Single-Lead Abnormal ECG Screening with a Lightweight Deep Learning Model for Wearable Health Monitoring

Estela Ribeiro[1], Quenaz Bezerra Soares[1], Douglas de Andrade de Almeida[1], Fabio Jatene[1], Marco Antonio Gutierrez[1]

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

{estela.ribeiro, quenaz.soares, douglas.andrade}@hc.fm.usp.br

{fabio.jatene, marco.gutierrez}@incor.usp.br

Abstract. *Electrocardiograms (ECGs) are essential for cardiac assessment, and the rising adoption of wearables has intensified the need for automated identification of abnormal ECGs on resource-constrained platforms. We evaluate LiteVGG-11, a lightweight CNN (20,365 parameters), for binary classification using Lead II signals. The model was trained on CODE15 and validated on PTB-XL and real-world wearable data, achieving AUCs of 0.736 and 0.773, respectively. Wearable validation revealed a decline in specificity due to artifacts and Lead II's spatial constraints. Findings indicate that LiteVGG-11 is feasible for preliminary screening in continuous health monitoring, despite limitations in identifying non-specific abnormalities that require multi-lead observation.*

1. Introduction

Electrocardiogram (ECG) remains one of the most widely used non-invasive tools for assessing cardiac function and diagnosing cardiovascular conditions. Traditionally, ECG interpretation relies on the analysis of 12-lead signals to detect abnormalities such as arrhythmia, myocardial infarction, and conduction disorders [Goldberger et al. 2018]. With the increasing availability of large ECG datasets and advances in deep learning, there has been a growing interest in developing automated systems capable of classifying ECG signals with high accuracy and scalability.

Despite these advancements, relatively few studies focus specifically on the binary classification of ECGs into normal and abnormal categories. Most existing approaches adopt multiclass or multilabel classification frameworks, often targeting a wide range of cardiac pathologies. These methodological differences, particularly in classification strategy, dataset composition, and input modality, limit the comparability of results across studies. For example, Bahuguna et al. (2023) proposed a method that converts ECG signals into scalograms and employs a VGG16-based autoencoder to detect abnormalities in simulated ECG signals, achieving high sensitivity (99.49%) and specificity (99.40%). Zhu et al. (2022) used a convolutional neural network to classify abnormalities using 12-lead ECGs from the PTB-XL dataset, reporting an F1-score of 0.902. Similarly, Dias et al. (2023) applied DL to classify 12-lead ECG images into four categories, achieving an F1-score of 81.0% for the normal class using a private dataset. Although these studies report strong performance, they typically rely on full 12-lead inputs, which is not feasible for wearable applications.

The use of single-lead data, specifically Lead II, is of growing importance due to the proliferation of portable health technologies with significant size and power constraints. In this context, binary classification serves as an efficient tool for initial screening and early abnormality detection. Recent research introduced the LiteVGG-11 architecture, which employs depth-wise separable 1D convolutions and a reduced filter count to minimize parameters and computational costs. Clinical assessments have demonstrated its efficacy for real-time prediction in continuous wearable monitoring [Soares et al. 2024].

In this study, we evaluate the feasibility of using LiteVGG-11 for binary classification using exclusively Lead II data. The model was trained on the large-scale CODE15 dataset and externally validated against a real-world clinical dataset obtained from a single-lead wearable device.

2. Material and Method

2.1. Datasets

We employed three datasets in this study, summarized in Table 1. For all datasets, the "Normal" category served as the negative class, while all other findings were grouped into an "Abnormal" category for binary classification.

CODE15 Dataset: This subset of the Brazilian Telehealth Network CODE dataset [Ribeiro et al. 2021] includes 345,779 12-lead ECGs collected between 2010 and 2016. It features seven diagnostic classes: Normal (134,657 exams), First-degree AV block (1dAVb), Right and Left bundle branch blocks (RBBB and LBBB), Sinus bradycardia (SB), Sinus tachycardia (ST), and Atrial fibrillation (AFib). The latter six were consolidated into the Abnormal category.

PTB-XL Dataset: For external validation, we used 21,837 12-lead ECG recordings from the public PTB-XL dataset [Wagner et al. 2020]. To maintain a strict binary task, only recordings labeled exclusively as NORM (9,083) were selected as Normal. Records with mixed labels (NORM alongside other superclasses, here named as "Others") or those labeled as Myocardial Infarction (MI), ST/T changes (STTC), Conduction Disturbances (CD), or Hypertrophy (HYP) were categorized as Abnormal (12,754).

Wearable ECG Dataset: To evaluate performance under real-world conditions, 24-hour continuous ECGs were collected from 56 participants (28 with confirmed AFib) using the MD Sensor, a certified Class IIa medical device compliant with the EU Medical Device Regulation 2017/745 (Certificate CR-03-1229-813-23 02 / FI-MF- 000024281) [Soares et al. 2024]. The device operated at 128 Hz in a modified Lead II configuration (reference on the right shoulder and positive on the left abdomen). An experienced cardiologist annotated 120 random 30-second windows per subject, which were re-segmented into 10-second segments for a total of 20,160 signals. The final distribution consists of 11,634 Normal, 4,632 AFib, 1,503 Other Abnormalities, and 2,391 Artifacts (8,526 Abnormal instances).

2.2. Preprocessing

Following the preprocessing approach outlined in Soares et al. (2024), we extracted only the Lead II signals from the 12-lead CODE15 and PTB-XL datasets and resampled them

Table 1. Summary of the datasets used in this study.

ECG Dataset	Source	Total Signals	Number of Classes	Normal ECGs
CODE 15 (12-lead)	Public	345,779	7	134,657
PTB-XL (12-lead)	Public	21,837	5	9,083
Wearable (1-lead)	Private	20,160	4	11,634

to 128 Hz to align with the MD sensor's sampling rate. Signals were standardized to a 10-second duration by zero-padding shorter signals and truncating longer ones, resulting in a consistent 1280x1 vector shape. Baseline drift and low-frequency noise were reduced using a 1 Hz high-pass Butterworth filter (second-order), followed by a 40 Hz low-pass Butterworth filter (second-order) to retain diagnostically relevant frequencies. Finally, each signal was normalized to zero mean and unit variance for consistent model input.

2.3. Deep Learning Model

We employed the LiteVGG-11 architecture [Soares et al. 2022], a lightweight 1D-CNN optimized for resource-constrained wearable devices. To maximize computational efficiency, standard convolutions were replaced with depthwise separable convolutions, significantly reducing the parameter count while maintaining feature extraction capabilities. Efficiency is further enhanced by a smooth filter growth rule, reduced hidden units in the dense layers, and the use of global max pooling instead of flattening. As illustrated in Figure 1, the model comprises 11 main layers and 20,365 parameters.

Training was conducted using the Adam optimizer (learning rate: 5×10^{-3}) and binary cross-entropy loss with logits for up to 200 epochs. To mitigate class imbalance, we applied a weighted loss strategy based on inverse class frequencies. The training process utilized a batch size of 128, with 20% of the data reserved for validation. We employed early stopping with a 50-epoch patience and a learning rate reduction factor of 0.2 if the validation loss plateaued for 30 consecutive epochs. In this study, abnormal ECGs were defined as the positive class, while normal ECGs constituted the negative class.

2.4. Evaluation

Model performance was assessed using Accuracy, Sensitivity, Specificity, Precision, F1-score, and the Area Under the ROC Curve (AUC). For development, the CODE15 dataset

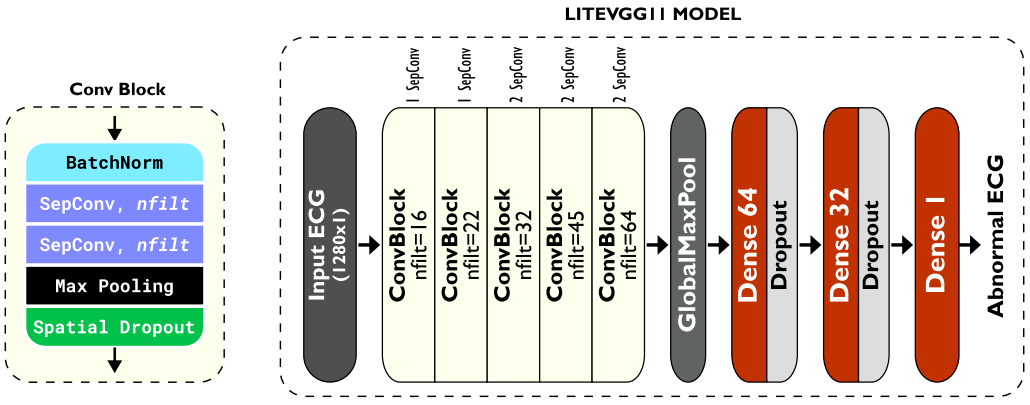

Figure 1. Architecture of the LiteVGG-11 model.

was partitioned into 80% for training/validation and 20% for testing, using stratified sampling to preserve the "Normal" class proportion. The evaluation focused on misclassifications, specifically abnormal segments incorrectly predicted as normal, to identify which pathological classes were most prone to error. Generalizability and robustness were further tested via external validation on the PTB-XL and a private wearable dataset. For the wearable data, performance was compared across two scenarios, including and excluding signal artifacts, to determine the model's resilience to real-world noise and its practical applicability for continuous monitoring.

3. Results and Discussion

The performance of the LiteVGG-11 model across the three datasets is summarized in Table 2. The model achieved moderate performance on the CODE15 test set with an AUC of 0.736 and demonstrated a degree of generalizability when applied to the standardized signals of the PTB-XL dataset (AUC: 0.773, Accuracy: 0.788). However, the transition to real-world wearable data highlighted significant deployment challenges. While sensitivity remained high (> 0.94), specificity experienced a sharp decline from 0.519 in artifact-free segments to 0.365 when signal artifacts were present. This substantial drop in specificity suggests that environmental noise significantly compromises the model's utility. To address this, the implementation of a simple signal quality filter could be a viable strategy to discard unreliable segments before classification, thereby reducing false alarms.

A detailed examination of subclass-wise false negative rates (FNR), detailed in Table 3, reveals a critical diagnostic boundary. In the CODE15 test set, the model performed effectively on well-defined rhythm and conduction disorders, specifically achieving low FNR for LBBB (0.003) and AFib (0.005). Similar trends were observed in the wearable dataset, where the model maintained high sensitivity for AFib (FNR: 0.005) and correctly identified most artifacts (FNR: 0.007). Nevertheless, performance degraded notably within the heterogeneous "Others" category (FNR: 0.349 in CODE15; 0.222 in PTB-XL) and for specific pathologies in the PTB-XL dataset, such as Myocardial Infarction (FNR: 0.119) and Hypertrophy (FNR: 0.118).

These performance discrepancies stem from the inherent physical and diagnostic constraints of utilizing a single lead (Lead II). As a unidimensional vector oriented at approximately $+60^\circ$ in the frontal plane, Lead II is effectively "blind" to electrical activity propagating perpendicularly to its axis or occurring within the horizontal plane. Consequently, ischemic events located in the anterior or lateral walls, which are typically associated with occlusions in the left anterior descending or circumflex arteries, may not

Table 2. Performance metrics for the LiteVGG-11 model.

Dataset	Sensitivity	Specificity	Precision	F1-Score	AUC	Accuracy
CODE15 test set	0.709	0.763	0.824	0.763	0.736	0.730
PTB-XL dataset	0.864	0.682	0.792	0.827	0.773	0.788
Wearable dataset						
Without Artifacts	0.945	0.519	0.509	0.662	0.732	0.666
With Artifacts	0.975	0.365	0.529	0.686	0.670	0.623

produce detectable changes in this single vector. Furthermore, without the spatial context provided by precordial leads such as V1 and V6, it is clinically unfeasible to accurately differentiate between various types of bundle branch blocks or identify complex axis deviations.

Given these findings, LiteVGG-11 is best positioned as a preliminary screening tool rather than a definitive diagnostic solution. While its high sensitivity (reaching 0.975 in the presence of artifacts) ensures that critical rhythm events like AFib are rarely missed, its low specificity and the resulting potential for false positives necessitate constant clinician oversight. Furthermore, while the architecture is optimized for resource-constrained environments, the current study is limited by the absence of a direct performance comparison with established lightweight models such as MobileNet or EfficientNet-lite. Future work will focus on benchmarking against these architectures to validate the structural advantages of LiteVGG-11 and exploring methods to enhance robustness against real-world artifacts.

Table 3. Subclass-wise false negative rates in the binary normal/abnormal classification task, for each test dataset.

CODE15-Test							
	AFib	ST	SB	LBBB	RBBB	1dAVb	Others
	0.005	0.009	0.017	0.003	0.010	0.053	0.349
PTB-XL					**Wearable**		
CD	HYP	MI	STTC	Others	AFib	Others	Artifacts
0.051	0.118	0.119	0.066	0.222	0.005	0.110	0.007

4. Conclusion

This study validated LiteVGG-11 as a lightweight solution for binary ECG screening using exclusively Lead II data. Trained on the CODE15 dataset and validated across PTB-XL and wearable recordings, the model demonstrates the feasibility of identifying well-defined rhythm disorders in resource-constrained environments. However, its performance in wearable settings is significantly challenged by signal artifacts, which reduce specificity and may limit its practical utility without further noise mitigation. Analysis of misclassifications confirms that while the model achieves exceptionally low false negative rates for critical pathologies like AFib, errors are primarily concentrated in subtle or diffuse abnormalities that are inherently difficult to detect with a unidimensional lead. This reinforces the model's role as a preliminary screening tool that requires clinician oversight rather than a definitive diagnostic solution.

To further its clinical utility, future work will focus on benchmarking LiteVGG-11 against established lightweight architectures, such as MobileNet or EfficientNet-lite, to better quantify its structural advantages. Additionally, subsequent research will prioritize enhancing robustness to signal artifacts and exploring transfer learning strategies to adapt the model to the diverse real-world signal variations encountered in continuous wearable monitoring.

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".

References

Bahuguna, R., Upadhyay, S., Aditi, Kumar, V., Saini, A., and Jain, A. (2023). Normal and abnormal ecg signal classification using deep learning. In *2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, pages 645–649.

Dias, F. M., Ribeiro, E., Moreno, R. A., Ribeiro, A. H., Samesima, N., Pastore, C. A., Krieger, J. E., and Gutierrez, M. A. (2023). Artificial intelligence-driven screening system for rapid image-based classification of 12-lead ecg exams: A promising solution for emergency room prioritization. *IEEE Access*, 11:121739–121752.

Goldberger, A. L., Goldberger, Z. D., and Shvilkin, A. (2018). Chapter 5 - the normal ecg. In Goldberger, A. L., Goldberger, Z. D., and Shvilkin, A., editors, *Goldberger's Clinical Electrocardiography (Ninth Edition)*, pages 32–40. Elsevier, 9th edition.

Ribeiro, A. H., Paixao, G. M., Lima, E. M., Horta Ribeiro, M., Pinto Filho, M. M., Gomes, P. R., Oliveira, D. M., Meira Jr, W., Schon, T. B., and Ribeiro, A. L. P. (2021). Code-15%: a large scale annotated dataset of 12-lead ecgs.

Soares, Q. B., Andrade, D. A., Ribeiro, E., Verardino, R. G. S., Reis, T. C., Samesima, N., Monteiro, R., Jatene, F. B., and Gutierrez, M. A. (2024). 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)*, pages 1–4.

Soares, Q. B., Monteiro, R., Jatene, F. B., and Gutierrez, M. A. (2022). A Lightweight Unidimensional Deep Learning Model for Atrial Fibrillation Detection. In *2022 Computing in Cardiology (CinC)*, volume 498, pages 1–4. IEEE.

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., and Schaeffter, T. (2020). PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154.

Zhu, J., Lv, J., and Kong, D. (2022). CNN-FWS: A Model for the Diagnosis of Normal and Abnormal ECG with Feature Adaptive. *Entropy*, 24(4).