Respiration Rate Estimation via Smartwatch-based Photoplethysmography and Accelerometer Data: A Transfer Learning Approach

KIANOOSH KAZEMI*, Department of Computing, University of Turku, Finland
IMAN AZIMI, Department of Computer Science, University of California, Irvine, USA
PASI LILJEBERG, Department of Computing, University of Turku, Finland
AMIR M. RAHMANI, Department of Computer Science, University of California, Irvine, USA

Respiration Rate (RR) is a biomarker for several illnesses that can be extracted from biosignals, such as photoplethysmogram (PPG) and accelerometers. Smartwatch-based PPG signals are more prone to noise interference, particularly within their lower frequency spectrum where respiratory data is embedded. Therefore, existing methods are insufficient for extracting RR from PPG data collected from wrists reliably. Additionally, accelerometer sensors embedded in smartwatches capture respiration-induced motion and can be integrated with PPG signals to improve RR extraction. This paper proposes a deep learning-based model to extract RR from raw PPG and accelerometer signals captured via a smartwatch. The proposed network combines dilated residual inception module and Multi-Scale convolutions. We propose a pre-trained foundation model for smartwatch-based RR extraction and apply a transfer learning technique to enhance the generalizability of our method across different datasets. We test the proposed method using two public datasets (i.e., WESAD and PPG-DaLiA). The proposed method shows the Mean Absolute Error (MAE) of 2.29 and 3.09 and Root Mean Squared Errors (RMSE) of 3.11 and 3.79 across PPG-DaLiA and WESAD datasets, respectively. In contrast, the best results obtained by the existing methods are an MAE of 2.68, an RMSE of 3.5 for PPG-DaLiA, an MAE of 3.46, and an RMSE of 4.02 for WESAD datasets.

CCS Concepts: • **Computing methodologies** → **Neural networks;Ubiquitous and mobile computing systems and tools**; • **Applied computing** → **Health informatics**.

Additional Key Words and Phrases: PPG, Respiratory Rate, Transfer Learning, Convolutional Neural Network, Wearable Devices

ACM Reference Format:
Kianoosh Kazemi, Iman Azimi, Pasi Liljeberg, and Amir M. Rahmani. 2025. Respiration Rate Estimation via Smartwatch-based Photoplethysmography and Accelerometer Data: A Transfer Learning Approach. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 1, Article 7 (March 2025), 24 pages. https://doi.org/10.1145/3712280

1 Introduction

Respiratory Rate (RR) refers to the amount of breaths taken per minute by a person. Adults typically breathe 12 to 20 breaths per minute at rest [1]. Healthcare professionals use RR as one vital sign to determine a patient's clinical status [38]. There are several physiological conditions associated with deviations from the standard RR,

*Corresponding author.

Authors' Contact Information: Kianoosh Kazemi, kianoosh.k.kazemi@utu.fi, Department of Computing, University of Turku, Turku, Finland; Iman Azimi, Department of Computer Science, University of California, Irvine, Irvine, California, USA, azimii@uci.edu; Pasi Liljeberg, Department of Computing, University of Turku, Turku, Finland, pasi.liljeberg@utu.fi; Amir M. Rahmani, Department of Computer Science, University of California, Irvine, Irvine, California, USA, a.rahmani@uci.edu.

ACM 2474-9567/2025/3-ART7
https://doi.org/10.1145/3712280

including anxiety, hypoxia, hypercapnia, respiratory acidosis to name a few [9]. In numerous studies, respiration rate has been shown to be an important indicator of cardiac arrest and deterioration of the patient [11, 12]. Thus, it is essential to monitor RR to evaluate the health condition of patients in medical facilities and home settings.

Recent studies have proposed using physiological signals such as Photoplethysmograms (PPG) and Electrocardiograms (ECG) in addition to inertial measurement unit (IMU) sensor-based signals (i.e., Accelerometers (ACC) and Gyroscopes (GYR)) to estimate respiratory rate [4, 6, 9, 22, 45]. The existing biosignals-derived respiration rate methods can be classified into three main categories: Waveform Analysis, conventional Machine Learning (ML) and Deep Learning (DL) techniques.

Waveform analysis approaches are commonly used for RR estimation in biomedical signals (e.g., ECG, PPG, and ACC). There are several steps involved in waveform analysis, including filtering, signal quality evaluation, time/frequency domain analysis, respiratory-induced waveform extraction, and fiducial point extraction [22, 37]. In this method, the respiratory-dominated signals are first extracted from the PPG signal. Then, the respiratory-induced signal pass through Time/Frequency domain techniques for RR estimation. Time domain techniques commonly use an adaptive threshold for peak and trough detection and subsequently RR estimation [50]. On the other hand, the frequency domain analysis involves identifying breathing frequencies from a power spectrum using mainly fast Fourier transforms (FFTs) [18] and auto-regression analysis [44]. Moreover, multiple RR estimate methods can be fused to improve the robustness of a final RR estimate [21]. In spite of their ease of implementation and speed, these methods are sensitive to noise and need manual tuning and handcrafted rules customized for specific patient populations.

In addition, conventional ML methods are proposed for RR estimation. These approaches utilize temporal and spectral features extracted from PPG and respiratory-induced signals as training data for ML models aimed at predicting RR. For example, in [45], frequency domains and morphological features were extracted from PPG and respiration-induced signals and fed to different learning models, including Random Forest, Gradient Boosting, and Support Vector Machine for RR prediction. In another study [57], the feature extraction step involves segmentation, filtering for motion artifact removal, and extraction of time and frequency-domain features. Then, these extracted features were used to train an extreme gradient boosting regression model to estimate RR.

In recent advancements, deep learning methods have emerged for the estimation of RR [4, 6, 64]. These techniques extract temporal and spatial information from raw PPG signals and yield RR estimates. For example, in [4], Cycle Generative Adversarial Networks (GANs) were used for the generation of respiration signals. They harness the ability of Cycle GAN to translate the PPG signals into respiration signals, equipped with an optimization function that uses the error between the reconstructed and ground truth signals to ensure the reconstructed signals' precision. For automatic and accurate RR estimation, a deep learning approach was introduced in [6]. RR estimation from raw PPG signals was carried out with an end-to-end learning model that used the ResNet block as a backbone and fully connected layers. To decrease the computation time, in [64], a lightweight end-to-end deep learning RR estimator was proposed. They leverage depthwise separable convolutions to extract high-dimensional features from PPG signals and fully connected layers to convert the extracted features from the convolution blocks to RR.

Existing ML and DL algorithms have been designed to estimate RR from PPG signals. However, the majority of these techniques fail when PPG has a low signal-to-noise ratio, leading to inaccurate extracted RRs. These methods have been developed and applied to PPG signals acquired in a stationary position (i.e., with minimal or no movement) via, for example, a pulse oximeter worn by a patient in an intensive care unit [23, 44]. Furthermore, the existing RR estimation methods are designed for fingertips-based PPG signals. Choosing the measurement site is an important factor in the quality of the PPG signal collected. Wrist-based PPG signals, in contrast to their finger-based counterparts, are more susceptible to noise and motion artifacts, especially in the low-frequency range associated with RR [19]. Accordingly, the estimation of RR using wrist-based PPG signals is a more challenging task compared to estimation via finger-based PPG signals.

Wearable technology, specifically smartwatches and fitness trackers, has become increasingly popular in recent years. Sensors embedded in these devices, such as PPGs and inertial measurement units, enable efficient and ubiquitous acquisition of a wide range of physiological signals. Studies have shown that these signals can provide useful information about respiration-induced motion [29, 58], even though they are prone to noise caused by wrist movements during daily routines in free-living environments. There is a need for a noise-resilient RR estimation method for wrist-based devices that integrates PPG and IMU-based signals. This integration will leverage the temporal and spatial information of these signals to achieve a robust and noise-resistant RR estimation.

Moreover, the existing methods lack the ability to generalize across different devices/datasets. In other words, when the existing methods are tested with a new dataset having a different distribution compared to the trained dataset, the methods' performance drops. This issue might arise due to using different wearable devices with different sensors' quality and specifications, resulting in variances in the collected signals. This highlights a critical limitation in their ability to generalize across various real-world scenarios. We believe that a foundation model designed to be resilient to noise and capable of adapting to diverse data distributions is necessary to address this issue. Such a model could be trained across a broader range of biosignal data, incorporating both stationary and active conditions, to ensure robustness and transferability across different datasets, and user contexts. Additionally, leveraging transfer learning techniques enables a model with a large learning capacity to be trained on smaller datasets, whereas training from scratch typically requires substantial amounts of data and time.

In this paper, we introduce an end-to-end explainable RR estimation framework based on multi-scale CNN for smartwatches' PPG and tri-axial ACC signals. The network incorporates dilated convolution layers, which contribute to a larger receptive field, improving the efficiency of time series processing. We developed a pre-trained foundation model trained on a large volume of data collected from smartwatches in everyday settings and then applied a transfer learning approach to boost the model's accuracy for other datasets. The pre-trained model is built using raw PPG and ACC signals from a Health Monitoring dataset collected from 28 individuals in free-living conditions. Then, the pre-trained model is fine-tuned on the target dataset. We assess our method with two public benchmark datasets comprised of wrist-based PPG and ACC signals. The proposed RR estimation method's performance is evaluated compared to existing state-of-the-art methods using different evaluation metrics. Furthermore, we investigate the level of contribution of input signals to the RR estimation using the Deep Shapley additive explanations (Deep-SHAP) method. In summary, the paper's contributions are manifold as follows:

- Proposing multi-scale convolution incorporated with residual inception networks for RR extraction from wrist-based physiological signals.
- Proposing a foundation model for smartwatch-based RR extraction.
- Evaluating the proposed model using wrist-based data collected in free-living conditions.
- Assessing the proposed model's performance in terms of MAE, RMSE, accuracy, mean bias, and confidence interval and correlation compared to six state-of-the-art methods.
- Releasing the foundation model and the Python implementation [1] of the model as open source for the community.

Section 2 outlines the related work of this research. The proposed method is presented in Section 3. The experimental setup and the dataset are outlined in Section 4. The proposed method is compared to other published methods in Section 5. Finally, Section 6 concludes the paper. A preliminary version of this work has been reported in [26].

[1] https://github.com/kazemikianoosh/RR_Estimation.git

2 RELATED WORK

We delve into the state-of-the-art methods developed for respiratory rate estimation utilizing biomedical signals, such as ECG, PPG, and IMU-based signals (i.e., ACC). These methods can mainly be categorized into three main groups: Waveform Analysis techniques, Conventional Machine Learning approaches, and Deep Learning methodologies.

2.1 Waveform Analysis Techniques

The waveform analysis method is commonly used for respiration rate estimation, leveraging the inherent respiratory modulations present in the PPG signals. Waveform analysis involves a multi-stage process, starting with the extraction of respiratory signals dominated by modulation patterns. This extraction is achieved through either filter-based techniques, which attenuate non-respiratory frequencies, or feature-based techniques, involving beat-by-beat feature measurements. Then, the extracted respiratory signals are processed using Time/Frequency domain techniques to estimate the RR [9].

Breathing rate modulates PPG waveforms in three ways [37]: 1) Baseline Wander (BW) is the PPG signal's envelope, affected by changes in intrathoracic pressure. This modulation is extracted using filter-based approaches, suppressing the non-respiratory frequency components in the PPG signal, such as Bandpass filter [33], Empirical Mode Decompositions [41] and wavelet transforms [13]. 2) Amplitude Modulation (AM) [30] is associated with variation in peak-trough level, reflecting stroke volume during inhalation. 3) Frequency Modulation (FM) is the variation in successive peak intervals, manifesting the heart rate fluctuations during the respiratory cycle. Feature-based techniques are primarily utilized for extracting AM and FM-dominated respiratory signals by detecting the fiducial points and, subsequently, a variety of features, such as peak and trough amplitude, morphological variations, PPG pulse widths, and onset [30].

Following respiratory-induced signal extraction, breathing rate can be derived from these signals using different processing algorithms such as time or frequency domain and feature extraction techniques. Time-domain methods identify individual breaths and estimate the breathing rate based on mean breath duration within a defined window, using adaptive thresholding to detect peaks and troughs in respiratory signals. For example, Schäfer et al. [50] proposed a method that involves band-pass filtering, identifying local extrema, and using a threshold-based approach to detect valid respiratory cycles from PPG signal. In their method, prior to detecting the peaks and troughs in the PPG-derived respiratory signals, these signals are detrended, and eventually, a heuristic rule is applied to estimate the RR.

On the other hand, frequency-domain techniques analyze respiratory signal spectra, often employing the FFT and auto-regressive modeling to locate the dominant breathing frequency. For instance, Nicholas et al. [22] introduced an RR estimation approach that combines PPG and ACC-derived respiratory rates. They extracted surrogate respiration waveforms from various PPG modulations (i.e., AM, BW, FM) and employed time-frequency spectrum analysis and spectral peak tracking for ACC-based RR estimation. The fusion of PPG- and ACC-based RR estimates was performed in the last step based on their quality. Another study [21] presented an RR estimation method using ACC and GYR, involving steps such as ACC normalization, filtering for relevant information isolation, sensor data aggregation, and frequency-domain-based RR estimation. Pimentel et al. [44] also employed several Auto-Regressive (AR) models with varying orders to identify the breathing frequency in three respiratory-dominated modulations (BW, AM, FM) extracted from PPG data. Their method includes signal quality assessment in the preprocessing stage to identify the artifacts and low-quality segments in the signal. They then compiled the frequency spectrum of each AR model, considering all possible model orders and modulated signals, into a single spectrum representing the median amplitude. Ultimately, they estimated the respiratory rate by converting the frequency with the highest amplitude into breaths per minute. Their approach was evaluated on two datasets

obtained from diverse clinical environments, achieving mean absolute errors (MAE) of 4.0 bpm with 32-second window size.

Although waveform analysis methods are cost-effective and simple, they require parameters to be tuned manually, experience parameter optimization, and handcrafted rules customized for specific patient populations, causing a lack of robustness to the presence of noise. Additionally, these techniques are susceptible to spurious breath detection, particularly in the presence of noise and abnormal respiratory signal morphology.

2.2 Conventional Machine Learning

Traditional Machine Learning techniques are developed for estimating respiratory rates. These methods are supplied with temporal and spectral features derived from PPG and respiratory-induced signals. For example, Kapil et al. [45] derived frequency domains (such as Fourier spectrum, power spectral density, and Spectral entropy) and morphological features (e.g., autocorrelation, autoregression coefficient, Hjorth complexity, and peak-to-valley amplitude) from PPG, ACC, gyroscope and different respiration-induced signal modalities. Then, they developed machine learning models, including Random Forest, Gradient Boosting, and Support Vector Machine to predict RR. In [57], an ML-based approach for RR estimation in PPG signals was proposed using the Gradient Boosting model. Their method includes segmentation and filtering procedures to remove motion artifacts. Subsequently, several time and frequency-domain features – such as peak-to-peak interval, peak-valley amplitude, and PPG power-spectral-density-based features – were extracted from pre-processed PPG signals. These extracted feature vectors were employed to train a regression machine learning model, specifically the extreme gradient boosting algorithm. In another study [32], an approach is proposed to calculate the RR using a two-step process involving a filter and an extractor. First, a random forest model as a filter is applied to detect and reject inaccurate sensor data from the smartwatch's ACC and gyroscope. Data that passes this filter is then processed by a CNN model, which extracts the RR by identifying breathing patterns within the multi-channel time series data.

These methodologies are accurate under conditions of high signal quality. Nevertheless, their robustness and accuracy drop in the presence of noise and motion artifacts, leading to compromised RR estimation. Consequently, these approaches prove inadequate for applications in wearable monitoring, particularly in scenarios where users engage in diverse physical activities.

2.3 Deep Learning

Recently, deep learning methods have been developed for RR estimation. These techniques effectively capture both temporal and spatial information from 1-D time series and automatically estimate RR from raw biosignals (i.e., PPG) [4, 6, 64]. Aqajari et al. [4] utilized Cycle GANs to generate respiration signals. They harnessed the capability of Cycle GANs to translate PPG signals into respiration signals and integrated the reconstruction error between the generated and actual signals into an optimization function to ensure precise RR estimation. In another study [14], a deep learning PPG-based RR estimation using a convolutional encoder-decoder model is introduced. The model processes PPG signals through three convolutional layers to extract respiratory features, and then uses three transposed convolutional layers to reconstruct these features into a respiratory waveform that resembles a standard capnogram. The model is trained to minimize the mean squared error between the generated waveform and a reference capnogram signal. In the final phase, the FFT is applied to the output waveform to determine the dominant respiratory frequency, which is then converted into breaths per minute. While the aforementioned studies are an end-to-end deep learning model, they require feature engineering to extract RR from the respiration waveform. Another study in [6] introduced a deep-learning approach for automated and accurate RR estimation. They proposed an end-to-end learning model based on the ResNet block as a foundational structure, supplemented with fully connected layers to estimate RR directly from raw PPG signals. In [64], a

deep learning-based method was proposed for RR estimation from PPG signal with low computation cost. The structure comprised two components: a feature extraction module using three convolution blocks with depthwise separable convolutions and a prediction module with three dense layers, resulting in a scalar RR estimation output after processing the original PPG signal. Chin et. al. [10], proposed a deep learning approach that integrates CNN and Long Short-Term Memory (LSTM) networks. In this method, PPG signals are classified as either inhalation or exhalation based on their amplitude relative to a zero baseline. The CNN network extracts both spatial features, followed by the LSTM layer, to capture temporal and long-term dependencies in the data. The model is trained to distinguish between inhalation and exhalation phases, with the RR calculated from the number of detected inhalations. In addition, a more accurate method is proposed in [42] called RRWaveNet. This method employs three modules: multi-scale convolution to capture features at various resolutions, deep spatial-temporal residual blocks for further processing, and a respiratory rate estimator. RRWaveNet leverages transfer learning to improve accuracy on lower-quality datasets. This approach enables accurate, subject-independent RR estimation from raw PPG signals.

Deep learning methods offer promising results and show potential superiority over conventional waveform analysis and traditional machine learning techniques in wearable-derived data as they automatically extract complex features and learn intricate patterns from raw data. However, these studies are based on PPG signals collected in the intensive care unit, whereas in real-life applications (such as wearable monitoring), PPG signals are more prone to noise and motion artifacts, impairing their methods' accuracy and robustness.

2.4 Wearable-based Devices for RR Extraction

Recently, a variety of wrist-worn wearable devices, such as Fitbit [61], Apple [60], and Garmin [62] Smartwatch, have introduced to the market to monitor RR. These wrist-worn wearable devices incorporate different sensors and algorithms to estimate the RR. For instance, both Fitbit and Garmin Smartwatch employ an algorithm based on heart rate variability (HRV) to derive RR [63]. This method involves calculating the interbeat interval (IBI) from the PPG signal collected from the Smartwatch. The IBI, which is the foundation of HRV, is mapped to create a visual pattern of peaks and valleys—decreasing IBI indicates inhalation while increasing IBI reflects exhalation. By counting these peaks, each representing one breath, the devices can determine the average respiratory rate per minute.

The Apple Watch, on the other hand, estimates respiratory rate using its built-in accelerometer, which detects vibrations and movements to calculate the number of breaths per minute [29]. Other wearables, such as the OURA Ring [63], also track RR using a similar algorithm to that of Fitbit and Garmin. However, these commercially available devices are primarily designed to monitor RR during sleep, as in challenging monitoring conditions (i.e., the presence of physical activity), they fail to accurately pinpoint the RR.

2.5 Clinical Applications

Studies show that RR is a fundamental vital sign with broad clinical applications, as it provides essential insights into a patient's overall health status [8, 39]. RR measurement is crucial in diagnosing and monitoring sleep apnea [39], particularly obstructive sleep apnea (OSA). OSA is characterized by repeated episodes of airflow cessation (apnea) or reduction (hypopnea) during sleep, which can disrupt sleep and lead to various health issues. Accurate RR measurement helps identify these events by detecting changes in breathing patterns. Effective RR monitoring enables timely diagnosis and treatment, reducing the risk of complications like cardiovascular disease and improving the quality of life for patients with sleep apnea.

In the context of cardiac care, RR measurement is a vital predictor and monitoring tool for adverse cardiac events [39] . An elevated resting RR is strongly linked to an increased risk of cardiac arrest and other serious cardiac conditions. Studies indicate that RR is one of the most accurate vital signs for predicting cardiac arrest

[17, 54] , with higher values indicating greater risk. Continuous RR monitoring can provide early warnings of cardiac decompensation, allowing for prompt medical intervention. Techniques such as PPG-derived respiration are used to monitor RR in patients with heart conditions, both in clinical settings and through wearable devices for remote monitoring. This continuous monitoring is particularly valuable for patients at risk of out-of-hospital cardiac arrest, as it enables early detection and management of potential cardiac events, ultimately improving patient outcomes and reducing mortality rates.

Early Warning Score (EWS) is a crucial tool in clinical settings for identifying patients at risk of deterioration, and RR monitoring plays an important role in this process [55] . RR is a sensitive indicator of physiological distress, often changing before other vital signs, such as blood pressure and heart rate. Incorporating RR into an EWS system allows clinicians to take timely actions, potentially preventing adverse outcomes and improving patient survival. Continuous assessment of RRs is crucial for healthcare professionals to identify impending conditions such as sepsis, respiratory failure, and cardiac arrest. Automated systems for continuous RR monitoring can significantly reduce the burden on medical staff and can be complemented by an automated alert system for deviations from normal ranges.

3 Method

We develop a deep learning-based RR estimation method for smartwatches-based PPG and ACC signals. Our proposed method consists of two major phases (See Fig. 1). First, we develop the pre-trained model using PPG and ACC signals collected from smartwatches worn by individuals in free-living conditions. The pre-training phase employs a Multi-Scale Residual CNN-based model as the backbone network, selected to effectively capture features at multiple scales. Subsequently, we fine-tune the pre-trained model using a small portion of target datasets. In the following, we delve into the detailed description of each phase.

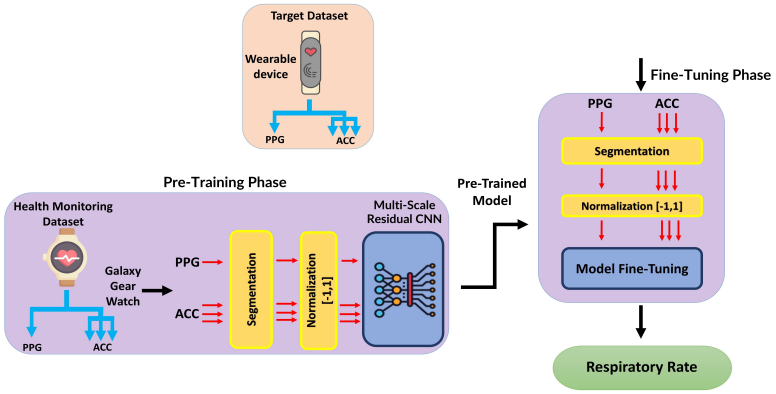


Fig. 1. Proposed RR estimation pipeline including pre-training phase and fine-tuning phase.

3.1 Pre-training Phase

3.1.1 Model Architecture. Our foundation model architecture includes two major components: 1) Segmentation and Normalization and 2) Multi-Scale Residual CNN. A view of the architecture is illustrated in Fig. 2. The Segmentation and Normalization are applied to raw PPG and tri-axial ACC before feeding them into the deep learning model. The proposed deep learning model is trained and subsequently used in the fine-tuning phase.

Segmentation and Normalization: Respiratory signals and their noise levels exhibit non-stationary characteristics in collected PPG and ACC signals, with both signal of interest and noise vary throughout monitoring. To

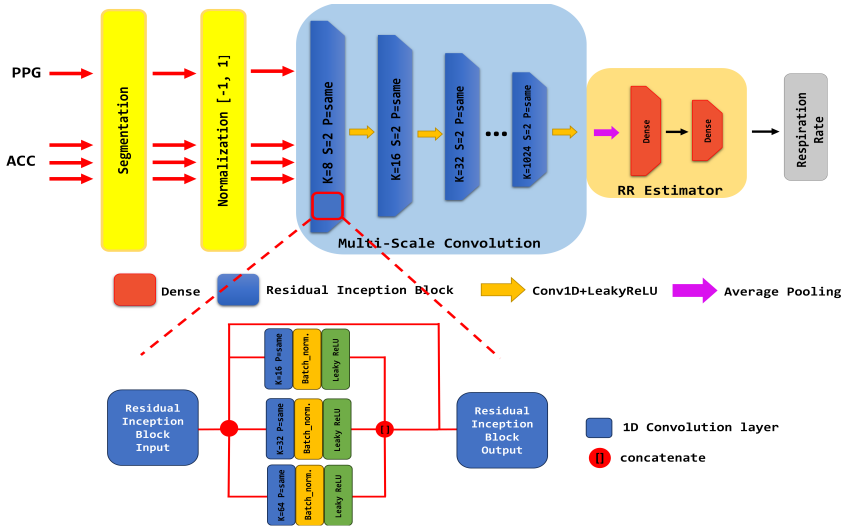


Fig. 2. Deep learning architecture including the Segmentation, Normalization, Multi-Scale Residual Convolutions and RR estimator. The Multi-Scale Residual Convolutions module receives normalized PPG and tri-axial ACC segment to capture spatial and temporal features using residual convolution, with kernel size (k), stride (s), and padding (p).

address this, we divide the input signals into (quasi) stationary segments, assuming each segment maintains a consistent respiration rate and noise level. The segment length should be sufficient to maintain a balance between enabling accurate waveform analysis and retaining the (quasi) stationary property. The literature suggests that a suitable duration for estimating respiratory rate and respiration signals ranges from 16 to 90 seconds [23, 44, 50]. Therefore, in this study, we repeat our experiments with different lengths of raw PPG and tri-axial ACC to examine the effects of window lengths. Three window sizes (i.e., 16, 32, and 64 seconds) are tested for RR estimation. the 32-second segment yields the best performance empirically compared to other window sizes as it has a sufficient amount of data and resolution for accurate waveform analysis and fewer artifacts than longer window sizes. Additionally, prior to training and testing the proposed deep network model, the PPG and tri-axial ACC signals are scaled to [-1, 1].

Multi-Scale Residual CNN: To estimate RR, We employ and customize a deep neural network, presented [26], to derive RR from the signals. This method consists of two distinct modules: a Multi-Scale Residual Convolution module and an RR estimator module. An overview of our model architecture is shown in Fig. 2. We outline these two modules in the following.

1) Multi-Scale Convolution: Inspired by the multi-scale convolution technique found in the grid modules of various inception networks [59], we incorporate a dilated residual inception module into our proposed model. This incorporation effectively captures signal features at different resolutions and significantly mitigates the issue of vanishing gradients [20]. Furthermore, the dilated residual inception blocks offer larger receptive fields without substantially increasing the number of parameters. Our approach consists of three parallel branches, each guiding the input signals through separate channel-wise convolutional layers before concatenating them at the end. Each convolutional layer employs varying filter sizes, facilitating the determination of suitable weights for each convolutional resolution, thus discerning the frequency of recurring features present in the input signals. This information is subsequently fused for downstream modules. Integration of the dilated residual inception block enhances feature extraction robustness, leading to a more compact architecture.

This module is enhanced by the inclusion of Batch Normalization, a Leaky ReLU activation layer with a slope of 0.2, and 1D convolution layers. Each 1D convolution layer utilizes a kernel size of three and a stride of two. The initial convolution layer has a filter size of 8, which progressively increases by a factor of two until it reaches 1024. For more efficient training, we opt for strided convolution for downsampling over max-pooling [56]. This downsampling procedure decreases input size while simultaneously increasing the filter count for each layer by a factor of two until reaching 1024 filters for subsequent levels.

The Multi-Scale Convolution is fed with the raw PPG and the tri-axial ACC signals. Our model inputs can be expressed as: $X = (x^{(11)}, x^{(12)}, x^{(13)}, x^{(14)}, RR_r^{(11)}), (x^{(21)}, x^{(22)}, x^{(23)}, x^{(24)}, RR_r^{(21)}), \ldots, (x^{(n1)}, x^{(n2)}, x^{(n3)}, x^{(n4)}, RR_r^{(n1)})$, where $(x^{(i)}, RR_r^{(i)} \in R^n)$ and $RR_r^{(i)}$ is the reference RR, $x^{(i1)}$ is the PPG signal and $x^{(i2)}, x^{(i3)}, x^{(i4)}$ are the tri-axial ACC signals. The Multi-Scale Convolution downsamples the input signals $x(i)$ to obtain a compressed feature vector $Z^{(i)}$. The compressed feature vector is represented by (1):

$$Z^{(i)} = B_1\left(x^{(i)}; \theta_1\right) \tag{1}$$

Where the θ_1 is the parameters in the multi-Scale convolution denoted as B1.

2) Respiratory Rate Estimator: To estimate RR values, we utilize a fully connected layer. The convolutional layers' output undergoes global average pooling, transforming the downsampled features from the residual blocks into a one-dimensional format. Following this, the data passes through two sets of activation layers and a fully-connected layer, reducing the input to dimensions of 64 and then 1, which represents the estimated RR.The dense layers can be expressed as:

$$RR_e^{(i1)} = B_2\left(z^{(i)}; \theta_2\right) \tag{2}$$

where the θ_2 is the parameters in the dense layer denoted as B2. The optimization process for the architecture minimizes the $SmoothL$ loss between the estimated values $RR_e^{(ni)}$ and the reference values $RR_r^{(ni)}$. The loss function is defined as follows:

$$L(X) = \sum_{i=1}^{m} SmoothL(RR_{diff}) \tag{3}$$

$$RR_{diff} = RR_r^{(i1)} - RR_e^{(i1)} \tag{4}$$

$$SmoothL_1(RR_{diff}) = \begin{cases} 0.5(RR_{diff}^2) & abs(RR_{diff}) < 1 \\ abs(RR_{diff}) - 0.5 & \text{otherwise} \end{cases} \tag{5}$$

3.1.2 Pre-training Dataset. We use a dataset obtained as part of a Health Monitoring study [36] to train the model. Participants were instructed to wear Samsung Gear Sport smartwatches [49] on their non-dominant wrists and a Shimmer3 device [53] on their chests for continuous 24-hour data collection as shown in Fig. 3. During data collection, participants were engaged in their typical daily activities to track their vital signs, physical activity and sleep data. Data collection took place in southern Finland between July and August 2019. A total of 28 healthy participants, both male and female, were recruited with exclusion criteria to ensure homogeneity.

Chest-based accelerometers are proven to be robust and noise-resistant RR estimation methods [16, 51]. The Shimmer3 device recorded chest tri-axial ACC data, used as the reference signal for respiratory rate extraction, at a sampling frequency of 512 Hz. The device has sufficient internal storage and battery capacity, allowing uninterrupted data recording for a full 24 hours. The Samsung Gear Sport smartwatch was utilized to capture wrist PPG and tri-axial ACC signals using inertial measurement unit sensors at a frequency of 20 Hz [49]. The Samsung Gear Sport Watch weighs 67 g with the strap, which is lightweight and compact. This waterproof smartwatch runs Tizen, which is an open-source operating system and has three days of battery life.

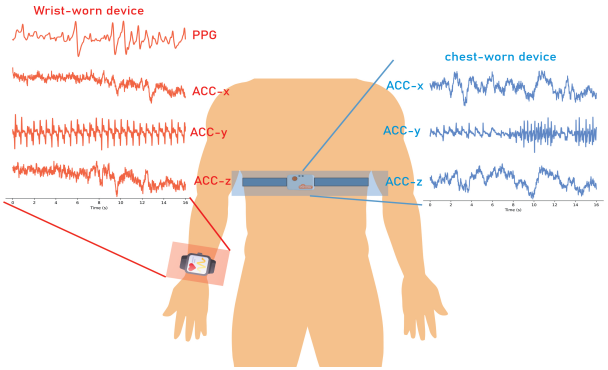


Fig. 3. We have used various data sources with different characteristics, including WESAD, PPG-DaLiA, and Health-Monitoring, which are noisy datasets obtained using wrist-worn sensors and chest-worn devices.

We upsampled the PPG and tri-axial ACC signals from the smartwatch to 64 Hz and downsampled the Simmer 3 data to 64 Hz to unify the sampling frequencies. To achieve this, we employed a conventional linear interpolation technique [43]. In this method, a line is fitted between each pair of data points. Then, based on the upsampling or downsampling rate, new data points are fitted on the line. These upsampled PPG and tri-axial ACC signals from the smartwatch are used for model training and testing. Additionally, the ground truth RR is extracted from the chest-based Shimmer3 device as labels.

Throughout the research, adherence was maintained to the Declaration of Helsinki and the Finnish Medical Research Act (No 488/1999). The Ethics Committee for Human Sciences at the University of Turku granted approval for this study (No 44/2019). Before providing their written consent, participants received both verbal and written information about the study. Participation was entirely voluntary, and individuals had the option to exit the study at any point if they chose to discontinue their involvement.

Ground truth Respiration Rate (RR) As mentioned previously, we acquire the ground truth RRs from ACC signals obtained through Shimmer3 in the Health Monitoring dataset. These ground truth RR values serve as labels for the pre-trained model's training phases. The respiration rates are calculated using tri-axial ACC signals, according to Sun et al. [58]. The approach includes three steps: 1) a pre-processing step designed to eliminate noise from ACC while preserving fluctuations associated with respiration, 2) a Fast Fourier Transform technique to derive the frequency of respiratory rates, and 3) a multi-axis fusion strategy to enhance estimation accuracy. The fusion strategy employs a Kalman filter to integrate RR estimates obtained from the three axes (X, Y, and Z), taking into account historical data.

3.1.3 Pre-training. In the pre-training phase, we use the Health Monitoring dataset collected from smartwatches to train the multi-scale residual CNN network. The model is fed with smartwatches' PPG and tri-axial ACC signal, and the labels are RR extracted from Shimmer 3 devices. The model is trained across 200 epochs with 30 steps per epoch. Early stopping technique is implemented to prevent overfitting. Moreover, to enhance training success, the Cosine Decay learning rate scheduler was used. Additionally, we use the Health Monitoring dataset for pre-training due to its larger data volume compared to existing public datasets. This dataset, collected from smartwatches worn in free-living conditions over 24 hours, provides a robust foundation for model training. To ensure a fair comparison with other studies, we select publicly available datasets specifically for the evaluation and testing phases.

3.2 Fine Tuning

In the fine-tuning phase, the target dataset's signals are first segmented and normalized to be fitted to the pre-trained model. Then, the pre-trained model is fine-tuned on the target dataset signals by extending the training using the same model's configuration. The 2 Dense layers in the Respiratory Rate Estimator part and the last 4 layers of 1D convolutions in the Multi-Scale Residual Convolution are re-trained for fine-tuning. Through this knowledge-transferring method, we are able to leverage the high-quality Health Monitoring dataset to be more confidently applied to a smaller dataset (i.e., target dataset) with a lower-quality signal.

4 EXPERIMENTAL SETUP

This section outlines the description of target datasets and model fine-tuning details. Then, we briefly describe four existing state-of-the-art methods, including deep learning and conventional machine learning approaches, as well as the waveform analysis approach. Finally, the evaluation metrics used for comparison are described.

4.1 Target Datasets

In this section, we outline two public [46, 52] benchmark datasets which will be used for our model's evaluation. A portion of the datasets are used for model fine-tuning. We specifically chose these datasets because they were collected using wrist-worn wearable devices in everyday life settings, incorporating a variety of motions and artifacts. However, the other existing datasets, such as BIDMC [44], CapnoBase [25], and MIMIC -II/III [48] waveform datasets have been collected in stationary environments with minimal or no motion, primarily from fingertips. To enhance the robustness of our proposed method against motion artifacts, we opted for the PPG-DaLiA and WESAD datasets. Additionally, our pre-training dataset (i.e., Health Monitoring study) was collected using smartwatches. The signal types in this pre-training dataset closely resemble those found in the PPG-DaLiA and WESAD datasets, further aligning with our research objectives. Furthermore, the collected PPG and tri-axial ACC from wrist-worn devices in the PPG-DaLiA and WESAD dataset are used for model training and testing as well as the model pre-training phase is conducted using the data collected by smartwatches in the Health Monitoring dataset.

4.1.1 WESAD. The dataset focuses on the detection of stress and affect, distinguishing between different affective states, such as neutral, stress, and amusement. The dataset comprises recordings from both wrist-worn (containing PPG and tri-axial ACC at 64 and 32 Hz, respectively) and chest-worn devices (containing tri-axial ACC at 700 Hz). The data were collected from 15 individuals (including 12 male and 3 female subjects), aged between 24 and 35 (mean age of 27.5 ± 2.4), each contributing data for approximately 100 minutes. The participants were healthy graduate students from the University of Siegen, Germany, with height (cm) mean (standard deviation) of 177.6 (±6.51) and weight mean (std) of 73.1 (±9.92). WESAD primarily captures data during sedentary activities. The ground truth RRs are extracted from the chest-worn respiration using the method explained in [50] . This method identifies local extremes (maxima and minima) in the respiratory signal, and eliminates irrelevant oscillations by applying a threshold threshold based on the amplitude difference. The remaining extrema correspond to complete respiratory cycles, allowing for the calculation of the average breathing duration and thus the respiration rate. This dataset is described in more detail in [52].

4.1.2 PPG-DaLiA. This dataset comprises multimodal signals from 15 participants (seven males and eight females, aged 30.60 ± 9.59 years) engaged in various real-life activities. The dataset includes chest-recorded three-axis chest-recorded acceleration (700 Hz), wrist-recorded PPG, tri-axial ACC signals (64 Hz), and subject information (age, gender, height (cm) of 175.25 ± 8.27, weight (kg) of 69.0 ± 12.35, skin color, and fitness level, indicating the subject's sports participation frequency). The dataset was collected under comprehensive and representative daily life conditions and includes reference RR intervals aligned with R peak positions for 8-second intervals,

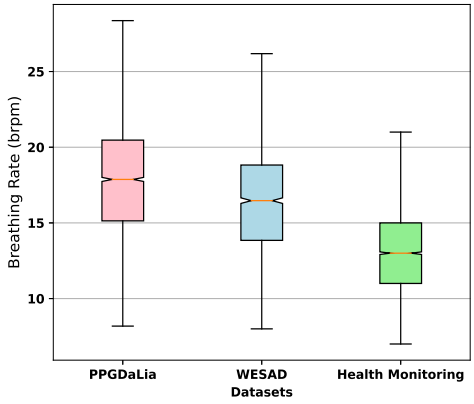


Fig. 4. Distribution of RR for PPG-DaLiA, WESAD, and Health Monitoring datasets.

sampled every two seconds. The data collection protocol involved eight distinct daily life activities such as sitting, stair climbing, table soccer, cycling, car driving, lunch breaks, walking, and work. For each subject, data collection took around 150 minutes, resulting in a total of approximately 36 hours of data in the PPG-DaLiA dataset. Similarly, the ground truth RR extracted by the chest-worn devices using the proposed method in [50]. The detailed information is described in [46].

Due to the difference in the sampling frequency of the data collected in these two datasets, all the PPG and ACC signals are resampled to 64 Hz to unify the sampling frequencies. During the resampling process, linear interpolation [43], a method commonly used for signal frequency conversion, was used.

Moreover, the distribution of ground truth RR in each dataset is plotted in Fig. 4. This figure illustrates the discrepancies in RR distributions across datasets, notably the higher RR observed in PPG-DaLiA due to medium and vigorous activities during data collection. Importantly, for Health monitoring datasets, the method outlined in [58] is used to extract the ground truth RR from the tri-axial ACC signals recorded by the chest bands. While the Ground truth RR for out target datasets (i.e., PPG-DaLiA and WESAD) are extracted from the chest-worn respiration using the method explained in [50].

4.2 Model Fine-Tuning and Testing

The proposed pre-trained model is fine-tuned and assessed using the wrist-worn wearable PPG and ACC signals of the two target datasets separately. In total, PPG-DaLiA datasets comprise 15 participants and the WESAD dataset has 15 subjects in which the PPG and ACC signals are collected by the wrist-worn wearable.

Note that an inter-patient test is conducted in fine-tuning phase. In other words, the dataset for fine-tuning and testing do not share the data of the same patients. This test aims to ensure the model is generalizable and avoids data leakage between training and testing. The whole training phase (including pre-training and fine-tuning) employs 2,300,000 32-second segments of PPG ACC data from 42 participants, including 28 subjects (2,000,000 32-second segments) from the Health Monitoring dataset for pre-training and seven distinct subjects each from the WESAD and PPG-DaLiA datasets (30,000 32-second segments) for fine-tuning. For the testing dataset, we randomly selected 15 subjects—7 from WESAD and 8 from PPG-DaLiA—ensuring that they were kept separate during the inter-patient phase. In total, the testing dataset comprises 30,000 32-second segments from these 15 participants. This approach ensures a rigorous and unbiased evaluation of the model's performance across different subjects.

We fine-tuned the model on a Linux machine equipped with an AMD Ryzen Threadripper 2920X 12-Core processor, an NVIDIA TITAN RTX GPU, and 126GB of RAM. Tensorflow (version 2) and Keras API framework on Python were used for model development. The fine-tuning experiment involved training the model for 100 epochs, with 60 steps per epoch. To prevent overfitting, an early stopping technique was implemented. During the pre-training and fine-tuning process, the early stopping technique monitor the validation loss to decide when to halt further training. Moreover, the *patience* parameter is set at 8, which allowed the model to continue training for up to 8 additional epochs after the validation loss stopped improving. The Adam optimizer was used for pre-training and fine-tuning, paired with the Huber loss function to provide a balance between mean squared error and mean absolute error, particularly useful in reducing the impact of outliers. Additionally, the Cosine Decay learning rate scheduler was employed to optimize the training process and achieve superior results.

4.3 Base-line Methods

The proposed method is compared with six existing approaches, including waveform analysis, conventional machine learning and deep learning, for extracting RR from PPG and ACC signals. All the state-of-the-art models are trained and tested on the two target datasets (i.e., WESAD and PPG-DaLiA). The target datasets are passed through an inter-patient test, where the training and testing datasets consist of entirely separate individuals. More importantly, the test datasets for the baseline methods are the same as the test dataset used for the proposed method evaluation. In the following, a brief description of the six RR-based estimation methods is provided.

4.3.1 Waveform Analysis Technique.

A-P Synthetise: A hybrid method was introduced in [22] for estimating respiration rate (RR) from both PPG and ACC signals. The algorithm involved two independent steps for deriving RR estimates. First, a PPG-based RR algorithm was developed to extract surrogate respiratory signals, such as AM, FM, and BW, from the PPG signal and estimate RR from the reciprocal of the average of the most recent 30 inner breath intervals. Second, an ACC-based RR algorithm was introduced to project the three orthogonal ACC waveforms to the first principal axes and implement FFT spectrum and peak tracking to derive RR estimates. Final RR values were calculated by combining the PPG and ACC estimates considering the quality of the estimates. The final estimate would be discarded if both methods provided low-quality estimates. It is important to highlight that in this method, there are no machine learning algorithms involved, which means that the number of parameters is simply zero.

4.3.2 Conventional Machine Learning.

Smart Fusion (SF): SF is a machine learning-based RR estimation method enabled by PPG signals [45]. PPG-based respiratory modalities signals (i.e., AM, FM, and BW) are first extracted. Subsequently, frequency-based features such as Fourier spectrum, power spectral density, and Spectral entropy are computed from each respiratory-induced modality signal. Besides these frequency-based features, morphology-based features (i.e., autocorrelation, autoregression coefficient, Hjorth complexity, and peak-to-valley amplitude features) are also extracted to enhance the method's accuracy. The frequency and morphology-based driven features are then fused using three machine learning models (support vector machines, Random Forest, and Gradient Boosting). The predicted absolute errors from these models are utilized to determine weights for each modality. These weights are calculated using an expression that considers the predicted absolute error and a constant value τ, empirically set at 5.

4.3.3 Deep Learning.

CycleGAN: Aqajari et al. [4] employed a cycle-generative adversarial network to reconstruct respiration signals from PPG signals, leveraging the reconstruction ability of the network. Initially, an image translator was employed to convert the PPG and respiration signals into images. Then, the images were used for training the model. Furthermore, the optimization function incorporated adversarial losses, cycle consistency losses, and RR

losses to ensure that the reconstructed signal preserved the respiratory characteristics. Three types of convolutions were used in this approach: stride-2 convolutions, residual blocks, and fractionally-scaled convolutions for the generative networks and 70×70 PathGANs for the discriminator networks. Moreover, BreathMetrics [40], an open-source library designed for respiration signal analysis, was used to identify peak values and estimate the RR.

CNN: Bian et al. [6] employed a CNN-based model to estimate RRs using raw PPG signals. The proposed network uses ResNet blocks as the backbone architecture followed by fully-connected layers in order to predict the RR. The ResNet architecture is designed using the specifications in the literature [6], including 1D convolution layers activated by ReLU. The model takes a PPG signal as input, represented as a vector with dimensions (2048,1) – equivalent to a 32-second segment sampled at 64 Hz. Training the model involves experimenting with three distinct learning rates as reported in [6], with the most optimal outcome achieved at a learning rate of 10^{-5}.

RRWaveNet [42] algorithm employs a sophisticated architecture consisting of three main modules: multi-scale convolution, deep spatial-temporal residual blocks, and a respiratory rate estimator. The multi-scale convolution module captures features from PPG signals at different resolutions using three parallel convolutional layers with varying filter sizes. The convolution layers are followed by a batch normalization layer, a ReLU activation layer, and a max-pooling layer. The deep spatial-temporal residual blocks further process these features through eight residual blocks with increasing filter sizes. Finally, the respiratory rate estimator uses global average pooling followed by fully connected layers squeezing the input size of 128, 64 and 1 to predict the RR. The model is trained using mean squared error loss and optimized with the AdaBelief optimizer.

Multi-Scale CNN [26] leverage convolutional neural networks to extract RR from PPG, ACC, and gyroscope (GYR) signals captured via smartwatches. The method involves segmenting the input signals into 32-second windows and filtering out segments corrupted by motion artifacts. Independent Component Analysis is then applied to extract respiration-related signals from ACC and GYR data. These signals, along with the PPG data, are normalized and fed into a multi-scale residual CNN. The CNN consists of a dilated residual inception module and 1D convolutional layers to capture temporal features at different resolutions. Each convolution layer is augmented with Batch Normalization and a Leaky ReLU activation layer. The 1D convolution layer's filter size is set at 8 in the initial layer and increases by a factor of two until it reaches 1024. The extracted features are then processed by fully connected layers to estimate the RR. The model is trained using a SmoothL1 loss function to minimize the difference between the estimated and reference RR values.

4.4 Evaluation Measures

The proposed method is evaluated and compared with the performance of other existing methods using 6 different performance matrices: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), Accuracy, BlandAltman method [47], and correlation coefficient. These metrics serve to quantify the disparity between the predicted RR and the actual RR values.

RR_e^i and RR_r^i refer to the estimates and reference respiratory rates, respectively, and N represents the total number of RR samples.

MAE represents the average absolute differences between the predicted values and the actual values.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|RR_e^i - RR_r^i| \tag{6}$$

RMSE is a statistic that measures the standard deviation of the prediction error, or residuals, whose distance from the regression line is the residuals.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} |RR_e^i - RR_r^i|^2}{N}} \tag{7}$$

Bland Altman[15] assesses how well the estimated RR agrees with the ground truth RR. Based on the differences between the estimation and actual values, Bland-Altman provides mean bias and 95% confidence intervals (CI). Moreover, in Bland Altman's analysis, the absolute value of the mean bias, irrespective of its sign, shows the wellness of prediction in alignment with the reference values, indicating the method's precision and reliability. **Correlation Coefficient (R)** determines the closeness of two variables (prediction and ground truth) to each other.

Accuracy [24]: To determine accuracy, we used Normalized Root Mean Square Error (NRMSE), defined as:

$$Accuracy = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(RR_e^i - RR_r^i\right)^2}}{\frac{1}{N}\sum_{i=1}^{N} RR_r^i} \tag{8}$$

In addition to emphasizing the magnitude of the variance, the NRMSE penalizes large errors when the estimator is unbiased.

5 RESULT AND DISCUSSION

This section outlines the performance and effectiveness of the proposed method for RR estimation. Section 5.1 compares the proposed method with the stat-of-the-art methods. Next, the impact of different activity types on RR estimation is explained. Following that, in Section 5.3, the impact of signal quality in terms of signal-to-noise ratio (SNR) on RR estimation is evaluated. Finally, Section 5.4 describes the association of PPG and ACC signals on RR estimation using an explainable AI method.

5.1 Performance Comparison

We evaluate the proposed approach's performance using the test dataset explained in section 4.2, including raw PPG and ACC from two benchmark datasets (i.e., WESAD and PPG-DaLiA). In total, the data from 15 users (i.e., 7 users of the WESAD dataset and 8 users from PPG-DaLiA) are included in the test dataset. Table 1 summarizes the proposed method's performance compared to the other six baseline approaches. As shown in Table 1, our proposed method approach demonstrates superior performance compared to existing methods for RR estimation across two datasets considering MAE and RMSE, with values of 2.29 and 3.11 for the WESAD dataset and 3.09 and 3.79 for the PPG DaLiA dataset, respectively. Moreover, the Multi-Scale CNN method achieved the second lowest MAE and RMSE of 2.68 and 3.5 in the PPG DaLiA dataset, while the MAE and RMSE in the WESAD dataset are 5.13 and 5.67, respectively, which is higher than Smart Fusion and CycleGAN. RRWaveNet also performs better in the WESAD dataset with MAE and RMSE of 3.38 and 4.16, respectively, followed by Multi-Scale CNN and Smart Fusion. However, the A-P Synthetise method failed to perform accurate and obtained the highest MAE and RMSE of 8.20 and 9.93 in the PPG DaLiA dataset and 6.21 and 8.52 in the WESAD dataset, respectively. This difference is due to that the other methods are based on machine learning methods, which do not need handcrafted features, and they deliver robust performance across the datasets.

In terms of accuracy, our method outperforms others, achieving an accuracy of 88.4% on the PPG DaLiA dataset, compared to 82.8% for Multi-Scale CNN and 81.1% for RRWaveNet. Similarly, when tested on the WESAD dataset, our method achieves an accuracy of 85.4%, surpassing RRWaveNet and Multi-Scale CNN, which have accuracies

Table 1. The Comparison of Proposed Method with Other Existing Methods: Mean Absolut Error (MAE), Root Mean Square Error (RMSE), Accuracy and data type used in each method.

Method	Dataset						Data Type
	PPG DaliA			WESAD			
	MAE	RMSE	Accuracy	MAE	RMSE	Accuracy	
Proposed	**2.29**	**3.11**	**88.4**	**3.09**	**3.79**	**85.4**	PPG, ACC
Multi-Scale CNN	2.68	3.5	82.8	3.46	4.02	79.7	PPG, ACC, Gyroscope
CNN	2.94	3.88	78.5	5.37	6.48	72.5	PPG
CycleGAN	4.65	5.61	74.3	5.13	5.67	73.8	PPG
A-P Synthetise	8.20	9.93	66.5	6.21	8.52	61.7	PPG, ACC
Smart Fusion	3.68	4.90	77.7	4.48	5.29	76.9	PPG
RRWaveNet	3.12	3.83	81.1	3.38	4.16	80.9	PPG

The corresponding method with the highest MAE, RMSE and Accuracy (in each row) is presented in bold type.

of 80.9% and 79.7%, respectively. Additionally, as shown in Table 1, the A-P Synthetise method performs poorly in estimating RR with the accuracy of 66.5% and 61.7% in the PPG DaLiA and WESAD datasets, respectively.

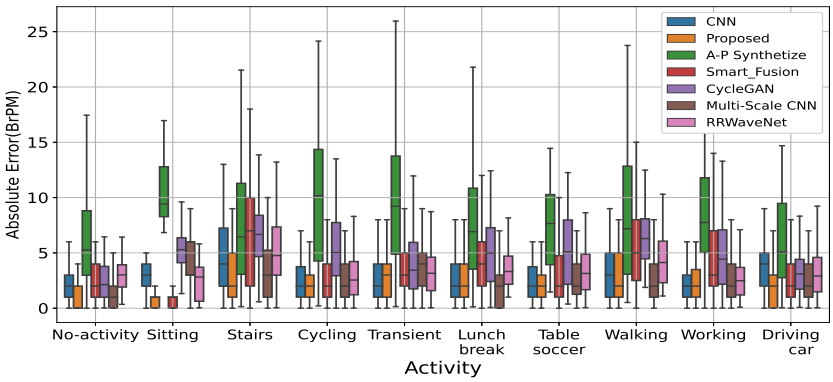
5.2 Impacts of Activities

We assess the performance of the methods across diverse activity types. The estimation of RR in the context of different activity types holds significant importance due to the fluctuation in RR rates during medium and high-intensity activities. These variations render the collected signals susceptible to motion artifacts, thereby intensifying the challenge associated with accurate RR estimation.

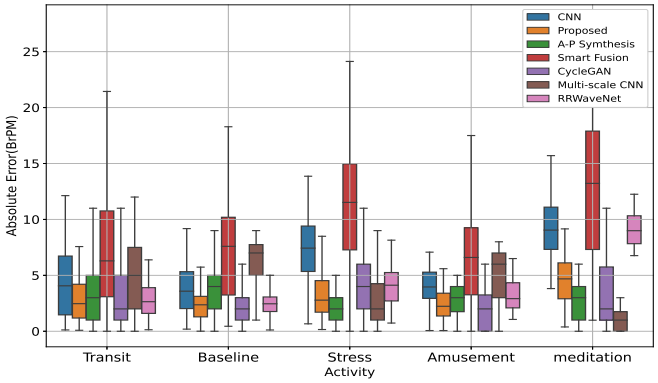
The performance of all the methods across different activity types in two datasets (i.e., PPG-DaLiA and WESAD) is shown in Fig. 5. As shown in Fig. 5(a), the proposed approach exhibited a lower error distribution and median error across different activities compared to other methods. In contrast, CycleGAN, A-P Synthetise, and Smart Fusion methods showed higher median absolute errors and broader error distribution. For example, during soccer and walking, both our method and the Multi-Scale CNN yielded MAE below 2.7 brpm (breathing rate per minute), while in CycleGAN, A-P Synthetise, and Smart Fusion methods exhibited MAE values exceeding 3 brpm. Likewise, in cycling, our method and the Multi-Scale CNN achieved MAE values below 3.53, whereas CycleGAN, A-P Synthetise, and Smart Fusion methods recorded MAE values surpassing 6.46. It is noteworthy that an exception was observed in the case of stairs, where the Multi-Scale CNN demonstrated a better MAE of 2.65 brpm compared to our proposed method's MAE of 2.81 brpm.

The error distribution analysis for the WESAD dataset is depicted in Fig. 5(b). Across various activities, our proposed method exhibited a lower median error in most cases, except during meditation, where the Smart Fusion method achieved a median absolute error of approximately 3.2, while our method showed a higher median absolute error of 4.8. However, compared with the Smart Fusion approach, our proposed method demonstrated a narrower error distribution (lower third quartile value). Furthermore, during the baseline state, our method obtained slightly higher MAE than the CycleGAN method (i.e., 2.46 and 2.40 brpm, respectively).

Table 2 and 3 also present a comprehensive quantitative comparison. As shown in Table 2, the proposed method exhibits the smallest mean bias (i.e., 0.28) followed by the Multi-Scale CNN and RRWaveNet methods with a marginally higher mean bias of 0.62 and 0.67, respectively. Similarly, our method achieves the highest correlation coefficient (i.e., 0.84) across different activities, and the Multi-Scale CNN method obtained a correlation of (0.82), prior to the RRWaveNet method with a correlation of (0.81) exhibiting a higher degree of agreement between the



(a) PPG-DaLiA Dataset, The activity types



(b) WESAD Dataset, The activity types

Fig. 5. Box plot analysis of Methods vs. Absolute Error for (a) PPG-DaLiA Dataset and (b) WESAD Dataset. The absolute error is calculated for different types of activities using the estimated RR and the reference RR.

estimated and the reference RR in both methods. On the other hand, the A-P Synthetise method obtained the highest mean bias and relatively low correlation, 5.46 and 0.67, respectively. Furthermore, the confidence interval of the proposed method (i.e., [-6.53, 7.10]) and RRWaveNet in the second palace (i.e., [-7.40, 7.55]) were relatively narrower than that of the CNN and MultiScale CNN methods (i.e., [-6.28, 9.30] and [-6.92, 8.18], respectively), indicating the lowest level of uncertainty in RR estimation between these methods. Comparatively, existing methods such as CNN, A-P Synthetise, CycleGAN, and Smart Fusion exhibit higher mean biases, ranging from 1.51 to 5.46, and lower correlation coefficients, ranging from 0.77 to 0.67, indicating their inferior accuracy in comparison to our approach.

Similarly, in the WESAD dataset, our proposed method maintained a low mean bias of 0.30 and a higher correlation coefficient of 0.87, followed by the RRWaveNet and MultiScale CNN methods with mean bias of 0.98 and 0.79 and correlation coefficient and 0.78 and 0.76, respectively, highlighting their consistency and accuracy across diverse activities. When compared to other methods, our approach consistently outperforms them, as they exhibit higher mean biases, ranging from -2.14 to 5.24, and lower correlation coefficients, ranging from

Table 2. The summary of mean bias, the 95% confidence interval, correlation coefficient, and activity-based MAE in **PPG-DaLiA** dataset for all methods.

PPG DaLiA	Mean Bias	Confidence Interval	Correlation	P-value	MAE							
					stairs	cycling	walking	baseline	driving	working	lunch	soccer
Proposed	**0.28** [2]	(-6.53, **7.10**)	**0.84**	< 0.01	2.81	**2.90**	**2.42**	**0.94**	**2.37**	**1.89**	**2.87**	**2.4**
CNN	1.51	(-6.28, 9.30)	0.75	< 0.01	3.34	4.86	2.64	2.40	2.52	3.96	3.02	2.53
A-P Synthetise	5.46	(-7.23, 18.16)	0.67	< 0.01	8.14	7.70	8.45	6.67	9.94	6.10	9.86	8.07
CycleGAN	3.71	(**-4.51**, 11.95)	0.77	< 0.01	6.48	6.46	4.99	2.59	5.25	3.42	4.13	5.33
MultiScale CNN	0.62	(-6.92, 8.18)	0.82	< 0.01	**2.65**	3.53	2.63	1.36	2.57	2.35	3.77	2.76
Smart Fusion	1.48	(-8.47, 11.43)	0.74	< 0.01	5.60	6.68	4.77	2.36	2.76	2.72	3.32	3.0
RRWaveNet	0.67	(-7.40, 7.55)	0.81	< 0.01	3.63	5.14	2.68	2.99	3.09	3.16	3.26	3.3

0.64 to 0.74. Moreover, our method obtained a confidence interval of [-6.50, 8.08] with a slight difference with RRWaveNet with a confidence interval of [-5.48, 9.15], lower than the Smart Fusion and CycleGAN methods with a confidence interval of [-6.60, 8.93] and [-6.70, 8.87], respectively.

5.3 Impacts of Dataset Characteristics

In this section, we indicate the effect of motion artifacts and hand motions on signal quality and, subsequently, on RR estimation. PPG signals are susceptible to interference from noise and motion artifacts, leading to a notable degradation in the quality of the acquired PPG signals. These noises mainly stem from user hand movements and environmental factors (e.g., ambient light) collected alongside PPG during monitoring. A number of factors impact the noise level in a PPG signal, including the quality of the sensor, the sensor setup (e.g., electric current) and the intensity of the activity [3, 5].

We utilize SNR as a metric to assess the data quality of the evaluation datasets (i.e., PPG-DaLiA and WESAD). The SNR was obtained using the multi-step method described in [31], starting with transforming the signal to the frequency domain using the Fast Fourier transform, computing the power spectral density, detecting the fundamental frequency as the maximum peak, and identifying harmonics as local maxima. After removing the harmonics, the power of the remaining signal, assumed to be noise, is compared to the power of the fundamental frequency.

Fig. 6 shows the calculated SNR for each window of PPG-DaLiA and WESAD datasets. The SNR values for the PPG-DaLia dataset range from -7.6 to 19.4, while for the WESAD dataset, the SNR spans from -7.9 to 11.1. This discrepancy reveals that the PPG-DaLia dataset exhibits a broader SNR range, indicating a wider spectrum of signal quality compared to the WESAD dataset. Moreover, the trend in Fig. 6 indicates that higher data quality led to lower absolute errors, while relatively poor SNR values correlated to greater absolute errors. As indicated, our proposed method consistently exhibits the lowest errors throughout the entire SNR range, indicating its robust

[2]The corresponding method with the highest Mean Bias, Confidence Interval, Correlation, and MAE for each activity (in each row) is presented in bold type.

Table 3. The summary of mean bias, the 95% confidence interval, correlation coefficient, and activity-based MAE in **WESAD** dataset for all methods.

WESAD	Mean Bias	Confidence Interval	Correlation	P-value	MAE				
					transit	baseline	stress	meditation	amusement
Proposed	**0.30**	(-6.50, 8.08)	**0.87**	< 0.01	**2.86**	2.46	**3.17**	**2.48**	**4.51**
CNN	-2.14	(-9.61, **5.32**)	0.64	< 0.01	5.4	5.77	4.22	4.70	4.46
A-P Synthetise	5.24	(-9.73, 20.21)	0.67	< 0.01	5.04	4.92	8.20	4.59	10.29
CycleGAN	1.08	(-6.70, 8.87)	0.73	< 0.01	4.49	3.74	6.96	4.58	9.14
MultiScale CNN	0.79	(-7.78, 8.39)	0.76	< 0.01	3.00	**2.40**	4.06	3.37	6.16
Smart Fusion	1.16	(-6.60, 8.93)	0.74	< 0.01	4.62	3.64	5.65	3.66	4.70
RRWaveNet	0.98	(**-5.48**, 9.15)	0.78	< 0.01	3.34	2.72	4.56	2.68	5.49

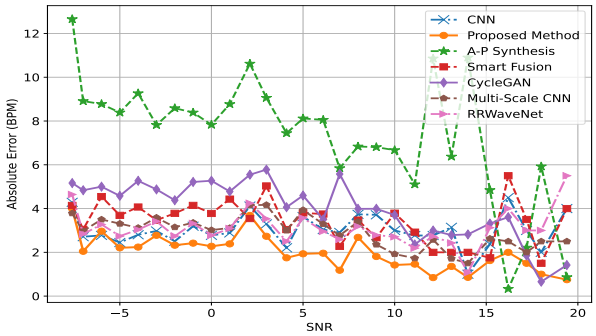
performance. In contrast, A-P Synthetise and CycleGAN display fluctuating absolute error rates, indicating higher sensitivity to the signal quality for estimating RRs and highlighting their susceptibility to noisy data. In addition, Multi-Scale CNN and RRWaveNet, while stable, exhibit a slightly higher error compared to the Proposed method.

5.4 Explanation of Proposed Method Effectiveness with SHAP

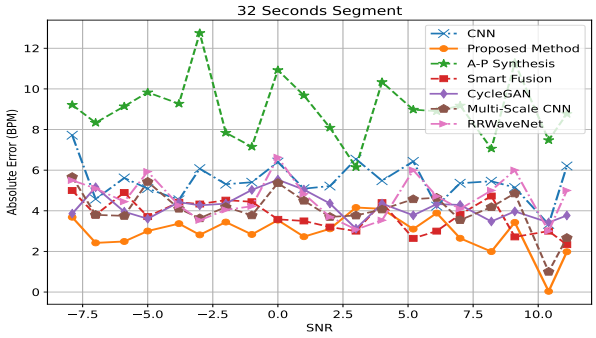
In this section, we explore the effect of input signals (i.e., PPG and tri-axial ACC) on RR estimation. To achieve this, we conduct SHapley Additive exPlanations (SHAP) [35] analysis to obtain comprehensive insights into the proposed predictive model. SHAP value indicates the association of each input signal (i.e., PPG and ACC signals) on the final prediction (i.e., RR estimation). To determine feature attribution, SHAP excludes input features and evaluates their effect on the prediction. Features that considerably impact the prediction yield a higher SHAP score. In implementing SHAP explanations, we utilized the Deep SHAP (i.e., available in the official SHAP package [35]). We calculated the mean SHAP values across each signal for all the test samples to assess how each input signal affects the RR estimation. Table 4 summarizes the mean SHAP values calculated for each signal in two benchmark PPG-DaLiA and WESAD datasets. As shown, the PPG signal has higher values in both datasets, indicating PPG has the highest impact on RR estimation, followed by the y-axis ACC. This information is particularly useful when aiming to reduce the model's complexity for computational or energy efficiency reasons. For example,in scenarios where energy or resource constraints necessitate a more lightweight model, the SHAP values indicate that certain signals, such as the x-axis accelerometer, can be excluded without significantly compromising performance. This approach enables the retention of high accuracy while optimizing the model to meet practical constraints. This includes factors such as energy consumption and computational capacity, making the model more suitable for deployment in real-world applications.

5.5 Discussion

The proposed approach leveraged the advantages of transfer learning to gain knowledge from the pre-training dataset to improve the model's performance on out-of-distribution samples in the target dataset. This indicates the superiority of transfer learning over training models from scratch, reinforcing its effective role in enhancing



(a) PPG-DaLiA Dataset, The activity types



(b) WESAD Dataset, The activity types

Fig. 6. Absolute Error for different methods based on the signal SNR for (a) PPG-DaLiA Dataset and (b) WESAD Dataset. The absolute error is calculated using the estimated RR and the reference RR.

Table 4. The mean shap value of PPG and tri-axial ACC signals in WESAD and PPG-DaLiA datasets.

SHAP value	PGG	x-axis ACC	y-axis ACC	z-axis ACC
PPG-DaLiA Dataset	4.8	1.1	4.06	3.8
WESAD Dataset	4.7	1.9	4.0	2.8

the RR estimation process. This approach enables the model to align its knowledge with the target dataset's characteristics, allowing it to make accurate predictions despite differences in data distribution.

The baseline methods encountered challenges, particularly when dealing with lower-quality target datasets and a diverse distribution of RR. For instance, the introduced model in [26], faces challenges in transferring knowledge from one dataset to another, given the differences in dataset RR distributions, leading to deliver higher MAE and RMSE. The existing methods also underestimate the datasets varying RR distributions due to factors like data demographic disparities or environmental conditions, making them unfit to adapt to these shifts. In addition, while other methods such as [4, 6, 22] have been tested on the dataset that has been acquired in a stationary position, our method demonstrates robustness across different types of activities, showcasing its

practical utility in real-life scenarios where users engage in various movements. Moreover, our proposed method is an end-to-end architecture, requiring no handcrafted features, whereas other methods such as [22, 45] are mainly based on time-domain and frequency domain analysis techniques. Another advantage of the proposed methods is the generation of robust results when applied to a wide range of datasets with a different range of signal quality. The proposed method is a deep learning-based model that not only produces the lowest error across all datasets but also calculates the RR straight from raw PPG and ACC signals (end-to-end).

Furthermore, in this work, we aimed to propose a foundation model for RR estimation. Unlike other work [4, 6, 42] , where the described methods are often constrained to theoretical implementation or specific use cases, our foundation model is publicly accessible and designed to be easily fine-tuned by researchers and practitioners. This allows for its application to a wide range of datasets and tailored to specific needs, providing a flexible and practical resource for various research and real-world settings. Furthermore, our approach is augmented with the integration of tri-axial ACC signals alongside PPG, which enhances the model's robustness, particularly in situations where PPG signals may be compromised due to motion or noise.

In addition, while our method leverages transfer learning to adapt a pre-trained model for RR estimation on new datasets, domain adaptation could also play a crucial role in handling domain shifts between datasets [27]. Domain adaptation focuses on aligning data distributions between the source and target domains, which can be beneficial in cases where data from different wearable devices exhibit significant variability due to sensor differences. In our study, we employ transfer learning to enhance the generalizability of our respiration rate estimation model across different datasets. This approach allows us to fine-tune a pre-trained model on a smaller target dataset, thereby improving performance despite variations in data distribution. Future work could explore the integration of domain adaptation techniques to further improve model generalization across various wearable devices and environments.

From both clinical [28, 34] and research [7] perspectives, RR estimation typically allows for an acceptable error margin of ±3 to ±4 breaths per minute, especially in critical diagnostic tools like the Early Warning Score. In this study, our method achieved 80.34% of measurements within ±4 breaths per minute, outperforming other methods such as Smart Fusion (62.29%), CycleGAN (48.03%), RRWaveNet (71.38%), CNN (69.83%), MultiScale CNN (73.83%), and A-P Synthetise (41.47%). This demonstrates that our approach yields the highest proportion of accurate results within ±4 breaths per minute. Compared to clinical studies like that of Latten et al., [28] which reported 78.2% within ±4 breaths per minute in controlled, low-noise settings, our results are equally competitive. However, unlike these studies, our data were collected during varying levels of activity, where motion artifacts and noise posed significant challenges. Despite these conditions, our method remained robust, showcasing its applicability in real-world, dynamic environments where noise is a prominent factor.

5.5.1 Limitation. Our evaluation was limited to datasets of healthy subjects. However, studies [2] showed that various health conditions– such as hyperventilation, and sleep apnea– might affect the accuracy of RR estimation methods. Our future work in this direction will consider improving the generalizability of the proposed method. To this end, we will train, fine-tune, and test the proposed model using data from individuals with health conditions and from diverse demographics. Another limitation is that the datasets used in this study contain data from a relatively short time period. Datasets with longer durations of data could provide more comprehensive insights into the performance and reliability of the proposed method, especially over extended periods and under varying conditions. In future work, we plan to incorporate longer-duration datasets to further evaluate and enhance the model's robustness.

6 Conclusion

This paper proposed a smartwatch-based RR estimation method enabled by raw PPG and tri-axial ACC. The proposed method comprised two parts. In the first part, a CNN-based network, including Multi-scale convolutions

augmented with dilated residual inception, 1D convolution layers, and dense layers, was trained using the Health Monitoring dataset to build the pre-trained model. The pre-trained model was presented and was fine-tuned on target datasets in the second part. The model was evaluated on two benchmark datasets (i.e., PPG-DaLiA and WESAD) collected with wrist-worn devices. We performed several experiments to compare the efficacy and accuracy of our model with six existing state-of-the-art methods. Upon performance, our method outperforms other methods, obtaining the lowest MAE and RMSE of 2.29 and 3.11 on PPG-DaLiA and 3.09 and 3.79 on the WESAD dataset, respectively, while the MultiScale CNN achieved MAE and RMSE of 2.68 and 3.5 on PPG-DaLiA and 5.13 and 5.67 on the WESAD dataset. Moreover, the performance of the methods across different activity types was assessed, where our method showed the smallest mean bias (i.e., 0.28 and 0.30) on both datasets according to the Bland-Altman method, as well as a narrower absolute error distribution compared to the baseline methods. Furthermore, the performance of the proposed method was studied from data quality perspective using the SNR metric, revealing that our model obtained the lowest error throughout the entire SNR range. We also showed that higher data quality led to lower absolute errors, while relatively poor SNR values correlated to greater absolute errors. Finally, by using the SHAP method, we demonstrated that PPG and y-axis ACC signals had more contribution in the final RR estimation.

Acknowledgments

This work was partially supported by the Finnish Foundation for Technology Promotion and the Nokia Foundation.

References

[1] 2021. my.clevelandclinic.org. Available: https://my.clevelandclinic.org/health/articles. Accessed March 2023.
[2] Al-Halhouli et al. 2021. Clinical evaluation of respiratory rate measurements on COPD (Male) patients using wearable inkjet-printed sensor. *Sensors* 21, 2 (2021), 468.
[3] Arman Anzanpour et al. 2020. Edge-assisted control for healthcare internet of things: A case study on ppg-based early warning score. *ACM Transactions on Internet of Things* 2, 1 (2020), 1–21.
[4] Seyed Amir Aqajari et al. 2021. An end-to-end and accurate ppg-based respiratory rate estimation approach using cycle generative adversarial networks. In *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 744–747.
[5] Behnam Askarian, Kwanghee Jung, and Jo Woon Chong. 2019. Monitoring of heart rate from photoplethysmographic signals using a Samsung Galaxy Note8 in underwater environments. *Sensors* 19, 13 (2019), 2846.
[6] Dayi Bian et al. 2020. Respiratory rate estimation using PPG: a deep learning approach. In *42nd international conference of the IEEE engineering in Medicine & Biology Society (EMBC)*. IEEE, 5948–5952.
[7] Martine JM Breteler, Eline J KleinJan, Daan AJ Dohmen, Luke PH Leenen, Richard van Hillegersberg, Jelle P Ruurda, Kim van Loon, Taco J Blokhuis, and Cor J Kalkman. 2020. Vital signs monitoring with wearable sensors in high-risk surgical patients: a clinical validation study. *Anesthesiology* 132, 3 (2020), 424–439.
[8] Bartosz Bujan, Tobit Fischer, Sarah Dietz-Terjung, Aribert Bauerfeind, Piotr Jedrysiak, Martina Große Sundrup, Janne Hamann, and Christoph Schöbel. 2023. Clinical validation of a contactless respiration rate monitor. *Scientific Reports* 13, 1 (2023), 3480.
[9] Peter Hand others Charlton. 2017. Breathing rate estimation from the electrocardiogram and photoplethysmogram: A review. *IEEE reviews in biomedical engineering* 11 (2017), 2–20.
[10] Wee Jian Chin, Ban-Hoe Kwan, Wei Yin Lim, Yee Kai Tee, Shalini Darmaraju, Haipeng Liu, and Choon-Hian Goh. 2024. A Novel Respiratory Rate Estimation Algorithm from Photoplethysmogram Using Deep Learning Model. *Diagnostics* 14, 3 (2024), 284.
[11] Simon John Rankin Cooper et al. 2014. Respiratory rate records: the repeated rate? *Journal of clinical nursing* 23, 9-10 (2014), 1236–1238.
[12] Michelle A Cretikos et al. 2008. Respiratory rate: the neglected vital sign. *Medical Journal of Australia* 188, 11 (2008), 657–659.
[13] Guo Dan, Zhijian Li, and Huijung Ding. 2015. A mother wavelet selection algorithm for respiratory rate estimation from photoplethysmogram. In *World Congress on Medical Physics and Biomedical Engineering, June 7-12, 2015, Toronto, Canada*. Springer, 962–965.
[14] Harry J Davies and Danilo P Mandic. 2023. Rapid extraction of respiratory waveforms from photoplethysmography: A deep corr-encoder approach. *Biomedical Signal Processing and Control* 85 (2023), 104992.
[15] Nurettin Özgür Doğan. 2018. Bland-Altman analysis: A paradigm to understand correlation and agreement. *Turkish journal of emergency medicine* 18, 4 (2018), 139–141.
[16] Emer P Doheny et al. 2020. Estimation of respiration rate and sleeping position using a wearable accelerometer. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 4668–4671.

[17] Michael Dommasch, Daniel Sinnecker, Petra Barthel, Alexander Müller, Ralf J Dirschinger, Alexander Hapfelmeier, Katharina M Huster, Karl-Ludwig Laugwitz, Marek Malik, and Georg Schmidt. 2014. Nocturnal Respiratory Rate Predicts Non–Sudden Cardiac Death in Survivors of Acute Myocardial Infarction. *Journal of the American College of Cardiology* 63, 22 (2014), 2432–2433.

[18] Serj Haddad et al. 2021. Ppg-based respiratory rate monitoring using hybrid vote-aggregate fusion technique. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 1605–1608.

[19] Vera Hartmann et al. 2019. Quantitative comparison of photoplethysmographic waveform characteristics: Effect of measurement site. *Frontiers in physiology* 10 (2019), 198.

[20] Kaiming He et al. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[21] Javier Hernandez et al. 2015. Biowatch: estimation of heart and breathing rates from wrist motions. In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. IEEE, 169–176.

[22] Nicholas Huang et al. 2021. Novel continuous respiratory rate monitoring using an armband wearable sensor. In *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 7470–7475.

[23] Walter Karlen et al. 2013. Multiparameter respiratory rate estimation from the photoplethysmogram. *IEEE Transactions on Biomedical Engineering* 60, 7 (2013), 1946–1953.

[24] Walter Karlen, Heng Gan, Michelle Chiu, Dustin Dunsmuir, Guohai Zhou, Guy A Dumont, and J Mark Ansermino. 2014. Improving the accuracy and efficiency of respiratory rate measurements in children using mobile devices. *PloS one* 9, 6 (2014), e99266.

[25] Walter Karlen, M Turner, Erin Cooke, Guy Dumont, and J Mark Ansermino. 2010. CapnoBase: Signal database and tools to collect, share and annotate respiratory signals. In *2010 Annual meeting of the society for technology in anesthesia*. Society for Technology in Anesthesia, 27.

[26] Kianoosh Kazemi, Iman Azimi, Pasi Liljeberg, and Amir M Rahmani. 2023. Robust CNN-based Respiration Rate Estimation for Smartwatch PPG and IMU. In *Proceedings of the 2023 10th International Conference on Bioinformatics Research and Applications*. 94–100.

[27] Wouter M Kouw and Marco Loog. 2018. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806* (2018).

[28] Gideon HP Latten, Michelle Spek, Jean WM Muris, Jochen WL Cals, and Patricia M Stassen. 2019. Accuracy and interobserver-agreement of respiratory rate measurements by healthcare professionals, and its effect on the outcomes of clinical prediction/diagnostic rules. *PLoS One* 14, 10 (2019), e0223155.

[29] JeeEun Lee and Sun K Yoo. 2020. Respiration rate estimation based on independent component analysis of accelerometer data: Pilot single-arm intervention study. *JMIR mHealth and uHealth* 8, 8 (2020), e17803.

[30] Jin Li et al. 2010. Comparison of respiratory-induced variations in photoplethysmographic signals. *Physiological measurement* 31, 3 (2010), 415.

[31] Jiyang Li et al. 2022. Enhancement of Remote PPG and Heart Rate Estimation with Optimal Signal Quality Index. In *2022 IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 1–4.

[32] Daniyal Liaqat, Mohamed Abdalla, Pegah Abed-Esfahani, Moshe Gabel, Tatiana Son, Robert Wu, Andrea Gershon, Frank Rudzicz, and Eyal De Lara. 2019. WearBreathing: Real world respiratory rate monitoring using smartwatches. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–22.

[33] L G Lindberg, H Ugnell, and PÅ Öberg. 1992. Monitoring of respiratory and heart rates using a fibre-optic sensor. *Medical and Biological Engineering and Computing* 30 (1992), 533–537.

[34] Feike J Loots, Irma Dekker, Ruo Chen Wang, Arthur RH van Zanten, Rogier M Hopstaken, Theo JM Verheij, Paul Giesen, and Marleen Smits. 2022. The accuracy and feasibility of respiratory rate measurements in acutely ill adult patients by GPs: a mixed-methods study. *BJGP open* 6, 4 (2022).

[35] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[36] Milad Asgari Mehrabadi et al. 2020. Sleep tracking of a commercially available smart ring and smartwatch against medical-grade actigraphy in everyday settings: instrument validation study. *JMIR mHealth and uHealth* 8, 11 (2020), e20465.

[37] David J Meredith et al. 2012. Photoplethysmographic derivation of respiratory rate: a review of relevant physiology. *Journal of medical engineering & technology* 36, 1 (2012), 1–7.

[38] R. Mølgaard et al. 2016. Effectiveness of respiratory rates in determining clinical deterioration: a systematic review protocol. *JBI Database of Systematic Reviews & Implementation Reports* (2016), 19–27.

[39] Andrea Nicolò, Carlo Massaroni, Emiliano Schena, and Massimo Sacchetti. 2020. The importance of respiratory rate monitoring: From healthcare to sport and exercise. *Sensors* 20, 21 (2020), 6396.

[40] Torben Noto et al. 2018. Automated analysis of breathing waveforms using BreathMetrics: a respiratory signal processing toolbox. *Chemical senses* 43, 8 (2018), 583–597.

[41] Christina Orphanidou. 2017. Derivation of respiration rate from ambulatory ECG and PPG using ensemble empirical mode decomposition: Comparison and fusion. *Computers in biology and medicine* 81 (2017), 45–54.

[42] Pongpanut Osathitporn, Guntitat Sawadwuthikul, Punnawish Thuwajit, Kawisara Ueafuea, Thee Mateepithaktham, Narin Kunaseth, Tanut Choksatchawathi, Proadpran Punyabukkana, Emmanuel Mignot, and Theerawit Wilaiprasitporn. 2023. RRWaveNet: A Compact End-to-End Multiscale Residual CNN for Robust PPG Respiratory Rate Estimation. *IEEE Internet of Things Journal* 10, 18 (2023), 15943–15952.
[43] J Anthony Parker, Robert V Kenyon, and Donald E Troxel. 1983. Comparison of interpolating methods for image resampling. *IEEE Transactions on medical imaging* 2, 1 (1983), 31–39.
[44] Marco AF Pimentel et al. 2016. Toward a robust estimation of respiratory rate from pulse oximeters. *IEEE Transactions on Biomedical Engineering* 64, 8 (2016), 1914–1923.
[45] Kapil Singh Rathore et al. 2022. MRNet-A Deep Learning Based Multitasking Model for Respiration Rate Estimation in Practical Settings. In *2022 IEEE 10th International Conference on Serious Games and Applications for Health (SeGAH)*. IEEE, 1–6.
[46] Attila Reiss et al. 2019. Deep PPG: Large-scale heart rate estimation with convolutional neural networks. *Sensors* 19, 14 (2019), 3079.
[47] Robert H Riffenburgh and Daniel L Gillen. 2020. *Statistics in medicine*. Academic press.
[48] Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. 2011. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. *Critical care medicine* 39, 5 (2011), 952–960.
[49] Samsung. 2021. Samsung Gear Sport Smartwatch. Retrieved on January 2022, from https://www.samsung.com/global/galaxy/gear-sport/.
[50] Axel Schäfer and Karl W Kratky. 2008. Estimation of breathing rate from respiratory sinus arrhythmia: comparison of various methods. *Annals of Biomedical Engineering* 36 (2008), 476–485.
[51] Fons Schipper et al. 2021. Estimation of respiratory rate and effort from a chest-worn accelerometer using constrained and recursive principal component analysis. *Physiological Measurement* 42, 4 (2021), 045004.
[52] Philip Schmidt et al. 2018. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*. 400–8.
[53] Shimmer. 2021. Shimmer Device Specification. Available from: https://www.shimmersensing.com/products/shimmer3-ecg-sensor#applications-tab. Accessed December 2021.
[54] Daniel Sinnecker, Michael Dommasch, Petra Barthel, Alexander Müller, Ralf J Dirschinger, Alexander Hapfelmeier, Katharina M Huster, Karl-Ludwig Laugwitz, Marek Malik, and Georg Schmidt. 2014. Assessment of mean respiratory rate from ECG recordings for risk stratification after myocardial infarction. *Journal of Electrocardiology* 47, 5 (2014), 700–704.
[55] Gary B Smith, David R Prytherch, Paul Meredith, Paul E Schmidt, and Peter I Featherstone. 2013. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* 84, 4 (2013), 465–470.
[56] Jost Tobias Springenberg et al. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).
[57] Simon Stankoski et al. 2022. Breathing rate estimation from head-worn photoplethysmography sensor data using machine learning. *Sensors* 22, 6 (2022), 2079.
[58] Xiao Sun et al. 2017. Sleepmonitor: Monitoring respiratory rate and body position during sleep using smartwatch. *ACM on interactive, mobile, wearable and ubiquitous technologies* (2017), 1–22.
[59] Christian Szegedy et al. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.
[60] Apple Team. 2023. Review your sleeping respiratory rate. https://support.apple.com/sr-rs/guide/watch/. [Accessed: (2024)].
[61] FitBit Team. 2022. Breathing Rate. https://dev.fitbit.com/build/reference/web-api/breathing-rate/ [Accessed: (2024)].
[62] Garmin Team. 2020. How Does My Garmin Device Measure Respiration Rate? https://support.garmin.com/en-US/. [Accessed: (2024)].
[63] Oura Team. 2020. How Accurate Is Oura"s Respiratory Rate? https://ouraring.com/blog/how-accurate-is-ouras-respiratory-rate/. [Accessed: (2024)].
[64] Xu Wang et al. 2023. A Lightweight Neural Network Based Respiratory Rate Estimation Approach Using PPG Signal. In *6th International Conference on Electronics Technology (ICET)*. IEEE, 1446–1449.