Digital Signal and Image-based ECG Classification and Its Performance by Modern Residual Convolutional Networks

Ľubomír Antoni[2], Erik Bruoth[2], Peter Bugata[1], Peter Bugata Jr[1], Dávid Gajdoš[1], Dávid Hudák[1], Vladimíra Kmečová[1], Manohar Gowdru Shridhara[2], Monika Staňková[1], Gabriela Vozáriková[2], Ivan Žežula[2]

[1] VSL Software, a.s., Košice, Slovakia
[2] Pavol Jozef Šafárik University in Košice, Faculty of Science, Košice, Slovakia

Abstract

In The George B. Moody PhysioNet Challenge 2024, we developed a method for classifying electrocardiograms (ECGs) captured from images or paper printouts. Our prediction model consists of two neural networks. The first network predicts the image's rotation angle for its reverse rotation. Thus, the second network processes the modified image and predicts 11 possible labels for the ECG. In both cases, we applied the modern residual convolutional network ConvNeXt. We compared the performance of the image-based prediction model with that of a model for digital signal classification. As a reference model, we used the 1D variant ResNet-50. By evaluating the image processing model on our training set using 5-fold cross-validation, we obtained a macro F1 score of 0.8496 for multilabel classification. In the competition's official round, our CeZIS team achieved a Challenge score of 0.645 on the hidden test set (ranked 4th).

1. Introduction

The George B. Moody PhysioNet Challenge 2024 (Challenge) [1, 2] focused on developing algorithms for digitizing and classifying electrocardiograms (ECGs) captured from images or paper printouts.

For the participants of the Challenge, a synthetic ECG image generator [3] was provided, which, based on the digital ECG signal, generates ECG images with various distortions that resemble artifacts on paper ECGs. Due to the availability of several large digital ECG databases, we used the generator to create a large training set of synthetic ECG images. On this set, we trained a classification model that predicted the presence of 11 labels corresponding to diagnostic classes. At the same time, we compared the performance of the created model with a model directly processing a digital ECG signal on the same set of ECGs with the same labels.

2. Methods

This chapter describes the training data, labels, and classification models used.

2.1. Training Data

As training data, we used several freely available databases containing short diagnostic 12-lead ECGs. We consider the use of multiple databases important, as individual databases differ in the ECG devices used, post-processing, and digitization methods and may also reflect geographical differences. We selected four databases from the training data for the PhysioNet Challenge 2021 [4], to which we added databases from the Hefei Cup [5] and Shandong Provincial Hospital (SPH) [6].

Individual ECGs were captured as a digital signal with a frequency of 500 Hz. Diagnostic codes were assigned to each ECG as part of the metadata, the scope and semantics of which differed from database to database [7]. We selected only those ECGs assigned an original diagnostic code related to one of the eleven diagnostic classes used in the multilabel classification in the Challenge. Table 1 lists the source databases with the number of all and the number of used ECGs.

Table 1. Source datasets for the training set.

Dataset	# used	# all
PTB-XL	19 039	21 799
SPH	21 133	25 770
Hefei Cup	35 254	44 142
Ningbo	32 300	34 905
Chapman/CUSPH	10 063	10 646
CPSC2018	5 949	6 877
Total	**123 738**	144 139

To obtain images for training, we used the synthetic

ECG image generator supplied as part of the Challenge. We generated 20 images for each ECG with random paper rotations and various artifacts in different formats (10, 5, and 5 images in a 4-, 2-, and 1-column design, respectively). Due to the large amount of data (the training set had more than 2.4 million images), we resized the generated images so that their width was 600 px (keeping the aspect ratio).

2.2. Labels

In the Challenge's classification task, 11 labels corresponding to diagnostic classes were defined. The exact definition of individual labels was not given; only a program was provided that mapped the SCP codes used in the PTB-XL database to individual labels.

We modified the provided mapping algorithm using multi-valued logic. In PTB-XL, the probability of occurrence of the given diagnosis was given for the diagnostic SCP codes. If there were SCP codes corresponding to the label on the ECG, but all of them had a stated probability of less than or equal to 50%, we set the label value to 0.5. If any of the corresponding SCP codes had a stated probability greater than 50% or was without a stated probability, we kept the value 1. The most cases with a value of 0.5 were for the labels Old MI, HYP, and STTC on the ECG from the PTB-XL database (in other the databases did not list probabilities of diagnoses, and values of 0.5 did not occur).

In addition to the values 1, 0.5, and 0, we also used a special value NA for the labels in cases where:

- In the source database, the relevant diagnoses were not evaluated at all (e.g., myocardial infarctions).
- Collision original codes were indicated on the ECG, e.g., the ECG was labeled as normal and had a serious diagnosis code.
- The ECG had no code assigned to the diagnostic class, but there was a possible diagnosis symptom code, e.g., the presence of code for "Prolonged PR interval" for CD class (attempt to reduce underlabeling)

We extended the mapping algorithm to other databases as well. Table 2 shows the frequencies of label values in our training set.

During training the models, we did not take the NA values into account when calculating the loss function (we applied the value 0.5 similarly to using Mixup). When evaluating the metric on the validation or test set, we used only label values 0 and 1.

2.3. Model for ECG images

Our prediction model for ECG images consists of two neural networks. The first network predicts the image's rotation angle for its reverse rotation. Subsequently, the second neural network processes the modified image and predicts the probabilities of individual labels.

Table 2. Frequencies of label values in the training set.

Label	1	0.5	0	NA
NORM	30 819	209	85 036	7 674
Acute MI	179	17	85 907	37 635
Old MI	2 892	2 101	79 810	38 935
STTC	36 859	484	80 928	5 467
CD	17 370	55	102 864	3 449
HYP	13 214	1 084	103 388	6 052
PAC	4 307	2	119 431	0
PVC	5 101	0	118 637	0
AFIB/AFL	13 663	0	104 891	5 184
TACHY	15 781	2	101 932	6 023
BRADY	26 363	0	91 357	6 018

In both cases, we used the modern residual convolutional network ConvNeXt [8], whose architecture was optimized based on experience from visual transformers. One of the differences compared to classic convolutional networks is the use of larger kernels (7×7 versus 3×3). Thanks to the greatly expanded receptive field of the neurons, we were able to use the given network without changing the architecture even for images with a higher resolution of 600×512 compared to the commonly used lower resolution, e.g., 224×224.

We trained the neural network for rotation prediction separately on the training set with ECG images randomly rotated between -15 and +15 degrees. After training for 3 epochs, we achieved a mean absolute error of about 0.08 degrees on the test fold. The network accurately identified the rotation angle, and the error rarely exceeded 1 degree.

The neural network for multilabel classification was also trained separately. Before entering the network, we rotated the image based on the rotation angle recorded when the image was generated. However, we added a random error in the range of -2 to 2 degrees to the angle to simulate possible errors of the first network.

Another difference between the ConvNeXt network and the classic ResNet [9] is the different number of virtual channels in network layers. In the ConvNeXt Tiny variant we used, there were only 768 channels in the final layer, while up to 2 048 channels in the comparable ResNet-50. On the other hand, the final two-dimensional map was larger due to the higher resolution of the input image, namely 18×16, instead of the standard 7×7. Considering that some anomalies appear only in some places of the ECG and that sometimes it is useful to compare a local value (e.g., the length of the RR interval) with the global average, we doubled the number of final latent factors by

applying not only GlobalAveragePooling but also GlobalMaxPooling layer to the final feature map (Figure 1).

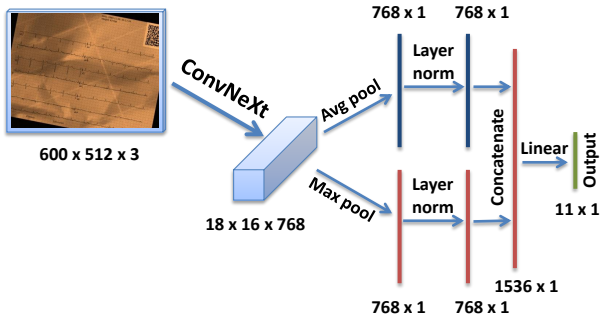

Figure 1. Architecture of the image processing network.

We initialized the network with weights pre-trained on ImageNet (available directly in the PyTorch ConvNeXt implementation). We trained the first epoch with frozen weights except for the final classification layer. After that, the macro F1 reached less than 0.23, which means that the latent factors pre-trained on ImageNet alone are not sufficient for ECG classification.

Subsequently, we trained four epochs (i.e., 80 images were processed for each ECG) with a One-Cycle learning rate [10] schedule (max LR $= 10^{-4}$). We used Random Erasing for data augmentation, Weight Decay, Stochastic Depth, and Label Smoothing for regularization.

We trained with the AdamW optimizer and Weighted Cross Entropy loss function, setting the weights of the positive class for individual labels so that Recall was slightly higher than Precission (due to the assumption of underlabeling).

2.4. Model for digital signals

For digital signal processing, we used a 1D variant of the ResNet-50 network with a width of ¼ , i.e., the number of virtual channels was reduced to ¼ , and there were 512 latent factors in the final layer. The network architecture and training process were similar to our PhysioNet Challenge 2021 solution [7].

Signal segments with 4 608 time steps and 12 channels corresponding to individual leads were the input to the network. The network was initialized with random weights and trained for 80 epochs with a One-Cycle learning rate schedule (max LR $= 0.01$). During training, the same optimizer and Weighted Cross Entropy loss function with the same positive class weights as in the image processing model were used.

As data augmentation, random fixed-length slices, lead rescaling, lead dropout, and periodic slice cutout were utilized. Weight Decay and Label Smoothing regularization techniques were also applied.

3. Results

Both models were evaluated via 5-fold CV on the large training set described in Chapter 2.1. The assignment of ECGs to folds was stratified by source database and label values. We used the macro F1 score as a metric to evaluate the multilabel classification analogously to the Challenge. As seen from Table 3, the performance of models was surprisingly similar.

Table 3. Performance comparison of the models.

Model	Macro F1	mean	std
Image processing model		0.8496	0.0020
Signal processing model		0.8518	0.0050

We also analyzed the performance of the models through F1 scores separately for each label (Figure 2). While the performance of both models is very similar for all labels, the given models differ significantly in the nature of the input data, the number of parameters, and the computational complexity, as shown in Table 4.

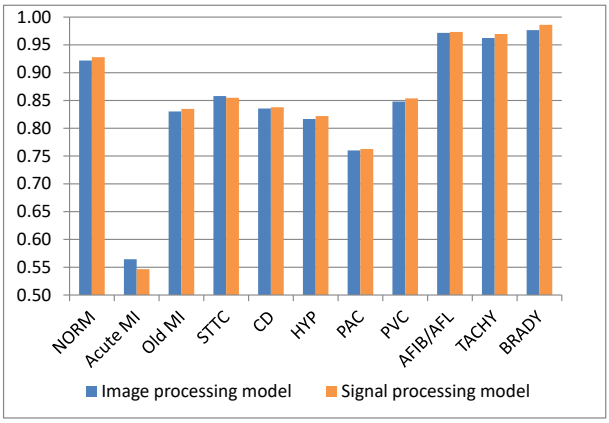

Figure 2. Comparison of the performance of the models by individual labels using F1 score.

Table 4. Characteristics of the models.

Model input	Input shape	Params	Training
Image	$600 \times 512 \times 3$	27.8 mil	46.3 h
Signal	$4\,608 \times 12$	1.2 mil	2.7 h

During the official round of the Challenge, the image processing model was evaluated on a hidden test set, with additional hidden image sets used in the final evaluation. The results are shown in Table 5.

Table 5. F-measure of our team's model on the hidden data for the classification task.

Task	Score	Rank
Digitization	-	-
Classification	F-measure: 0.645	4/16

4. Discussion and Conclusions

While the signal processing model receives a fully digital signal on all leads at a high sampling frequency of 500 Hz, the second model processes an image with a lower resolution (approx. 5 000 time steps in the signal vs. 600 pixels for the image width). In addition, images contain various distortions and often provide only a partial view of some leads. Nevertheless, the image processing model demonstrated comparable performance, which testifies to the high tuning of 2D convolutional networks for image processing. However, these images were synthetic and in real practice there is greater variability in ECG imaging.

The signal processing model has more details of the ECG signal available but is also more prone to overfitting. Our experiment shows possible reserves in data augmentation methods and the architecture and training of neural networks processing digital physiological signals. On the other hand, these models have better starting points for further improving and surpassing the performance of image processing models even with fewer parameters and lower computational complexity.

The created models show large differences in prediction accuracy for individual labels. Rhythm labels such as AFIB/AFL, BRADY, and TACHY have the best performance. The Acute MI label was the worst, probably due to very few positive examples in the training set. The performance of some other labels, especially PAC, was not quite satisfactory.

The Challenge score achieved on the hidden test set is significantly lower than the results on the training set. The reason may be differences in the generation and subsequent processing of images, significantly worse results were mainly for black-and-white scans. We consider the main cause to be the different semantics of the diagnostic classes on our training set and on the test data. Clarification of the label semantics would help a more objective comparison of individual Challenge solutions.

Acknowledgments

This work was supported by ERDF EU grant and by the Slovak Ministry of Economy under the contract No. ITMS313012S703 and by Erasmus+ grant no. 2023-1-PL01-KA220-HED-000166765.

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 2024;51:1–4.

[2] Reyna MA, Deepanshi, Weigle J, Koscova Z, Campbell K, Shivashankara KK, Saghafi S, Nikookar S, Motie-Shirazi M, Kiarashi Y, Seyedi S, Clifford GD, Sameni R. 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.

[3] Shivashankara KK, Deepanshi, Shervedani AM, Clifford GD, Reyna MA, Sameni R. ECG-Image-Kit: A Synthetic Image Generation Toolbox to Facilitate Deep Learning-Based Electrocardiogram Digitization. Physiological Measurement may 2024;45(5):055019. URL https://dx.doi.org/10.1088/1361-6579/ad4954.

[4] Reyna MA, Sadr N, Perez Alday EA, Gu A, Shah A, Robichaux C, Rad AB, Elola A, Seyedi S, Ansari S, Ghanbari H, Li Q, Sharma A, Clifford GD. Will Two Do? Varying Dimensions in Electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021. Computing in Cardiology 2021 2021;48:1–4.

[5] Hefei Hi-tech Cup ECG Intelligent Competition. https://tianchi.aliyun.com/competition/entrance/231754/introduction. Accessed 2021-08-15.

[6] Liu H, Chen D, Chen D, Zhang X, Li H, Bian L, Shu M, Wang Y. A Large-Scale Multi-Label 12-Lead Electrocardiogram Database with Standardized Diagnostic Statements. Scientific Data 2022;9(1):272. ISSN 2052-4463.

[7] Antoni L, Bruoth E, Bugata P, Bugata Jr P, Gajdoš D, Horvát S, Hudák D, Kmečová V, Staňa R, Staňková M, Szabari A, Vozáriková G. Automatic ECG Classification and Label Quality in Training Data. Physiological Measurement jun 2022;43(6):064008. URL https://dx.doi.org/10.1088/1361-6579/ac69a8.

[8] 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 (CVPR). jun 2022; 11976–11986.

[9] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; 770–778.

[10] Smith LN, Topin N. Super-Convergence: Very Fast Training of Residual Networks Using Large Learning Rates. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. 2019; 1100612.

Address for correspondence:

Dávid Hudák
VSL Software, a.s., Lomená 8, 040 01 Košice, Slovakia
hudak@vsl.sk