

Applying Pre-Trained Deep Learning Models for Multi-Label Classification of Realistic and Noisy Electrocardiogram Images

Max A Kowalchuk¹, Swati Gupta², Navchetan Awasthi^{1,3}

¹ University of Amsterdam, Amsterdam, The Netherlands

² VIB KU Leuven Center for Cancer Biology, Belgium

³ Amsterdam UMC, Amsterdam, The Netherlands

Abstract

ECG is an indispensable tool for diagnosing cardiovascular diseases because it can quickly and non-invasively monitor heart rhythms. Recent research has demonstrated that innovative AI approaches have great potential to advance ECG interpretation. Despite these advances, widespread application of such techniques is hindered by several factors such as digital copies of ECGs remain unavailable in many resource-limited areas. Additionally, ECGs scanned from paper contain noise such as shadows or creases, which are often not accounted for in digital approaches. Furthermore, contemporary research is concentrated on detecting single conditions or diseases, despite the possibility of their co-occurrence. To address these shortcomings, this work investigates the application of pretrained CNNs by fine-tuning popular models such as AlexNet and ResNet separately for multi-label classification of realistic and noisy ECG images. The results underscore the potential for the improvement of image-based models and suggest that future research should target this approach to provide a scalable method for interpreting ECG images. Our Team name is ECG_UVA and got the classification F-measure of 0.516, achieving a rank of 5/16 on the leaderboard on the hidden dataset in the official phase of the 'Digitization and Classification of ECG Images: George B. Moody PhysioNet Challenge 2024'.

1. Introduction

Heart disease is one of the world's leading cause of death. Detection of heart rhythm abnormalities is critical in efforts to treat heart disease. The electrocardiogram (ECG) has proven to be an indispensable pre-screening tool for cardiovascular diseases (CVDs) [1]. ECGs are widely used to observe heart behavior, mainly due to their non-invasive nature, reliability, and minimal risks. With the advancement of technology and artificial intelligence (AI) techniques, much innovation has been made in the

medical field, including the automatic interpretation and classification of ECG waveforms. In some studies, deep learning methods even rival or surpass individual cardiologist performance when diagnosing ECG signals [2]. These advancements are both impressive and promising; however, they are not without exclusion or constraint. Firstly, the majority of research published on the topic regards digital ECG representations, referring to raw signal data. Digital ECG representations are not yet used universally. In many areas, printed-out paper ECGs or digital images remain the standard, especially in the global south. In addition, remote regions, where digital ECGs are often not available, are more likely to suffer from a scarcity of expert or specialist medical opinion [3]. As such, direct image-based methods of classification could best fill the needs of many areas [4]. Secondly, contemporary research is concentrated on diagnosing single cardiovascular conditions, while multiple may exist in a single ECG. A few studies such as [5] demonstrate that models that are able to predict more than one label can rival models that predict only a single label. Digitizing and interpreting ECGs is challenging both in terms of hardware and computation, demanding powerful resources and therefore further hindering its global accessibility. Hence, *there is a need for low-cost multi-label classification of realistic paper ECG image representations*, which is the focus of this work.

2. Related research

The field of automatic ECG interpretation is rapidly developing. The majority of research in this area is on some form of CNN, but recent advances have shifted focus to the utilization of hybrid-CNN methods and other models including LSTMs (Long Short Term Memory).

The work done by [2] is one such study that achieves impressive performance with a deep 1D CNN. Utilizing a large, novel dataset of raw ECG signals, annotated by at least 6 cardiologists, it achieves an F-score of 0.837 on the classification of 12 rhythm classes. In the domain of two-dimensional approaches, [6] emphasizes the viability

of image-based deep CNNs in clinical ECG diagnosis. The study demonstrates through grad cams that a pre-trained CNN, fine-tuned to ECG images, is capable of detecting meaningful features and patterns that a cardiologist might interpret similarly. By employing numerous thresholding techniques on input data, their ImageNet pre-trained VGG16 architecture achieves a near-perfect AUPRC when classifying some individual pathologies.

ECGs in practice are often either printed or scanned as images, due to habit or the inaccessibility of resources. This limits the utility of signal-based models. Therefore, this work focuses on a 2D image-based approach. While less common than 1D methods, 2D techniques have seen varied research. The task is to classify ECGs into 1 or more of 11 different categories, one of which is the "normal" class.

3. Materials and Methods

3.1. Dataset Preprocessing

This work utilizes the PTB-XL annotated dataset [7], ECG-Image-Kit: A Synthetic Image Generation Toolbox to Facilitate Deep Learning-Based Electrocardiogram Digitization[8]. The data was constructed from the large, publicly available PTB-XL electrocardiography dataset [7]. This dataset contains 21799 clinical 12-lead ECG signals, annotated by at least one and sometimes two cardiologists.

From the PTB-XL time-series data, synthetic images can be generated using ECG-Image-Kit [8]. Images generated from signal data originally have a size of 2200×1700 pixels. This volume of data [9] can be difficult for a model to process and can drive up the complexity, especially for heavier models [10]. Thus, the images used in this research are always cropped to remove the top and subsequently resized. The image sizes used vary across different models, as lighter models can accommodate larger images. In order to simultaneously remedy the class imbalance in the PTB-XL dataset, ECGs with rarer conditions are separated and more synthetic images are generated specifically for these uncommon labels. Several image sizes were tested to find a balance between training time/cost and image detail, eventually ending on a shape of 1065 by 560 pixels.

3.2. Model architecture and training

CNNs are the most commonly used Deep Neural Network approach [11], particularly in tasks performed on visual data, such as classification, segmentation, or object detection. A CNN is constructed of several different layers, which the input data passes through sequentially. The outputs of the final layer form the model's final decision(s). For the task of classification, this final output comes in the form of a list of probabilities for each class being present in

the image. From these probabilities, a final classification can be inferred.

In this work, multiple models and training variations were utilized and tested. The Adam optimizer was employed for all model training [12]. Unless mentioned otherwise, a starting learning rate of 0.0001 and a batch size of 40 were applied, training for 60 epochs with early stopping. With a patience of 10, the model stops training if the validation loss has not improved in the last 10 epochs. The model weights with the best performance on the validation set are then saved. All training is performed on GPUs provided by the Snellius supercomputer system.

The final model has a VGG16 architecture introduced by [13] with ImageNet pre-trained weights and some slight adjustments. The final fully connected layer of the model is replaced with a dropout layer and a classification layer with 11 outputs, one for each class. Because this is a multi-label problem, the classification layer is succeeded by a sigmoid activation [14], rather than the usual softmax. This ensures every class is evaluated separately, and multiple labels can be obtained simultaneously. If the final activation were a softmax, all label probabilities would sum to 1.0, meaning only one class is finally predicted. Another model using Resnet-50 architecture from [15] is also trained and evaluated for this task because it is a deeper model, but less computationally demanding. Similar modifying steps are taken for this model, such as pre-training on ImageNet [16], replacing the last classification layer, and assigning a final sigmoid activation function.

The binary cross-entropy loss function was used during training [17]. The model must assign a probability between 0 and 1 to every class for every data sample. This essentially transforms the problem into one of 11 binary classifications. For the same reason, a sigmoid activation function is employed, rather than a softmax, which would be more applicable to a multiclass problem where each sample has only one label.

The aforementioned models are very proficient on their own. However, they are significantly heavier than the simple CNN used as the baseline. ResNet50 has approximately 25.6 M trainable parameters and VGG16 has around 138 M. To decrease training time and computational load, improve performance, and enhance generalization, transfer learning is employed in this work.

The pre-trained models are trained on ImageNet images that have dimensions of 224×224 pixels. Consequently, when preprocessing ECG data for transfer learning, it is resized to match this same 224×224 shape. Several combinations of frozen layers were tested: specifically freezing no layers, freezing all extraction/convolutional layers, and freezing most extraction/convolutional layers. The latter configuration is used for the remainder of the research, though more experimentation is possible in this area. Un-

freezing some feature layers is likely effective for this case due to there being enough training data to learn some more specific features during extraction.

3.3. Evaluation measurements

This work employs a range of evaluation metrics, including (macro) F-measure, Precision, Recall, Perfect Accuracy, Capture Accuracy, Hamming Accuracy, and Specificity [18]. The most comprehensive metric and the one used by Physionet is the F-measure. Specifically, the macro-F-measure is used in the case of multi-label data. For the multi-label case, the individual F-measures of all classes are averaged. The challenge opts for the macro-averaging method, where all the classes are viewed as equally important, regardless of appearance frequency.

4. Results and Discussion

The model that achieved the highest performance is the VGG16-based model, which proved capable of recognizing patterns in new, unseen data. It managed to attribute multiple classes to ECGs that exhibit more than one condition, whilst achieving an overall macro-F1-score of 0.61 on the validation dataset. Of the held-out subset of the training set, 51% of ECG images were classified perfectly, meaning all 11 conditions (including normal) were correctly predicted to be present or absent. In terms of catching present conditions, the model achieves a capture accuracy of 63%, meaning that in 63% of ECGs, all present conditions were correctly identified. In addition, the hamming accuracy was 92%. This is the accuracy per label on all labels in the test set. Note that this is expected to be very high, as the correctly predicted negative/absent labels also contribute to the majority of this metric.

The model further reaches a precision of 0.653 and a recall/sensitivity of 0.593, meaning the model is slightly on the conservative side when it comes to predicting positive instances. Seeing as the recall is the proportion of positive occurrences caught by the model, one would expect it to be higher than the capture accuracy of 63%, because the catching accuracy requires all present labels to be caught. The reason this is not the case is that 0.593 is the macro-recall also used to calculate the macro-F-measure, hence the class imbalance is weighing it down. The micro recall, precision, and F-measure are 0.708, 0.728, and 0.717 respectively. Lastly, the model has a specificity of 0.951. This is the proportion of correct negative predictions in relation to the total negative cases. Similar to the hamming accuracy, this is high because individual classes are absent most of the time. As is expected with higher precision, the model makes very few false positive errors compared to false negatives. The submitted model received an F-measure of 0.516 on the Physionet validation data. Table

1 shows an overview of model measurements. The baseline model from The George B. Moody PhysioNet Challenge 2024 [1, 19, 20] achieves an F1-score of 0.323. The F-measure of 0.613 dropped to 0.516 when tested by Physionet (Table 2). This demonstrates that, although model performance declines on new data, the model can still adapt with a degree of success to real and unpredictable data.

4.1. Future work and improvements

There is a clear correlation between class frequency and performance, highlighting the remaining negative effects of class imbalance in the training process. Despite the additional image generation for rarer classes, the data remains unbalanced. There are several ways this could be improved with more time and attention spent on the challenge. Firstly, more data could be generated to balance out the dataset. Generating augmented data with Image Kit is a lengthy and expensive process. Given more time and resources, images could be properly generated in proportion to their class frequencies. Performance on perfectly evenly represented data could be worth testing. Perfectly even data for this challenge might not be desirable though, as the normal class being equally common as Acute MI for example is not very indicative of realistic applications. Regardless, this work has shown that making the dataset more equitable results in model improvement.

Other methods of handling class imbalance that are worth exploring further in this task are using specific class weights or further investigating Focal loss. During this research, class weights were not utilized. Because the test set the challenge uses is not public, it is unclear whether class frequencies, which are used to determine these weights, are similar across datasets. Class weights are still usable in the setting of this challenge, as they could be coded to adapt to the properties of the Physionet training set during their evaluation.

5. Conclusions

This work evaluates the various pre-trained deep learning methods for classifying realistic multi-label ECG images. Specifically, it focused on techniques utilizing pre-trained convolutional neural networks, although there exist valid and promising strategies outside of this architecture. This work has found that CNNs are capable of identifying multiple simultaneous abnormalities in distorted, realistic images of electrocardiograms. The best model trained in this work achieved an F-measure of 0.613 on our private set. Further analysis indicates that this score could be significantly higher with some improvements such as class imbalance. In spite of its limitations, there are a range of reasons for people to specifically require this form of tech-

Task : Classification	F-measure*	Rank	Precision*	Recall*	Perfect Acc	Capture Acc	Hamming Acc	Specificity
VGG16	0.613		0.653	0.593	0.509	0.627	0.924	0.951
Resnet-50	0.562		0.571	0.564	0.461	0.610	0.915	0.943
Resnet-50-Focal-Loss	0.545		0.607	0.514	0.466	0.580	0.915	0.948
Challenge Baseline	0.323							

Table 1. Model scores on various evaluation metrics. Perfect accuracy refers to the correct classification of every label for an ECG while catching accuracy refers to correctly labeling all present conditions, and hamming accuracy is the accuracy on all labels combined. For Focal loss $\alpha = 0.5$ and $\gamma = 2.0$ were used. * In metrics where averaging is necessary for multi-label assessment, macro-averaging is applied.

Task	Score	Rank
Classification	0.516	5/16

Table 2. Model F-measure score using the VGG16 model on th physionet validation dataset.

nology, such as paper ECG tradition, low budgets, and/or limited access to specialists.

References

- [1] Reyna MA, Deepanshi, Weigle J, Koscova Z, Elola A, Seyedi S, Campbell K, Clifford GD, Sameni R. Digitization and Classification of ECG Images: The George B. Moody PhysioNet Challenge 2024. *Computing in Cardiology* 2024;51:1–4.
- [2] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, Ng AY. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine* 2019; 25(1):65–69.
- [3] Schopfer DW. Rural health disparities in chronic heart disease. *Preventive Medicine* 2021;152:106782. ISSN 0091-7435. Behavior Change, Health, and Health Disparities 2021: Rural Addiction and Health.
- [4] Gupta U, Paluru N, Nankani D, Kulkarni K, Awasthi N. A comprehensive review on efficient artificial intelligence models for classification of abnormal cardiac rhythms using electrocardiograms. *Heliyon* Feb 2024;10(5).
- [5] Rawi AA, Elbashir MK, Ahmed AM. Deep learning models for multilabel ecg abnormalities classification: A comparative study using tpe optimization. *Journal of Intelligent Systems* 2023;32(1):20230002.
- [6] Ao R, He G. Image based deep learning in 12-lead ecg diagnosis. *Frontiers in Artificial Intelligence* 2023;5:1087370.
- [7] Wagner P, Strodthoff N, Bousseljot RD, Samek W, Schaeffer T. Ptb-xl, a large publicly available electrocardiography dataset, Nov 2022.
- [8] Shivashankara KK, Deepanshi, Mehri Shervedani A, Clifford GD, Reyna MA, Sameni R. Ecg-image-kit: a synthetic image generation toolbox to facilitate deep learning-based electrocardiogram digitization. *Physiological Measurement* 2024;.
- [9] Reyna MA, Deepanshi, Weigle J, Koscova Z, Campbell K, Shivashankara KK, Saghaifi S, Nikookar S, Motie-Shirazi M, Kiarashi Y, Seyedi S, 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. *arXiv preprint arXiv240916612* 2024;.

- [10] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks 2012;25.
- [11] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JA, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. *Medical Image Analysis* 2017;42:60–88. ISSN 1361-8415.
- [12] Ruder S. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv160904747* 2016;.
- [13] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv14091556* 2014;.
- [14] Zhang ML, Zhou ZH. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 2014;26(8):1819–1837.
- [15] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *CoRR* 2015;abs/1512.03385.
- [16] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009; 248–255.
- [17] Shannon CE. A mathematical theory of communication. *The Bell system technical journal* 1948;27(3):379–423.
- [18] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Information Processing Management* 2009;45(4):427–437. ISSN 0306-4573.
- [19] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation* 2000;101(23):e215–e220.
- [20] Deepanshi Shivashankara KK CGRMSR. Ecg-image-kit: A toolkit for synthesis, analysis, and digitization of electrocardiogram images 2024;.

Address for correspondence:

Navchetan Awasthi

Faculty of Science, Mathematics and Computer Science, Informatics Institute, University of Amsterdam, Amsterdam, 1090 GH, The Netherlands
n.awasthi@uva.nl