of all parameters.

6.3.4 SDNN Models. Table 4 summarizes the overall performance of the proposed models for the SDNN feature. Moreover, Figure 7(d) shows the classification accuracy, f1-score, and ROC area of the DL models for the test set. It can be seen that all the 2D DL models trained for SDNN have a promising performance. The *1D CNN* model is the best classifier among all DL models with an accuracy of 96.01%, f1-score of 96.97%, and AUC of 97.08%. Therefore, the reliability of PPG signal w.r.t SDNN can be assessed with high confidence using a one-dimensional CNN classifier, a smaller and simpler model than the 2D models.

6.3.5 LF/HF Models. The performance of the proposed models for LF/HF is shown in Table 4 (validation set) and in Figure 7(e) (test set). As indicated, the 2D DL models trained for LF/HF parameter have a promising performance; however, *1D CNN* model outperforms the other DL models with an accuracy of 97.71%, f1-score of 97.71% and AUC of 97.71%. As a result, the assessment of the reliability of PPG signal w.r.t. LF/HF can be performed with high confidence using a one-dimensional CNN classifier that is less complex compared to the two-dimensional models.

Figure 8, represents the ROC curves of the best model for each HR-HRV metric along with the AUC of the corresponding feature. It can be seen that different features have different characteristic in terms of PPG signal reliability. The *1D CNN* performs exceptionally well classifying all HR-HRV features models into reliable and unreliable classes. In addition, *MobileNetV2* did also a great job for all HR-HRV models to classify them into reliable and unreliable classes. Our results show a promising performance for the proposed CNN-based approaches, through which a binary decision is delivered to indicate PPG signal quality among five different features, HR, AVNN, RMSSD, SDNN, and LF/HF ratio in a real-time manner.

6.4 Comparison with State-of-the-art Methods

The results in Section 6.3 show that the *1D CNN* model outperforms the other DL models in terms of performance w.r.t to the HR and HRV features. Therefore, the proposed 1D CNN model is selected for the comparison with the state-of-the-art models. As outlined in Section 2, there is a broad variety of PPG signal quality assessment methods in the literature. We compared our proposed method with six different methods. First, a rule-based method [40] is selected for comparison to distinguish low-quality signals, leveraging hierarchical decision rules combined with simple features, such as absolute amplitude, threshold crossing rate, and autocorrelation function features. Second, we compare the proposed method with Support Vector Machine, K nearest neighbors, and decision trees algorithms as supervised machine learning algorithms [28]. Furthermore, we compare our proposed method with an unsupervised method using elliptical envelope [30]. Finally, our proposed method is compared with Xception [31] as a deep learning approach.

Table 5. Comparison between the Proposed Method and the State-of-the-art PPG Quality Assessment Methods on Test Set

Label	HR			AVNN			RMSSD			SDNN			LF/HF		
Metric / model	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
Rule Based [40]	61.19	77.28	67.80	63.08	72.44	68.35	52.13	59.95	61.81	63.08	75.44	71.35	67.65	76.15	65.70
KNN [26]	87.56	92.20	79.66	92.47	95.71	80.69	**93.70**	88.62	92.01	85.78	83.60	85.40	79.56	86.23	72.38
SVM [26]	78.70	88.08	50.00	86.84	92.95	50.00	89.54	79.52	84.46	78.39	75.84	78.24	72.42	84.00	50.00
DT [26]	78.70	88.08	50.00	86.84	92.95	50.00	89.54	79.52	84.46	78.39	75.84	78.24	72.42	84.00	50.00
Eliptical Envelope [30]	71.54	81.94	57.39	78.89	87.85	53.64	67.35	41.28	59.33	61.76	56.54	61.19	62.54	74.14	53.10
Xception [31]	89.30	90.28	85.80	89.54	90.44	85.35	81.03	70.95	77.81	88.54	90.44	85.35	85.09	88.15	84.70
Proposed Method	**95.64**	**96.12**	**97.71**	**96.71**	**97.68**	**97.71**	91.43	**91.48**	**93.59**	**94.02**	**94.98**	**95.02**	**94.82**	**95.22**	**95.31**

In this table, any bold value represents higher value in each column.

Table 5 shows the performance of the proposed method and the aforementioned state-of-the-art algorithms. The state-of-the-art quality assessment algorithms are only defined for the PPG signal itself and HR feature. However, our proposed method can assess quality of the PPG signals based on HR and HRV features using separate models. To make a fair comparison, the labels of the state-of-the-art algorithms are created using our automatic annotated labeling method described in Section 4.4.4. The proposed method (with the 1D CNN architecture) results in more accurate signal quality assessment compared to the state-of-the-art. As indicated in the table, the rule-based and the unsupervised algorithms had the lowest overall performance. However, supervised traditional machine learning methods, such as KNN, showed promising results for AVNN and RMSSD. The accuracy of KNN for the RMSSD feature was slightly better than the proposed method. However, for the other features/metrics, the proposed method performs considerably better. The Xception algorithm, as a deep learning approach, had better performance, in general, than the traditional supervised machine learning methods. It was able to pursue a stable performance around 90%. However, the proposed method outperformed the Xception algorithm. However, the rule-based method [40] is the only algorithm that requires no training phase.

6.5 Limitations and Future Works

This study is limited to HR and a couple of HRV parameters. Future work should include the quality assessment of PPG signal w.r.t. to other frequency domain HRV parameters such as LF power, HF power, and total power and nonlinear domain HRV features such as SD1 and SD2 (Standard Deviation of major and minor axis in poincare plot) and sample entropy. These HRV features might have potential use cases in various healthcare applications.

The length of the data collection is also another limitation of this study. In this study we collected the data for a period of 24 hours. Future work can be a long-term data collection for such signal quality assessment studies.

Another limitation of our study is that the results does not generalize to unhealthy individuals, as the data was collected from healthy individuals and models were trained with no prior information about abnormality in PPG signals. There are studies in the literature showing that the reliability of PPG signals from wearable devices may be different for different population groups [82, 83]. For instance, patients with cardio vascular diseases have abnormal PPG morphology with irregular heart beats [47]. This abnormality results in different PPG parameters, i.e., HR and HRV features compared to the healthy individuals [84]. As a result, further investigation is required in future work to perform PPG quality assessment of unhealthy subjects.

7 CONCLUSION

PPG is a non-invasive optical method utilized in various wearable devices, enabling home-based health monitoring systems, to continuously acquire vital signs such as HR and HRV. PPG is highly susceptible to motion artifacts and environmental noise. Low-quality PPG signals negatively impact the accuracy of the extracted health parameters, leading to inaccurate decision-making. Different studies introduced approaches considering morphological features of the PPG to determine decision rules for designing PPG quality assessment systems. Such methods can be only applied on HR detection application, since having an acceptable shape is insufficient in measuring HRV parameters. We proposed DL-based PPG quality assessment methods using CNNs for (1) HR, (2) AVNN, (3) RMSSD, (4) SDNN, and (5) LF/HF ratio. We utilized our customized 1D CNN and three 2D deep neural networks architectures. We presented an automatic annotation method for each HR-HRV parameter, where the PPG quality was labeled automatically as "reliable" or "unreliable" by comparing against an ECG signal, as the baseline method for HR and HRV measurements. We evaluated our trained models using 210 hours of PPG data collected from a home-based health monitoring application on an independent test dataset. Our results indicated that the proposed 1D CNN method outperforms the other 2D CNN approaches. The accuracy of the customized 1D CNN was 95.63%, 96.71%, 91.42%, 96.01%, and 97.71 for the HR, AVNN, RMSSD, SDNN, and LF/HF ratio, respectively. In addition, we compared the performance of our proposed method with six different state-of-the-art signal quality assessment methods. Our results indicate that the proposed method performs considerably better (at least 6% improvement in all performance metrics) than the existing algorithms. We also provided a portable and open source model implemented in Python for the community to be integrated into their solutions.

REFERENCES

[1] J. Gubbi et al. 2013. Internet of Things (IoT): A vision, architectural elements, and future directions. *Fut. Gener. Comput. Syst.* 29, 7 (2013), 1645–60.

[2] R. Mieronkoski et al. 2017. The Internet of Things for basic nursing care—A scoping review. *Int. J. Nurs. Stud.* 69 (2017), 78–90.

[3] M. M. Baig and H. Gholamhosseini. 2013. Smart health monitoring systems: An overview of design and modeling. *J. Med. Syst.* 37, 2 (2013), 9898.

[4] Iman Azimi, Olugbenga Oti, Sina Labbaf, Hannakaisa Niela-Vilén, Anna Axelin, Nikil Dutt, Pasi Liljeberg, and Amir M Rahmani. 2019. Personalized maternal sleep quality assessment: An objective IoT-based longitudinal study. *IEEE Access* 7 (2019), 93433–93447.

[5] Milad Asgari Mehrabadi, Iman Azimi, Fatemeh Sarhaddi, Anna Axelin, Hannakaisa Niela-Vilén, Saana Myllyntausta, Sari Stenholm, Nikil Dutt, Pasi Liljeberg, and Amir M Rahmani. 2020. Sleep tracking of a commercially available smart ring and smartwatch against medical-grade actigraphy in everyday settings: Instrument validation study. *JMIR Mhealth Uhealth* 8, 10 (2 Nov. 2020), e20465. DOI: http://dx.doi.org/10.2196/20465

[6] J. Allen. 2007. Photoplethysmography and its application in clinical physiological measurement. *Physiol. Meas.* 28, 3 (2007), 1–39.

[7] T. Tamura et al. 2014. Wearable photoplethysmographic sensors—Past and present. *Electronics* 3, 2 (2014), 282–302.

[8] Sumit Majumder, Tapas Mondal, and M Jamal Deen. 2017. Wearable sensors for remote health monitoring. *Sensors* 17, 1 (2017), 130.

[9] A. Pantelopoulos and N. G. Bourbakis. 2010. A survey on wearable sensor-based systems for health monitoring and prognosis. *Trans. Syst. Man Cyber. C* 40, 1 (2010), 1–12.

[10] N. Constant et al. 2015. Pulse-glasses: An unobtrusive, wearable HR monitor with Internet-of-Things functionality. In *Proceedings of the IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks*. IEEE, Cambridge, MA, USA, 1–5.

[11] Delaram Amiri, Arman Anzanpour, Iman Azimi, Marco Levorato, Amir M. Rahmani, Pasi Liljeberg, and Nikil Dutt. 2018. Edge-assisted sensor control in healthcare IoT. In *Proceedings of the IEEE Global Communications Conference (GLOBECOM'18)*. IEEE, 1–6. DOI: http://dx.doi.org/10.1109/GLOCOM.2018.8647457

[12] Delaram Amiri, Arman Anzanpour, Iman Azimi, Marco Levorato, Pasi Liljeberg, Nikil Dutt, and Amir M. Rahmani. 2020. Context-aware sensing via dynamic programming for edge-assisted wearable systems. *ACM Trans. Comput. Healthcare* 1, 2 (2020), 1–25.

[13] Seyed Amir Hossein Aqajari, Rui Cao, Amir Hosein Afandizadeh Zargari, and Amir M. Rahmani. 2021. An end-to-end and accurate PPG-based respiratory rate estimation approach using cycle generative adversarial networks. In *Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'21)*. IEEE, 744–747.

[14] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 2302–2310.

[15] C. Zong and R. Jafari. 2015. Robust heart rate estimation using wrist-based PPG signals in the presence of intense physical activities. In *Proceedings of the IEEE Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'15)*. IEEE, 8078–8082.

[16] H. Han et al. 2007. Development of real-time motion artifact reduction algorithm for a wearable photoplethysmography. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'07)* IEEE, Lyon, France, 1538–1541.
[17] Simhadri Vadrevu and M. Sabarimalai Manikandan. 2019. Real-time PPG signal quality assessment system for improving battery life and false alarms. *IEEE Trans. Circ. Syst. II: Expr. Briefs* 66, 11 (2019), 1910–1914.
[18] M. Elgendi. 2016. Optimal signal quality index for photoplethysmogram signals. *Bioengineering* 3, 4 (2016), 21.
[19] Jiajia Song, Dan Li, Xiaoyuan Ma, Guowei Teng, and Jianming Wei. 2019. PQR signal quality indexes: A method for real-time photoplethysmogram signal quality estimation based on noise interferences. *Biomed. Sign. Process. Contr.* 47 (2019), 88–95.
[20] Hyeon Seok Seok, Sangjin Han, Junyung Park, Donggeun Roh, and Hangsik Shin. 2018. Photoplethysmographic pulse quality assessment methods based on similarity analysis. In *Proceedings of the Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS'18) and 19th International Symposium on Advanced Intelligent Systems (ISIS'18).* IEEE, 350–353.
[21] X. Sun et al. 2012. Assessment of photoplethysmogram signal quality using morphology integrated with temporal information approach. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'12).* IEEE, 3456–3459.
[22] Gabriele B. Papini, Pedro Fonseca, Xavier L. Aubert, Sebastiaan Overeem, Jan W. M. Bergmans, and Rik Vullings. 2017. Photoplethysmography beat detection and pulse morphology quality assessment for signal reliability estimation. In *Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'17).* IEEE, 117–120.
[23] Arlene John, Barry Cardiff, and Deepu John. 2020. A generalized signal quality estimation method for IoT sensors. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS'20).* IEEE, 1–5.
[24] Shahnawaz Alam, Shreyasi Datta, Anirban Dutta Choudhury, and Arpan Pal. 2017. Sensor agnostic photoplethysmogram signal quality assessment using morphological analysis. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services.* Association for Computing Machinery, New York, NY, 176–185.
[25] Yue Zhang and Junjun Pan. 2017. Assessment of photoplethysmogram signal quality based on frequency domain and time series parameters. In *Proceedings of the 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI'17).* IEEE, 1–5.
[26] Tania Pereira, Kais Gadhoumi, Mitchell Ma, Rene Colorado, Kevin J. Keenan, Karl Meisel, and Xiao Hu. 2018. Robust assessment of photoplethysmogram signal quality in the presence of atrial fibrillation. In *Proceedings of the Computing in Cardiology Conference (CinC'18)*, Vol. 45. IEEE, 1–4.
[27] Elyas Sabeti, Narathip Reamaroon, Michael Mathis, Jonathan Gryak, Michael Sjoding, and Kayvan Najarian. 2019. Signal quality measure for pulsatile physiological signals using morphological features: Applications in reliability measure for pulse oximetry. *Inf. Med. Unlocked* 16 (2019), 100222.
[28] Tania Pereira, Kais Gadhoumi, Mitchell Ma, Liu Xiuyun, Ran Xiao, Rene A. Colorado, Kevin J. Keenan, Karl Meisel, and Xiao Hu. 2019. A supervised approach to robust photoplethysmography quality assessment. *IEEE J. Biomed. Health Inf.* 45 (2019), 1–4.
[29] Q. Li and G. D. Clifford. 2012. Dynamic time warping and machine learning for signal quality assessment of pulsatile signals. *Physiol. Meas.* 33 (2012), 1491–1501.
[30] Aysan Mahmoudzade, Iman Azimi, Amir M. Rahmani, and Pasi Liljeberg. 2021. Lightweight photoplethysmography quality assessment for real-time IoT-based health monitoring using unsupervised anomaly detection. *Proced. Comput. Sci.* 184 (2021), 140–147.
[31] Tania Pereira, Cheng Ding, Kais Gadhoumi, Nate Tran, Rene A. Colorado, Karl Meisel, and Xiao Hu. 2019. Deep learning approaches for plethysmography signal quality assessment in the presence of atrial fibrillation. *Physiol. Meas.* 40, 12 (2019), 125002.
[32] Jessica Torres Soto and Euan Ashley. 2020. DeepBeat: A multi-task deep learning approach to assess signal quality and arrhythmia detection in wearable devices. arXiv:2001.00155. Retrieved from https://arxiv.org/abs/2001.00155
[33] Emad Kasaeyan Naeini, Iman Azimi, Amir M. Rahmani, Pasi Liljeberg, and Nikil Dutt. 2019. A real-time PPG quality assessment approach for healthcare Internet-of-Things. *Proc. Comput. Sci.* 151, C (2019).
[34] Choon-Hian Goh, Li Kuo Tan, Nigel H. Lovell, Siew-Cheok Ng, Maw Pin Tan, and Einly Lim. 2020. Robust PPG motion artifact detection using a 1-D convolution neural network. *Comput. Methods Progr. Biomed.* 196 (2020), 105596.
[35] Donggeun Roh and Hangsik Shin. 2021. Recurrence plot and machine learning for signal quality assessment of photoplethysmogram in mobile environment. *Sensors* 21, 6 (2021), 2188.
[36] Philip Schmidt et al. 2018. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI'18).* Association for Computing Machinery, New York, NY.
[37] Christina Orphanidou, Timothy Bonnici, Peter Charlton, David Clifton, David Vallance, and Lionel Tarassenko. 2014. Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring. *IEEE J. Biomed. Health Inf.* 19, 3 (2014), 832–838.
[38] Gangireddy Narendra Kumar Reddy, M. Sabarimalai Manikandan, and NVL Narasimha Murty. 2020. On-device integrated PPG quality assessment and sensor disconnection/saturation detection system for IoT health monitoring. *IEEE Trans. Instrum. Meas.* 69, 9 (2020).
[39] Konstantin Tyapochkin, Evgeniya Smorodnikova, and Pavel Pravdin. 2019. Smartphone PPG: Signal processing, quality assessment, and impact on HRV parameters. In *Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'19).* IEEE, 4237–4240.
[40] Simhadri Vadrevu and M. Sabarimalai Manikandan. 2019. A new quality-aware quality-control data compression framework for power reduction in IoT and smartphone PPG monitoring devices. *IEEE Sens. Lett.* 3, 7 (2019), 1–4.

[41] K. Li et al. 2012. Onboard tagging for real-time quality assessment of photoplethysmograms acquired by a wireless reflectance pulse oximeter. *IEEE Trans. Biomed. Circ. Syst.* 6, 1 (2012), 54–63.
[42] Monalisa Singha Roy, Rajarshi Gupta, and Kaushik Das Sharma. 2020. Photoplethysmogram signal quality evaluation by unsupervised learning approach. In *Proceedings of the IEEE Applied Signal Processing Conference (ASPCON'20)*. IEEE, 6–10.
[43] Emad Kasaeyan Naeini, Sina Shahhosseini, Ajan Subramanian, Tingjue Yin, Amir M. Rahmani, and Nikil Dutt. 2019. An edge-assisted and smart system for real-time pain monitoring. In *Proceedings of the IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE'19)*. IEEE, 47–52.
[44] Vega Pradana Rachim and Wan-Young Chung. 2016. Wearable noncontact armband for mobile ECG monitoring system. *IEEE Trans. Biomed. Circ. Syst.* 10, 6 (2016), 1112–1118.
[45] Michael T. Petterson, Valerie L. Begnoche, and John M. Graybeal. 2007. The effect of motion on pulse oximetry and its clinical significance. *Anesthes. Analges.* 105, 6 (2007), S78–S84.
[46] P. Cheang and P. Smith. 2003. An overview of non-contact photoplethysmography. *Electr. Syst. Contr. Div. Res.* 1, 1 (2003).
[47] Rollin McCraty and Fred Shaffer. 2015. Heart rate variability: New perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk. *Glob. Adv. Health Med.* 4, 1 (2015), 46–61.
[48] Mingzhe Jiang, Riitta Mieronkoski, Amir M Rahmani, Nora Hagelberg, Sanna Salanterä, and Pasi Liljeberg. 2017. Ultra-short-term analysis of heart rate variability for real-time acute pain monitoring with wearable electronics. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM'17)*. IEEE, 1025–1032.
[49] Lizawati Salahuddin, Jaegeol Cho, Myeong Gi Jeong, and Desok Kim. 2007. Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. In *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 4656–4659.
[50] Sandrine Devot, Anna M. Bianchi, Elke Naujoka, Martin O. Mendez, Andreas Braurs, and Sergio Cerutti. 2007. Sleep monitoring through a textile recording system. In *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2560–2563.
[51] X. F. Teng and Y. T. Zhang. 2003. Study on the peak interval variability of photoplethysmogtaphic signals. In *Proceedings of the IEEE EMBS Asian-Pacific Conference on Biomedical Engineering*. IEEE, 140–141.
[52] Nandakumar Selvaraj, Ashok Jaryal, Jayashree Santhosh, Kishore K. Deepak, and Sneh Anand. 2008. Assessment of heart rate variability derived from finger-tip photoplethysmography as compared to electrocardiography. *J. Med. Eng. Technol.* 32, 6 (2008), 479–484.
[53] M. Bolanos, H. Nazeran, and E. Haltiwanger. 2006. Comparison of heart rate variability signal features derived from electrocardiography and photoplethysmography in healthy individuals. In *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, New York, NY, 4289–4294.
[54] Musa Sesay, Georges Robin, Patrick Tauzin-Fin, Oumar Sacko, Edouard Gimbert, Jean-Rodolphe Vignes, Dominique Liguoro, and Karine Nouette-Gaulain. 2015. Responses of heart rate variability to acute pain after minor spinal surgery: Optimal thresholds and correlation with the numeric rating scale. *J. Neurosurg. Anesthesiol.* 27, 2 (2015), 148–154.
[55] Arto J. Hautala, Jaro Karppinen, and Tapio Seppänen. 2016. Short-term assessment of autonomic nervous system as a potential tool to quantify pain experience. In *Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'16)*. IEEE, 2684–2687.
[56] A. John Camm, Marek Malik, J. Thomas Bigger, Günter Breithardt, Sergio Cerutti, R. J. Cohen, Philippe Coumel, E. L. Fallen, H. L. Kennedy, R. E. Kleiger, et al. 1996. Heart rate variability: Standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. *Circulation.* 93, 5 (1996), 1043–1065.
[57] Ahyoung Choi and Hangsik Shin. 2017. Photoplethysmography sampling frequency: Pilot assessment of how low can we go to analyze pulse rate variability with reliability? *Physiol. Meas.* 38, 3 (2017), 586.
[58] Y. LeCun and Y. Bengio. 1998. Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, MA, 255–258.
[59] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. Retrieved from https://arxiv.org/abs/1409.1556
[60] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 770–778.
[61] Nagarajan Ganapathy, Ramakrishnan Swaminathan, and Thomas M. Deserno. 2018. Deep learning on 1-D biosignals: A taxonomy-based survey. *Yearbk. Med. Inf.* 27, 01 (2018), 098–109.
[62] Shimmer. ECG Sensor Development Kit, Wearable ECG Sensor, Wireless ECG. Retrieved January 2019 from https://www.shimmersensing.com/products/ecg-development-kit
[63] Samsung. Samsung Gear Sport Smartwatch. Retrieved March 2020 from https://www.samsung.com/global/galaxy/gear-sport/
[64] I. W. Selesnick and C. S. Burrus. 1998. Generalized digital Butterworth filter design. *IEEE Trans. Sign. Process.* 46, 6 (1998), 1688–1694.

[65] Juho Laitala, Mingzhe Jiang, Elise Syrjälä, Emad Kasaeyan Naeini, Antti Airola, Amir M. Rahmani, Nikil D. Dutt, and Pasi Liljeberg. 2020. Robust ECG R-peak detection using LSTM. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*. Association for Computing Machinery, New York, NY, 1104–1111.
[66] Jiapu Pan and Willis J. Tompkins. 1985. A real-time QRS detection algorithm. *IEEE Trans. Biomed. Eng.* 32, 3 (1985), 230–236.
[67] Ivaylo I. Christov. 2004. Real time electrocardiogram QRS detection using combined adaptive threshold. *Biomed. Eng. Online* 3, 1 (2004), 28.
[68] Pat Hamilton. 2002. Open source ECG analysis. In *Computers in Cardiology*. IEEE, 101–104.
[69] Willem A. H. Engelse and Cees Zeelenberg. 1979. A single scan algorithm for QRS-detection and feature extraction. *J. Biomed. Sci. Eng.* 1, 1 (1979), 37–42.
[70] André Lourenço, Hugo Silva, Paulo Leite, Renato Lourenço, and Ana L. N. Fred. 2012. Real time electrocardiogram segmentation for finger based ECG biometrics. In *Biosignals*. Semantic Scholar, NA, 49–54.
[71] Jarosław Piskorski and Przemysław Guzik. 2005. Filtering poincare plots. *Comput. Methods Sci. Technol.* 11, 1 (2005), 39–48.
[72] Paul van Gent, Haneen Farah, Nicole van Nes, and Bart van Arem. 2019. HeartPy: A novel heart rate algorithm for the analysis of noisy signals. *Transport. Res. Part F: Traffic Psychol. Behav.* 66 (2019), 368–378.
[73] Ramshur. ECG Viewer. Retrieved March 2020 from https://github.com/jramshur/ECG_Viewer/
[74] Tania Pereira, Pedro R. Almeida, Joao P. S. Cunha, and Ana Aguiar. 2017. Heart rate variability metrics for fine-grained stress level assessment. *Comput. Methods Progr. Biomed.* 148 (2017), 71–80.
[75] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*. PMLR, 448–456.
[76] Zhiguang Wang and Tim Oates. 2015. Imaging time-series to improve classification and imputation. In *Proceedings of the 24rth International Joint Conference on Artificial Intelligence*. AAAI Press, 3939–3945.
[77] Eamonn J. Keogh and Michael J. Pazzani. 2000. Scaling up dynamic time warping for datamining applications. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, 285–289.
[78] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Los Alamitos, CA, 4510–4520.
[79] M. Abadi et al. 2015. Tensorflow: Large-Scale Machine Learning on Heterogeneous Systems. Retrieved from https://www.tensorflow.org/
[80] D. Kinga, Jimmy Ba Adam, et al. 2015. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, Vol. 5, San Diego, CA, 6 page.
[81] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manage.* 45, 4 (2009), 427–437.
[82] Jesse D. Cook, Michael L. Prairie, and David T. Plante. 2018. Ability of the multisensory jawbone UP3 to quantify and classify sleep in patients with suspected central disorders of hypersomnolence: A comparison against polysomnography and actigraphy. *J. Clin. Sleep Med.* 14, 5 (2018), 841–848.
[83] Jesse D. Cook, Michael L. Prairie, and David T. Plante. 2017. Utility of the fitbit flex to evaluate sleep in major depressive disorder: A comparison against polysomnography and wrist-worn actigraphy. *J. Affect. Disord.* 217 (2017), 299–305.
[84] Jay Chen, Stephen L. Wasmund, and Mohamed H. Hamdan. 2006. Back to the future: The role of the autonomic nervous system in atrial fibrillation. *Pacing Clin. Electrophysiol.* 29, 4 (2006), 413–421.

Received 18 June 2021; revised 8 March 2023; accepted 14 July 2023