Biomedical Signal Processing and Control 112 (2026) 108688

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Assessing the impact of signal quality on heart rate detection from long-term clinical wrist PPG under varying cardiac rhythms

Yangyang Zhao [a,*], Olli Lahdenoja [a], Jonas Sandelin [a], Sepehr Seifizarei [a], Arman Anzanpour [a], Joonas Lehto [b], Joel Nuotio [b], Jussi Jaakkola [b], Arto Relander [b], Tuija Vasankari [b], Juhani Airaksinen [b], Tuomas Kiviniemi [b], Matti Kaisti [a], Tero Koivisto [a]

[a] University of Turku, Department of Computing, Vesilinnantie 5, Turku, 20014, Finland
[b] Heart Center, Turku University Hospital, Kiinamyllynkatu 4–8, Turku, 20520, Finland

ARTICLE INFO

Keywords:
Photoplethysmography (PPG)
Atrial fibrillation (AF)
Signal quality assessment (SQA)
Heart rate estimation
Motion artifacts
Machine learning
Wearable monitoring
Clinical validation

ABSTRACT

Reliable heart rate (HR) detection is essential for long-term cardiac monitoring, particularly in hospitalized patients with complex conditions. Due to its optical and non-invasive nature, photoplethysmography (PPG) is inherently susceptible to motion artifacts and noise. These challenges intensify under arrhythmic conditions such as atrial fibrillation (AF), where signal distortions may blur the boundary between poor-quality segments and pathological rhythms, potentially impairing downstream tasks like HR estimation. This study developed a signal quality assessment (SQA) algorithm designed for this high-risk clinical population and evaluated its robustness through HR estimation. We collected 24-hour synchronous PPG and electrocardiogram (ECG) recordings from 49 hospitalized cardiac patients, with all PPG segments manually annotated for quality. External validation was conducted using the MIMIC-IV dataset. To avoid dependence on specific segment lengths or classifier types, we assessed SQA performance using seven machine learning models and four segmentation lengths. The SQA framework was then applied to HR estimation to evaluate clinical utility. We implemented a Standard Deviation of Successive Differences (SDSD)-based peak filtering method and compared it with an autocorrelation-based approach under different cardiac rhythm conditions. Threshold tuning in both SQA classification and SDSD filtering was conducted to explore the balance between data usability and reliable HR estimation. The proposed model achieved an AUROC of 96.1% (Sinus Rhythm (SR) + AF), with 90.6% on MIMIC-IV. Predicted SQA labels closely matched manual annotations, with mean absolute error (MAE) differences of 0.08 bpm (SR+AF), 0.25 bpm (SR), 0.62 bpm (AF), and 0.53 bpm (MIMIC-IV). SDSD reduced MAE by 46.57% for SR+AF, 41.67% for SR, and 49.69% for AF, further demonstrating the effectiveness of integrating SQA into HR estimation workflows.

1. Introduction

Long-term monitoring of cardiac health is particularly important for elderly individuals, as it enables early detection and timely intervention of cardiovascular diseases [1,2]. Reliable heart rate (HR) measurement plays a central role in such monitoring. Traditionally, electrocardiography (ECG) has been the clinical standard for continuous cardiac monitoring, and numerous studies have focused on ECG-based classification and detection of heart conditions [3,4].

In recent years, with the rapid development of wearable technologies, photoplethysmography (PPG) has emerged as a promising alternative to ECG [5,6]. Compared to ECG, PPG offers easier wearability and lower cost, making it more suitable for long-term monitoring in daily life. In addition to its primary use in HR estimation, PPG enables the detection of cardiac arrhythmias such as atrial fibrillation (AF) [5–7], as well as monitoring of other physiological parameters. PPG is an optical technique that measures blood volume changes by

* Corresponding author.
E-mail addresses: yazhao@utu.fi (Y. Zhao), olanla@utu.fi (O. Lahdenoja), jojusan@utu.fi (J. Sandelin), sepehr.seifizarei@utu.fi (S. Seifizarei), armanz@utu.fi (A. Anzanpour), jojuleh@utu.fi (J. Lehto), joel.nuotio@utu.fi (J. Nuotio), jussi.jaakkola@utu.fi (J. Jaakkola), arnire@utu.fi (A. Relander), tuija.vasankari@tyks.fi (T. Vasankari), juhani.airaksinen@tyks.fi (J. Airaksinen), tuomas.kiviniemi@tyks.fi (T. Kiviniemi), mkaist@utu.fi (M. Kaisti), tejuko@utu.fi (T. Koivisto).

https://doi.org/10.1016/j.bspc.2025.108688
Received 22 April 2025; Received in revised form 11 August 2025; Accepted 18 September 2025

analyzing variations in light absorption in superficial vessels. It is commonly implemented in pulse oximeters, wrist-worn smartwatches, and other wearable sensors, enabling real-time cardiovascular tracking in both clinical and home environments [8].

However, compared to ECG, PPG signals are highly susceptible to disruptions caused by motion artifacts, poor perfusion, sensor misplacement, and individual physiological variability. These challenges are particularly evident during AF episodes, where the irregular pulse morphology and rhythm fluctuations introduce additional distortion [9]. The resulting signal degradation can increase false alarm rates, leading to unnecessary clinical interventions, and subsequently compromise downstream tasks such as HR estimation. This problem becomes more severe in remote or ambulatory healthcare scenarios, where continuous monitoring is required but signal quality can vary significantly due to changes in patient activity and environmental conditions [10].

To investigate these challenges, we introduced an signal quality assessment (SQA) algorithm tailored to the clinical characteristics of this high-risk cohort and validated its clinical applicability by evaluating its effectiveness in supporting downstream HR estimation tasks. We collected 24-h continuous ECG and PPG recordings from elderly patients hospitalized at Turku University Hospital due to recurrent AF or other cardiac conditions. Most participants had a prior history of heart disease and were re-admitted for related symptoms. Our SQA algorithm extracts a set of handcrafted features, including time-domain, frequency-domain, and template-matching-based heart rate variability (HRV) indicators. We then perform feature selection and evaluate the performance across seven different ML models and four segment lengths, ensuring the robustness of the approach beyond any specific model or window size. To address the impact of signal quality on HR estimation, we additionally developed a filtering strategy based on the standard deviation of successive RR interval differences (SDSD). This approach excludes unreliable segments and enhances HR estimation accuracy. In parallel, we implemented an autocorrelation-based [11] HR estimation method to serve as a comparative baseline. Furthermore, we analyzed different thresholding strategies for both SQA classification and RR interval filtering, aiming to identify a practical balance between data retention and estimation reliability under real-world clinical conditions. In summary, our main contribution are as follows:

1. We propose a machine learning (ML)-based SQA framework, tailored to the clinical characteristics of a high-risk hospitalized cohort. The framework integrates handcrafted feature extraction, feature selection, and evaluation across seven classifiers and four segment lengths, demonstrating robust performance across different cardiac rhythms.
2. We apply the predicted SQA outputs to the downstream task of HR estimation, validating the reliability of the SQA model from a reverse perspective.
3. We introduce a filtering strategy based on the SDSD to exclude noisy segments and improve estimation accuracy. In addition, we implement and evaluate an autocorrelation-based HR estimation method as a comparative baseline.
4. We systematically analyze different thresholding strategies for both SQA and SDSD filtering, enabling a tunable trade-off between signal coverage and HR accuracy in adaptive monitoring scenarios.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 describes the dataset and devices used in the study. Section 4 outlines the signal processing pipeline, including segmentation, feature extraction, and HR estimation methods. Section 5 presents the experimental results and comparative analyses. Section 6 discusses key findings and study limitations, and Section 7 concludes the study.

2. Related works

PPG SQA is essential for ensuring reliable HR and HRV estimation. While effective SQA can enhance HR estimation [12], recent studies have shown that even research-grade wearables struggle with HRV accuracy during active conditions [13]. This highlights the need for robust SQA methods, from early rule-based approaches to current ML and deep learning (DL) solutions.

2.1. Rule-based approaches for PPG SQA

Early studies on PPG SQA were mostly rule-based. For instance, Selvaraj et al. [14] proposed a MNA detection method based on statistical features such as kurtosis and Shannon entropy, using fixed thresholds to classify 60-s PPG segments as clean or contaminated. In a related work, Elgendi et al. [15] conducted a short-term study on 40 healthy adults under heat-stress conditions, and compared eight different signal quality indices (SQIs) to identify the most suitable metrics for PPG signal assessment. Vadrevu et al. [16] and Reddy et al. [17] both proposed real-time, on-device PPG quality assessment methods targeting long-term wearable health monitoring in IoT settings. Vadrevu's system used low-complexity features and a six-level hierarchical decision rule, validated on data from both public databases and lab recordings. Reddy's approach employed a first-order predictor coefficient (FOPC), amplitude, and saturation features with a five-level rule system.

2.2. Machine learning approaches for PPG SQA

Several studies have explored ML-based PPG SQA under long-term, free-living conditions, focusing on healthy populations. Pradhan et al. [8] collected 24-h wrist PPG and accelerometer data from 26 subjects (including 10 elderly participants) using Empatica E4 devices during natural daily activities. A random forest model was trained on nine handcrafted features. Moscato et al. [18] also utilized Empatica E4 devices to collect 24-h data from 31 healthy adults and proposed two classifiers trained on 19 features to support both HR estimation and morphological analysis. Feli et al. [19] presented an energy-efficient, semi-supervised method designed for real-time use on wearable devices, using PPG signals from 46 healthy adults over one day. Five computationally efficient features were selected. In contrast to long-term studies on healthy individuals, some works have targeted short-term PPG quality assessment in clinical settings. Li and Clifford [20] used beat-level features—such as template correlation, DTW distance, and clipping ratio—with an MLP classifier to assess 6-s ICU PPG segments from 104 adult patients in MIMIC-II. Pereira et al. [21] focused on arrhythmic contexts, developing a supervised model using 42 signal features. Data were collected from 13 elderly neuro-ICU patients and over 3700 general ICU cases, covering both fPPG and rPPG.

2.3. Deep learning approaches for PPG SQA

Recent advances in deep learning have led to various methods for assessing PPG signal quality, particularly on short-term data collected from healthy subjects. Chen et al. [22] proposed a SQA framework using STFT spectrograms of 10-s PPG segments, followed by ResNet18 classifiers. Similarly, Chatterjee et al. [12] developed a two-stage pipeline combining quantum pattern recognition (QPR)-based image encoding with a lightweight CNN for 10 s PPG quality classification. Sivanjaneyulu et al. [23] used raw PPG signals as input to a CNN model for IoMT applications, achieving up to 99.99% accuracy on noise-added signals. Roh and Shin [24] using recurrence plot–based images of beat-level segments (from 5-min healthy recordings), enabling a simple two-layer CNN to achieve 97.5% balanced accuracy without complex preprocessing. In addition, Sawangjai et al. [25] introduced an attention-based GAN with dual discriminators to remove motion artifacts from 10-s PPG segments without relying on auxiliary sensors. For short-term ICU settings, Liu et al. [26] proposed a lightweight hybrid model using 5-s segments and multiscale time-state images, achieving real-time performance on datasets including ICU patients. In contrast, long-term monitoring in healthy individuals was addressed by Naeini et al. [27], who assessed the reliability of 5-min PPG segments over 24 h using wearable data, focusing on HR and HRV consistency. For long-term ICU data, Shin [28] trained a CNN on large-scale MIMIC-III data, achieving 97.8% accuracy in classifying 5-s segments.

2.4. SQA validation via heart rate estimation

However, these methods did not incorporate downstream evaluations, such as HR estimation, to provide reverse validation of their SQA models, particularly for long-term AF monitoring in clinical settings. In some prior studies, HR estimation has been included in SQA research, either as a means to generate signal quality labels or for indirect validation. For example, Orphanidou et al. [29] and Khan et al. [30] defined signal quality based on whether HR could be reliably extracted, but neither quantitatively assessed how their SQA methods improved HR estimation accuracy. In contrast to studies that used HR only for label generation, some works further validated SQA effectiveness through downstream HR estimation tasks. Sukor et al. [31] developed a three-class SQA method based on morphological features of the PPG waveform and validated its effectiveness in signal cleaning through HR estimation experiments. Their study used manual annotations supported by ECG R-peak information to label each pulse and compared HR extracted from the cleaned PPG signal with a reference ECG. Additionally, Xu et al. [32] embedded the SQA module into a real-time HR estimation system and tested different quality assessment mechanisms under wearable and edge computing scenarios. Their results demonstrated that incorporating SQA could significantly reduce HR fluctuations and instability. However, these studies did not quantitatively analyze the impact of SQA under different cardiac conditions. In comparison, our study not only uses SQA for segment classification but also quantifies its impact on HR estimation under both sinus rhythm (SR) and AF, using long-term PPG data from hospitalized elderly patients with clinically verified AF episodes.

3. Data collection

3.1. Dataset I

The Dataset I measurements were conducted at the Heart Center, Turku University Hospital, Finland, in accordance with the Declaration of Helsinki guidelines. The study was approved by the Ethical Committee of the Hospital District of Southwest Finland, and written informed consent was obtained from all participants. All patient data were fully anonymized prior to analysis, and no identifiable information was stored or processed. Data transfer from the Philips wrist band to the hub was secured via Bluetooth 4.0 LE using 128-bit AES-CCM encryption, and hub-to-cloud communication was transmitted over an encrypted 4G cellular network. This research is part of the Moore4Medical (M4M) EU project, which aims to advance remote health monitoring through wearable-based sensing and signal analysis.

Dataset I consists of synchronous 24-h ECG and PPG recordings from 49 cardiac patients undergoing Holter monitoring as part of routine clinical care. All PPG signals were manually annotated for quality assessment. The dataset includes 35 patients with SR and 14 with AF, totaling 1009.21 h of data collected between September 2022 and August 2023. To ensure a balanced dataset not biased by rhythm distribution, we divided the data into training and testing sets. The training set includes 17 SR and 8 AF recordings, referred to as DSI-train-SR and DSI-train-AF, respectively. The test set comprises 18 SR and 6 AF recordings, referred to as DSI-test-SR and DSI-test-AF. These four subsets together constitute Dataset I. Background information and a summary of participants' clinical history are provided in Table 1.

3.2. Dataset II

This dataset includes 14 randomly selected PPG and ECG records, representing 83.2 h in total, from a total of 200 records in the MIMIC-IV Waveform database, which is part of the MIMIC-IV database. The MIMIC-IV database comprises de-identified electronic health records for patients admitted to the Beth Israel Deaconess Medical Center. It contains high-resolution physiological signals and measurements from critically ill patients. Our analysis is based on version 0.1.0 of the MIMIC-IV Waveform database. The database was published on PhysioNet by [33,34] and adheres to strict standards to protect patient privacy.

Table 1
Baseline characteristics and medical history of the study population: Total subjects N = 49 (26 Males).

Baseline characteristics		
Characteristic	Train set	Test set
Age mean (SD)	67.7 (13.15)	74.2 (11.17)
Height mean (SD)	170.6 (9.9)	169.2 (10.33)
Weight mean (SD)	86.4 (21.29)	84.1 (18.13)
BMI mean (SD)	29.4 (5.88)	29.5 (6.5)
Average recording time	20.2 (2.0)	16.1 (7.0)
Participants' medical history		
Medical condition	(%) Train set	(%) Test set
Hypertension	72	60
Dyslipidemia	40	40
Heart failure (HFrEF and HFpEF)	36	20
Coronary artery disease	32	60
Myocardial infarction	12	20
Diabetes mellitus type 2	20	20
Chronic lung disease	12	40
No relevant medical history	4	0

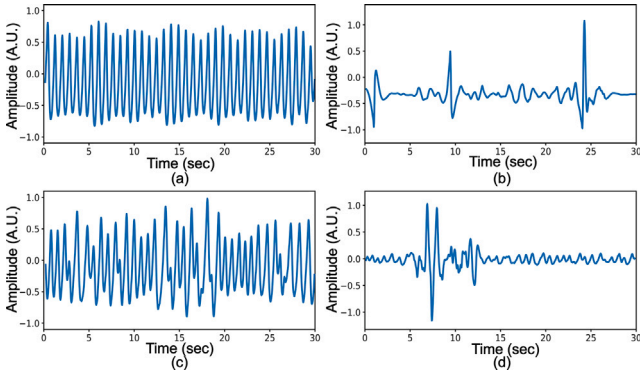

Fig. 1. Good PPG segment and bad PPG segment. (a) Good for SR; (b) Bad for SR; (c) Good for AF; (d) Bad for AF.

3.3. Quality label rules and mapping

We proceed with the annotation in three steps. Firstly, to assist annotators to recognize signals better, we applied a three-order Butterworth bandpass filter within the 0.5 Hz to 8 Hz frequency range, followed by the visualization of the filtered signal. Subsequently, a good signal must meet the following criteria: (1) It has a clear waveform with easily identifiable peaks and distinguishable systolic and diastolic phases; (2) Peaks must be identifiable even in irregular waveforms; (3) It does not exhibit random noise or artifacts; (4) It can include short bursts of high amplitude within an overall good signal. Additionally, these conditions must be met for at least a continuous 10 s. Fig. 1 shows some typical example signals in the Dataset I.

To verify the reliability of manual annotation, we validated it by plotting the signals and annotations together. Inaccurate labels are re-annotated through multiple observations. As shown in Fig. 2, a complete subject's signal graph is depicted, with a pink background indicating good quality and a white background indicating bad quality signal. The distribution of durations based on quality across training and test sets is shown in Table 2.

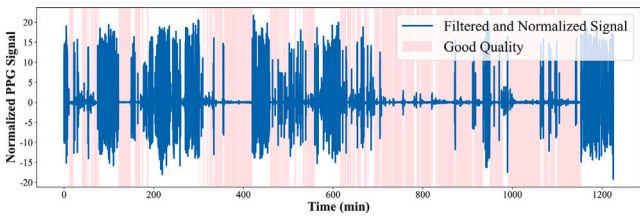

Fig. 2. An example of signal quality assessment (SQA). Filtered PPG waveform including noise (dark blue) is shown. The pink-colored area means a 'Good' manually annotated interval.

Table 2
Duration (h) and percentage of PPG signals categorized by quality (good/bad), with used datasets.

Type	Dataset	Good	Bad	Total
Train	SR+AF	270.6 h (56.8%)	205.9 h (43.2%)	476.5 h
	SR	174.2 h (54%)	148.3 h (46%)	322.5 h
	AF	96.4 h (62.6%)	57.6 h (37.4%)	154.0 h
Test	SR+AF	311.6 h (58.5%)	221.1 h (41.5%)	532.8 h
	SR	266.9 h (60.2%)	176.5 h (39.8%)	443.4 h
	AF	44.7 h (50.06%)	44.6 h (49.94%)	89.3 h
Transform	MIMIC	60.0 h (72.1%)	23.2 h (27.9%)	83.2 h

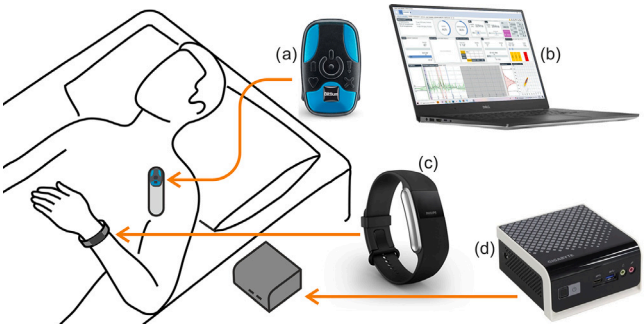

Fig. 3. (a) Device for ECG: Bittium Faros; (b) Bittium Cardiac Navigator software; (c) Device for PPG: Philips Datalogger (PDL) wristband device; (d) MSX gateway.

3.4. Devices for data collection

3.4.1. Bittium Faros device for ECG data collection

Our dataset encompasses simultaneous PPG and ECG data. Fig. 3 illustrates the entire data collection equipment and the locations of the sensors.

As shown in Fig. 3(a), the Bittium Faros™ 360 is a compact wireless ECG sensor. The device can act either as an online ECG monitor or as an offline Holter monitor. In our setup, we used the device in Holter mode recording a single channel ECG signal with a sampling rate of 125 Hz. The Faros device integrates seamlessly with the Bittium Cardiac Navigator software, which provides millisecond-precise rhythm label output for ECG data and extracts RR intervals from the detected heartbeats as shown in Fig. 3(b). In the measurement preparation phase, the Faros Manager software assists in managing the data collection settings. The sensor is wearable via quick-fix electrodes as shown in Fig. 3.

3.4.2. Philips Datalogger for PPG data collection

The Philips Datalogger (PDL) is a wrist-worn PPG sensor and accelerometer sensor datalogger that records PPG signal at a sampling frequency of 32 Hz shown in Fig. 3(c). The PDL's PPG sensor operates in a reflective mode using two green LED lights. The PDL collects PPG signal in its internal memory and transfers data to a gateway device via Bluetooth Low Energy. Shown in Fig. 3(d), the gateway device is an ultra-compact PC configured to receive data from the PDL and send it to a cloud server via 3G connection. The PDL wristband has sufficient battery to operated continuously for more than one day. An instance of an online InfluxDB database stores ECG and PPG signals, making exporting both signals with a common synchronized timestamp feasible.

4. Algorithm overview

Fig. 4 illustrates the overall framework, which includes the proposed SQA algorithm and its clinical validation through the downstream task of HR estimation. While the SQA module focuses on segment-wise quality classification, the subsequent HR estimation phase serves to assess the practical impact of SQA under different cardiac rhythm conditions. In the training phase, preprocessing steps, including bandpass filtering, normalization, segmentation, and peak detection. SQA is performed using feature extraction and selection, with five key features. ML models are trained in a leave-one-subject-out (LOSO) manner to classify signal quality. In the testing phase, the trained SQA model is used to predict signal quality labels, which are then integrated into the downstream HR estimation pipeline. HR is computed under three conditions: (1) No Filter, where HR is directly estimated from raw PPG signals; (2) SDSD-filtered, where implausible peaks are excluded based on the SDSD; and (3) Autocorrelation-based, where HR is derived by identifying the dominant periodicity via autocorrelation analysis. Each approach is evaluated using both predicted and manually annotated quality labels to assess the impact of SQA and filtering strategies on HR estimation accuracy. All HR estimates are benchmarked against ECG-derived reference values.

4.1. Pre-processing

4.1.1. Filtering and normalization

We employed a third-order Butterworth bandpass filter to remove noise outside the frequency range of 0.5 to 8 Hz, followed by normalizing the signal to have a mean of zero and a standard deviation of one, in preparation for further analysis.

4.1.2. Segmentation

We divided the filtered signals into discrete, non-overlapping segments, specifically into durations of 30 s, 60 s, 120 s, and 360 s for different experiments. Our definition of segment quality states that a signal segment is labeled as 'good' if more than 80% of the signal is of good quality.

4.2. Feature extraction

In this section, we selected 15 features from time domain, frequency domain, and template matching method. The descriptions of the 15 features are presented in Table 3. Following this, a detailed description of five selected features follows.

4.2.1. Absolute Signal-to-Noise Ratio (ASNR)

The Absolute Signal-to-Noise Ratio (ASNR) quantifies the quality of a filtered PPG signal, defined as the ratio of the variance of its absolute values ($\sigma^2_{\text{abs(signal)}}$) to the variance of the signal itself (σ^2_{signal}), expressed as:

$$\text{Absolute Signal-to-Noise Ratio} = \frac{\sigma^2_{\text{abs(signal)}}}{\sigma^2_{\text{signal}}} \quad (1)$$

This metric emphasizes the signal's overall strength relative to its variability.

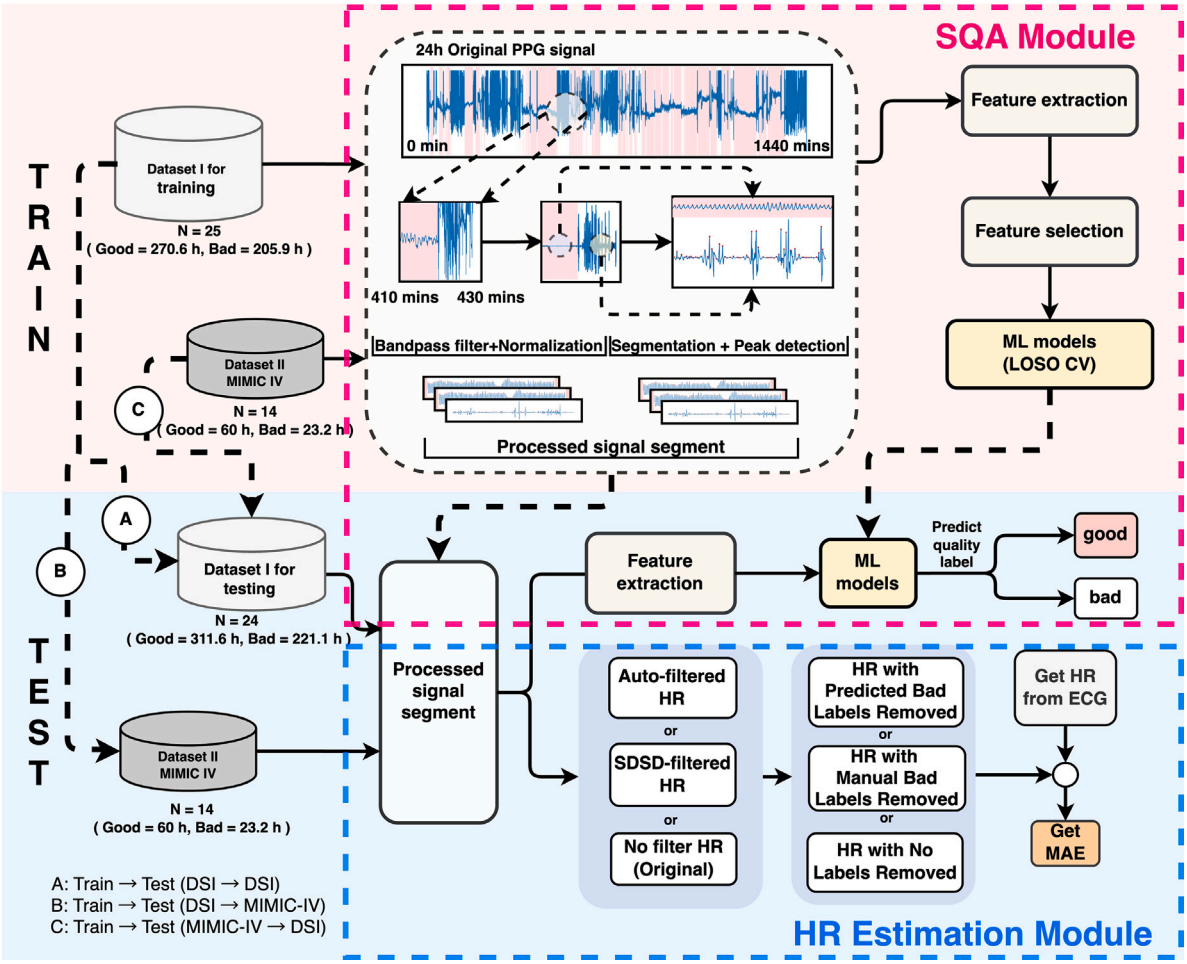

Fig. 4. Overview of the proposed framework for PPG SQA and its validation through downstream heart rate (HR) estimation. The process is divided vertically into a training phase (top) and a testing phase (bottom), and horizontally into two functional modules: the SQA Module (enclosed by a pink dashed box) and the HR Estimation Module (enclosed by a blue dashed box). In the testing phase, predicted quality labels from the trained SQA model are used to evaluate its effectiveness through downstream HR estimation. Dataset I (clinical recordings) and Dataset II (MIMIC-IV) are used in three train–test configurations (A–C). Raw PPG signals are preprocessed through filtering, normalization, segmentation, and peak detection. The SQA module involves feature extraction, feature selection, and machine learning (ML) model training using leave-one-subject-out (LOSO) cross-validation. HR is evaluated against ECG-derived HR, and performance is measured using mean absolute error (MAE).

Table 3
Summary of SQI features in PPG signal analysis.

Feature domain	SQI	Description	Ref.
Time domain	Skewness index	Assesses symmetry in PPG signal distributions, indicating possible corruption	[35,36]
	Kurtosis index	Evaluates the prominence of distribution tails, reflecting signal noise and outliers	[14,37]
	Shannon entropy	Quantifies signal unpredictability by measuring distributional deviations	[14,38]
	Zero cross rate	Counts axis oscillations to assist in state differentiation	[15,39]
	Absolute signal-to-noise ratio	Compares signal variance against noise, highlighting clarity	[40]
	Signal consistency range	Determines central variability by the 25th to 75th percentile range	[19]
	Cardiac pulse power	Computes the sum of squared amplitudes for signal integrity	[19]
	Lempel–Ziv complexity	Assesses signal predictability via patterns in a binary sequence	[41,42]
Frequency domain	Std of PSD	Indicates variability in power distribution across frequencies	[43]
	Relative power SQI	Measures power in critical bands to reflect signal integrity	[44]
Template matching	Waveform correlation index	Measures average similarity of waveforms to a reference template	[31]
	Euclidean distance	Computes mean distance between waveforms and a reference	[31]
	Linear resampling	Correlation of time-normalized waveforms with a reference	[20]
	Clipping quality index	Detects saturation by checking non-clipped percentage of beats	[20]
	Dynamic correlation index	Uses historical waveform data for correlation-based indexing	[20,45]

Note: The five features that have been selected for final use are highlighted in bold.

4.2.2. Cardiac pulse power

Cardiac Pulse Power in PPG signal analysis is an important indicator for assessing signal integrity [19]. It is defined as the area under the squared amplitude of the cardiac pulse.

$$\text{Cardiac pulse power} = N \sum_{n=1}^{N} \left| x_{sp}(n) \right|^2 \quad (2)$$

In this formula (2), $x_{sp}(n)$ represents the value of the signal at the nth sample of one cardiac pulse, and N is the number of samples in the cardiac pulse.

4.2.3. Lempel–Ziv complexity

The Lempel–Ziv complexity, originally proposed by Lempel and Ziv in 1976 [41]. This method involves converting the signal into a binary sequence based on its median, followed by a transformation into a string. The complexity is calculated by identifying new patterns within the string, with the final value normalized between 0 and 1 [42].

4.2.4. Euclidean distance

The Euclidean distance initially extracts and aligns individual waveforms to generate a standard template waveform by averaging these aligned waveforms. The Euclidean distance between each waveform and this template is calculated, the average of these distances assesses the overall signal consistency, where lower distances indicate a closer match, reflecting higher signal quality [31].

4.2.5. Linear resampling

The linear resampling method uses the same method as the Euclidean distance to obtain the template, and then resamples each waveform by linear stretching or compression to match the length of the template. Linear resampling is determined by averaging the correlation coefficients between the resampled waveforms and the template [20].

4.3. Heart rate detection

4.3.1. Peak detection and filtering in PPG signals

The Automatic Multi-scale Peak Detection (AMPD) algorithm [46] was applied for peak detection. This algorithm uses a multiscale technique to detect local maxima in the signal.

To ensure the effectiveness of feature analysis based on detected peaks, it is necessary to verify the accuracy of peaks detected in PPG signals, particularly when compared with ECG data. Considering that approximately 41.2% of the signals are of bad quality, these segments fail to provide accurate HR information. To filter out unreliable peaks from the data, we designed a filter targeting implausible HRs. This filter initially employed a median filter with a kernel size of 7 to smooth the data, reducing the impact of noise and outliers. Subsequently, we applied SDSD metric based filter for beat intervals. We experimentally set the threshold for SDSD as 0.35. In the filtered data for testing, we further excluded HRs below 40 or above 200.

4.3.2. Autocorrelation method for HR estimation

In the process of calculating HR from PPG signals using the autocorrelation method [11], we begin by employing a 5-s window to calculate the signal's autocorrelation at the zero-lag position, where there is always an initial peak due to the signal's perfect match with itself at this point. The first significant peak following the zero-lag peak reveals the most prominent repetitive pattern in the signal, namely, the heartbeat cycle. By measuring the time distance between these peaks, the HR can be directly calculated.

4.3.3. ECG as reference

For Dataset I, we use the intervals between heartbeats generated by the Bittium Cardiac Navigator software, and for Dataset II, we perform peak detection using the Pan-Tompkins [47] method to identify cardiac peaks, followed by HR calculation. To ensure consistent HR processing standards for both PPG and ECG data, we applied a median filter with a kernel size of 7 to smooth the data after obtaining the HR from ECG. Additionally, we excluded HR below 40 or above 200. We assessed the difference between HR from PPG and ECG using the MAE, which is the average absolute difference between HRs from both methods across all segments.

4.4. Machine learning pipeline

4.4.1. LOSO-CV and test

We applied LOSO-CV to Dataset I (for train) and tested the models on both Dataset I (for test) and Dataset II. Additionally, we conducted LOSO-CV exclusively on Dataset II and tested the trained model on Dataset I (for test). We divided the Dataset I (for train) into three subsets: one containing all SR and AF data, one containing only SR data, and one containing only AF data. Similarly, we divided the test data into three subsets: SR + AF, SR, and AF-only datasets. We conducted testing on each of these datasets separately. The data were segmented into 30 s, 60 s, 120 s, and 360 s and trained using a Logistic Regression (LR) model to examine which segment length yields the best performance and to ensure that the method does not depend on a specific window length. For Dataset II, since the focus was on cross-dataset validation, we only evaluated 60 s segments.

To verify the robustness and general applicability of the proposed features across different ML model types, we evaluated seven ML models and performed LOSO-CV: LR (max 1000 iterations), AdaBoost (100 estimators), GBDT (100 estimators, learning rate 1.0, max depth 1), KNN (K = 5), SVC (linear kernel, auto gamma), RF (100 estimators, max depth 5), and DT (max depth 5). All models were trained using the full feature set and tested across different segment lengths. The LOSO-CV required excluding segments from the training data that came from the same subject as the test data.

4.4.2. Feature selection

We employed the RLScore library [48] for feature selection, using the Greedy RLS algorithm, which is based on Regularized Least Squares (RLS). The algorithm starts with an empty feature set and iteratively adds the feature that yields the greatest improvement in the RLS objective. It uses an efficient matrix update strategy to avoid retraining the model at each step. We applied this method to select five features from the initial candidate pool for the SQA classification task.

5. Results

In this section, we present the results of three experiments. **Experiment I** evaluated five signal quality features across four datasets (SR + AF, SR, AF, and MIMIC-IV) and four segment lengths, using seven ML models. We also included a model interpretability analysis to investigate how different features contribute to classification performance. **Experiment II** served as a clinical validation of the proposed SQA method by applying it to the downstream task of HR estimation. We compared PPG-derived HR with ECG-derived HR under various conditions, including the use of the SQA model and two filtering approaches: SDSD and autocorrelation. **Experiment III** further investigated the impact of SDSD-based filtering and ML-based thresholds on HR estimation accuracy (MAE) and signal coverage.

Table 4
LOSO-CV AUC scores for LR model using five selected features, trained on SR+AF, SR, and AF data over different segment lengths.

Feature	Length	SR+AF	SR	AF	Mean
ASNR	30 s	90.7	92.6	88.2	90.5
	60 s	90.1	91.5	88.1	89.9
	120 s	88.0	89.3	86.1	87.8
	360 s	82.2	84.1	79.0	81.8
Euclidean distance	30 s	81.6	81.6	79.1	80.8
	60 s	83.5	83.6	81.2	82.8
	120 s	84.1	84.2	81.7	83.3
	360 s	83.8	84.0	81.2	83.0
Cardiac pulse power	30 s	76.2	73.8	76.0	75.3
	60 s	80.2	78.0	80.6	79.6
	120 s	82.3	80.5	82.5	81.8
	360 s	83.4	82.1	83.2	82.9
Linear resampling	30 s	85.2	89.8	78.0	84.3
	60 s	86.4	91.0	79.7	85.7
	120 s	86.6	91.1	80.6	86.1
	360 s	85.5	90.8	79.1	85.1
Lempel–Ziv complexity	30 s	65.5	61.7	68.0	65.1
	60 s	65.9	62.2	67.9	65.3
	120 s	66.5	61.8	66.4	64.9
	360 s	63.8	60.9	63.1	62.6
All features	30 s	93.0	94.4	90.7	92.7
	60 s	93.8	95.0	91.5	93.4
	120 s	93.9	95.1	91.8	93.6
	360 s	93.3	94.9	90.2	92.8
All classifiers' mean	30 s	92.3	93.8	89.7	92.0
	60 s	93.4	94.6	90.8	92.9
	120 s	93.4	94.5	90.7	92.9
	360 s	92.6	93.9	88.1	91.5

5.1. Experiment I: Quality estimation

5.1.1. Train and LOSO-CV

The purpose of Experiment I was to verify the performance of the SQA model designed based on this study. Table 4 shows the performance of the LR model in classifying signal quality. The absolute ASNR was the top feature for 30 and 60 s segments, indicated by the highest Area Under the Receiver Operating Characteristic values scores (AUROCs). The cardiac pulse power improved in its classification ability as the segment length increased, peaking at 360 s. Linear resampling consistently aided in SR data classification across all segment lengths, especially at 120 s. However, the Lempel–Ziv complexity underperformed in the SQA across all lengths, indicating its limited utility in LR models. The 'All features' part indicates that using a combination of all features yielded the best results. We also calculated the average AUROC for seven models using all combined features, and included these averages in Table 4 to compare model performance, it is evident that all models performed best on datasets containing only SR data, and poorer on AF datasets. Moreover, across all datasets, the segments of 60 and 120 s yielded a relatively stable performance, whereas the longest segments of 360 s fared the worst.

5.1.2. Test and cross-site data validation

Table 5 summarizes the performance (AUROC and MCC) of seven ML models trained on SR + AF, SR-only, AF-only, and MIMIC datasets using 60 s segments, and tested on SR + AF, SR, AF, and external MIMIC test sets. Due to space constraints, only 60 s AUROC-based results are shown here; results for other segment lengths (30 s, 120 s, 360 s) and the full set of metrics are provided in the Supplementary Material (Tables S1–S4).

According to Table 5, vertical comparisons across test sets show that SR testing consistently yields the highest AUROC, with SR-trained models achieving up to 95.41% AUROC (96.2% for RF). Horizontally, across different training datasets, SR + AF training outperforms SR-only and AF-only models on DS I. Comparing the seven classifiers, AUROC was generally stable across segment lengths and test subsets, except for the MIMIC dataset, where values ranged from 77.03% to 84.73%. As shown in Table S1, AUROCs on the MIMIC test set at 30 s ranged from 79.4% to 90.6%.

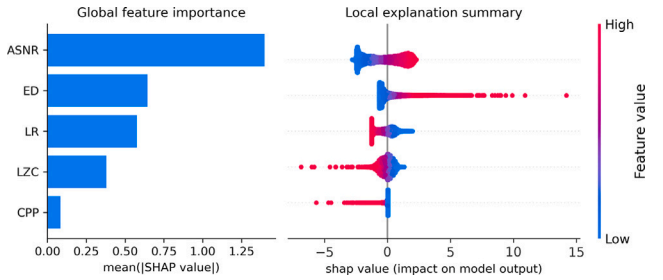

Fig. 5. SHAP summary plot showing global feature importance (left) and local feature impact (right) for the selected features. ASNR: Absolute Signal-to-Noise Ratio; ED: Euclidean Distance; LR: Linear Resampling SQI; LZC: Lempel–Ziv Complexity; CPP: Cardiac Pulse Power. The left panel shows the mean absolute SHAP value for each feature, indicating overall importance. The right panel shows the distribution of SHAP values per feature, with color representing the feature value (red = high, blue = low).

In terms of segment length, shorter segments generally yielded slightly better classification performance. The 30-s segments achieved the highest overall AUROC, followed by 60 s, 120 s, and 360 s. For instance, in Table S1, the average AUROC on the SR + AF test set decreased from 95.01% (30 s) to 93.84% (360 s), while on the MIMIC dataset it dropped from 86.06% to 80.69%. As shown in Table S4, models trained on the MIMIC dataset achieved an average AUROC of 88.24% on the SR-only test set, 87.01% on the SR + AF test set, and 80.44% on the AF-only test set.

Finally, the MCC results followed trends similar to AUROC, reinforcing the consistency of model predictions across different training–testing combinations. Higher MCC values were observed when training and testing conditions were similar, whereas cross-dataset evaluations, especially involving the MIMIC dataset, yielded lower MCCs, highlighting the domain shift between datasets.

5.1.3. Model interpretability

To better interpret the prediction mechanism of the proposed SQA model, we utilized SHapley Additive exPlanations (SHAP) [49] to assess both global and local feature contributions. As illustrated in Fig. 5, the global importance plot (left) shows that ASNR is by far the most influential feature, followed by Euclidean Distance, Linear Resampling, Lempel–Ziv Complexity, and Cardiac Pulse Power. These features respectively reflect signal-to-noise characteristics, morphological deviation from a reference waveform, temporal consistency, signal complexity, and frequency content—all crucial for reliable PPG quality assessment. The local explanation summary (right) reveals how specific feature values impact the model's output. Each point represents a SHAP value for a single prediction, with color indicating the feature's actual value (red = high, blue = low). For ASNR, higher values (in red) strongly increase the likelihood of high-quality classification, while lower ASNR values reduce it—consistent with physiological expectations. Similarly, larger Euclidean distances from the template (also red) push predictions toward poor quality, highlighting the model's sensitivity to waveform distortion. In contrast, features such as CPP and LZC show smaller but still interpretable effects. In addition, the confusion matrices (Supplementary Fig. S1) show consistently high classification accuracy across test sets, with slightly more false negatives under noisier conditions.

We inspected representative cases from the supplementary error analysis figures (Supplementary Figs. S2 and S3). The observations

Table 5
Quality classification results on the test set using a combination of the five selected features. 'SR+AF', 'SR', 'AF', and 'MIMIC' in the first row represent models trained on the respective datasets, while the same labels in the first column indicate on which test set the model is evaluated.

Test	60 s	Train: SR+AF		Train: SR		Train: AF		Train: MIMIC		Mean	
	Model	AUROC	MCC	AUROC	MCC	AUROC	MCC	AUROC	MCC	AUROC	MCC
SR+AF	LR	94.9 (94.7–95.1)	0.755	95.0 (94.7–95.2)	0.769	94.6 (94.4–94.9)	0.719	92.1 (91.8–92.4)	0.682	94.15	0.731
	Ada	95.6 (95.3–95.7)	0.772	95.4 (95.1–95.6)	0.769	95.3 (95.1–95.5)	0.753	83.8 (83.3–84.2)	0.471	92.53	0.691
	GBDT	95.6 (95.4–95.8)	0.771	94.6 (94.4–94.9)	0.763	95.3 (95.1–95.5)	0.750	87.4 (87.0–87.7)	0.606	93.23	0.723
	SVC	94.9 (94.7–95.1)	0.759	94.9 (94.6–95.1)	0.770	94.7 (94.5–94.9)	0.739	91.9 (91.6–92.2)	0.669	94.10	0.734
	KNN	92.8 (92.5–93.1)	0.745	93.0 (92.7–93.3)	0.752	92.2 (91.9–92.4)	0.709	89.7 (89.3–90.0)	0.674	91.93	0.720
	RF	95.7 (95.4–95.9)	0.776	95.7 (95.5–95.9)	0.778	95.1 (94.9–95.3)	0.747	93.5 (93.3–93.8)	0.656	95.00	0.739
	DT	95.1 (94.8–95.3)	0.759	95.1 (94.9–95.4)	0.763	94.4 (94.1–94.6)	0.745	70.7 (70.1–71.3)	0.509	88.83	0.694
	Mean	94.94	0.762	94.81	0.766	94.51	0.736	87.01	0.610	92.82	0.719
SR	LR	95.4 (95.1–95.6)	0.764	95.5 (95.3–95.7)	0.783	95.1 (94.9–95.4)	0.717	92.8 (92.5–93.1)	0.715	94.70	0.745
	Ada	96.0 (95.8–96.2)	0.784	95.9 (95.7–96.1)	0.786	95.7 (95.5–96.0)	0.755	85.2 (84.8–85.7)	0.505	93.20	0.708
	GBDT	96.0 (95.8–96.2)	0.781	95.3 (95.0–95.5)	0.778	95.7 (95.5–96.0)	0.752	90.0 (89.6–90.3)	0.656	94.25	0.742
	SVC	95.3 (95.1–95.6)	0.769	95.4 (95.2–95.6)	0.786	95.2 (94.9–95.4)	0.744	92.6 (92.3–92.9)	0.704	94.63	0.751
	KNN	93.4 (93.1–93.7)	0.754	93.9 (93.6–94.2)	0.771	92.5 (92.2–92.8)	0.710	91.2 (90.8–91.5)	0.702	92.75	0.734
	RF	96.1 (95.8–96.3)	0.785	96.2 (96.0–96.4)	0.796	95.4 (95.2–95.6)	0.749	95.0 (94.7–95.2)	0.694	95.68	0.756
	DT	95.5 (95.3–95.7)	0.771	95.7 (95.5–95.9)	0.784	94.8 (94.5–95.0)	0.750	70.9 (70.3–71.6)	0.528	89.23	0.708
	Mean	95.39	0.773	95.41	0.783	94.91	0.643	88.24	0.643	93.49	0.710
AF	LR	92.9 (92.1–93.6)	0.700	92.7 (91.9–93.4)	0.683	92.5 (91.7–93.2)	0.696	88.5 (87.6–89.6)	0.481	91.65	0.640
	Ada	93.7 (93.0–94.3)	0.707	93.0 (92.3–93.7)	0.661	93.4 (92.7–94.0)	0.718	75.0 (73.6–76.3)	0.185	88.78	0.568
	GBDT	93.8 (93.1–94.4)	0.712	91.7 (91.0–92.4)	0.670	93.5 (92.8–94.1)	0.709	72.8 (71.5–74.1)	0.252	87.95	0.586
	SVC	92.9 (92.1–93.6)	0.699	92.5 (91.7–93.2)	0.677	92.5 (91.8–93.2)	0.694	88.3 (87.3–89.3)	0.456	91.55	0.632
	KNN	89.7 (88.8–90.6)	0.687	88.1 (87.1–89.0)	0.638	90.1 (89.2–90.9)	0.675	81.1 (80.0–82.1)	0.505	87.25	0.626
	RF	93.9 (93.2–94.5)	0.723	93.7 (93.1–94.3)	0.667	93.6 (93.0–94.3)	0.715	86.5 (85.4–87.5)	0.419	91.93	0.631
	DT	93.1 (92.3–93.7)	0.688	92.7 (91.9–93.3)	0.642	92.4 (91.6–93.1)	0.691	70.9 (69.5–72.6)	0.368	87.28	0.597
	Mean	92.86	0.703	92.1	0.663	92.57	0.700	80.44	0.381	89.49	0.611
MIMIC	LR	83.1 (81.9–84.4)	0.507	84.8 (83.7–85.9)	0.527	85.0 (83.9–86.2)	0.333	–	–	84.30	0.456
	Ada	87.3 (86.1–88.3)	0.545	78.9 (77.7–80.1)	0.417	83.9 (82.8–85.2)	0.476	–	–	83.37	0.479
	GBDT	85.9 (84.7–87.0)	0.523	79.9 (78.6–81.3)	0.358	83.3 (82.1–84.6)	0.490	–	–	83.03	0.457
	SVC	83.5 (82.3–84.7)	0.526	86.1 (85.1–87.2)	0.529	84.6 (83.4–85.7)	0.397	–	–	84.73	0.484
	KNN	77.9 (76.4–79.4)	0.481	74.1 (72.7–75.6)	0.363	72.5 (71.2–73.9)	0.411	–	–	74.83	0.418
	RF	80.9 (79.5–82.2)	0.574	85.8 (84.7–86.9)	0.543	76.6 (75.1–78.2)	0.488	–	–	81.10	0.535
	DT	85.3 (84.1–86.5)	0.522	67.1 (65.5–68.5)	0.262	78.7 (77.4–80.0)	0.359	–	–	77.03	0.381
	Mean	83.41	0.525	79.53	0.428	80.66	0.422	–	–	81.20	0.459

indicate that longer segment lengths can lead to less precise quality boundaries, while shorter segments may provide finer granularity. Occasional mismatches between predictions and manual annotations suggest potential labeling inaccuracies.

5.2. Experiment II: HR estimation

5.2.1. HR estimation using PPG and ECG reference

To assess the impact of signal quality on HR estimation, we compared HR derived from original and quality-filtered PPG signals against reference ECG HR. For the PPG-derived HR, we compared two different algorithms for cleaning: one using SDSD for peak removal, and the other directly obtaining HR using autocorrelation without peak detection. Additionally, the original data without any filtering was also used for comparison. Fig. 6 provides a comparative scatterplot of HR data derived from ECG and PPG. It includes scatterplots of the original HR data as well as scatterplots of the HR data after the removal of unreasonable peaks and labels predicted to be of bad quality. According to Fig. 6, the left figure illustrates a higher concentration of SR samples, whereas AF samples are more scattered. After removing unreasonable peaks and predicted bad-quality labels, as shown in the right figure, the number of AF samples significantly decreases.

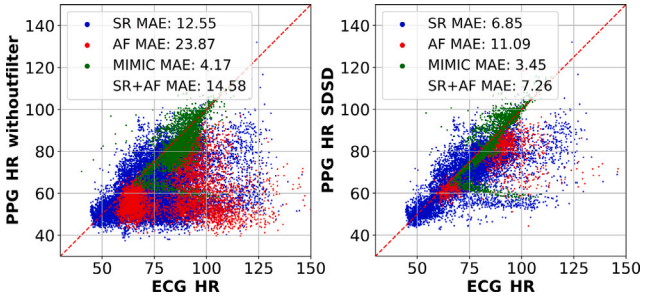

Fig. 6. HR comparison between ECG and PPG on Dataset I (for test): (left) without filtering; (right) after removing predicted bad label segments and applying the SDSD filter to eliminate unreasonable peaks.

5.2.2. Metrics for HR from PPG vs. ECG

We compared PPG-derived HR with ECG-derived HR under different conditions. The model was trained on the SR + AF dataset with a 60-s segment length using logistic regression (LR) under consistent experimental conditions. As summarized in Table 6, three key findings emerged from this analysis: (1) Comparison of HR filters: The SDSD filter consistently produced the lowest MAE across all datasets. Specifically, applying the SDSD filter reduced MAE from 14.58 bpm to 7.79 bpm (−46.57%) for SR + AF, from 12.55 bpm to 7.32 bpm (−41.67%) for SR, from 23.87 bpm to 12.01 bpm (−49.69%) for AF, and from 4.17 bpm to 3.98 bpm (−4.56%) for MIMIC-IV. (2) Effect of SQA-based label removal: We compared HR estimates obtained after removing low-quality signals identified by the ML-based SQA model versus those removed using manually labeled signal quality. The difference in MAE between predicted and manually labeled signal quality was only 0.08 bpm for SR + AF, 0.25 bpm for SR, 0.62 bpm for AF, and 0.53 bpm for MIMIC-IV. (3) Comparison across rhythm conditions: Independent analysis of SR, AF, and combined SR + AF datasets showed that MAE was lowest for SR, highest for AF, and intermediate for SR + AF. Using an autocorrelation-based method resulted in a small MAE reduction of 1.30% for SR + AF and 4.30% for SR compared to the original unfiltered data. However, this approach was more inaccurate in AF, leading to a 24.09% increase in MAE for AF.

Table 6
MAE of HR derived from ECG vs. PPG across different sub-datasets and filter types on the test set, reported with 95% confidence intervals (bpm).

HR filter type	Dataset	Filter type		
		No filter (95% CI)	SDSD (95% CI)	Autocorrelation (95% CI)
Original	SR+AF	14.58 (14.41–14.76)	7.79 (7.65–7.93)	14.39 (14.15–14.61)
	SR	12.55 (12.39–12.72)	7.32 (7.18–7.46)	12.01 (11.81–12.21)
	AF	23.87 (23.33–24.43)	12.01 (11.49–12.59)	29.62 (28.82–30.43)
	MIMIC	4.17 (3.99–4.36)	3.98 (3.25–4.67)	9.07 (8.66–9.45)
Predicted labels	SR+AF	9.47 (9.28–9.67)	7.26 (7.10–7.41)	10.28 (10.05–10.54)
	SR	8.26 (8.07–8.44)	6.85 (6.70–7.02)	8.38 (8.17–8.60)
	AF	17.46 (16.69–18.21)	11.09 (10.49–11.71)	27.16 (25.81–28.47)
	MIMIC	3.75 (3.44–3.81)	3.45 (3.23–3.67)	7.38 (7.06–7.74)
Manually labeled	SR+AF	9.55 (9.35–9.74)	7.14 (6.98–7.30)	10.08 (9.83–10.33)
	SR	8.01 (7.84–8.20)	6.71 (6.55–6.87)	7.90 (7.69–8.09)
	AF	18.08 (17.36–18.84)	11.04 (10.46–11.71)	26.76 (25.58–27.82)
	MIMIC	3.22 (3.01–3.44)	3.44 (3.12–3.59)	5.51 (5.18–5.88)

To assess the statistical significance of these improvements, we conducted three sets of paired t-tests: (1) unfiltered vs. SDSD-filtered HR, (2) unfiltered vs. HR after predicted-quality-based filtering, and (3) unfiltered vs. SDSD with predicted label filtering. Across all datasets, all differences in MAE were statistically significant ($p < 0.001$), confirming the robustness of the observed improvements.

5.3. Experiment III: Balancing coverage and MAE in HR predictions

This experiment assesses the impact of thresholds for SDSD and ML on the coverage and MAE of HR predictions, comparing ML model predictions and SDSD filtering across various health labels (SR, AF, MIMIC). Coverage is defined as the proportion of data remaining after filtering. Additionally, this experiment examines the effects of varying SDSD and ML thresholds (0.2, 0.5, 0.7) on the coverage and MAE of HR predictions. These ML thresholds derive from cutoff values in an LR model based on five selected features.

Fig. 7(a) demonstrates that the MAE across different health labels and processing techniques increases with coverage, particularly in AF data where MAE is significantly higher than in SR and MIMIC. ML predictions show a sharp increase in MAE values at low to medium coverage in AF data. In contrast, SDSD filtering demonstrates more stable MAE growth in AF data. Both techniques gradually increase MAE in different health labels as coverage rises, whereas the SDSD points showing a more uniform distribution across the coverage.

Fig. 7(b) illustrates that the SDSD filtering under different ML thresholds shows a general trend of increasing MAE with coverage. MAE values are consistently lower with the use of ML compared to without it. Moreover, lower ML thresholds are associated with reduced MAE. Within the 0%–50% coverage range, MAE remains stable across all ML thresholds. Beyond this range, MAE values differ by threshold: the red points, representing the lowest threshold - are below green, followed by blue, with all three being lower than the black stars, which indicate no ML usage.

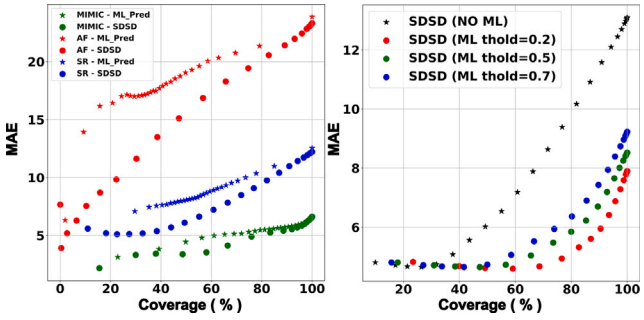

Fig. 7. (a) Coverage vs. MAE across health labels (SR and AF from DS I; MIMIC from DS II): Evaluating ML predictions and SDSD filtering ; (b) Comparative analysis: Coverage and MAE across combined health data with SDSD filtering and varying ML thresholds.

6. Discussion

We proposed a SQA method specifically designed for PPG data collected from real hospitalized patients, whose average age was approximately 70 years and who often presented with complex cardiac histories and prolonged AF episodes. Compared to previous studies, our dataset features more pronounced signal variability and pathological rhythms, presenting greater challenges for remote and long-term monitoring. To validate the practical value of the proposed SQA method, we further evaluated its impact on a downstream clinical task—HR estimation.

Based on the experimental results reported in Experiment I, shorter segment lengths consistently led to better classification performance. This suggests that shorter segments are less affected by rhythm changes and noise, especially in cross-dataset scenarios. Although longer segments provide more temporal context, they may introduce more variability, reducing model stability. In addition, the consistent trends across multiple classifiers support the reliability of the extracted features. However, performance degradation observed on the MIMIC dataset highlights the challenges posed by real clinical data, such as higher noise levels and distributional differences. Models trained on SR + AF data outperformed those trained on SR or AF alone, emphasizing the importance of rhythm diversity in training. Interestingly, MIMIC-trained models also achieved competitive results on SR-only data, further demonstrating the generalizability of the selected features.

For HR estimation, we followed FDA and ANSI/AAMI guidelines defining clinically acceptable accuracy as within ±10% or ±5 bpm [50]. As shown in Fig. 7, ML-based SQA combined with SDSD filtering maintained MAE below 5 bpm at quality thresholds of 0.2–0.5, retaining 35%–45% of the signal. In contrast, SDSD filtering alone required coverage below 25% to achieve the same accuracy, underscoring the value of ML-based SQA in improving both usability and reliability for clinical monitoring.

We also compared HR estimates derived from autocorrelation with those obtained using the AMPD peak detection algorithm. The consistency between the two approaches suggests both are viable for HR estimation. However, accurate HR extraction from AF segments remains particularly challenging. This underscores the need for effective filtering strategies to exclude noisy or irregular segments. Choosing an appropriate SDSD threshold is key: while a higher threshold increases peak detection, it can also degrade MAE. Experiment III demonstrated that SDSD filtering is effective at removing anomalous peaks and, when combined with ML-based predictions, further improves HR estimation accuracy. Finding the right balance between SDSD filtering and ML-based thresholds is essential for maximizing performance.

Finally, we observed that overall MAE results were affected by a small number of individual outliers with large estimation errors. These cases may be linked to factors such as suboptimal ECG patch placement. In some patients, the patches were positioned toward the left side of the chest due to interference with other sensors (as illustrated in Fig. 3), potentially contributing to discrepancies observed between Dataset I and the MIMIC dataset.

Limitations and Future Work

This study has several limitations. First, performance degradation in cross-dataset evaluations reflects domain shifts from device heterogeneity and acquisition environments, and generalization to home-based wearable data with greater motion artifacts and rhythm variability remains challenging. Second, although we used interpretable ML models, we did not explore deep learning architectures or accelerometer-derived motion features, which may improve quality assessment in noisy settings. Third, the SDSD filtering threshold was not rhythm-adapted, potentially limiting effectiveness across diverse rhythms; adaptive strategies warrant exploration. Fourth, inter-annotator agreement for manual SQA labels was not assessed, potentially affecting model performance. Finally, while HR estimation accuracy improved, the clinical implications of residual errors—particularly for rhythm classification or AF burden estimation—require further study.

7. Conclusion

In this study, we collected 24-h ECG and PPG data from 49 hospitalized cardiac patients, manually annotating all PPG segments for signal quality. Additionally, we included 14 recordings from MIMIC-IV to assess model generalization. We introduced an SQA method for PPG, evaluating it across seven ML models and four segmentation lengths, and explored the impact of signal quality on HR estimation under different cardiac rhythms. By integrating SQA with HR estimation, we demonstrated that filtering poor-quality signals improves HR accuracy, with the SDSD-based filter significantly reducing MAE across all datasets. Our findings highlight the necessity of incorporating SQA into HR detection pipelines to enhance the reliability of PPG-based monitoring, especially in clinical settings with arrhythmic patients.

CRediT authorship contribution statement

Yangyang Zhao: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Olli Lahdenoja:** Writing – review & editing, Supervision. **Jonas Sandelin:** Writing – review & editing, Formal analysis. **Sepehr Seifizarei:** Writing – review & editing. **Arman Anzanpour:** Visualization. **Joonas Lehto:** Data curation. **Joel Nuotio:** Data curation. **Jussi Jaakkola:** Data curation. **Arto Relander:** Data curation. **Tuija Vasankari:** Data curation. **Juhani Airaksinen:** Writing – review & editing, Data curation. **Tuomas Kiviniemi:** Data curation. **Matti Kaisti:** Writing – review & editing, Supervision. **Tero Koivisto:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was funded by the Moore4Medical project under grant agreements H2020-ECSEL-2019-IA-876190 and 7215/31/2019, and by the ITEA project RM4HEALTH under grant agreement 8139/31/2022. Additionally, the study was conducted as part of the CARE-DETECT Part I clinical trial (ClinicalTrials.gov ID: NCT05351775), initiated on April 12, 2022.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.bspc.2025.108688.

Data availability

Data will be made available on request.

References

[1] Abduljabbar S. Ba Mahel, Alexander N. Kalinichenko, Classification of arrhythmia using parallel channels and different features, in: 2024 Conference of Young Researchers in Electrical and Electronic Engineering (ElCon), IEEE, 2024, pp. 1007–1010.
[2] Abduljabbar S. Ba Mahel, Mehdhar S.A.M. Al-Gaashani, Reem Ibrahim Alkanhel, Dina S.M. Hassan, Mohammed Saleh Ali Muthanna, Ammar Muthanna, Ahmed Aziz, A multi-scale deep learning framework combining MobileViT-ECA and LSTM for accurate ECG analysis, IEEE Access 13 (2025) 85473–85492.
[3] Abduljabbar S Ba Mahel, Fahad Mushabbab G Alotaibi, Zenebe Markos Lonseko, Nini Rao, XABH-CNN-GRU: Explainable attention-based hybrid CNN-GRU model for accurate identification of common arrhythmias, J. Electron. Sci. Technol. (2025) 100322.
[4] Zhijun Xiao, Caiyun Ma, Yantao Xing, Chenxi Yang, Menglong Hao, Jianqing Li, Chengyu Liu, Atrial fibrillation monitoring based on noncontact capacitive ECG using an integrated microhumidity fabric electrode-sheet sensing scheme, IEEE Trans. Instrum. Meas. 72 (2023) 1–11.
[5] Amjed S Al Fahoum, Ansam Omar Abu Al-Haija, Hussam A Alshraideh, Identification of coronary artery diseases using photoplethysmography signals and practical feature selection process, Bioengineering 10 (2) (2023) 249.
[6] Amjed Al Fahoum, Ahmad Al Omari, Ghadeer Al Omari, Ala'a Zyout, Development of a novel light-sensitive PPG model using PPG scalograms and PPG-NET learning for non-invasive hypertension monitoring, Heliyon 10 (21) (2024).
[7] Amjed S. Al-Fahoum, Awad Al-Zaben, Waseem Seafan, A multiple signal classification approach for photoplethysmography signals in healthy and athletic subjects, Int. J. Biomed. Eng. Technol. 17 (1) (2015) 1–23.
[8] Nikhilesh Pradhan, Sreeraman Rajan, Andy Adler, Evaluation of the signal quality of wrist-based photoplethysmography, Physiol. Meas. 40 (6) (2019) 065008.
[9] Christina Orphanidou, Signal quality assessment in physiological monitoring: state of the art and practical considerations, 2017.
[10] Junyung Park, Hyeon Seok Seok, Sang-Su Kim, Hangsik Shin, Photoplethysmogram analysis and applications: An integrative review, Front. Physiol. 12 (2022) 808451.
[11] Tero Hurnanen, Eero Lehtonen, Mojtaba Jafari Tadi, Tom Kuusela, Tuomas Kiviniemi, Antti Saraste, Tuija Vasankari, Juhani Airaksinen, Tero Koivisto, Mikko Pänkäälä, Automated detection of atrial fibrillation based on time–frequency analysis of seismocardiograms, IEEE J. Biomed. Heal. Inform. 21 (5) (2016) 1233–1241.
[12] Tamaghno Chatterjee, Aayushman Ghosh, Sayan Sarkar, Signal quality assessment of photoplethysmogram signals using quantum pattern recognition technique and lightweight cnn module, in: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, IEEE, 2022, pp. 3382–3386.
[13] Mohammadamin Sinichi, Martin J. Gevonden, Lydia Krabbendam, Quality in question: Assessing the accuracy of four heart rate wearables and the implications for psychophysiological research, Psychophysiology 62 (2) (2025) e70004.
[14] Nandakumar Selvaraj, Yitzhak Mendelson, Kirk H Shelley, David G Silverman, Ki H Chon, Statistical approach for the detection of motion/noise artifacts in photoplethysmogram, in: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2011, pp. 4972–4975.
[15] Mohamed Elgendi, Optimal signal quality index for photoplethysmogram signals, Bioengineering 3 (4) (2016) 21.
[16] Simhadri Vadrevu, M. Sabarimalai Manikandan, Real-time PPG signal quality assessment system for improving battery life and false alarms, IEEE Trans. Circuits Syst. II: Express Briefs 66 (11) (2019) 1910–1914.
[17] Gangireddy Narendra Kumar Reddy, M Sabarimalai Manikandan, NVL Narasimha Murty, On-device integrated PPG quality assessment and sensor disconnection/saturation detection system for IoT health monitoring, IEEE Trans. Instrum. Meas. 69 (9) (2020) 6351–6361.
[18] Serena Moscato, Stella Lo Giudice, Giulia Massaro, Lorenzo Chiari, Wrist photoplethysmography signal quality assessment for reliable heart rate estimate and morphological analysis, Sensors 22 (15) (2022) 5831.
[19] Mohammad Feli, Iman Azimi, Arman Anzanpour, Amir M Rahmani, Pasi Liljeberg, An energy-efficient semi-supervised approach for on-device photoplethysmogram signal quality assessment, Smart Heal. 28 (2023) 100390.
[20] Qiao Li, Gari D. Clifford, Dynamic time warping and machine learning for signal quality assessment of pulsatile signals, Physiol. Meas. 33 (9) (2012) 1491.

[21] Tania Pereira, Kais Gadhoumi, Mitchell Ma, Xiuyun Liu, Ran Xiao, Rene A Colorado, Kevin J Keenan, Karl Meisel, Xiao Hu, A supervised approach to robust photoplethysmography quality assessment, IEEE J. Biomed. Heal. Inform. 24 (3) (2019) 649–657.
[22] Jianzhong Chen, Ke Sun, Yi Sun, Xinxin Li, Signal quality assessment of PPG signals using STFT time-frequency spectra and deep learning approaches, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, IEEE, 2021, pp. 1153–1156.
[23] Yalagala Sivanjaneyulu, M Sabarimalai Manikandan, Srinivas Boppu, Cnn based ppg signal quality assessment using raw ppg signal for energy-efficient ppg analysis devices in internet of medical things, in: 2022 International Conference on Artificial Intelligence of Things, ICAIoT, IEEE, 2022, pp. 1–6.
[24] Donggeun Roh, Hangsik Shin, Recurrence plot and machine learning for signal quality assessment of photoplethysmogram in mobile environment, Sensors 21 (6) (2021) 2188.
[25] Phattarapong Sawangjai, Narongrid Seesawad, Theerawit Wilaiprasitporn, Removal of motion artifacts from the PPG signal using attentive generative adversarial networks with dual discriminator, IEEE Trans. Instrum. Meas. 74 (2025) 1–10.
[26] Jian Liu, Shuaicong Hu, Ya'nan Wang, Qihan Hu, Daomiao Wang, Cuiwei Yang, A lightweight hybrid model using multiscale Markov transition field for real-time quality assessment of photoplethysmography signals, IEEE J. Biomed. Heal. Inform. 28 (2) (2023) 1078–1088.
[27] Emad Kasaeyan Naeini, Fatemeh Sarhaddi, Iman Azimi, Pasi Liljeberg, Nikil Dutt, Amir M Rahmani, A deep learning–based PPG quality assessment approach for heart rate and heart rate variability, ACM Trans. Comput. Heal. 4 (4) (2023) 1–22.
[28] Hangsik Shin, Deep convolutional neural network-based signal quality assessment for photoplethysmogram, Comput. Biol. Med. 145 (2022) 105430.
[29] C. Orphanidou, G. Howells, A. Iliodromitis, S. Jones, Signal quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring, IEEE J. Biomed. Heal. Inform. 19 (3) (2015) 832–838.
[30] Asad Khan, Arne Leijon, Mojtaba Etemad, Optimized signal quality assessment for photoplethysmogram signals using feature selection, IEEE Access 9 (2021) 130150–130163.
[31] J. Abdul Sukor, S.J. Redmond, N.H. Lovell, Signal quality measures for pulse oximetry through waveform morphology analysis, Physiol. Meas. 32 (3) (2011) 369.
[32] Chuanrong Xu, Shuying Liu, Huan Hu, Chen Shen, Minghui Hao, Yuxuan Liu, Zai Luo, Design of a real-time photoplethysmogram signal quality checker for wearables and edge computing, Sensors 21 (9) (2021) 3047.
[33] B. Moody, S. Hao, B. Gow, T. Pollard, W. Zong, R. Mark, MIMIC-IV waveform database (version 0.1.0), 2022, http://dx.doi.org/10.13026/a2mw-f949, PhysioNet.
[34] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, H Eugene Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220.
[35] Rajet Krishnan, Balasubramaniam Natarajan, Steve Warren, Two-stage approach for detection and reduction of motion artifacts in photoplethysmographic data, IEEE Trans. Biomed. Eng. 57 (8) (2010) 1867–1876.
[36] Andrew D. Ho, Carol C. Yu, Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects, Educ. Psychol. Meas. 75 (3) (2015) 365–388.
[37] David Ruppert, What is kurtosis? An influence function approach, Amer. Statist. 41 (1) (1987) 1–5.
[38] Claude Elwood Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[39] Chi-hau Chen, Signal Processing Handbook, vol. 51, CRC Press, 1988.
[40] Saeed V. Vaseghi, Advanced Digital Signal Processing and Noise Reduction, John Wiley & Sons, 2008.
[41] Abraham Lempel, Jacob Ziv, On the complexity of finite sequences, IEEE Trans. Inform. Theory 22 (1) (1976) 75–81.
[42] Yue Zhang, Junjun Pan, Assessment of photoplethysmogram signal quality based on frequency domain and time series parameters, in: 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI, IEEE, 2017, pp. 1–5.
[43] Peter Welch, The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms, IEEE Trans. Audio Electroacoust. 15 (2) (1967) 70–73.
[44] Mohamed Elgendi, Ian Norton, Matt Brearley, Derek Abbott, Dale Schuurmans, Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions, PloS One 8 (10) (2013) e76585.
[45] Xuxue Sun, Ping Yang, Yuan-Ting Zhang, Assessment of photoplethysmogram signal quality using morphology integrated with temporal information approach, in: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2012, pp. 3456–3459.
[46] Felix Scholkmann, Jens Boss, Martin Wolf, An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals, Algorithms 5 (4) (2012) 588–603.
[47] Jiapu Pan, Willis J. Tompkins, A real-time QRS detection algorithm, IEEE Trans. Biomed. Eng. (3) (1985) 230–236.
[48] Tapio Pahikkala, Antti Airola, RLScore: regularized least-squares learners, J. Mach. Learn. Res. 17 (1) (2016) 7803–7807.
[49] Christoph Molnar, Interpretable Machine Learning, third ed., 2025.
[50] U.S. Food and Drug Administration, Pulse oximeters – premarket notification submissions [510(k)] guidance for industry and FDA staff, 2013.