# A Two-Phase Multilabel ECG Classification Using One-Dimensional Convolutional Neural Network and Modified Labels

Ľubomír Antoni[2], Erik Bruoth[2], Peter Bugata[1], Peter Bugata Jr.[1], Dávid Gajdoš[1], Šimon Horvát[2],
Dávid Hudák[1], Vladimíra Kmečová[1], Richard Staňa[2], Monika Staňková[1], Alexander Szabari[2],
Gabriela Vozáriková[2]

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

## Abstract

*Within PhysioNet/Computing in Cardiology Challenge 2021, we developed a two-phase method of automatic ECG recording classification. In the first phase, we pre-trained a model on a large training set with our proposed mapping of original labels to the SNOMED codes, using three-valued labels. To solve the multilabel binary classification task, we used a deep convolutional neural network, which is a 1D variant of the popular ResNet50 network. In the second phase, we performed fine-tuning for the Challenge metric and conditions. In the official round, our team CeZIS obtained the Challenge metric score of 0.717, 0.680, 0.703, 0.702, and 0.681 on the 12-lead, 6-lead, 4-lead, 3-lead, and 2-lead validation datasets, respectively.*

## 1. Introduction

The PhysioNet/CinC Challenge 2021 [1] addressed the issue of automated approaches for classifying cardiac abnormalities from ECG recordings (abbreviation ECGs will be applied in the following). Due to the increasing availability of smaller, lower-cost, and easier-to-use devices, the Challenge required the development of an algorithm that can classify 12-lead, 6-lead, 4-lead, 3-lead, and 2-lead ECGs in order to compare its performance on different numbers of leads.

## 2. Methods

In this paper, we present a method based on a deep 1D convolutional neural network that processes a raw ECG signal and predicts the probabilities of individual labels occurrence on the ECG recording. We solved this multilabel binary classification task in two phases:

- Phase 1 – creating a base model using data from different sources with unified label semantics.
- Phase 2 – fine tuning the model for PhysioNet/CinC Challenge 2021 conditions.

### 2.1. Training Data

A quality deep learning model requires a sufficient amount of training data. The public training set supplied for the Challenge is relatively large and originates from several source databases [1]. The inclusion of data from multiple sources is very important as they differ in the used tools, the scanning conditions, the methods of post-processing and digitization, and may also reflect geographical differences.

Table 1. Structure of used training data.

| # | Dataset | N_Samples |
|---|---|---|
| 1 | CPSC2018 | 6 877 |
| 2 | Georgia | 10 334 |
| 3 | PTB-XL | 16 094 |
| 4 | Ningbo | 34 905 |
| 5 | Chapman/CUSPH | 10 646 |
| 6 | Hefei | 44 142 |
|  | **Total** | **122 998** |

In our training data (described in Table 1), we excluded the INCART dataset containing very long signals, the PTB dataset containing signals from a small number of patients and the CPSC-Extra dataset. Moreover, we omitted ECGs from the PTB-XL dataset that were not validated by human. On the contrary, we added data from the Hefei Cup [2] containing a large number of ECGs with a relatively wide range of labels.

### 2.2. Labels

In the provided training data, the labels from the original datasets were mapped to the SNOMED-CT codes. In our solution, we omitted the Brady (bradycardia) and BBB (bundle branch block) labels due to their unclear semantics and the LPR (prolonged PR interval) label due to the small number of positive labels.

The two-valued logic for the provided labels seems to be insufficient. For example, all ECGs originated from the CPSC2018 dataset have the SB (sinus bradycardia) label set to 0 in the Challenge training set, since SB was not evaluated in CPSC2018. However, almost 700 ECGs from them indicate the sinus bradycardia symptoms.

Therefore, instead of the provided binary values, we used our own mapping based on the following three-valued logic:

- Value 1 – the diagnosis/anomaly occurs in the ECG,
- Value 0 – the diagnosis/anomaly certainly does not occur in the ECG,
- NA value – it is not known whether the diagnosis/ anomaly occurs in the ECG.

Moreover, we set the NA values in cases of inconsistency with other labels or with calculated parameters (e.g., heart rate).

For some labels, the ambiguous semantics causes a serious problem. For example, the NSR (normal sinus rhythm) label has different semantics for the ECGs originated from the CPSC2018, Georgia, and PTB-XL datasets. Because its semantics for undisclosed test data is unknown, it is not clear how to set up the prediction model.

## 2.3. Neural Network Architecture

As a backbone of our solution, specifically, we implemented the deep convolutional network by 1D variant of the ResNet50 [3]. All ECGs in our training set have a sampling rate of 500 Hz and the network inputs slices of length 4608 timesteps, which are gradually shortened to 18 final timesteps.

The backbone maps the ECGs to a fixed-dimensional space of latent variables. Since the latent space dimension of the standard ResNet50 is 2048, we use width ¼ corresponding to only 512 latent factors. We apply 1D kernel of size 5, our network contains approximately 1.2 million parameters. Compared to the 2D variant with more than 25 million parameters, our 1D variant has a significantly smaller number of parameters, and requires less memory and computing power.

The overall architecture is presented in Figure 1. The input layer is followed by the FlowMixup [4] layer, which adds random convex combinations of samples to the original batch. This doubles the size of the input batch. FlowMixup also allows to mix samples on layers deeper in the network, but in our solution the use of manifold mixup on one or two deeper layers has not yet led to better results.

Finally, the obtained latent factors are placed on the unit sphere using $\ell_2$ normalization and then a simple linear layer is added. In the output layer, the neurons correspond to the individual labels and output their probabilities.

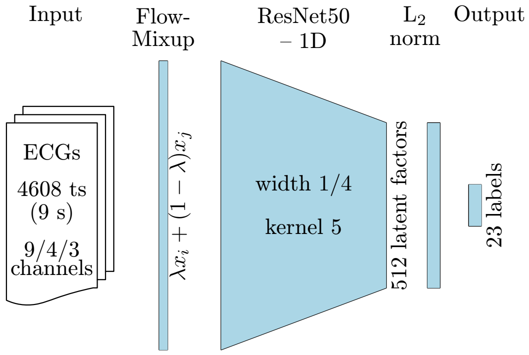

Figure 1. Neural network architecture.

## 2.4. Phase 1 – Base Model Training

In network training, the weighted binary cross entropy loss function (BCE) was used. We set the weights of positive and negative classes separately for each label (the NA class always had a weight of 0). Due to the class imbalance, we usually set the weight of the negative class to 1 and the weight of the positive class from 2 to 4.

Although the original classification task was defined for 5 different lead configurations, we trained only 3 models listed in Table 2. The leads used in models are highlighted in bold (the augmented leads were excluded). The models are ordered according to the number of actually used leads downward.

Table 2. Trained models.

| Model | Available/Used leads | N_used |
|---|---|---|
| 12-lead | **I**, **II**, **III**, aVR, aVL, aVF, **V1-V6** | 9 |
| 4-lead | **I**, **II**, **III**, **V2** | 4 |
| 6-lead | **I**, **II**, **III**, aVR, aVL, aVF | 3 |

We applied the Discrete Wavelet Transformation (DWT) to remove the baseline wandering [5] with level 9 and wavelet family db8. All signals were converted to the values in millivolts. Moreover, we extensively used data augmentation such as the random fixed-length slices, a lead rescaling, a lead dropout, and a periodic slice cutout.

To train and evaluate our models, we employed 10-fold cross-validation using eight folds as the training set, the ninth as the validation set, and the tenth as the held-out test set. We selected the best model based on the achieved micro F2-score on the validation set. We preferred the F2 metric because the error in identifying the diagnosis is considered more serious than a false alarm. The result of the training was an ensemble model composed of 10 neural networks.

We trained the model using the AdamW optimizer with a weight decay of 0.0005, a batch size of 128 (doubled by Flow-Mixup), and the OneCycle learning rate schedule [6] total of 100 epochs for each fold.

The proposed neural networks process an ECG recording slice of approximately 9 seconds in length. However, we need to predict longer ECGs, as well. Therefore, for validation and test data, we apply several slices evenly distributed in the ECG recording. For each slice, we predict the respective labels, and we obtain the resulting prediction using the maximum of the predicted probabilities.

## 2.5. Phase 2 – Fine Tuning for Challenge

The main result of Phase 1 is the extraction of quality latent factors that allow the prediction of individual labels in ECGs. In Phase 2, these latent factors are fixed, i.e., the network weights obtained in Phase 1 are frozen. Then, another output layer is added that produces label probabilities optimized for the Challenge metric.

The Challenge metric is a generalized version of the Accuracy measure, in which, even in the case of an inaccurate classification, partial reward is included when predicting a medically similar label. The reward amount for the combination of the predicted and true label is defined in the scoring matrix $\mathbf{W}$. If the vector $\mathbf{p}$ contains the predicted classes from the set $\{0, 1\}$ and the vector $\mathbf{t}$ includes the true classes for a sample, then the non-normalized metric for that sample can be expressed as follows:

$$M = \frac{\mathbf{p}^\intercal \mathbf{W} \mathbf{t}}{\sum \left( \mathbf{p}^\intercal (\mathbf{1} - \mathbf{t}) + \mathbf{t} \right)} \,. \tag{1}$$

We used the relaxed version of the metric $M$. For the vector $\mathbf{p}$ containing the predicted probabilities of individual labels from the interval $(0, 1)$, we defined a loss function $L = -\log(M)$ with the average value for all samples in the batch having at least one true label $\left(\sum \mathbf{t} > \mathbf{0}\right)$. The loss function is not used separately but only in combination with BCE in an appropriate ratio (from $1:1$ to $1:10$ in favor of the loss function $L$). BCE, unlike the loss $L$, is non-zero even in cases where all true classes are 0, and thus allows learning from such cases.

In this way, we obtained a novel output layer optimized for the Challenge metric. This layer was trained on our proposed three-valued labels, however, the validation and test sets in the Challenge are evaluated against the original binary labels. Since they have different semantics for the CPSC2018 and Georgia datasets (Section 2.2), we added two additional linear layers, one for CPSC2018 and one for Georgia. Each of them was trained separately on the corresponding dataset with the original labels.

When predicting on a test set, it is necessary to decide for each sample which of the three additional network outputs to apply. For this purpose, we used a discriminator network similar to the label prediction network, which solves the classification into three classes – CPSC2018, Georgia and Other datasets. We trained the discriminator on a large multi-source training set separately without mixup and any pre-processing that unifies the signals. Note that for the label prediction network training, it is recommended to remove the signal properties allowing to identify the source dataset. Then the network learns general rules and not the specifics of a particular dataset.

## 3. Results

The 10-fold cross-validation applied in Phase 1 provided F2-scores for individual labels in the large multi-source training set. The results allow us to compare how well the models predict individual labels and what are the differences in the prediction quality between the models with different numbers of leads.

Table 3. Comparison of F2-score for individual models.

| | **12-lead** | **4-lead** | **6-lead** | **6 vs.12** |
|---|---|---|---|---|
| SB | 0.9936 | 0.9933 | 0.9935 | 99.99% |
| STach | 0.9861 | 0.9862 | 0.9861 | 100.00% |
| AF | 0.9475 | 0.9467 | 0.9470 | 99.95% |
| PR | 0.9342 | 0.9274 | 0.9152 | 97.97% |
| CRBBB | 0.9194 | 0.9034 | 0.8865 | 96.42% |
| LBBB | 0.8979 | 0.8949 | 0.8810 | 98.12% |
| NSR | 0.8867 | 0.8643 | 0.8494 | 95.79% |
| PVC | 0.8598 | 0.8567 | 0.8446 | 98.23% |
| LAD | 0.8513 | 0.8555 | 0.8567 | 100.63% |
| RAD | 0.8222 | 0.8305 | 0.8207 | 99.82% |
| IAVB | 0.8052 | 0.8032 | 0.7933 | 98.52% |
| SA | 0.8051 | 0.8065 | 0.8051 | 100.00% |
| PAC | 0.7870 | 0.7815 | 0.7786 | 98.93% |
| TAb | 0.7748 | 0.7487 | 0.7337 | 94.70% |
| AFL | 0.7716 | 0.7290 | 0.7209 | 93.43% |
| TInv | 0.7629 | 0.7343 | 0.7202 | 94.40% |
| LAnFB | 0.7619 | 0.7574 | 0.7583 | 99.53% |
| LQRSV | 0.6989 | 0.6153 | 0.5586 | 79.93% |
| IRBBB | 0.6580 | 0.5489 | 0.3446 | 52.37% |
| QAb | 0.6414 | 0.6335 | 0.5272 | 82.20% |
| LQT | 0.5816 | 0.5604 | 0.5310 | 91.30% |
| PRWP | 0.5169 | 0.4040 | 0.1640 | 31.73% |
| NSIVCB | 0.4184 | 0.4092 | 0.3717 | 88.84% |

Table 3 shows the results using the F2-score metric. The labels are ordered according to the performance of the 12-lead model. In addition to the F2-score values for the three models used, the relative performance of the 6-lead model to the 12-lead model is given in the last column.

The values in the last column indicate whether the information from the three limb leads allows the same accurate prediction of the individual labels as the information from the twelve leads. Labels can be divided into four groups according to the percentage decrease in prediction quality differentiated by colors: dark green to 0.5%, light green from 0.5% to approximately 2%, orange from 3% to 10%, and red above 10%.

It remains to verify that the prediction quality does not decrease even after omitting lead III.

Table 4 contains Challenge score on the validation set obtained during the official round of the competition. The prediction for the 2-lead dataset was performed using the 6-lead model after calculating lead III from leads I and II. A similar procedure was used for the 3-lead dataset using the 4-lead model. The scores for the 2-lead dataset with calculated lead III and for the 6-lead dataset are very similar. The same is true for the 3-lead and 4-lead datasets.

Table 4. Challenge score on validation and test data.

| Dataset | Validation | Test | Rank |
|---|---|---|---|
| 12-lead | 0.717 | | |
| 4-lead | 0.703 | | |
| 3-lead | 0.702 | | |
| 6-lead | 0.680 | | |
| 2-lead | 0.681 | | |

## 4. Discussion and Conclusions

The achieved results show that it is enough to consider three of the five examined lead configurations. Since there is a linear relationship between the limb leads, the missing limb lead can be calculated using the other two without affecting the accuracy of the prediction.

For 2-lead configuration, several observations were made:

1. There is almost no decrease in accuracy for rhythm labels (SB, STach, SA, AF) and heart axis deviation labels (LAD, RAD).
2. There is a group of various labels (LAnFB, LBBB, IAVB, PVC, PAC, PR) where the decrease in accuracy is small (up to 2%).
3. For some labels (PRWP, LQRSV), the insufficiency of two leads was expected, but the presence of precordial leads also appears to be important for the identification of several other labels.

The created prediction models have poor results in the prediction of several labels at the bottom of Table 3, even with the availability of 12 leads. We suppose that the main limiting factor in obtaining better models is the quality of the labels and the inconsistency of their semantics across different data sources.

We believe that a service for the automatic identification of cardiac abnormalities in the ECGs, based on artificial intelligence, which can be constantly improved with the growing amount of processed data, could be useful 1) for rapid diagnostics in the absence of a cardiologist, 2) for early detection of heart disease in preventive examinations, 3) for additional automatic verification of all ECG findings generated in hospitals, 4) for self-diagnostics in connection with wearable electronics devices, 5) as an aid in teaching medics. As our contribution in this direction, we decided to temporarily make our solution publicly available on the web [7].

## Acknowledgments

This work was supported by ERDF EU grant and by the Slovak Ministry of Economy under the contract No. ITMS313012S703 and by the Slovak Grant Agency for Science (VEGA) under contract 1/0177/21.

## References

[1] Reyna M, Sadr N, Gu A, Perez A, Erick A, Liu C, Seyedi S, Shah A, Clifford G. Will two do? varying dimensions in electrocardiography: the physionet - computing in cardiology challenge 2021, 2021. URL https://doi.org/10.13026/jz9p-0m02.

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

[3] 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.

[4] Chen J, Yu H, Feng R, Chen DZ, Wu J. Flow-mixup: Classifying multi-labeled medical images with corrupted labels. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Los Alamitos, CA, USA: IEEE Computer Society, dec 2020; 534–541.

[5] Zhang D. Wavelet approach for ecg baseline wander correction and noise reduction. In IEEE Engineering in Medicine and Biology 27th Annual Conference. 2005; 1212–1215.

[6] 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; .

[7] Cordelia. https://cordelia.vsl.sk:8443. Accessed 2021-09-15.

Address for correspondence:

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