# Image-Based Electrocardiogram Classification Using Pre-trained ConvNext

Felipe M Dias[1,2], Estela Ribeiro[1], Quenaz B Soares[1,2], Jose E Krieger[1], Marco A Gutierrez[1,2]

[1]Heart Institute, Clinics Hospital, University of Sao Paulo Medical School, Brazil
[2]Polytechnique School, University of Sao Paulo, Brazil

## Abstract

*As part of the Digitization and Classification of ECG Images: The George B. Moody PhysioNet Challenge 2024, we developed a deep learning-based approach to classify electrocardiogram (ECG) images. Our methodology utilized the ConvNext architecture pre-trained on ECG images from CODE15 dataset, along with our private image-based ECG datasets, InCor-AMB and InCor-EMG. We fine-tuned the model on the PTB-XL dataset to classify 11 specific abnormalities. The training strategy included cross-entropy loss minimization, class-weight adjustments, and the use of the AdamW optimizer over 30 epochs, with early stopping to mitigate overfitting. A 5-fold cross-validation was employed, and the final model was an ensemble of the best-performing models from each fold. Our model achieved an F1-score of 0.698 $\pm$ 0.022 on internal 5-fold cross-validation. On the challenge's test set, our team, AIMED, achieved an F1-score of 0.817, securing 1st place in the classification task of the PhysioNet Challenge 2024. This study highlights significant advancements in the integration of digital technologies in clinical settings, with the potential to improve ECG diagnostic accuracy worldwide.*

## 1. Introduction

Automatic electrocardiogram (ECG) classification is a rapidly evolving research area with significant potential to aid physicians in ECG interpretation and diagnosis. Although most studies focus on time-series ECG signals due to their public availability, clinical practice predominantly relies on ECG images or printouts, creating a gap between research and real-world application. This discrepancy limits the practical utility of time-series-based approaches in real-world healthcare settings. The 2024 George B. Moody PhysioNet Challenge aimed to address this gap by inviting teams to develop automated, open-source tools that 1) digitize and 2) classify ECGs from images or paper printouts [1–3]. Our approach specifically targets the classification of ECG images.

We developed a model to classify ECG images using a ConvNext architecture [4]. The model was pre-trained on ECG images from the CODE15 dataset [5], artificially generated images using ECG-Image-Kit [6, 7], and two private datasets containing real ECG exam images: InCor-AMB [8,9] and InCor-EMG. Subsequently, we fine-tuned the model on the PTB-XL [10,11] challenge training dataset.

## 2. Methods

Our framework, illustrated in Figure 1, is based on the ConvNeXt neural network architecture [4] pre-trained on ImageNet dataset. We employed transfer learning on ECG-specific datasets to address the semantic gap between ImageNet and ECG image data. The model was fine-tuned on the PTB-XL dataset using 5-fold cross-validation, resulting in an ensemble of the five derived models. Further details are provided in the subsections that follow.

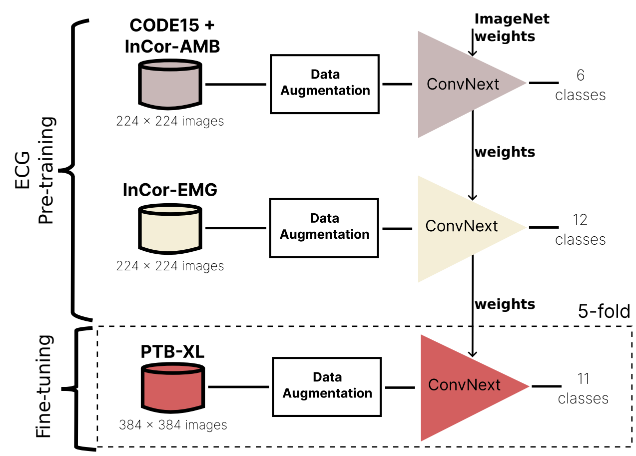


Figure 1: Overview of our proposed approach.

### 2.1. Data

In this work, we used two publicly available datasets, CODE15 [5] and PTB-XL [10, 11], and two private datasets, InCor-AMB [8, 9] and InCor-EMG. Table 1 provides a description of the datasets used.

 ISSN: 2325-887X DOI: 10.22489/CinC.2024.398

Table 1: Summary of employed ECG datasets: CODE15, InCor-AMB, InCor-EMG, and PTB-XL.

| **Exam Classes** | **CODE15** | **InCor-AMB** | **InCor-EMG** | **PTB-XL** |
|---|---|---|---|---|
| First-degree AV block (1dAVb) | 5,716 (1.7%) | 8,882 (8.8%) | - | - |
| Right bundle branch block (RBBB) | 9,672 (2.8%) | 9,195 (9.2%) | 1,563 (8.4%) | - |
| Left bundle branch block (LBBB) | 6,026 (1.7%) | 6,131 (6.1%) | 1,406 (7.6%) | - |
| Sinus bradycardia (SB/BRADY) | 5,605 (1.6%) | 4,560 (4.5%) | - | 637 (2.9%) |
| Atrial fibrillation (AF) | 7,033 (2.0%) | 9,064 (9.0%) | 2,303 (14.4%) | - |
| Sinus tachycardia (ST/TACHY) | 7,544 (2.2%) | 907 (0.9%) | - | 867 (4.0%) |
| Normal ECG (NORM) | - | - | 2,820 (15.2%) | 9,514 (43.6%) |
| Premature ventricular complex (PVC) | - | - | 2,569 (13.9%) | 1,143 (5.2%) |
| Paroxysmal supraventricular tachycardia / Sinus tachycardia (PSVT / ST) | - | - | 1,469 (7.9%) | - |
| Pacemaker | - | - | 1,228 (6.6%) | - |
| ST-segment elevation | - | - | 1,083 (5.8%) | - |
| Current of Injury | - | - | 755 (4.1%) | - |
| Atrial tachycardia/ Atrial Flutter | - | - | 483 (2.6%) | - |
| 2nd/3rd/Adv AV block | - | - | 400 (2.2%) | - |
| Ventricular tachycardia/Ventricular fibrillation (VT/VF) | - | - | 99 (0.5%) | - |
| Acute myocardial infarction (Acute MI) | - | - | - | 209 (1.0%) |
| Old myocardial infarction (Old MI) | - | - | - | 5,310 (24.4%) |
| ST/T changes (STTC) | - | - | - | 5,235 (24.0%) |
| Conduction disturbances (CD) | - | - | - | 4,898 (22.5%) |
| Hypertrophy (HYP) | - | - | - | 2,649 (12.2%) |
| Premature atrial complex (PAC) | - | - | - | 398 (1.8%) |
| Atrial fibrillation / Atrial Flutter (AFIB/AFL) | - | - | - | 1,570 (7.2%) |
| **Total Exams** | 341,794 | 100,450 | 18,517 | 21,799 |

The CODE15 and PTB-XL datasets consist of unidimensional time-series data, which were converted into image format using the ECG-Image-Kit library [6,7] with default parameters.

The InCor-AMB dataset [8,9], obtained from an ambulatory setting, includes ECG image exams collected from the Heart Institute (InCor). The dataset was filtered to include only patients over 18 years of age with associated medical reports. For this study, we selected six specific classes (1dAVb, RBBB, LBBB, SB, ST, and AF) from the original 56, to align with the CODE15 dataset.

The InCor-EMG dataset comprises 18,517 ECG image exams collected from InCor's emergency department. These exams initially lacked associated reports. A multi-step annotation process was conducted to generate the reports, involving 19 cardiologists who reviewed the data in pairs, with a third cardiologist consulted in cases of disagreement.

## 2.2. Model

To classify the ECG images, we employed the ConvNeXt convolutional neural network [4] as the backbone of our method, utilizing a pre-training approach. In deep learning-based computer vision tasks, it is common to use pre-trained models, typically initialized with ImageNet weights, and then fine-tune them for specific tasks. However, using ImageNet for pre-training in ECG image classification introduces a significant semantic gap, as ImageNet images differ vastly from ECG data. To address this, we first performed transfer learning using the CODE15 and InCor-AMB datasets. This step allowed the model to adapt its learned features to those specific to ECG images, enhancing its ability to recognize relevant patterns associated with different cardiac conditions. After the transfer learning phase, we further fine-tuned the model using the InCor-EMG dataset. In the final stage, we fine-tuned the model on the challenge's training dataset, PTB-XL, employing 5-fold cross-validation to ensure robust performance, resulting in an ensemble of the five models.

## 2.3. Pre-processing

The ECG images were resized to 224 x 224 pixels during the pre-training steps and to 384 x 384 during the fine-tuning of the PTB-XL dataset. The images, initially in RGB format, were converted to grayscale.

## 2.4. Training Strategy

Table 2 describes the training strategy adopted in our work. We also employed several data augmentation strategies, including *RandomBrightnessContrast*, *GaussNoise*, *ShiftScaleRotate*, *ElasticTransform*, and *GridDistortion*, using the *albumentations* library to perform. The parameters for each augmentation were set heuristically.

Table 2: Training hyperparameters.

| **Hyperparameter** | **Value** |
|---|---|
| Optimizer | AdamW |
| Loss Function | BCE |
| Batch size | 16 |
| Epochs | 30 |
| Learning rate | Smith et al [12] |
| Early stop | 7 epochs |

Given the highly imbalanced distribution of classes in the PTB-XL dataset, as described in Table 1, we employed a weighted version of the binary cross-entropy (BCE) loss. The weights were calculated as the total number of exams divided by the total number of exams in each class, assigning higher weights to less-represented classes. Specifically, we used weights of 2.3, 104.3, 4.1, 4.2, 4.4, 8.2, 54.8, 19.1, 13.9, 25.1, and 34.2 for the classes NORM, Acute MI, Old MI, STTC, CD, HYP, PAC, PVC, AFIB/AFL, TACHY, and BRADY, respectively. The learning rate was optimized using the algorithm proposed by Smith et al [12].

## 2.5. Computational environment

We built our model using *PyTorch*, integrated with the *Lightning* framework, and leveraged *scikit-learn* for additional utilities. The training was conducted on a machine powered by an NVIDIA RTX 3080 GPU with 10GB of VRAM. To enhance efficiency and reduce memory consumption, we used mixed precision training.

# 3. Results

Table 3 presents the F1-scores for all classes obtained from our internal 5-fold validation. Additionally, the performance of our team - AIMED - on the hidden data from the challenge (classification task) is shown in Table 4

Table 3: F1-score performance results from the internal 5-fold cross-validation for each class.

| | **F1 Score** |
|---|---|
| AFIB/AFL | $0.915 \pm 0.012$ |
| NORM | $0.862 \pm 0.005$ |
| PVC | $0.837 \pm 0.006$ |
| TACHY | $0.803 \pm 0.026$ |
| CD | $0.726 \pm 0.036$ |
| STTC | $0.713 \pm 0.011$ |
| Old MI | $0.634 \pm 0.049$ |
| PAC | $0.610 \pm 0.058$ |
| HYP | $0.609 \pm 0.027$ |
| BRADY | $0.565 \pm 0.117$ |
| Acute MI | $0.409 \pm 0.052$ |
| **Macro Average** | $0.698 \pm 0.022$ |

Table 4: F-measure of our team's model on the hidden data for the classification task.

| **Task** | **Score** | **Rank** |
|---|---|---|
| Classification | F-measure: 0.817 | 1/16 |

Additionally, our model was evaluated using ECG exams in different scenarios: mobile phone images, scans of printouts of ECG exams on clean paper, stained paper, deteriorated paper, and computer screenshots [3]. Table 5 summarizes the model's performance in each scenario.

Table 5: Macro F1-score results for different real-world scenarios.

| **Scenario** | **F1 Score** |
|---|---|
| Color scans of clean papers | 0.833 |
| BW scans of clean papers | 0.796 |
| Mobile photos of clean papers | 0.644 |
| Mobile photos of stained papers | 0.636 |
| Mobile photos of deteriorated papers | 0.711 |
| Color scans of deteriorated papers | 0.830 |
| BW scans of deteriorated papers | 0.775 |
| Screenshots of computer monitor | 0.636 |
| **Average** | 0.733 |

# 4. Discussion

In this study, we introduce a method for classifying ECG images based on a ConvNeXt network. We began with a

ConvNext network pre-trained on the ImageNet database and applied transfer learning on ECG images from various datasets to bridge the gap between these images and our target dataset, PTB-XL. This step allowed the model to better adapt to ECG data. Finally, we performed fine-tuning on the PTB-XL dataset for optimal classification performance.

Table 3 shows the results of the internal training using PTB-XL with 5-fold cross-validation. Some classes showed strong performance in terms of F1-score, such as AFIB/AFL and NORM, with average F1-scores of 0.915 and 0.862, respectively. However, other classes like Acute MI and BRADY exhibited lower performance, with F1-scores of 0.409 and 0.565, respectively. The macro F1-score obtained in our internal validation was $0.698 \pm 0.022$.

When evaluating the hidden test dataset of the challenge (Table 4), we used an ensemble of the five models obtained from the 5-fold cross-validation, achieving a macro F1-score of 0.817, surpassing our internal validation result. This improved performance is likely due to differences between our internal validation set and the challenge's hidden test dataset. The hidden dataset likely contains examples that our model ensemble was particularly adept at classifying, resulting in a higher score.

When our model is evaluated in various real-world scenarios, such as real scans and photos of ECG exams, it continues to demonstrate strong performance (Table 5). However, there is a notable drop in performance, particularly with mobile phone photos and screenshots of computer monitors, which warrants further investigation.

## 5. Conclusion

Our study highlights the effectiveness of deep learning techniques for automated ECG image classification, offering a promising tool for enhancing diagnostic accuracy in clinical practice. Our pre-training strategy on diverse ECG datasets and fine-tuning on the PTB-XL database yielded a macro F1-score of 0.817 on the hidden test set. This performance earned us first place in the classification task of the competition, underscoring the strength of our approach.

## Acknowledgments

This research was supported by Foxconn Brazil and Zerbini Foundation as part of the research initiative titled "Machine Learning in Cardiovascular Medicine".

## References

[1] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[2] Reyna MA, Deepanshi, Weigle J, Koscova Z, Elola A, Seyedi S, et al. Digitization and Classification of ECG Images: The George B. Moody PhysioNet Challenge 2024. Computing in Cardiology 2024;51:1–4.

[3] Reyna MA, Deepanshi, Weigle J, Koscova Z, Campbell K, Shivashankara KK, et al. ECG-Image-Database: A dataset of ECG images with real-world imaging and scanning artifacts; a foundation for computerized ECG image digitization and analysis, 2024. URL `https://arxiv.org/abs/2409.16612`.

[4] Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022; 11976–11986.

[5] Lima EM, Ribeiro AH, Paixão GM, Ribeiro MH, Pinto-Filho MM, Gomes PR, et al. Deep neural network-estimated electrocardiographic age as a mortality predictor. Nature communications 2021;12(1):5117.

[6] Shivashankara KK, Deepanshi, Shervedani AM, Reyna MA, Clifford GD, Sameni R. ECG-Image-Kit: a synthetic image generation toolbox to facilitate deep learning-based electrocardiogram digitization. Physiological Measurement 2024;45:055019.

[7] Deepanshi, Shivashankara KK, Clifford GD, Reyna MA, Sameni R. ECG-Image-Kit: A Toolkit for Synthesis, Analysis, and Digitization of Electrocardiogram Images, January 2024. Online at: `https://github.com/alphanumericslab/ecg-image-kit`.

[8] Dias FM, Samesima N, Ribeiro A, Moreno RA, Pastore CA, Krieger JE, et al. 2d image-based atrial fibrillation classification. In 2021 Computing in Cardiology (CinC), volume 48. IEEE, 2021; 1–4.

[9] Dias FM, Ribeiro E, Moreno RA, Ribeiro AH, Samesima N, Pastore CA, et al. Artificial intelligence-driven screening system for rapid image-based classification of 12-lead ecg exams: A promising solution for emergency room prioritization. Ieee Access 2023;.

[10] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a large publicly available electrocardiography dataset. Scientific Data 2020;7:154.

[11] Strodthoff N, Mehari T, Nagel C, Aston PJ, Sundar A, Graff C, et al. PTB-XL+, a comprehensive electrocardiographic feature dataset. Scientific Data 2023;10:279.

[12] Smith LN. Cyclical learning rates for training neural networks. In 2017 IEEE winter conference on applications of computer vision (WACV). IEEE, 2017; 464–472.

Address for correspondence:

Felipe Meneguitti Dias
Heart Institute (InCor) - Av. Dr. Enéas Carvalho de Aguiar, 44 - Cerqueira César, São Paulo - SP, Brazil
f.dias@hc.fm.usp.br