A Framework for Task-Specific Signal Quality Assessment: A Case Study in Heart Rate Estimation

Aron B Syversen[1], Zhiqiang Zhang[1], David Jayne[2], Alexios Dosis[2], David C Wong[1]

[1]University of Leeds, Leeds, UK
[2]Leeds Teaching Hospitals NHS Trust, Leeds, UK

Abstract

Signal Quality Indices (SQIs) are essential for identifying usable ECG segments in long-term recordings. Most SQIs are designed for general-purpose use and do not account for the specific demands of feature extraction or downstream analyses. In this work, we introduce a task-specific SQI tailored to heart rate (HR) estimation. We develop a labelling strategy that uses a beat detector to classify 10-second ECG segments as Clean or Noisy based on whether the derived HR is within 10% of ground truth.

Using a combination of synthetic, semi-synthetic, and real-world ECG data, we trained and fine-tuned a 1D ResNet to classify segments accordingly. The model achieved F1 scores of 0.92 and 0.85 on internal test sets (PhysioNet 2014 and MIT-BIH Noise Stress Test), and generalised well to an external test set (TELE ECG), with an F1 score of 0.80. This framework presents an adaptable method for building SQIs that are aligned to specific clinical or analytical tasks, offering a more reproducible and targeted alternative to existing approaches.

1. Introduction

ECG recording is a fundamental investigation in cardiology. For many cardiovascular diseases, ECG measurement is considered the 'gold standard', offering critical information for both diagnosis and monitoring of heart conditions [1]. Specifically, this information includes anything from simple summary measures such as heart rate (HR), to information on heart function based on subtler morphological changes in the ECG.

Advances in wearable sensor technologies have enabled continuous recording of ECG in non-clinical environments. However, ECG recorded outside clinical settings—via chest or handheld devices—are susceptible to noise caused by motion artifacts, poor electrode contact, and environmental interference [2]. In noisy recordings, signal quality assessment is essential to avoid extracting false information. Manual inspection of signal quality is an impractical task for long-term recordings, prompting the need for automated Signal Quality Indices (SQIs).

Various SQI methods for ECGs have been proposed, from rule-based thresholds to machine learning classifiers [3]. These generally discriminate high- and low-quality signals classifying them either as acceptable or unacceptable, but are often general-purpose and not tailored to specific feature extraction or analysis. The quality required for signal analysis is task-dependent [4]. For example, detecting atrial fibrillation often requires investigation of smaller wave morphologies in the signal (e.g. p-wave), while HR estimation primarily relies on QRS detectability [5]. By generalising quality assessment across all types of ECG feature extraction, there is a risk of excluding clinically useful information or including misleading data, which in turn undermine clinicians trust in wearable monitoring.

To address these differences, we propose an SQI that is tailored to signal processing and analysis goals, rather than a 'one-size fits all approach'. In this paper, we select estimation of HR as the task, but the approach is generalisable to any task or processing pipeline in which appropriate training labels are available.

2. Methods

Our task-specific SQI follows a two stage approach. First, we label a large collection of training ECG segments as *'Clean'* or *'Noisy'* based on whether a specific beat detector can estimate HR within a defined tolerance of ground-truth (GT) HR. To deal with the limitation of accurate labels for noisy data, we combine real-world ECG data with synthetically generated data in which the GT beat annotation is known. Second, we use this relabelled data as input to a deep learning classifier that discriminates between *Noisy* and *Clean* ECGs, as shown in Figure 1.

The rest of this section outlines the datasets used for model development and testing, the process for generating labels and the pipeline for model training and evaluation. 10 seconds was selected as the segment length as input, as it is a common interval length used for SQIs [4].

2.1. Datasets

We use a combination of synthetic, semi-synthetic and real-world ECG datasets, using only single leads from each. To be included, datasets required manually labelled accurate beat annotations and some variation of noise:

- **Synthetic Data:** Generated using an open-source simulator with added noise (e.g., HR variability, white noise, power-line interference, motion artifacts) [6], providing diverse signals with known beat locations.
- **MIT-BIH Noise Stress Test (NST):** Semi-synthetic ECGs created by adding calibrated baseline wander, muscle, and electrode motion artefacts to clean signals, alternating with clean segments [7]. Additional augmentations increased variability.
- **PhysioNet 2014 Challenge Dataset:** 100 half-hour ECGs used for fine-tuning and evaluation, split into 10-second windows [8]. Augmentations were applied to improve generalisability.
- **TELE ECG (external validation):** 250 telehealth ECGs with dry electrodes; only annotated segments were used. Served solely for external validation [9].

2.2. Labelling process

We used the Neurokit2 detector for automatic beat detection, which is widely adopted in research [10]. The following process was implemented, as outlined in Figure 2:

1. Apply a 0.5–150 Hz bandpass filter to remove baseline wander and high-frequency noise, consistent with ECG preprocessing standards.
2. Detect beats using the NeuroKit detector.
3. Detected beats are matched against GT beat locations. If more than 50% of detected beats fall within 50 milliseconds of GT beats, the segment proceeds to the next step; otherwise, it is labelled as *Noisy*. This step prevents mislabelling segments as *Acceptable* when HR values coincided by chance, despite poor underlying beat detection.
4. HR is calculated from both the GT and detected beats. If the beat-derived HR is within 10% of the true HR, the segment is labelled *Clean*; otherwise, it is labelled as *Noisy*. 10% was selected based on previously reported standard thresholds for HR monitor accuracy [11].

2.3. Model Development

We implemented a 1D ResNet convolutional neural network (CNN)[1] to classify 10-second ECG segments as suitable or unsuitable for HR estimation, by labelling them as *Noisy* or *Clean*. The architecture consists of an initial convolutional layer with downsampling and batch normalisation, two residual blocks with increasing feature maps (32 and 64 filters respectively), each with skip connections and a final sigmoid output for binary classification.

Before input to the model, all signals were resampled to 500Hz and normalised using min-max scaling. The network was pre-trained for 80 epochs using the synthetic data before fine-tuning with the real-world data (MIT-BIH NST and the PhysioNet 2014 Augmented Challenge set) (Figure 1). The fine tuning set was split 60/20/20 for training, validation (threshold optimization), and testing, respectively. During fine-tuning, early layers were frozen for the initial epochs and later unfrozen. The model was trained using binary cross-entropy loss. Adam loss optimiser was used and a learning rate scheduler based on validation loss was used. The final decision threshold was tuned to optimise the F1 score on the validation set.

[1] Code available at: https://github.com/Syveraron/Task-specific-SQI.git

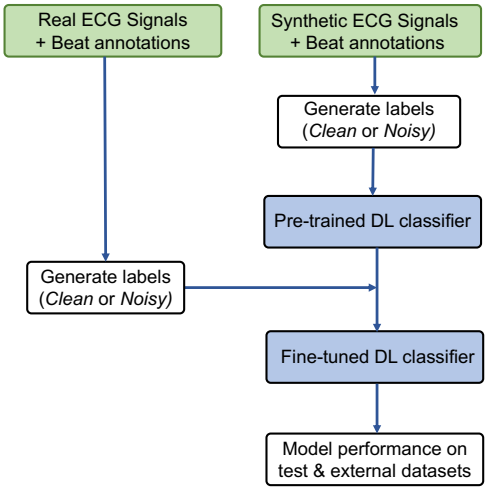

Figure 1. The framework for building a task-specific SQI for extracting accurate HR. Data input components are shown in green whilst model development is shown in blue.

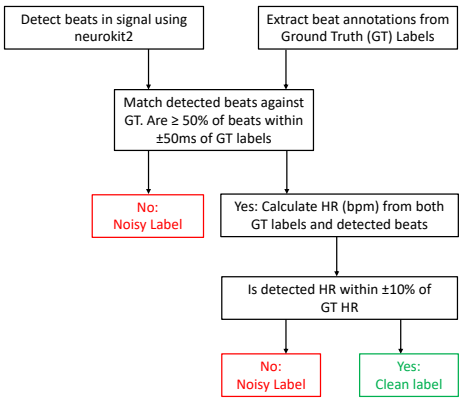

Figure 2. The process for assigning *Clean* or *Noisy* labels to signals, as described in 2.2.

3. Results

3.1. Dataset Overview

In Table 1 we report the number of 10-second ECG segments included from each dataset, along with their sampling frequencies and the proportion of segments labelled as *'Clean'*. Two large randomly generated synthetic datasets were used for pre-training with different sampling frequencies to match that of the real-world datasets. The Physionet 2014 and MIT-BIH NST were used for fine-tuning and evaluation while the TELE ECG dataset was held out entirely for external validation.

Table 1. Dataset composition, sampling frequency, and proportion of *Clean* signals

Dataset	Hz	Number of Signals	% Clean
Synthetic	500	40,000	37.4%
Synthetic 2	360	32,560	41.8%
MIT-BIH NST	360	804	39.2%
PhysioNet 2014	360	4,452	64.7%
TELE ECG	500	416	43.8%

3.2. Performance on Internal Test Sets

After training and fine-tuning, the model was evaluated on held-out partitions from the real-world and semi-synthetic datasets. A decision threshold of 0.38 was selected based on optimisation of the F1 score.

Across all internal test data, the model achieved an overall accuracy of 0.90 and an F1 score of 0.90 (Table 2). Performance was balanced across precision (0.90) and recall (0.90), suggesting the model could reliably distinguish between *Clean* and *Noisy* segments. When stratified by dataset, F1 scores were 0.91 for PhysioNet 2014 and 0.85 for MIT-BIH NST, indicating slightly stronger performance on the semi-controlled PhysioNet data but still robust performance in noisier conditions.

Table 2. Model performance on internal test sets.

Dataset	Acc.	Prec.	Rec.	F1
PhysioNet 2014	0.912	0.915	0.912	0.913
MIT-BIH NST	0.847	0.846	0.847	0.846

3.3. Performance on External Dataset

The model generalised well to the unseen TELE dataset, achieving an accuracy of 0.81, F1 score of 0.80, and AUROC of 0.86. Precision (0.81) was lower than recall (0.87), suggesting that the model is conservative in identifying *Clean* signals: it correctly detects most noisy signals but is more cautious when labelling segments as usable.

3.4. Incorrect Classification

To understand model behaviour, we examined incorrectly classified signals. Figure 3 shows two examples. Signals incorrectly labelled as *Noisy* often had small R-peaks and baseline drift, despite otherwise clean morphology. In contrast, some noisy signals incorrectly labelled as *Clean* had large R-peaks, which may have dominated the model's decision despite high-frequency noise that the Neurokit detector struggles with. This suggests the model relies heavily on R-peak prominence, potentially overlooking components such as higher-frequency noise.

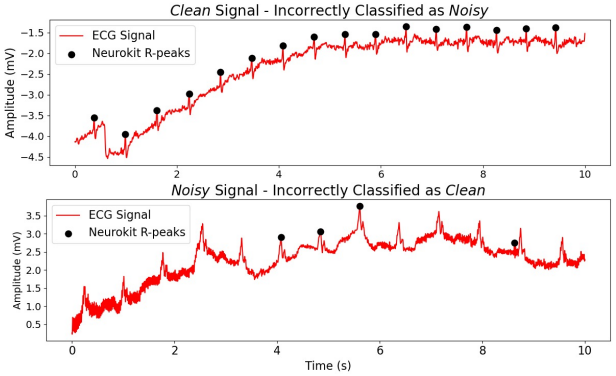

Figure 3. Example of incorrectly classified signals from the external dataset. Top: a *'Clean'* signal, with correctly detected R-peaks, incorrectly classified as *'Noisy'*. Bottom: a *'Noisy'* signal, with poor R-peak detection, misclassified as *'Clean'*.

4. Discussion

This project presents a task-specific framework to ECG signal quality assessment, aimed specifically at heart rate estimation using a defined beat detector. By aligning signal quality assessment with analysis goals, we move away from general-purpose SQIs towards a targeted, reproducible approach.

Despite using only a modest amount of real-world data for fine-tuning, our model achieved good performance on both internal and external test sets. Results show that pre-training a deep learning model on synthetic ECG signals simulating realistic noise provides a robust foundation for learning generalisable features. Given the difficulty of obtaining beat locations in noisy real-world data, synthetic signals were especially useful for HR estimation. Performance on edge cases suggests synthetic noise may not fully capture real signal variability, and augmenting real-world data may be key to bridging this gap. Encouragingly, the model generalised well to an unseen external dataset collected using different sensors (dry electrodes), indicating robustness to different sensor modalities.

A key benefit of this approach is its potential for real-world application in accurate ECG processing. In long-term ECG monitoring, manual inspection of signal quality is impractical and general purpose SQIs may not reliably segment signals for HR estimation. This SQI offers an automated way to flag periods unsuitable for HR estimation that could be deployed within clinical processing pipelines. In doing so, it may increase clinicians' trust in HR derived from wearable ECG's and support the safe integration of this data into clinical workflows.

More broadly, this work presents a generalisable pipeline for task-specific SQI development, linking signal quality assessment directly to the goals of the analysis. Previous research has developed processing pipelines tailored to HR estimation from ECGs, fusing multiple SQIs and HR from multiple ECG leads [12]. While also presenting a generalisable framework, our approach differs in its task-specific design, providing a modular framework specific to the signal processing pipeline itself. For example, it can be used with different beat detectors, stricter thresholds (e.g., $<$5% HR deviation), or extended to other physiological measurements such as respiratory rate or heart rate variability, and applied across sensor modalities.

A limitation of our work is that this considers a very simple pipeline where relatively simple preprocessing is employed. Alternative pipelines that include additional preprocessing can be used in our labelling process, and should be tested in future to demonstrate utility in more realistic scenarios.

To conclude, this work highlights the importance of task- and pipeline-specific signal quality labelling and offers a reproducible approach for future SQI development. Future work will implement a robust pre-processing strategy before benchmarking this SQI against existing general-purpose SQIs to determine comparative performance in labelling signals for accurate HR extraction.

Acknowledgments

This work was supported by UK Research and Innovation (UKRI) [CDT grant number EP/S024336/1]. We also acknowledge the contributors and patients whose data were made publicly available via PhysioNet.

References

[1] Stracina T, Ronzhina M, Redina R, Novakova M. Golden standard or obsolete method? review of ecg applications in clinical and experimental context. Frontiers in Physiology 4 2022;13:867033. ISSN 1664042X. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9082936/.

[2] Syversen A, Dosis A, Jayne D, Zhang Z. Wearable sensors as a preoperative assessment tool: A review. Sensors 1 2024;24:482. ISSN 14248220.

[3] Satija U, Ramkumar B, Manikandan MS. A review of signal processing techniques for electrocardiogram signal quality assessment. IEEE reviews in biomedical engineering 2 2018;11:36–52. ISSN 1941-1189. URL https://pubmed.ncbi.nlm.nih.gov/29994590/.

[4] Syversen AB, Zhang Z, Batty JA, Kaisti M, Jayne D, Wong DC. Assessment of ecg signal quality index algorithms using synthetic ecg data. Computing in Cardiology 2024;51.

[5] Nielsen JB, Kühl JT, Pietersen A, Graff C, Lind B, Struijk JJ, Olesen MS, Sinner MF, Bachmann TN, Haunsø S, Nordestgaard BG, Ellinor PT, Svendsen JH, Kofoed KF, Køber L, Holst AG. P-wave duration and the risk of atrial fibrillation: Results from the copenhagen ecg study. Heart Rhythm 9 2015;12:1887–1895. ISSN 1547-5271.

[6] Karhinoja K, Vasankari A, Sirkiä JP, Airola A, Wong D, Kaisti M. Flexible framework for generating synthetic electrocardiograms and photoplethysmograms 8 2024;URL https://arxiv.org/abs/2408.16291v1.

[7] Moody G, Muldrow W, Mark R. A noise stress test for arrhythmia detectors. Computers in Cardiology 1984; 11:381–384.

[8] Moody G, Moody B, Silva I. Robust detection of heart beats in multimodal data: The physionet/computing in cardiology challenge 2014 v1.0.0, 2014. URL https://physionet.org/content/challenge-2014/1.0.0/.

[9] Khamis H, Weiss R, Xie Y, Chang CW, Lovell NH, Redmond SJ. Tele ecg database: 250 telehealth ecg records (collected using dry metal electrodes) with annotated qrs and artifact masks, and matlab code for the unsw artifact detection and unsw qrs detection algorithms 2016;.

[10] Kristof F, Kapsecker M, Nissen L, Brimicombe J, Cowie MR, Ding Z, Dymond A, Jonas SM, Lindén HC, Lip GY, Williams K, Mant J, Charlton PH. Qrs detection in single-lead, telehealth electrocardiogram signals: Benchmarking open-source algorithms. PLOS Digital Health 8 2024;3:e0000538. ISSN 2767-3170. URL https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000538.

[11] Bae S, Borac S, Emre Y, Wang J, Wu J, Kashyap M, Kang SH, Chen L, Moran M, Cannon J, Teasley ES, Chai A, Liu Y, Wadhwa N, Krainin M, Rubinstein M, Maciel A, McConnell MV, Patel S, Corrado GS, Taylor JA, Zhan J, Po MJ. Prospective validation of smartphone-based heart rate and respiratory rate measurement algorithms. Communications Medicine 2022 21 4 2022;2:1–10. ISSN 2730-664X. URL https://www.nature.com/articles/s43856-022-00102-x.

[12] Li Q, Mark RG, Clifford GD. Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a kalman filter. Physiological Measurement 12 2007;29:15. ISSN 0967-3334.

Address for correspondence:

Aron B. Syversen
Sir William Henry Bragg Building, Woodhouse, Leeds LS2 9JT
scabs@leeds.ac.uk