Deep Learning Signal Quality Assessment for Continuous Wearable ECG Monitoring

Estela Ribeiro[1], Douglas A Almeida[1], Quenaz B Soares[1], Jaqueline J Pereira[1], Renata G S Verardino[1], Tallita C Reis[1], Nelson Samesima[1], Rosangela Monteiro[1], Fabio B Jatene[1] Marco A Gutierrez[1]

[1]Heart Institute (InCor) – Clinics Hospital University of Sao Paulo Medical School (HCFMUSP) – Sao Paulo – SP – Brazil

Abstract

Reliable electrocardiogram (ECG) signal quality estimation is crucial for intelligent healthcare systems and wearable monitoring. Ensuring signal fidelity prevents diagnostic errors and reduces false alarms or missed detections caused by noise and motion artifacts common in ambulatory recordings. We present a deep learning-based approach for automated ECG signal quality estimation using a fine-tuned convolutional neural network (CNN). The model is based on the LiteVGG-11 architecture, originally designed for atrial fibrillation detection, and adjusted through transfer learning to classify signal quality, retaining cardiac feature representations while learning to discriminate between clean and noisy segments.The model was trained and evaluated on a real-world clinical dataset from a wearable ECG patch, consisting of continuous recordings from two cohorts (56 participants over 24 hours and 117 participants over 7 days). Clinical experts provided ground truth annotations.The proposed model achieved a sensitivity of 0.8964 $\pm$ 0.0356, specificity of 0.9783 $\pm$ 0.0028, F1-score of 0.9112 $\pm$ 0.0263, AUC of 0.9374 $\pm$ 0.0174, and accuracy of 0.9597 $\pm$ 0.0041. These results demonstrate the effectiveness of deep learning for robust ECG signal quality estimation in wearable settings, enhancing the reliability of long-term cardiac monitoring and facilitating more accurate clinical interpretation.

1. Introduction

Electrocardiograms (ECG) have emerged as a widely used tool in the diagnosis and prevention of cardiovascular diseases [1], offering a straightforward and non-invasive method. However, ECG recordings are frequently susceptible to noise that can be related to electrode movement and muscle artifacts, among other factors, presenting a challenge to accurate analysis [2, 3].

The advancements in ECG devices have facilitated the development of prolonged ECG recording beyond traditional hospital environments. In this scenario, it is necessary to ensure that collected ECG data undergoes thorough examination to remove any low-quality instances before being forwarded for analysis, ensuring the reliability of the information contained in the ECG signal and preventing misinterpretation due to noise [2]. Therefore, ECG signal quality analysis plays an important role in the diagnosis and monitoring of cardiovascular conditions.

The PhysioNet/Computing in Cardiology Challenge 2011 (CinC2011) dataset [4], commonly referred to as the ECG quality dataset, has been a valuable resource for this field of ECG signal quality, and most of the works usually use this dataset. Clifford et al. [5] proposed feature-based SVM and MLP models to detect poor-quality ECGs using 12-lead ECGs or single-lead ECGs, achieving an accuracy (Acc) of 0.97 for the SVM model. Athif et al. [6] used a decision tree algorithm on features extracted from 12-lead ECG signals, achieving an Acc of 0.91. Zhang et al. [7] proposed an LSTM model on raw ECG signals, achieving an Acc of 0.93. Hermawan et al. [8] proposed a method based on wavelet decomposition and heuristic rules, achieving an Acc of 0.85. Liu et al. [9] proposed a CNN based on spectrograms of the ECG signals, achieving an Acc of 0.93. Finally, Kuetch et al. [3] compared 39 individual Signal Quality Indexes (SQIs)for classifying signals as good or bad, achieving AUCs $\geq$ 0.90.

Classical machine learning methods typically rely on a set of engineered features derived from well-established SQIs that can encapsulate various aspects of the ECG waveform [5, 10]. In contrast, convolutional neural networks (CNNs) have demonstrated remarkable success in diverse applications by automatically learning hierarchical representations from raw data. In the context of ECG signal quality, CNNs have the potential to capture intricate patterns and dependencies within the signal without the need for explicit feature engineering.

This paper proposes a CNN model for ECG signal quality estimation using a finetuning approach based on a model previously developed for atrial fibrillation (AFib)

detection in single-lead ECGs [11]. We evaluate our model on a real-world clinical dataset using a 5-fold cross-validation strategy and compare its performance with a dataset-specific training approach.

2. Material and Methods

2.1. Dataset

We conducted a clinical experiment to acquire continuous wearable ECG signals in a real-world clinical setting. Data were collected using a Medical Device (MD) sensor, a certified class IIa device for ECG acquisition compliant with the EU Medical Device Regulation 2017/745 (Certificate CR-03-1229-813-23 02 / FI-MF-000024281), configured with a 128 Hz sampling rate. The sensor was placed in a modified lead II configuration, with the reference electrode on the right shoulder and the positive electrode on the left abdomen. The experiment was divided into two independent phases:
1. Phase 1: Data were recorded from 56 subjects over a continuous 24-hour period [11].
2. Phase 2: Data were recorded from 117 subjects over a continuous 7-day period.

Following data acquisition, signals were segmented into non-overlapping 10-second windows. A subset was randomly selected for annotation consisting of 60 minutes of data for Phase 1, and 210 minutes of data for Phase 2.

An experienced cardiologist annotated all selected windows from this subset. In Phase 1, segments were classified into four categories: (i) Atrial Fibrillation, (ii) Normal, (iii) Other abnormalities, and (iv) Artifacts. In Phase 2, segments were categorized into three groups: (i) Atrial Fibrillation, (ii) Non-Atrial Fibrillation, and (iii) Artifacts. In total, Phase 1 resulted in 20,160 annotated 10-second segments, including 2,391 artifacts, while Phase 2 produced 145,008 annotated segments, including 37,271 artifacts. Table 1 summarizes these details.

2.2. Preprocessing

We applied the same preprocessing steps as in [11, 12]. To mitigate baseline drift and low-frequency noise, we used a second-order Butterworth high-pass filter with a 1 Hz cutoff frequency. This was followed by a second-order Butterworth low-pass filter with a 40 Hz cutoff to focus on the diagnostically relevant ECG frequency range. Each signal was then normalized to have zero mean and unit variance.

2.3. Deep Learning Model

We propose a finetuning approach to predict ECG signal quality. The overall framework is illustrated in Figure 1. For this, we used LiteVGG-11, a CNN originally designed to classify AFib from dII-lead ECG signals [11, 12]. LiteVGG-11 is a lightweight variant of VGGNet [13], optimized for efficient resource usage while maintaining classification performance. The original model was trained on the PhysioNet/CinC Challenge 2021 dataset (CinC2021), which contains 88,252 annotated 12-lead ECG recordings [14]. However, training was conducted using only the dII lead instead of all 12 leads.

For our finetuning approach, all layers were unfrozen, and the model was trained using the Adam optimizer with a learning rate of $5e^{-5}$. Training was performed for up to 200 epochs with early stopping (patience = 50), using binary cross-entropy loss. To address class imbalance, we applied a weighted loss strategy based on inverse class frequencies. The model was trained with a batch size of 128 and a validation split of 20%. Additionally, the learning rate was reduced by a factor of 0.2 if the validation loss plateaued for 30 epochs. All experiments were implemented in Python (version 3.11.9) using the Keras API (version 2.15.0) with a TensorFlow backend (version 2.15.0).

2.4. Evaluation

We compared three scenarios to evaluate our models:
1. (I) training only with Phase 1 data and testing with Phase 2 data;
2. (II) the opposite, training only with Phase 2 data and testing with Phase 1 data;
3. (III) a 5-fold cross-validation approach, training with all data from both phases, ensuring that data from the same subject didn't appear in both the train and the test set to avoid data leakage.

In terms of performance metrics, we considered accuracy (Acc), sensitivity (Se), specificity (Spe), F1-score (F1), and AUC-ROC (Auc) to evaluate our model.

3. Results

The performance metrics of our model across the three evaluation scenarios are summarized in Table 2.

Our model demonstrated good performance in all evaluation settings, particularly in the 5-fold cross-validation scenario. In Scenario I (training on Phase 1, testing on Phase 2), the model achieved 0.8719 sensitivity and 0.9671 specificity, with an overall accuracy of 0.9726. In Scenario II (training on Phase 2, testing on Phase 1), sensitivity dropped to 0.7696, while specificity remained high at 0.9765, resulting in an accuracy of 0.9519. When trained and evaluated on the combined dataset (Scenario III), the model achieved the best balance between sensitivity (0.8964 $\pm$ 0.0356) and specificity (0.9783 $\pm$ 0.0028), yielding an AUC of 0.9374 $\pm$

Table 1. Statistical Overview of our Clinical Experiment.

	# Subjects	Age	# Non-Artifacts	# Artifacts	# Total
Phase 1	56 (39 male)	46.8 ± 14.8	17,769	2,391	20,160
Phase 2	117 (52 male)	50.7 ± 12.6	107,737	37,271	145,008
Total			**125,506**	**39,662**	**165,168**

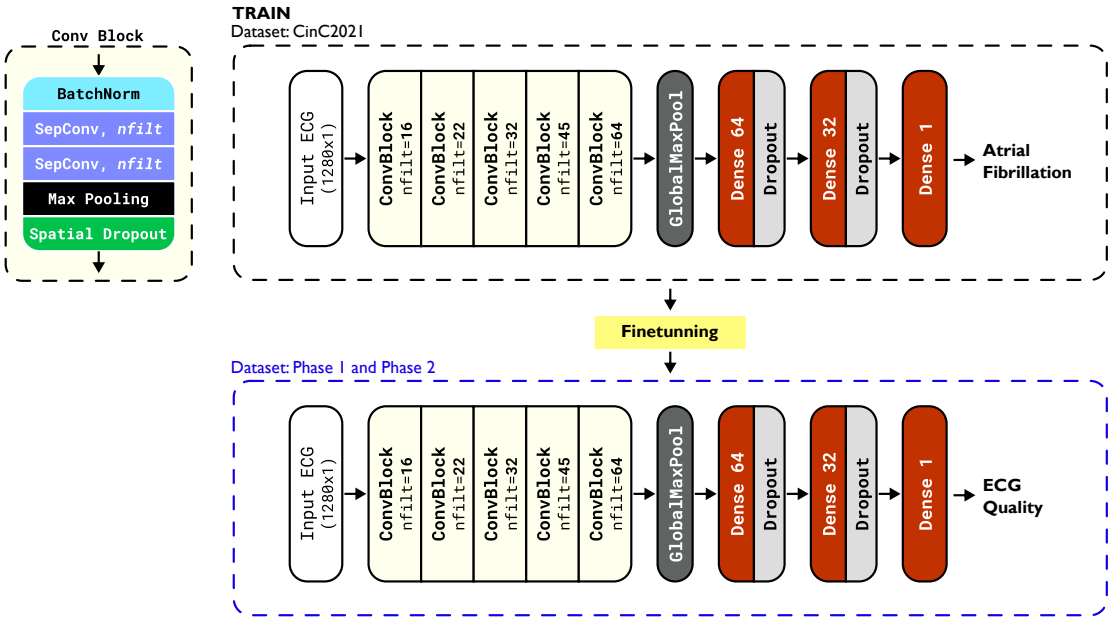

Figure 1. Overview of the proposed method for ECG signal quality estimation.

Table 2. Summary of Model's Performance for ECG quality signal Prediction.

Scenario	Sensibility	Specificity	F1-score	AUC	Accuracy
(I)	0.8719	0.9671	0.8865	0.9195	0.9426
(II)	0.7696	0.9765	0.7916	0.8730	0.9519
(III)	0.8964 ± 0.0356	0.9783 ± 0.0028	0.9112 ± 0.0263	0.9374 ± 0.0174	0.9597 ± 0.0041

0.0174 and an accuracy of 0.9597 ± 0.0041. These results demonstrate the effectiveness of our approach in developing a reliable model for continuous ECG signal quality monitoring in wearable devices.

4. Discussion

In this work, we developed a model for ECG signal quality estimation. Our model achieved good performance across all evaluation scenarios, with high sensitivity, specificity, and overall accuracy. The best performance was observed in the 5-fold cross-validation scenario (Scenario III), where the model leveraged the full dataset.

Our model was trained and evaluated on a dataset collected from a real-world clinical environment using wearable ECG sensors. This setting ensures that the model is well-suited for deployment in practical scenarios, particularly for continuous, long-term monitoring.

Most existing studies on ECG signal quality assessment rely on the CinC2011 dataset, which consists of 12-lead ECG recordings with global quality labels. However, our dataset was collected in a real-world wearable ECG setting, with annotations at the single-lead level. This fundamental difference prevents direct comparisons with prior state-of-the-art methods trained on CinC2011. Nevertheless, deep learning models such as LSTM [7] and CNN-based spectrogram analysis [9] reported accuracies of 0.93, while SQI-based methods achieved AUCs above 0.90 [3]. Our model, despite being trained on a different dataset, achieved similar or superior performance, with an AUC of 0.94 and accuracy up to 0.96 in cross-validation. These results indicate that our approach is competitive

with existing state-of-the-art models, even though it was trained on a dataset specifically collected from wearable ECG devices.

Despite its promising results, some limitations must be acknowledged. Our model was trained exclusively on dII lead signals, which may limit its generalizability to other ECG lead configurations. Future work should explore its applicability to multi-lead ECG systems. Moreover, while our dataset is clinically relevant, expanding it to larger and more diverse populations could further improve generalizability. Additionally, while our model performs well, incorporating hybrid approaches that combine deep learning with established SQI features could enhance interpretability and robustness.

5. Conclusion

This study demonstrates that a finetuned LiteVGG-11 model can effectively estimate ECG signal quality in a real-world clinical setting. Future work should focus on enhancing generalizability through dataset expansion and benchmarking against existing signal quality assessment frameworks.

Acknowledgements

This research received support from Lenovo, under Brazilian Informatics Law, and Zerbini Foundation as part of the research project "Telemonitoramento Remoto Assistido de Arritmias" (TRAdA).

References

[1] Thaler MS. The Only EKG Book You'll Ever Need. Philadelphia :Wolters Kluwer Health/Lippincott Williams and Wilkins, 2019.

[2] van der Bijl K, Elgendi M, Menon C. Automatic ecg quality assessment techniques: A systematic review. Diagnostics 2022;12(11).

[3] Kuetche F, Alexendre N, Pascal NE, Colince W, Thierry S. Signal quality indices evaluation for robust ecg signal quality assessment systems. Biomedical Physics Engineering Express aug 2023;9(5):055016.

[4] Kuetche F, Alexendre N, Pascal NE, Colince W, Thierry S. Sphysiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. Circulation Online 2000;101(23):e215–e220.

[5] Clifford GD, Behar J, Li Q, Rezek I. Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms. Physiological Measurement aug 2012; 33(9):1419.

[6] Athif M, Daluwatte C. Combination of rule based classification and decision trees to identify low quality ecg. In 2017 IEEE International Conference on Industrial and Information Systems (ICIIS). 2017; 1–4.

[7] Zhang J, Wang L, Zhang W, Yao J. A signal quality assessment method for electrocardiography acquired by mobile device. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2018; 1–3.

[8] Hermawan I, Ma'sum MA, Riskyana Dewi Intan P, Jatmiko W, Wiweko B, Boediman A, et al. Temporal feature and heuristics-based noise detection over classical machine learning for ecg signal quality assessment. In 2019 International Workshop on Big Data and Information Security (IWBIS). 2019; 1–8.

[9] LIU G, HAN X, TIAN L, ZHOU W, LIU H. Ecg quality assessment based on hand-crafted statistics and deep-learned s-transform spectrogram features. Computer Methods and Programs in Biomedicine 2021;208:106269. ISSN 0169-2607.

[10] Behar J, Oster J, Li Q, Clifford GD. Ecg signal quality during arrhythmia and its application to false alarm reduction. IEEE Transactions on Biomedical Engineering 2013;60(6):1660–1666.

[11] Soares QB, Andrade DA, Ribeiro E, Verardino RGS, Reis TC, Samesima N, et al. Clinical assessment of a lightweight cnn model for real-time atrial fibrillation prediction in continuous wearable monitoring. In 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2024; 1–4.

[12] Soares QB, Monteiro R, Jatene FB, Gutierrez MA. A lightweight unidimensional deep learning model for atrial fibrillation detection. In 2022 Computing in Cardiology (CinC), volume 498. IEEE, 2022; 1–4.

[13] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv14091556 2014;.

[14] Perez Alday EA, Gu A, J Shah A, Robichaux C, Ian Wong AK, Liu C, et al. Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. Physiological Measurement December 2020; 41(12):124003. ISSN 1361-6579.

Address for correspondence:

Estela Ribeiro

Av. Dr. Enéas Carvalho de Aguiar, 44, Sao Paulo, Brazil

estela.ribeiro@hc.fm.usp.br