Active Learning Approach for Clinical Noise Characterization in Long-Term Electrocardiogram Monitoring

Roberto Holgado-Cuadrado [1*], Francisco Manuel Melgarejo-Meseguer[2], José Luis Rojo-Álvarez[2] and Manuel Blanco-Velasco[1]

[1] Universidad de Alcalá, Madrid, Spain
[2] Universidad Rey Juan Carlos, Madrid, Spain

Abstract

Long-term monitoring (LTM) of electrocardiogram (ECG) can help identify intermittent arrhythmias that may not appear on shorter recordings, but it is often contaminated by noise, affecting their diagnostic utility. Recognizing clinically valid ECG parts can improve the performance of signal processing systems and reduce cardiologists' analysis time. This study computes the distortion of the ECG based on clinical readability rather than quantitative measures. Although we have proposed various machine learning systems to characterize clinical noise, we must expand our limited data to enhance performance. The main objective is to develop an active learning (AL) methodology to increase labeled data and improve our clinical noise classification models. Our experimental database comprises 8.467 excerpts of 5 seconds, labeled by a trained expert and a cardiologist, balanced across both clean and noisy categories. We adopt an AL scheme with a 1-D Convolutional Neural Network (CNN) based on an autoencoder, which iteratively incorporates new examples into the training set based on their class probability. The F1-score test curves illustrate that the AL scheme outperforms sample selection at random. This approach can refine the models by increasing labeled data for training, thus enhancing health professionals' confidence in medical decision support systems for clinical practice applications.

1. Introduction

Long-term monitoring (LTM) of ECG recordings involves collecting ambulatory signals that are acquired over an extended period during routine daily activities. This method enables the detection of cardiac pathologies that may not appear in shorter recordings [1]. However, due to their extended duration, some segments of the ECG are highly contaminated by various artifacts, and some portions of these recordings are invalid for diagnosis. Healthcare professionals analyze morphological changes in the heart rhythm throughout the monitoring period, spending approximately 10 to 15 minutes for each 24-hour recording. However, this task may be more time-consuming when the signal is difficult to analyze due to noise. Therefore, the automatic identification of invalid segments in the ECG is not only relevant to improve the performance of signal processing systems, but also to reduce the time required for the cardiologists' analysis.

Our alternative approach defines a clinical severity of noise criterion based on a set of rules in the reading and analysis procedure performed by a cardiologist [2, 3]. Our approach contrasts the traditional one, where noise is seen as an undesirable artifact appended to the ECG, assessing its severity quantitatively. Conversely, our approach is based on the readability of the ECG morphology, like cardiologists do, to discern whether any signal excerpt contains valuable information to provide a diagnosis. Thus, we introduced the criterion of clinical severity in [2], where we proved that our scale of clinical severity of noise is not related to the traditional approach. We then refer to it as *clinical noise* when defining its severity according to the clinical severity. Following this scale, we entirely labeled a database that reflects the common practice of clinician experts, and we tested different ML models in [3], showing the feasibility of classifying the different levels of clinical noise. We used several ML models based on feature engineering to distinguish between two classes: *clean*, when the whole morphology of the ECG is readable, *noise*, when the ECG lacks its pattern. We have also processed raw ECG signals and developed deep learning models, including scratch-designed and pre-trained convolutional neural networks (CNN), and architectures that enhance explainability in the decision-making process [4].

While our algorithms have shown promising results, they need to improve their performance, which can be achieved by incorporating new data to facilitate better training. However, the challenge lies in the difficulty of obtaining additional data conveniently labeled, as it is a labor-

Table 1: Distribution and size of the database by class and noise category.

Database			
Clean	4.206 (49.68%)	T0	2.093 (24.72%)
		T1	2.113 (24.96%)
Noisy	4.261 (50.32%)	T2	2.121 (25.05%)
		T3	2.140 (25.27%)

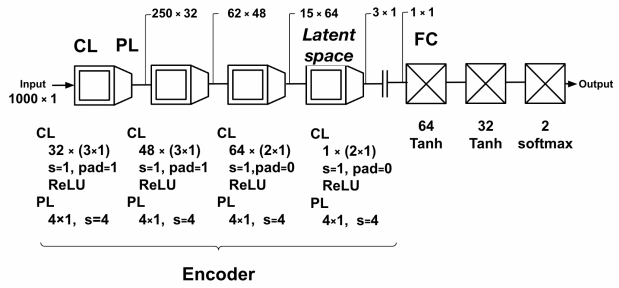

Figure 1: Architecture of the convolutional AE. The encoder is indicated with a brace, representing the set of layers responsible for compressing the input information into the latent space.

intensive, cumbersome, and expensive process guided by specific clinical criteria and traditionally done manually. To address this task, we propose the application of active learning (AL) techniques [5, 6], allowing our models to iteratively learn from previously acquired knowledge and feedback received when making new predictions. In this work, we explore an AL scheme to enhance clinical noise classification, aiming to increase our labeled database and refine our systems.

The structure of this paper is organized as follows: Section 2 introduces the database, Section 3 outlines the model and the AL scheme for clinical noise classification, Section 4 shows the results, and Section 5 offers the conclusions.

2. Database

In this study, we employ the data repository introduced in [2], comprising 6.5 hours of ECG recordings from 10 patients, collected using an External Event Recorder (EER) at a sampling rate of 200 Hz. EER devices enable continuous ambulatory ECG monitoring over extended periods, capturing significant changes in the ECG, including arrhythmias or artifacts. The data repository was labeled according to the procedure outlined in [2], involving an iterative process between a trained expert and a cardiologist, which led to the clinical severity score categorized as follows:

- *Noise-free* (T0) represents a noise-free segment, where all waves are clearly recognizable.
- *Low noise* (T1) presents some noise but with readable P and T waves and identifiable QRS complexes.
- *Moderate noise* (T2) represents a moderately noisy segment where only QRS complexes can be identified in at least three consecutive beats.
- *Hard noise* (T3) represents a segment with hardly recognizable or unrecognizable QRS complexes.
- *Other noise* (T4) represents segments containing calibration pulses or straight lines due to the complete absence of signal or amplifier saturation.

We adopted the same procedure as in [3,4], and we pose a binary problem where *clean* class gathers ECG with presenting full morphology (categories T0 and T1) and *noisy* class represents signal with no ECG morphology (categories T2 and T3). T4 (*other noise*) has been omitted as it can be easily identified by simple signal processing techniques.

To design the database for training learning methods, we randomly extract non-overlapping 5-s excerpts (1000 samples) from the data repository. Due to the imbalance in noise categories, we employ balancing strategies, which involve randomly undersampling ECGs from patients with an overrepresentation of noisy categories, and using specific overlap across underrepresented categories [4]. This process led to the creation of a balanced database, referred to as DB-2 in [4], that consists of 8.467 excerpts of 5 seconds from all ten patients, as detailed in Table 1.

3. Methods

3.1. Classifier

We employ a 1–D CNN based on an autoencoder (AE), designed from scratch. The proposed scheme, determined as convolutional AE (Figure 1), operates on the raw data and maps this information into a 3–D latent space. In the diagram, a double square represents a convolutional layer (CL), a trapezoid is a pooling layer (PL), a double vertical line stands for a global pooling layer, and a square with a cross inside is a fully connected layer (FC). Our convolutional AE model consists of four convolutional blocks followed by three fully connected layers. Each convolutional block consists of a 1–D convolution operation with batch normalization, and a leakyReLU activation function followed by a max-pooling layer to reduce the dimension at the output. The number of convolutional filters used in the first four blocks are 32, 48, 64, and 1, respectively, to enable the data visualization in a 3-D latent space. We apply max pooling with a size of 4 and a stride of 4 in each convolutional block. Zero–padding (pad) is used to process the incoming data and preserve the input size. After the last convolutional block, a global maxpooling layer is performed to flatten the extracted features. The fully con-

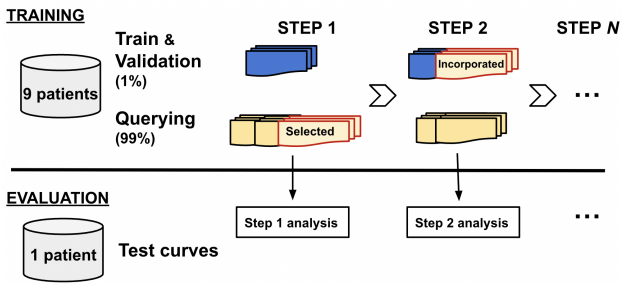

Figure 2: Scheme of the AL framework.

nected layers have 64 and 32 neurons, respectively, using leakyReLU activation. The last layer has two neurons corresponding to the two classes with softmax activation.

The network comprises a total of 13.638 weights. We use Min–Max normalization, and the optimal hyperparameters are determined through a grid search, wherein the learning rate is set at $\{10^{-3}, 10^{-4}, 10^{-5}\}$ and the mini–batch size is chosen from $\{128, 256, 512\}$. The best hyperparameter configuration entails the stochastic gradient descent algorithm with a momentum of 0.99 with a constant learning rate of 10^{-4}, a mini-batch size of 256, 200 epochs, and L2 regularization with a coefficient of 0.001 to mitigate overfitting.

3.2. AL scheme

AL is a labeling technique that involves an iterative process where the algorithm actively selects and incorporates new examples into the training set to improve the model's performance. This method is particularly useful when labeled data is scarce or expensive to obtain [5]. The proposed AL framework, illustrated in Figure 2, operates as follows. Initially, we used data from 9 out of 10 patients to train the AL scheme, partitioned into three data subsets: a training and validation set (1%) and a querying set (99%). The evaluation of the AL scheme is performed using a test set consisting of data from the remaining single patient. This patient permutation strategy, outlined in [3], mitigates the challenges of the limited availability of patients and the risk of intra-patient overfitting. We deploy an AL scheme per patient in the dataset included in the test set.

In the first iteration of the AL scheme, the convolutional AE is trained on a very small initial training set, and the best model is selected based on its performance on the validation set. The selected model then classifies samples from the querying set and actively selects a specific number of examples, specifically 30 samples, using margin sampling— a sample selection strategy that identifies samples whose predictions for any class are most uncertain, indicating proximity to the classification boundary. These selected samples are incorporated into the training set with their actual labels in the next iteration, being subsequently removed from the querying set. This iterative approach is executed until the querying set is empty. At each iteration of the AL scheme, the model performance is evaluated using the test set.

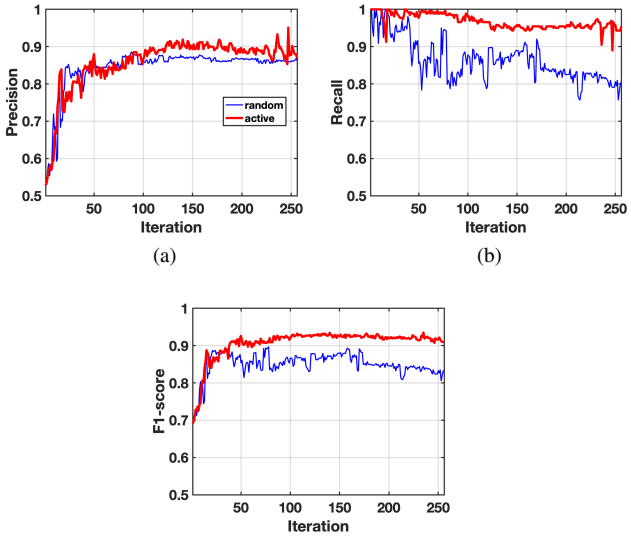

Figure 3: Performance curves for a) Precision, b) Recall, and c) F1-score, obtained for one permutation of 10 patients.

4. Results

Table 2 shows the overall performance of the convolutional AE [4]. The training and validation sets show better performance than the test set, suggesting an overestimation of performance when evaluating the same patients used in training. The higher standard deviation in the test set reflects limitations in generalization across patient permutations, highlighting the need for more training patients.

Figure 3 illustrates the precision, recall, and F1-score curves on the test set, generated by the convolutional AE within the AL scheme for a specific permutation. These curves are compared with a random sampling scheme that follows a similar framework but with sample selection at random. In the early iterations of the AL-as-random scheme, the learning model demonstrates a high detection

Table 2: Performance of the convolutional AE.

	Acc	Re	Pr	F1–score
Train	0.96 ± 0.02	0.97 ± 0.01	0.95 ± 0.03	0.96 ± 0.02
Validation	0.92 ± 0.02	0.94 ± 0.01	0.91 ± 0.03	0.93 ± 0.02
Test	0.83 ± 0.08	0.82 ± 0.13	0.83 ± 0.06	0.82 ± 0.09

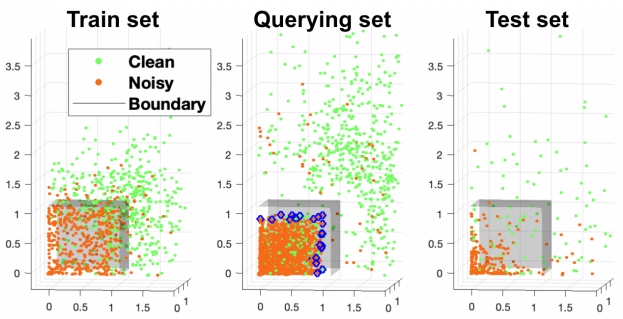

Figure 4: Representation of the latent space of the convolutional AE for the train, querying, and test set for one permutation of the 10 patients.

capability (Figure 3b), but this comes at the expense of low precision (Figure 3a). As more examples are incorporated into the model training, the number of false positives decreases, leading to improved precision. Figure 3c shows the F1-score curve, which is particularly insightful as it combines precision and recall into a single metric. It shows a progressive increase in F1-score across iterations. The performance achieved with the AL scheme outperforms that obtained with random sample selection across all metrics, including F1-score, recall, and precision. However, there are noticeable fluctuations in the curves over iterations, particularly pronounced in the random sampling scheme. This variability highlights the limited size of the database and the importance of techniques such as AL.

The model employed in the AL scheme, the convolutional AE, allows us to map the data into a latent space and observe how they are distributed across different phases of the training set. Figure 4 illustrates the latent space provided by the convolutional AE in the training, querying, and test sets, respectively, in a specific iteration of the AL scheme. By visualizing the latent space, we provide a qualitative explanation of the decisions made by the model and the AL scheme. The classification boundary, represented in gray, is determined using a domain detection technique within the resulting low-dimensional manifolds in the training latent space [7]. In the querying set, we can identify the samples (highlighted in blue circles) that lie closer to the boundary and are subsequently incorporated into the training set in the next iteration. Furthermore, in the test set, the model demonstrates effective generalization by separating classes for unknown ECG excerpts from a patient not included in the training set. This representation illustrates how different regions in the latent spaces are associated with the two noise categories, thereby enhancing the explainability of the model's decision-making for each patient.

5. Conclusions

We have developed an AL framework that integrates margin sampling with a convolutional AE model that provides interpretability in the decision-making process. We conclude that the proposed AL scheme can effectively refine the models by increasing the labeled data for training. This approach strengthens the confidence of health professionals in medical learning decision support systems, particularly in the detection of invalid parts in LTM ECG recordings, facilitating its application in clinical practice.

Acknowledgments

This work has been partially supported under research project grants EPU-INV/2020/002 from Community of Madrid, PIUAH23-IA-014 and PRE FPI-UAH-23 from the University of Alcalá, as well as PID2022-140786NB-C32 and PID2023-152331OA-I00 from MCIN/AEI/10.13039/501100011033.

References

[1] Jabaudon D, Sztajzel J, Sievert K, Landis T, Sztajzel R. Usefulness of ambulatory 7-day ECG monitoring for the detection of atrial fibrillation and flutter after acute stroke and transient ischemic attack. Stroke 2004;35(7):1647–1651.

[2] Everss-Villalba E, Melgarejo-Meseguer FM, Blanco-Velasco M, Gimeno-Blanes FJ, Sala-Pla S, Rojo-Álvarez JL, García-Alberola A. Noise maps for quantitative and clinical severity towards long-term ECG monitoring. Sensors 2017; 17(11):2448.

[3] Holgado-Cuadrado R, Plaza-Seco C, Lovisolo L, Blanco-Velasco M. Characterization of noise in long-term ECG monitoring with machine learning based on clinical criteria. Medical Biological Engineering Computing 2023;1–14.

[4] R. Holgado-Cuadrado C. Plaza-Seco LL, Blanco-Velasco M. A deep and interpretable learning approach for long-term ECG clinical noise classification. IEEE Transactions on Biomedical Engineering 2024;(Under review).

[5] Pasolli E, Melgani F. Active learning methods for electrocardiographic signal classification. IEEE Transactions on Information Technology in Biomedicine 2010;14(6):1405–1416.

[6] Hong S, Zhou Y, Shang J, Xiao C, Sun J. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. Computers in Biology and Medicine 2020;122:103801.

[7] Melgarejo-Meseguer FM, Lorenzo-Bleda A, Eduardo-Abbenante S, Gimeno-Blanes FJ, Everss-Villalba E, Muñoz-Romero S, Rojo-Álvarez JL, Ferriz RM. Anomaly detection from low-dimensional latent manifolds with home environmental sensors. IEEE Internet of Things Journal 2023;.

Address for correspondence:

Roberto Holgado Cuadrado. □ Signal Theory and Communicatrion Department, Universidad de Alcalá. □ roberto.holgado@uah.es