Self-Supervised ECG Foundation Models for Enhanced Detection of Paroxysmal Atrial Fibrillation

Sara Artal[1], Antonio Miguel[1], Juan Pablo Martínez [1,2], Julia Ramírez[1,2,3]

[1] Aragon Institute of Engineering Research, University of Zaragoza, Zaragoza, Spain
[2] Centro de Investigación Biomédica en Red, Biomateriales, Bioingeniería y Nanomedicina, Spain
[3] William Harvey Research Institute, Queen Mary University of London, London, United Kingdom

Abstract

Paroxysmal atrial fibrillation (PAF) is a common intermittent supraventricular arrhythmia that is challenging to detect. The electrocardiogram (ECG) is a low-cost, non-invasive tool suitable for its diagnosis. Recently, neural networks (NNs) have emerged as a promising approach for detecting PAF from ECG signals. However, their adoption remains limited by the scarcity of high-quality labeled data. This study evaluates the impact of key design choices in ECG foundation models (ECG-FMs) on representation quality and downstream performance for PAF detection under limited data conditions. First, a convolutional NN (CNN) was implemented for the detection of PAF from ECGs in sinus rhythm. Then, public ECG datasets were collected, yielding 1,784,071 ECG recordings. Using these data, six ECG-FMs were pre-trained in a self-supervised learning (SSL) manner via contrastive predictive coding (CPC) to assess the influence of dataset size (N), learning rate (Lr), batch size (BS), and data augmentation strategies. Incorporating representations from variant CPC-5 (larger N, lower Lr, larger BS, and augments) into a simplified CNN improved the area under the curve from 0.606 to 0.716 and raised specificity from 0.407 to 0.573. These results show that training design choices, particularly dataset size and Lr, are decisive for SSL performance.

1. Introduction

Atrial fibrillation (AF) is a major contributor to cardiovascular mortality [1]. It is an arrhythmia characterized by an irregular, often rapid, heart rate, which may occur intermittently, known as paroxysmal atrial fibrillation (PAF). Although invasive procedures can accurately diagnose PAF, they are not suitable for large-scale risk stratification, highlighting the need for effective non-invasive methods [2].

The electrocardiogram (ECG) is widely used for the early diagnosis of PAF. In recent years, neural networks (NNs) have enabled automated ECG analysis, identifying key morphological features associated with AF risk [3]. However, their performance remains limited, likely due to the need for large annotated datasets, which are costly and scarce. Additionally, these models are sensitive to class imbalance, as AF prevalence is only 2–4%, requiring even larger cohorts for sufficient statistical power [4].

To address these limitations, ECG foundation models (ECG-FMs), typically pre-trained on large and diverse unlabeled datasets using self-supervised learning (SSL), aim to learn general-purpose ECG representations that transfer effectively across downstream tasks [5]. In a previous work [6], we explored the use of Contrastive Predictive Coding (CPC) [7], a method that combines future prediction with a contrastive loss, and evaluated it as a frozen feature extractor to simplify downstream PAF detection models while assessing its impact on performance. However, this approach presents several limitations, including the lack of a systematic evaluation of key training hyperparameters that could further optimize its performance.

In this study, we address these gaps by evaluating six CPC-based models under different training configurations, varying dataset size (N), learning rate (Lr), batch size (BS), and the use of data augmentations, to identify the main factors affecting CPC representation quality.

2. Materials and Methods

2.1. Materials

In this project, datasets from public repositories (Table 1), including PhysioNet and Zenodo, were used.

2.2. ECG Preprocessing

All ECG signals were resampled to 500 Hz using a quadrature polyphase filter. Baseline noise was removed using a fourth-order Butterworth high-pass IIR filter with a cutoff frequency of 0.3 Hz.

Table 1. Databases used in this study. W=ECG lead I-like acquired from a wearable device.

Database	Patients	Recordings	Leads	Fs (Hz)	Duration
ICENTIA	11,000	542,157	1 (W)	250	∼70 min
ECG-AR	45,152	45,152	12	500	10 s
PTB-XL	18,869	21,799	12	500	10 s
GI2EC	15,742	10,292	12	500	5 or 10 s
PhysioNet17	8,528	8,528	1 (W)	300	9 to 60 s
CPSC	9,458	6,876	12	500	6 to 144 s
CPSC-Extra	9,458	3,453	12	500	8 to 98 s
MIMIC-IV	161,352	800,035	12	500	10 s
CODE-15%	233,770	345,779	12	400	10.24 s
AFP	200	300	2	128	30 or 5 min
Total	513,529	1,784,371	-	-	-

2.3. Supervised learning model

The aim of the supervised learning model is to detect PAF from ECG segments recorded in sinus rhythm.

We employed a convolutional NN (CNN) consisting of two main components: feature extraction and classification. As shown in Figure 1 (b), feature extraction uses five 1D convolutional layers with 128 hidden units, kernel size 3, and stride 2. This is followed by two bidirectional LSTM layers, able to capture relevant temporal dependencies. Furthermore, batch normalization (BN) is applied between convolutional layers to improve convergence, performance, and stability. ReLU is used as the activation function throughout the network, enabling the modeling of non-linear relationships in the data. For classification, a single fully connected layer outputs the probability that an input ECG belongs to the positive class (AF).

The model was trained on the AFP database (Table 1), using 300 ECG signals for training, corresponding to 100 subjects (50 with PAF and 50 without PAF), and 200 signals for testing from an additional 100 subjects. Training was performed for 100 epochs, with 300 samples per epoch. The Lr was initialized at 0.003 and decayed by 5% every 20 iterations to promote stable convergence. Data were processed in batches of 64 samples, where each input consisted of a randomly selected 7-second ECG segment.

Model performance was evaluated using specificity (Sp), the receiver operating characteristic (ROC) curve, and the area under the ROC curve (AUC), with sensitivity (Se) fixed at 0.750 to ensure minimum performance.

2.4. Self-supervised learning models

The developed CPC model (Figure 1 (a)) consists of two main components: a non-linear encoder and an autoregressive (AR) module. Given an input ECG signal, the encoder maps raw samples x_t into latent representations z_t [7]. In our implementation, the encoder comprises six 1D convolutional layers with 1024 hidden units, kernel size 3, and stride 2, with BN and ReLU activations applied between convolutional layers.

The latent sequence z_t is then processed by the AR module, implemented as a GRU layer [8], which aggregates past information into a context vector c_t. This vector summarizes the signal up to time t and is used to predict future representations. For each prediction horizon k (with $k = 8$), a prediction vector is generated from c_t using a prediction head composed of two convolutional layers with a ReLU in between. This can be interpreted as identifying the true continuation of the ECG in the latent space while distinguishing it from a set of negative distractors.

The SSL model is trained on multiple large-scale ECG datasets (Table 1), excluding AFP, which is reserved for the supervised model. Two pre-training configurations are defined: the first includes 637,757 lead I and wearable ECG signals from datasets such as ICENTIA (70-minute recordings), ECG-AR, PTB-XL, G12EC, PhysioNet17, CPSC, and CPSC-Extra. The second expands this to 5,036,013 signals by increasing ICENTIA data (10-minute recordings; 3,795,099 samples) and adding MIMIC-IV and CODE-15%. A separate test set of 1,000 signals, 500 from non-ICENTIA sources in the first cohort and 500 from MIMIC-IV, is used, with all test samples excluded from training to prevent data leakage.

Table 2. Summary of CPC training configurations. HP=hyperparameters

HP	CPC-0	CPC-1	CPC-2	CPC-3	CPC-4	CPC-5
N	637,757	5,036,013	637,757	637,757	637,757	5,036,013
Lr	0.001	0.001	0.0001	0.001	0.001	0.0001
BS	256	256	256	1024	256	1024
Augments	No	No	No	No	Yes	Yes

Furthermore, to evaluate the impact of dataset scale, optimization settings, and data augmentations, we define six CPC variants (Table 2). These configurations differ in training dataset size, Lr, effective BS, and the use of augmentations such as white Gaussian and EMG noise, amplitude scaling, high-/low-pass FIR filtering, baseline drift, power line interference, and amplitude/temporal quantization. For CPC-3 and CPC-5, large effective batch sizes (1024) are achieved via gradient accumulation using four mini-batches of size 256.

The models were trained for 3,000 epochs with 25,000 samples per epoch. Input signals consist of ECG segments of variable length, ranging from 3 to 7 seconds, allowing the model to generalize across different temporal scales.

2.5. Supervised learning models based on self-supervised features

The primary objective was to evaluate six CPC-based SSL models as frozen feature extractors for the supervised task using a transfer learning approach, where the SSL

model remained fixed and only the supervised model was trained. ECG signals from the AFP database were encoded into latent representations (z, c, and their concatenation, zc), which served as inputs to the supervised architecture.

Within the supervised baseline, we explored architectural variants by progressively reducing the number of convolutional layers, retaining bidirectional LSTM layers only when convolutional layers were present, and ultimately considering a single fully connected layer. Performance was compared against the supervised baseline and reported as the mean and standard deviation (SD) over five runs.

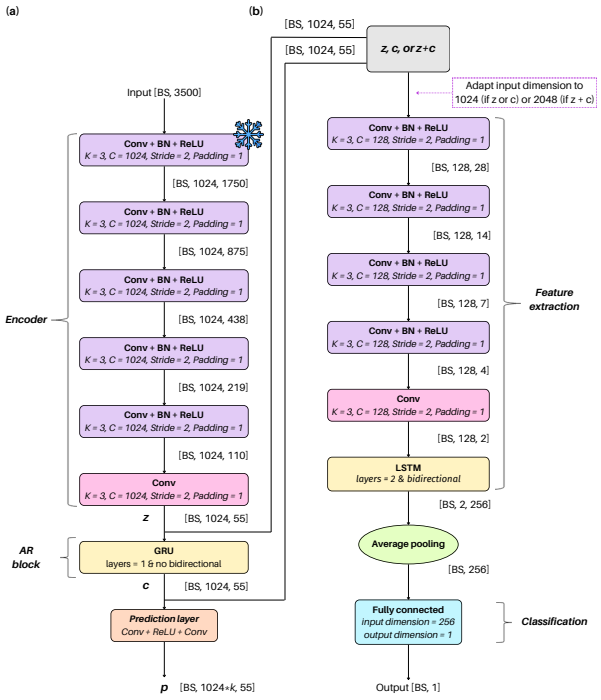

Figure 1. Overview of the proposed framework. (a) Self-supervised model architecture used as a frozen feature extractor. (b) Supervised model architecture with five 1D convolutional layers in the feature extraction stage.

3. Results

In the supervised PAF detection task, using the raw ECG signal (x) as input and a model with five convolutional layers, we obtained an AUC of 0.606 (0.034) and a Sp of 0.407 (0.060) at a fixed Se of 0.750 (Tables 3 and 4).

Results for supervised models built on self-supervised CPC features are also reported in these tables. Among all configurations, the best performance is achieved by combining the supervised architecture with the CPC-5 model. In this setup, variable zc is used as input, while the number of convolutional layers is reduced from five to one. This configuration yields a clear improvement over the baseline, reaching an AUC of 0.716 (0.023) and a Sp of 0.573 (0.048) at a Se of 0.750, while also showing a slight reduction in SD, indicating enhanced stability.

The CPC-2 model also provides notable enhancements, achieving competitive AUC (0.705 (0.020)) and Sp (0.568 (0.046)) values compared to the baseline. These results highlight the importance of the Lr, as better performance is obtained with a lower value. Both CPC-2 and CPC-5 use a Lr of 0.0001 instead of 0.001 (Table 2).

Overall, all CPC-based models outperform the purely supervised baseline. They also enable a substantial simplification of the supervised architecture, reducing it to a single convolutional layer without loss of performance. Moreover, even when reduced to a single fully connected layer, most CPC-based models achieve comparable or better results. CPC-4 is the only exception, showing no improvement over the baseline.

4. Discussion and Conclusions

In this study, we evaluated the performance gains obtained by leveraging different self-supervised pre-trained models as frozen feature extractors for PAF detection.

Our results indicate that training hyperparameters play a crucial role in the quality of CPC representations. Specifically, models trained with lower Lr (CPC-2 and CPC-5) achieve substantial improvements, suggesting more robust and transferable features. Dataset size also emerges as a critical factor, as CPC-1 shows notable gains over the supervised baseline. As expected, CPC-5, which combines all studied modifications, including a lower Lr and a larger dataset, achieves the best overall performance. While the individual effects of the considered factors can be partially assessed through CPC-1 to CPC-4, we did not perform a full factorial hyperparameter exploration. Therefore, potential interactions between dataset size, Lr, BS, and augmentations remain an important direction for future work.

Increasing BS (CPC-3) also enhances performance, although to a lesser extent. Notably, constructing batches from homogeneous data sources may be more beneficial than simply increasing BS, as it encourages the use of harder negative samples and reduces the risk of exploiting dataset-specific noise or acquisition artifacts [9]. However, not all configurations yield improvements. Only CPC-4 fails to outperform the baseline, suggesting that combining CPC-0 with augmentations may not suit this framework.

A key strength of this study lies in the scale and nature of the pre-training dataset, which comprises over 1.8 million single-lead ECG signals, including wearable-derived recordings (long-term chest-worn device recordings from the ICENTIA database and handheld device recordings from PhysioNet17). To the best of our knowledge, this is the first study to develop an ECG-FM pre-trained exclusively on lead I data, from both standard 12-lead ECGs and wearable recordings, enabling future validation in relevant

Table 3. AUC performance (mean ± SD) on the AFP dataset for PAF detection, comparing supervised and CPC-based self-supervised representations across convolutional depths. HP=hyperparameters, CL=convolutional layers.

Model	HP	Input	5 CL	4 CL	3 CL	2 CL	1 CL	0 CL
Supervised	–	x	**0.606 (0.034)**	0.529 (0.032)	0.531 (0.034)	0.588 (0.014)	0.598 (0.011)	0.502 (0.049)
CPC-0	Baseline	zc	0.653 (0.029)	0.626 (0.021)	0.624 (0.026)	0.641 (0.029)	**0.651 (0.011)**	0.602 (0.012)
CPC-1	Larger N	z	0.622 (0.030)	0.666 (0.010)	0.641 (0.021)	0.673 (0.032)	**0.680 (0.011)**	0.601 (0.007)
CPC-2	Lower Lr	c	0.675 (0.041)	0.709 (0.031)	0.687 (0.019)	0.699 (0.020)	**0.705 (0.020)**	0.656 (0.008)
CPC-3	Larger BS	z	0.601 (0.038)	0.633 (0.039)	0.623 (0.012)	0.610 (0.017)	**0.651 (0.016)**	0.625 (0.010)
CPC-4	Augments	c	0.489 (0.048)	0.525 (0.082)	0.438 (0.064)	0.493 (0.081)	0.519 (0.079)	0.498 (0.007)
CPC-5	All	zc	0.660 (0.030)	0.699 (0.031)	0.685 (0.019)	0.694 (0.015)	**0.716 (0.023)**	0.637 (0.007)

Table 4. Sp (Se=0.750) performance (mean ± SD) on the AFP dataset for PAF detection, comparing supervised and CPC-based self-supervised representations across convolutional depths. HP=hyperparameters, CL=convolutional layers.

Model	HP	Input	5 CL	4 CL	3 CL	2 CL	1 CL	0 CL
Supervised	–	x	**0.407 (0.060)**	0.325 (0.042)	0.325 (0.049)	0.350 (0.038)	0.361 (0.038)	0.316 (0.078)
CPC-0	Baseline	zc	0.457 (0.035)	0.409 (0.041)	0.475 (0.045)	0.450 (0.026)	**0.432 (0.062)**	0.400 (0.017)
CPC-1	Larger N	z	0.398 (0.044)	0.448 (0.069)	0.452 (0.018)	0.491 (0.036)	**0.473 (0.029)**	0.452 (0.022)
CPC-2	Lower Lr	c	0.489 (0.073)	0.527 (0.056)	0.536 (0.032)	0.557 (0.043)	**0.568 (0.046)**	0.427 (0.018)
CPC-3	Larger BS	z	0.332 (0.181)	0.455 (0.092)	0.395 (0.042)	0.414 (0.082)	**0.455 (0.016)**	0.455 (0.007)
CPC-4	Augments	c	0.045 (0.091)	0.091 (0.124)	0.043 (0.046)	0.041 (0.066)	0.343 (0.167)	0.136 (0.029)
CPC-5	All	zc	0.473 (0.069)	0.566 (0.074)	0.509 (0.053)	0.552 (0.034)	**0.573 (0.048)**	0.452 (0.013)

cohorts [10].

Another important finding is that informative SSL representations allow substantial simplification of the supervised architecture. In most cases, the model can be reduced to a single convolutional or fully connected layer while maintaining or improving performance (except for CPC-4). This reduces memory usage, accelerates training, and mitigates overfitting, although inference still requires evaluating both SSL and supervised models.

In conclusion, our results identify the CPC configurations that yield the most significant performance gains, particularly those using larger unlabeled datasets with lower Lr, which emerge as the most effective strategies. These findings highlight the sensitivity of SSL methods to training design choices and their value in scenarios with limited annotated data.

Acknowledgments

This work was supported by projects TED2021-130459B-I00, PID2022-140556OB-I00, PID2023-148975OB-I00, and CNS2023-143599 funded by the Spanish Ministry of Science and Innovation (MCIN) and FEDER, as well as the Biomedical Signal Interpretation and Computational Simulation Group T39-23R. S. A. acknowledges support from a Master's Thesis scholarship awarded by the Aragon Institute of Engineering Research and a predoctoral research fellowship funded by the Government of Aragón (2025-2029). J. R. acknowledges funding from the MCIN grant RYC2021-031413-I. Computations were carried out using the ICTS NANBIOSIS infrastructure (University of Zaragoza).

References

[1] Yoneda ZT, et al. Mortality among patients with early-onset atrial fibrillation and rare variants in cardiomyopathy and arrhythmia genes. JAMA cardiology 2022;7(7).

[2] Khurshid S, Healey JS, McIntyre WF, Lubitz SA. Population-based screening for atrial fibrillation. Circulation research 2020;127(1):143–154.

[3] Attia ZI, et al. An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. The Lancet 2019;394(10201):861–867.

[4] Liu H, Zhao Z, She Q. Self-supervised ecg pre-training. Biomedical Signal Processing and Control 2021; 70:103010.

[5] Han Y, Murino V, Liu X, Zhang X, Ding C. A systematic review on foundation models for electrocardiogram analysis: Initial strides and expansive horizons, 2025. URL https://arxiv.org/abs/2410.19877.

[6] Artal S, Martínez JP, Miguel A, Ramírez J. Improved detection of paroxysmal atrial fibrillation using an ecg-based semisupervised model. In Computing in Cardiology, volume 51. 2024; 1–4.

[7] van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

[8] Cho K, Van Merriënboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: Encoder–decoder approaches. In Proceedings of SSST-8. 2014; 103–111.

[9] Yang Z, et al. Batchsampler: Sampling mini-batches for contrastive learning in vision, language, and graphs. In Proceedings of the 29th Conference on Knowledge Discovery and Data Mining. 2023; 3057–3069.

[10] Ahluwalia N, et al. Patient-led smartwatch ecg monitoring after af ablation: A randomized trial. JACC 2026; 87(14):1817–1819.

Address for correspondence:

Sara Artal
c/Mariano Esquillor s/n, Lab 5.1.02, 50018, Zaragoza, Spain
sartal@unizar.es