Physiological Measurement

IPEM
Institute of Physics and
Engineering in Medicine

Label noise and self-learning label correction in cardiac abnormalities classification.

Label noise and self-learning label correction in cardiac abnormalities classification.

Cristina Gallego Vázquez[1], **Alexander Breuss**[1], **Oriella Gnarra**[1,3], **Julian Portmann**[2], **Antonio Madaffari**[4], **Giulia Da Poian**[1]

[1]Sensory-Motor Systems (SMS) Lab, Department of Health Sciences and Technology, ETH Zurich
[2]Department of Computer Science, ETH Zurich
[3]Sleep-Wake-Epilepsy-Center, Department of Neurology, Bern University Hospital (Inselspital)
[4] Cardiovascular Center, University Clinic for Cardiology, Bern University Hospital (Inselspital)

E-mail: `cristina.gallegovazquez@hest.ethz.ch`

January 2022

Abstract. *Objective.* Learning to classify cardiac abnormalities requires large and high-quality labeled datasets, which is a challenge in medical applications. Small datasets from various sources are often aggregated to meet this requirement, resulting in a final dataset prone to label noise due to inter- and intra-observer variability and different expertise. It is well known that label noise can affect the performance and generalizability of the trained models. In this work, we explore the impact of label noise and self-learning label correction on the classification of cardiac abnormalities on large heterogeneous datasets of electrocardiogram (ECG) signals. *Approach.* A state-of-the-art self-learning multi-class label correction method for image classification is adapted to learn a multi-label classifier for electrocardiogram signals. We evaluated our performance using 5-fold cross-validation on the publicly available PhysioNet/Computing in Cardiology (CinC) 2021 Challenge data, with full and reduced sets of leads. Due to the unknown label noise in the testing set, we tested our approach on the MNIST dataset. We investigated the performance under different levels of structured label noise for both datasets. *Main results.* Under high levels of noise, the cross-validation results of self-learning label correction show an improvement of approximately 3% in the challenge score for the PhysioNet/CinC 2021 Challenge dataset and an improvement in accuracy of 5% and reduction of the expected calibration error of 0.03 for the MNIST dataset. We demonstrate that self-learning label correction can be used to effectively deal with the presence of unknown label noise, also when using a reduced number of ECG leads.

1. Introduction

Cardiovascular diseases are the leading cause of death worldwide. In recent years, researchers from different fields have actively collaborated to develop new tools for the monitoring and early detection of cardiac abnormalities from electrocardiogram (ECG) signals (Jambukia, Dabhi, and Prajapati 2015). Thanks to the proliferation of new portable and wearable ECG recording devices, it is now possible to collect data as never before, making automated signal interpretation a fundamental requirement. Machine learning (ML), and in particular Deep Learning (DL), has been widely applied to achieve this task, showing high accuracy in ECG arrhythmia classification (Ebrahimi et al. 2020). Several DL models have been proven to be successful in detecting cardiac abnormalities from ECG recordings, such as restricted Boltzmann machines, stacked autoencoders, convolutional neural networks (CNNs), and deep belief networks (Mathews, Kambhamettu, and Barner 2018). DL overcomes the task of carefully selecting features by learning informative patterns from raw inputs using convolution operations. CNNs are currently the most widely employed models in the field of ECG classification (Ebrahimi et al. 2020). These studies primarily focused on identifying cardiac abnormalities using either 12-lead ECGs or a reduced number of leads, typically single-lead ECGs.

The PhysioNet/CinC 2021 Challenge aimed to explore the ability to achieve similar multi-class classification performance with 12-lead ECG and a reduced set of leads, motivated by the limited accessibility of 12-lead ECG devices.

The dataset provided for the challenge is a collection of annotated 12-lead ECG recordings from six sources collected in four different countries across three continents. Consequently, the dataset is prone to various types of label noise. Even when relying on experts, labeling cardiac abnormalities is not straightforward, and intra- and inter-rater variability is a common source of label noise. Additional bias noise can be introduced when different data sources are combined. For instance, slightly different rules are adopted to systematically assign a recording to two different classes in different datasets. When using DL models, label noise should be carefully considered since, despite being able to maintain high training performance under high levels of random label noise, they might lead to low generalization (Zhang et al. 2021).

Our team, SMS+1, participated in the PhysioNet/CinC 2021 Challenge, reaching the seventh position on the leader-board. Our approach aimed to deal with and gain a better understanding of two of the biggest challenges related to applying DL to ECG: the data imbalance and noisy nature of the labels arising from incorrectly labeled recordings (Hong et al. 2020).

In this work, we present an approach for the automatic classification of heart arrhythmias from an arbitrary number of leads without significant performance loss and is meant to provide better-calibrated predictions in the presence of label noise in the training data. In our previous work (Gallego Vázquez et al. 2021), we included a self-learning label correction module to our model, following the work from (Han, Luo, and X. Wang 2019) for multi-class image classification. The framework application to the PhysioNet/CinC 2021 Challenge dataset resulted in some challenges and issues that were not present in the original work. In particular, we investigate the impact of label noise on classification performance and the use of self-learning label correction in the context of ECG multi-class multi-label classification. Specifically, the main contribution of this paper is the adaptation of the self-learning label correction method to the *mutli-label* case, as well as the evaluation of its impact on both the overall performance and single classes under various label noise levels.

Given that the PhysioNet/CinC 2021 Challenge data used for testing during cross-validation is also affected by label noise, we test the proposed multi-label self-learning label correction on the MNIST data (Deng 2012), which provides a clean testing scenario. Furthermore, we include an adversarial condition in the PhysioNet/CinC 2021 Challenge dataset, reflecting real-world settings of label noise by adding structured noise only to the training labels. Finally, we performed a qualitative assessment of the correction method. We invited three trained cardiologists to review a subset of the corrected ECG recordings to understand whether the corrected labels were indeed mislabeled and correctly chosen by the newly selected class.

2. Related Work

2.1. Multi-lead vs reduced-lead classification

With the rise of technology and digital tools, telemedicine is becoming the future of health care. Anyone with a smartphone or wearable device can record different types of data and biosignals. Real-time monitoring of ECG would be a powerful diagnostic tool. Therefore, smaller and cheaper ECG devices than the standard 12-lead ECG are required. Recently, Sohn et al. 2020 reconstructed a 12-lead ECG from a 3-lead patch using a neural network. Other studies have investigated the possibility of arrhythmia classification using single-lead and 2-lead ECG (Liu, Cheng, and Lin 2013; Mathews, Kambhamettu, and Barner 2018; Yang et al. 2019). Kristensen et al. 2016 concluded that both 12- and 3-lead ECGs can be used to detect atrial fibrillation, reporting that the specificity and sensitivity are comparable. Lai, Zhou, and Trayanova 2021 have proposed a DL model that uses the optimal 4-lead subset, improving the generalizability and accuracy in ECG abnormality detection with respect to the whole 12-lead set. There is still limited evidence that reduced-lead ECGs are comparable to 12-lead ECGs as diagnostic tools (M. A. Reyna et al. 2021).

2.2. Noisy labels

Labeling medical data requires a certain level of expertise and is costly and time-consuming. Consequently, the risk of incorporating label noise increases with the size of the dataset. Merging datasets from different sources, where the different qualities of the labels differ between them, represents an additional source of label noise. Label noise has been the most frequently reported reason for the decrease in classification performance (Frenay and Verleysen 2013). Although it is known that many state-of-the-art ML classifiers can deal with random noise, this is not true for other sources of label noise.

Robust loss functions such as Ghosh, Kumar, and Sastry 2017 have been proposed to train models that are unaffected by label noise. However, these methods make the same assumptions regarding noise distribution, for example, simple uniform label noise.

There has been a large body of work on identifying and correcting mislabeled instances in image classification (Karimi et al. 2020). Time series data poses additional challenges, and nonrandom label noise present in real-world datasets has often been overlooked (Atkinson and Metsis 2021). A few approaches have been proposed to mitigate label noise in ECG signal classification. Most of these approaches rely on the identification of incorrectly labeled samples for their removal during training (Pasolli and Melgani 2015; Li and Cui 2019; Wu and Tian 2020; Stepien and Grzegorczyk 2017). Genetic optimization methods (Pasolli and Melgani 2015), cross-validation as an ensemble of machine learning classifiers to filter mislabeled instances (Li and Cui 2019; Wu and Tian 2020), and only keeping samples with high confidence of being correctly labeled (Stepien and Grzegorczyk 2017), are the approaches implemented for ECG data. Cleansing techniques have the drawback of increasing classifier bias and degrading accuracy, generalizing worse to label noise present in the test data (Atkinson and Metsis 2021). Wu and Tian 2020 have also implemented a semi-supervised clustering method to correct mislabeled training samples based on cross-validation and k-nearest neighbor (KNN) classification.

3. Methods

The components of the proposed approach are described in detail in this section, as illustrated in Figure 1. In particular, after describing the data preprocessing and classification architecture, in Section 3.3, we explain the self-learning label correction approach.

3.1. Dataset and preprocessing

The original training data consists of six publicly available datasets, with more than 88.253 12-lead ECG signals provided by the PhysioNet/CinC 2021 Challenge (M. A. Reyna et al. 2021). The official validation dataset consists of approximately 6.630 12-lead ECG signals and the official test dataset of 6.266 recordings. More than 100 labeled abnormalities are present in the complete dataset. The 30 classes considered by the challenge and their abbreviations are listed in Table **??**. The dataset is also characterized by class imbalance, with, for instance, sinus rhythm representing 22% of the labels and complete left bundle branch block appearing only in 0.16% of the recordings.

As a first step, instances not belonging to any of the 30 classes considered in the challenge (around 20%) are removed. Because recordings from separate hospitals and devices can have different sampling rates, we first resample each recording to 250 Hz. The signal duration varies between the datasets. During training and testing, signals are zero-padded to the maximum signal length in each batch before entering the model. A binary mask is added to the channels as an input to the model to identify the padded part of the signal.

3.2. Classification architecture

3.2.1. 1D-CNN The classification component of our model is consistent with the one proposed in (Gallego Vázquez et al. 2021) and is based on a series of convolution operations and two fully connected feedforward networks. We employ one-dimensional convolution operations, which are applied to the original ECG waveform segments and preprocessed as described in Section 3.1, to extract a latent space representation of the signals. A detailed summary of the network settings is presented in Table 1. The last two CNN layers before the FC networks reduce the number of features used to compute the similarity metrics during the label correction phase.

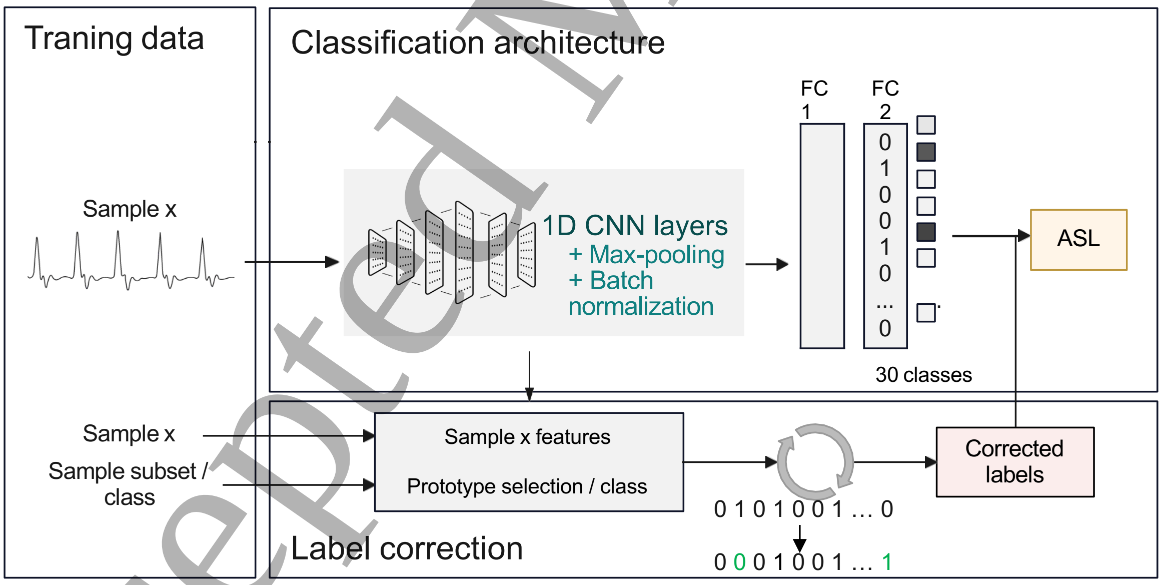

Figure 1: Architecture of the model: Training samples are first resampled before they are passed to a 1D CNN. The feature set obtained from the 1D CNN is used as an input to a label correction phase that iteratively estimates corrected labels during training by identifying prototypes of every class likely labeled correctly. The output of the label correction is then combined with the output after the convolution step. Adapted from (Gallego Vázquez et al. 2021)

Table 1: Deep Learning model settings.

Layer	In	Kernel	Stride/Padding	Out
CNN 1	leads	5	1/2	16
CNN 2	16	5	1/2	32
CNN 3	32	5	1/2	64
CNN 4	64	5	1/2	128
CNN 5	128	5	1/2	256
CNN 6	256	5	1/2	128
CNN 7	128	5	1/2	64
CNN 8	64	5	1/2	32
FC 1	32			128
FC 2	128			30

3.2.2. Asymmetric loss function To deal with the imbalanced nature of the dataset, we employ an asymmetric loss (ASL) for multi-label classification (Ben-Baruch et al. 2020), defined as:

$$\text{ASL} = \begin{cases} L_+ = (1-p)^{\gamma_+}\log(p) \\ L_- = (p_m)^{\gamma_-}\log(1-p_m) \end{cases}$$

where p is the network's output probability, and $p_m = \max(p-m, 0)$ denotes the shifted probability by a margin hyper parameter m.

This loss function contains two complementary asymmetric mechanisms that work differently on well-represented and under-represented samples and dynamically adjust the asymmetry levels throughout the training. It uses two focusing hyperparameters to modify the contribution of easy samples to the loss function (γ_+, γ_-), and hard thresholding via the probability margin m. In our work, we set the two focusing parameters to $\gamma_+ = 1$ and $\gamma_- = 3$, and the probability margin to $m = 0.2$.

3.3. Self-learning label correction

Once the model network has been trained for some epochs (n=2), label correction is performed in every training epoch until early stopping is reached. In each epoch, 16 class prototypes are selected from a random pool of samples from each class. Prototypes that present large density values have more similar samples from the same class around them and, therefore, have a high probability of being correctly labeled. However, if the chosen prototypes are those with the top highest density value, they are probably very close to each other. Therefore, we might miss other prototypes that are representative of the class. Both similarity and density values are considered for the selection of valid prototypes. Our approach differs from the original implementation (Han, Luo, and X. Wang 2019) in that each new prototype selected is compared only to the already selected prototypes and not to the complete sample pool. Prototypes are selected based on cosine similarity, which is defined as follows:

$$\cos(\mathbf{p}, \mathbf{s}) = \frac{\mathbf{ps}}{\|\mathbf{p}\|\|\mathbf{s}\|},$$

where **p** is the prototype and **s** is the new sample. A high cosine value indicates that a sample is closely related to the already selected prototype and is thus a good candidate for being a prototype for that class. To deal with multi-label classification, we use a modified correction criteria: First, we calculate a similarity value using the cosine similarity between the features of each training sample **s** and the features of the prototypes **p**. Then, we assign one or more corrected labels from the classes whose similarity between prototypes and sample is higher than a threshold set to 0.9. If no similarities exceed this threshold, we do not assign a new label to the sample and instead use the original label. During self-learning label correction, we use a weighted loss function with the original labels and the suggested corrected labels (Han, Luo, and X. Wang 2019):

$$\text{loss} = (1-\alpha) * \text{ASL}(y,t) + \alpha * \text{ASL}(y,c),$$

where y, t, and c are the predicted, original, and corrected labels, respectively, and ASL is an asymmetric loss function. The value of α determines the amount of label correction involved in the loss computation. In our case, we set α to 0.5.

3.4. Implementation details

The official evaluation metrics from the challenge (Alday et al. 2020; M. A. Reyna et al. 2021) consider that some misdiagnoses are less harmful than others. For this purpose, misdiagnoses are weighted differently, giving a final challenge score. We monitor the average challenge score during model training and use early stopping when the validation challenge score stops improving for four epochs.

Hyperparameters selection is not performed: the Adam optimizer is used with default parameters ($\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 10^{-8}$), the learning rate is set to 0.001, and batch size is set to 10. The complete model consists of 1.115.244 trainable parameters and it is trained on the PhysioNet/CinC 2021 Challenge datasets with no other external data sources. All algorithms were implemented in PyTorch 1.7.1 with Python 3.8.10.

4. Experiments

4.1. Multi-Label Correction Performance on MNIST

To compare the accuracy of the multi-label correction under different levels and types of noise starting from a clean dataset, we evaluated its performance on the MNIST (Deng 2012) digits data. The dataset comprised a training set of 60.000 samples and a test set of 10.000 samples. A multi-label version of the dataset was built by choosing several pairs of digits and pairing them into a single image. The labels corresponding to 10 binary variables were set to 1 if the corresponding digit was in the picture. As the next step, we trained a CNN with binary cross-entropy loss (baseline accuracy on MNIST 0.99%). The self-learning label correction method was then included in the training pipeline, and different types of noise were applied to the training and validation labels.

The test labels were not affected by the noise to provide a clean ground truth, which is not possible in the PhysioNet/CinC 2021 Challenge data. We consider *random* noise, for which we outnumbered labels sampled uniformly at random, and *bias* noise, for which labels were biased towards a random class. The level of label noise was varied from 10% to 30%. For the performance metrics, we used the model's accuracy and the percentage of correctly changed labels (CCL). When developing ML models for medical diagnosis, the confidence level of the model is also an important aspect to consider. Because of the unknown label noise present in the PhysioNet/CinC 2021 Challenge dataset, we investigated the reliability of the model with and without self-learning label correction on the MNIST dataset instead, using the calibration confidence to express the reliability of the model (Murphy and Epstein 1967). We computed the Expected Calibration Error (ECE) as proposed in (Naeini, Cooper, and Hauskrecht 2015) using Kuppersfrom et. al (Küppers et al. 2020) implementation. Since we have multiple binary classifiers, we computed ECE independently for each class and then calculated the average $\overline{\mathrm{ECE}}$.

4.2. Noise and model performance on ECG Data

To evaluate the performance of the label correction on the ECG data, we trained our model on 12-lead, 4-lead ("I", "II", "III", "V2"), and 2-lead ("I" and "II") recordings without label correction, to have a baseline, and with label correction. We followed a 5-fold cross-validation strategy, where the 4 folds of training data were further split into training (80%) and validation (20%).

Label correction has been shown to work well on datasets with approximately 20% noise (Lee et al. 2018). As the amount of noise present in the original dataset is unknown, we artificially added known label noise ranging from 10% up to 30%, only to the training and validation data. To make the noise nonrandom, we biased the labels by relabeling sample labels belonging to cardiac abnormalities to sinus rhythm (normal class). The same number of sinus rhythm samples was relabeled to different cardiac abnormalities. We quantified the label correction performance by comparing the number of artificial noisy labels that were detected and corrected back to the original labels.

As an evaluation metric, we favor the use of the challenge score, proposed in (M. A. Reyna et al. 2021). For completeness and future comparisons, we also report standard evaluation metrics such as Accuracy, Area Under Receiver Operating Characteristic (AUROC), Area Under Precision-Recall Curve (AUPRC), and the F-measure.

4.3. Visual assessment of label correction by experts

To qualitatively assess the performance of the self-learning label correction, we presented 50 corrected recordings (chosen randomly) to three different cardiologists and gave them the option to choose between the original labels, the corrected labels, or "neither". The order of the first two options was randomized to prevent bias. We then calculate the agreement between the three experts and the percentage of labels they would have also corrected among the presented recordings. Furthermore, we calculated the percentage of

Table 2: . Accuracy (Acc), average Expected Calibration Error ($\overline{\text{ECE}}$), and percentage of correctly corrected labels (CLL) for MNIST dataset under different types and levels of label noise on the training data. Baseline: CNN model, LC: CNN + self-learning Label Correction. Baseline model accuracy 0.99, $\overline{\text{ECE}}$= 0.006.

		Baseline		LC		
	Noise Level	Acc	$\overline{\text{ECE}}$	Acc	$\overline{\text{ECE}}$	CCL(%)
Random	10%	**0.99**	**0.01**	0.98	0.02	94
	20%	0.98	0.03	0.98	0.03	93
	25%	0.98	0.04	0.98	**0.03**	95
	30%	0.96	0.07	**0.98**	**0.04**	95
Bias	10%	0.98	0.02	0.98	0.02	95
	20%	0.98	0.03	0.98	0.03	95
	25%	0.97	0.04	0.97	**0.03**	95
	30%	0.92	0.07	**0.97**	**0.04**	93

recordings for which the expert would select the label, computed using the self-learning label correction method.

5. Results

5.1. Multi-Label Correction Performance MNIST

Table 2 shows the accuracy obtained from the test data of the MNIST dataset for the two different types of noise at four different levels (10%, 20%, 25 %, and 30%). We observed that CNN architectures are robust to uniform label noise in multi-label settings, with a constant accuracy of 0.99. However, a drop in performance was noticeable when the structured *bias* noise was introduced. While $\leq$ 20% added structured label noise does not seem to affect the accuracy or the ECE, when the noise is set to 30%, the self-learning label correction provides a more robust performance (LC: Acc = 0.97, ECE =0.04, Baseline: Acc = 0.92, ECE =0.07).

5.2. Noise and model performance on ECG Data

The challenge scores for different lead configurations and under different levels of label noise are shown in Figure 2 for the baseline model and self-learning label correction approach. The results of the 5-fold cross-validation for all metrics and all experiments are reported in Table S2.

When no additional noise is added to the training labels, the results suggest that the use of 12-lead ECG recordings provides slightly better overall classification results reported as (mean(std)): 0.73 (0.01), 0.72 (0.00), 0.71 (0.00) for the 12-, 4-, and 2-lead model, respectively. The confusion matrix in Figure 3 represents the percentage difference of the classification within each class for the 12-lead or 2-lead model. The results belong to one of the five test folds.

When using the original dataset with unknown label noise, the results of a 5-fold cross-validation show that the performance is slightly lower than the baseline, with a difference in challenge score of approximately 0.01. However, the test data used during cross-validation were affected by label noise. Thus, in our opinion, the results of the two methods are not comparable.

Figure 4 shows the number of diagnoses that were corrected during the training of the 12-lead model when no additional noise was added to the training labels. In total, 13% of the training samples were corrected. Figure S1 and Figure S2 show which classes were corrected to which classes for single diagnosis to single diagnosis correction. The most often changed class was from QAb to NSR. Figure S3 shows the distribution of recordings across different datasets and the percentage of corrected labels separated by dataset. The number of times a label is corrected corresponds to its presence in the dataset with some exceptions, such as SB, STach, and SA, which are some of the classes more represented with less corrected recordings. The number of times each label is corrected is also related to its label distribution within each dataset. Figure **??** shows the number of labels from recordings in each dataset that were corrected and chosen as correct. For a complete view of the number of recordings and labels per dataset, refer to Figure 4 in M. Reyna et al. 2022.

Finally, class-specific changes owing to the use of label correction are shown as the percentage difference in the confusion matrix in Figure 5. Independent of the number of

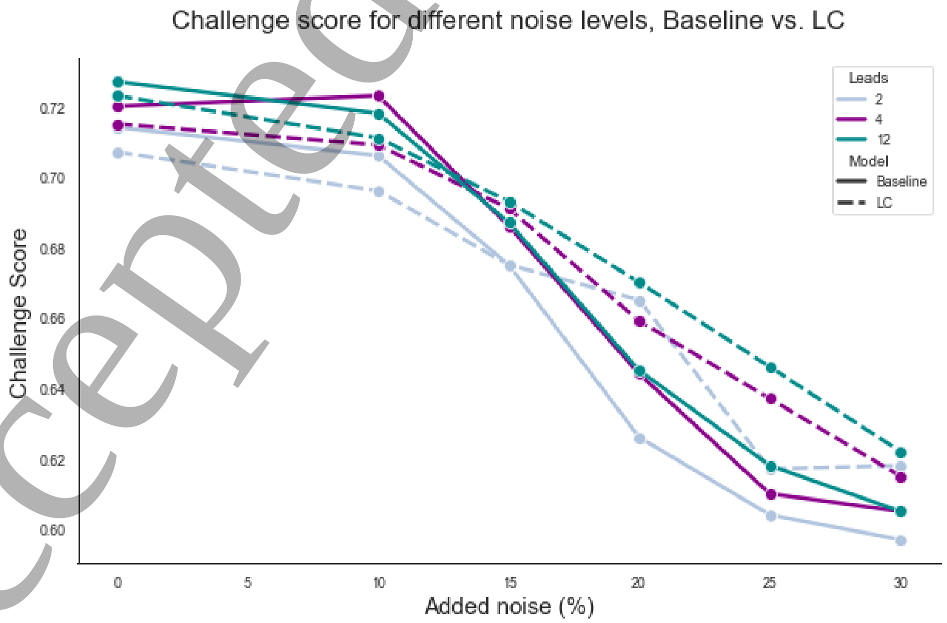

Figure 2: Cross-validation average challenge score at the different added bias noise levels, for the 12-, 4-, and 2-lead models, using the baseline approach (plain line) and with the self-learning label (LC) correction approach (dashed line).

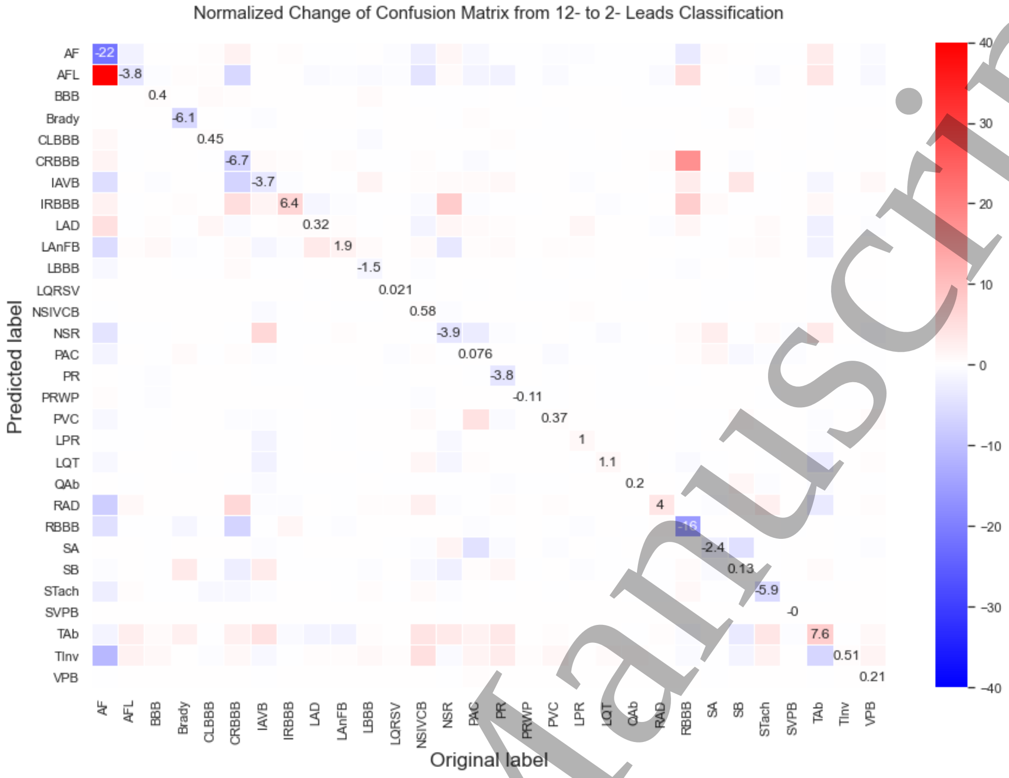

Figure 3: Change in the confusion matrix when reducing the 12-lead ECGs to 2-lead ECGs, normalized by the number of samples present in each of the 30 classes. Red results represent more samples predicted by the model with 12-lead ECGs. Blue results represent more samples predicted by the model with 2-lead ECGs. These results belong to one of the 5 test folds from the cross-validation.

leads used, Figure 5 shows that in the case of artificial bias noise added to the training labels, label correction outperformed the baseline model when the added noise level was equal to or higher than 20%. These results are consistent with those obtained when using the self-learning label correction on the MNIST data in the presence of structured *bias* noise.

5.3. Official challenge results

This section reports the official results obtained for the post-challenge submission. In contrast to the official challenge submission, we included the multiple-label self-learning label correction method. The official results obtained by re-submitting our code are listed in Table 3 and Table 4 for all test sets. The overall challenge score for the test set for the 12-, 6-, 4-, 3- and 2- models are: 0.48, 0.45, 0.15, 0.36 and 0.50, respectively.

A complete benchmark with challenge and post-challenge implementation can be found in (M. Reyna et al. 2022). Again, a direct comparison of challenge and post-challenge results does not reflect the performance change due to multi-label self-learning label correction, as the self-learning label correction was not used for the final official

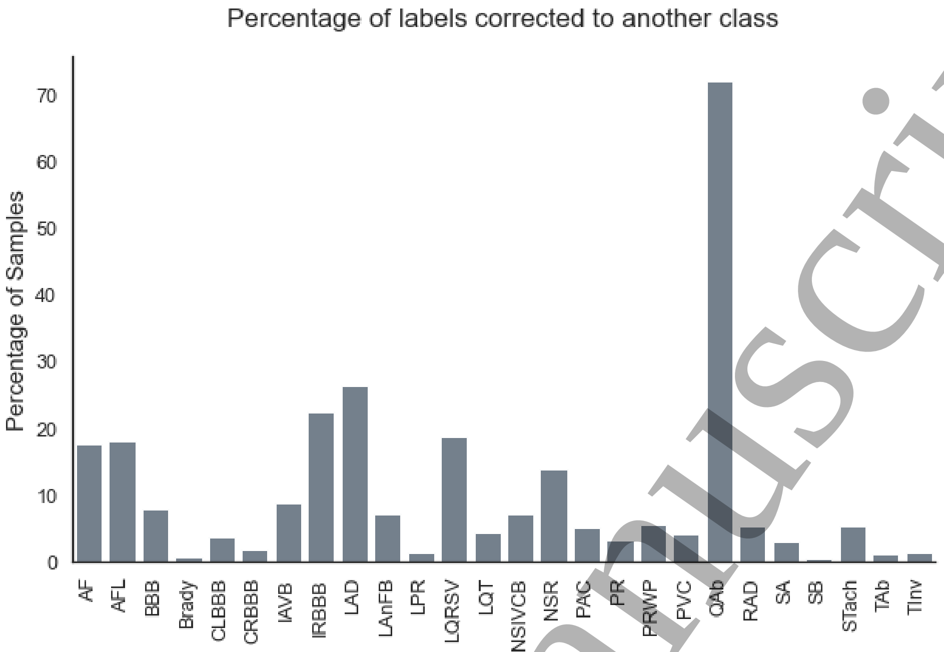

Figure 4: Percentage of samples per diagnosis that is corrected by the self-learning label correction algorithm to a different class. Recordings with single and multiple classes are considered. These results belong to one of the 5 test folds from the cross-validation.

challenge run.

Table 3: Official results of our model on the Challenge validation dataset for the 12-, 6- ('I', 'II', 'III', 'aVR', 'aVL', 'aVF'), 4- ('I', 'II', 'III', 'V2'), 3- ('I', 'II', 'V2'), and 2-lead ('I', 'II') models.

Leads-model	12	6	4	3	2
Accuracy	0.21	**0.27**	0.02	0.21	0.25
AUROC	0.86	**0.87**	0.69	0.84	**0.87**
AUPRC	0.35	**0.37**	0.14	0.32	0.35
Challenge	0.50	**0.52**	0.18	0.46	0.51
F-measure	0.31	**0.34**	0.12	0.31	0.31

5.4. Visual assessment of label correction by experts

When presented with a subset of corrected recordings, the experts independently identified that 52%, 66%, and 52% of the corresponding labels should have been corrected. In 60% of the cases, they agreed that a label had to be corrected. The results of the expert evaluations are presented in Table S4. They agreed on 38%, 34%, and 46% of the instances corrected by our model, respectively, while together they agreed on 34% of the presented recordings to be assigned to the wrong classes (with eight correctly corrected cases). When measuring the raters' agreement using Fleiss'

Kappa, we obtain a score of 0.88. Figure S4 and Figure S5 show two visual examples of recordings that were corrected by our approach, one correctly corrected and one wrongly corrected, respectively, according to the three experts.

6. Discussion

A 12-lead electrocardiogram is more expensive, bulky, and burdensome than a 2-lead electrocardiogram, so it is of particular interest to predict different cardiovascular diseases from 2-lead electrocardiogram signals.

It should be noted that the results from the undisclosed dataset were comparable among the different test sets. Compared to other team results, our model experienced a much smaller drop in scores from the validation to the test set (M. Reyna et al. 2022). This suggests that our approach can provide more robust results on the test data, with different characteristics and unseen during training. On the other hand, the 4-lead model shows a drop in performance metrics that we cannot explain or reproduce. Indeed, the results of the cross-validation for the 4-lead model do not show any significant reduction

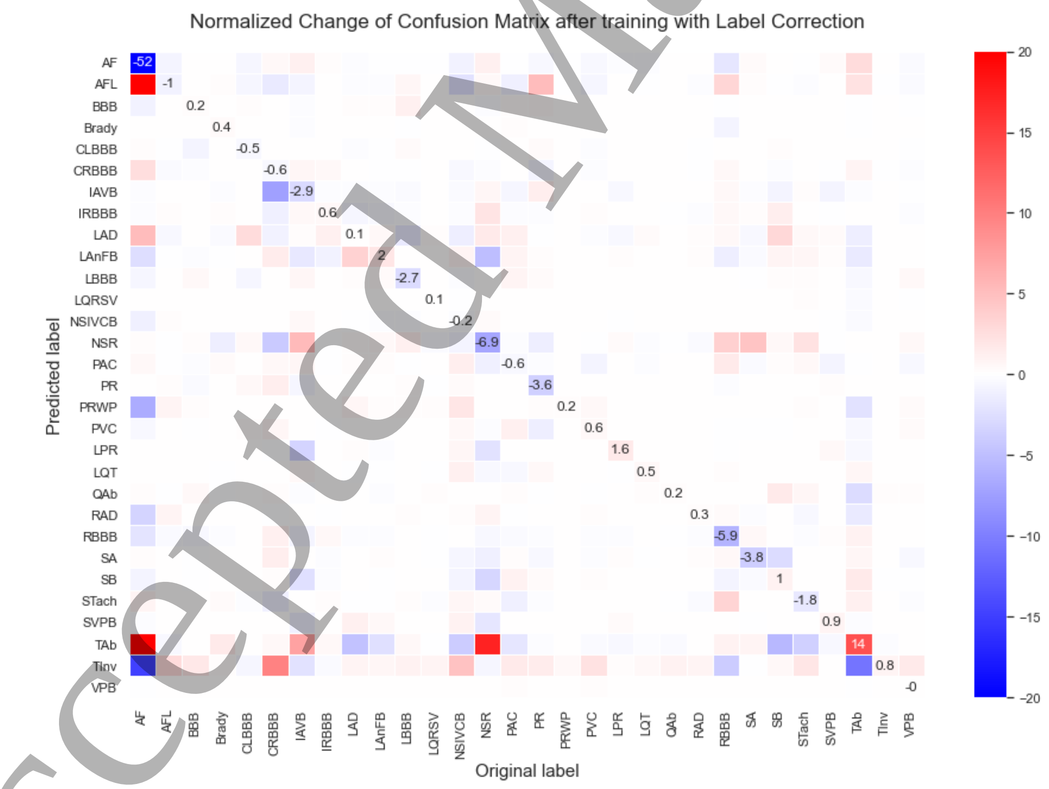

Figure 5: Change in the confusion matrix when including our label correction approach during the training, normalized by the number of samples present in each of the 30 classes. Results are shown for the case of the 12-lead model. Red results represent more samples predicted by the model without label correction. Blue results represent more samples predicted by the model with label correction.

Table 4: Official results of our model on the Challenge test datasets for the 12-, 6- ('I', 'II', 'III', 'aVR', 'aVL', 'aVF'), 4- ('I', 'II', 'III', 'V2'), 3- ('I', 'II', 'V2'), and 2-lead ('I', 'II') models. Best Challenge score per test-set in bold.

	CPSC2					G12EC				
Leads-model	12	6	4	3	2	12	6	4	3	2
Accuracy	0.35	0.41	0.04	0.28	0.37	0.16	0.22	0.01	0.18	0.21
AUROC	0.93	0.93	0.63	0.91	0.92	0.86	0.87	0.66	0.83	0.85
AUPRC	0.74	0.74	0.24	0.66	0.70	0.34	0.37	0.14	0.32	0.34
Challenge	0.60	**0.66**	0.01	0.56	0.61	0.48	**0.49**	0.20	0.44	**0.49**
F-measure	0.19	0.18	0.04	0.16	0.18	0.30	0.34	0.12	0.30	0.30

	Undisclosed					UMich				
Leads-model	12	6	4	3	2	12	6	4	3	2
Accuracy	0.23	0.31	0.01	0.12	0.28	0.21	0.26	0.02	0.19	0.26
AUROC	0.87	0.87	0.60	0.80	0.87	0.85	0.86	0.68	0.82	0.86
AUPRC	0.46	0.50	0.18	0.40	0.48	0.37	0.38	0.15	0.34	0.36
Challenge	0.47	0.39	0.07	0.22	**0.51**	0.47	0.46	0.17	0.39	**0.49**
F-measure	0.33	0.32	0.01	0.26	0.37	0.32	0.35	0.13	0.32	0.32

in performance.

The cross-validation results on available training data showed no substantial changes in accuracy (1%) when reducing the recording information to the 2-leads. The challenge score was 3% higher for the 12-lead model and the original dataset. Conversely, additional label noise reduces these differences. From Figure 3, we observe both improving and worsening in the classification of different classes between the 12- and 2-lead models. Noticeable changes are observed for AF and RBBB, where using 2-leads results in better classification. In addition, the classification of AFL also improves from the 12- to 2-lead model, with fewer AF samples misclassified as AFL for the 12-lead model.

When analyzing the performance in each class, we can see that AF, AFL, IAVB, NSR, PR, SA, and STach (accounting for 48% of samples in the dataset) classification improves with label correction. In contrast, the classification of LAnFB, LPR, SB, and TAb (accounting for 25% of samples in the dataset) worsened when performing label correction. It is not surprising that AF and AFL are very frequently incorrectly corrected, as shown in Figure S1. AF and AFL are very often misdiagnosed by physicians and belong to some of the most common and threatening cardiac conditions (Shiyovich et al. 2010). There have been some efforts in DL in recent years to train models that can distinguish between both (J. Wang 2021) because both signals look very similar to the human eye. Therefore, it is not surprising that a model that learns to differentiate between 30 classes will often mistake them. We included this information in our model; however, the results do not vary, likely because of the noisy test data. Nevertheless, most of the new AF labels correspond to recordings originally from Ningbo (the only dataset that does not include any AF label), and, as shown in Figure 5, fewer AF samples were misclassified as AFL when using label correction. This result indicates that our label correction technique can deal with heterogeneous databases with different labeling protocols.

Labels were only rarely corrected to the most represented classes (excluding AF and AFL), showing the robustness of our label correction towards class distribution. This implies that underrepresented labels that usually belong to less-frequent abnormalities are not corrected to the most prominent and better-known abnormalities. In order of presence, these classes are: NSR, SB, Tab, STach, and LAD. SA is the most chosen corrected label (33.3% of the time). It is a commonly encountered variation of NSR. Recordings from all datasets are corrected to SA in relation to the size of the datasets. From Figure S1, we can see that the majority of QAb (37%) were corrected to NSR. In Figure 5, we see that NSR classification is improved (6.9%) with label correction, while QAb classification almost does not experience any change (decrease of 0.2%). Moreover, results of the manual evaluation of recordings by the experts, reported in S4, highlight that the five QAb recordings identified as mislabeled by our method were mislabeled, according to the experts. We see also that the agreement on the newly assigned label does not always converge. These results indicate that the experts might mislabel QAb, and the self-learning label correction is identifying them and, in most cases, correctly correcting them.

Introducing additional noise only on training data shows that the use of self-learning label correction improves the performance of the underline CNN model, both for 12- and 2-lead. These results are consistent with the experiments on the MNIST dataset, where only a high level of structured noise made self-learning label correction necessary to increase the accuracy. This is consistent with the literature, as uniform random noise does not impact the underlying model to degrade the performance.

The challenge scores obtained during the official phase, in which self-learning label correction was disabled, are 0.52, 0.45, 0.50, 0.50, and 0.49 for twelve-, six-, four-, three-, and two-lead, respectively. As expected, on the official test set that contains unknown label noise, the implemented self-learning label correction seems to perform worst with respect to the same model without self-learning label correction. This is consistent with the experiments of the cross-validation reported in Figure 2. Nonetheless, the comparison is unfair: the performance assessment always favours models that agree with the uncorrected labels. Further analysis is needed, but for that a curated test set must be first assembled.

It is hard to estimate the actual level of label noise on the original data. Nevertheless, we have evidence that the self-learning label corrector detected wrong labels on the original data. The three experts agreed that only 24% of the presented recordings should not have been corrected, while the remaining 76% of the recordings are considered to present the wrong original label by at least one expert.

As previously mentioned, not having access to a clean dataset makes the estimation of the level of label noise a challenging task. According to our interpretation, the results show that self-learning label correction can change classification performance within a single class. As a next step, we would like to improve our model by considering these within-class improvements that might not be detected when computing the overall metrics.

7. Conclusion

Our approach demonstrates a consistent ability to detect various cardiac abnormalities on standard 12-lead ECGs and 2-lead ECGs. The framework's effectiveness was demonstrated using both a benchmark image classification dataset and six real-world ECG datasets. In the presence of a high level of label noise, label correction during training provides more robustness in favor of better classification accuracy. Self-learning label correction provides a valuable tool for leveraging large aggregated datasets, requiring limited supervision and allowing for different levels of labeling expertise. The proposed approach can effectively classify cardiac abnormalities using different sets of leads and simultaneously address the class imbalance and the presence of unknown and adversarial structured noise.

Acknowledgments

CGV and GDP are funded by the Swiss National Science Foundation (SNSF) grant PZ00P2_193291. We thank Dr. Gaspare Maranzano and Dr. Carlo Gigante for reviewing a subset of the corrected labels and providing feedback on the label correction from a cardiologist's point of view. We thank Prof. Dr. Robert Riener, Prof. Dr. med. Claudio L. Bassetti and Dr. Peter Wolf for their support.

References

Alday, E.A.P. et al. (Dec. 2020). “Classification of 12-lead ECGs: the PhysioNet - Computing in Cardiology Challenge 2020”. In: *Physiological measurement* 41.12, p. 124003.

Atkinson, G. and V. Metsis (June 2021). “A Survey of Methods for Detection and Correction of Noisy Labels in Time Series Data”. In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 479–493.

Ben-Baruch, E. et al. (Sept. 2020). “Asymmetric loss for multi-label classification”. In: *arXiv preprint arXiv:2009.14119*.

Deng, Li (Oct. 2012). “The MNIST database of handwritten digit images for machine learning research”. In: *IEEE Signal Processing Magazine* 29.6, pp. 141–142.

Ebrahimi, Z. et al. (Sept. 2020). “A review on deep learning methods for ECG arrhythmia classification”. In: *Expert Systems with Applications* X.7, p. 100033.

Frenay, B. and M. Verleysen (Dec. 2013). “Classification in the presence of label noise: a survey”. In: *IEEE transactions on neural networks and learning systems* 25.5, pp. 845–869.

Gallego Vázquez, C. et al. (Aug. 2021). “Two will do: Convolutional neural network with asymmetric loss, self-learning label correction, and hand-crafted features for imbalanced multi-label ECG data classification”. In: *2021 Computing in Cardiology (CinC)*, pp. 1–4.

Ghosh, A., H. Kumar, and P.S. Sastry (Dec. 2017). “Robust loss functions under label noise for deep neural networks”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1.

Han, J., P. Luo, and X. Wang (Oct. 2019). “Deep self-learning from noisy labels”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5138–5147.

Hong, S. et al. (July 2020). “Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review”. In: *Computers in Biology and Medicine* 122, p. 103801.

Jambukia, S.H., V.K. Dabhi, and H.B. Prajapati (Mar. 2015). “Classification of ECG signals using machine learning techniques: A survey”. In: *2015 International Conference on Advances in Computer Engineering and Applications*, pp. 714–721.

Karimi, D. et al. (Oct. 2020). “Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis”. In: *Medical Image Analysis* 65, p. 101759.

Kristensen, A.N. et al. (May 2016). “The use of a portable three-lead ECG monitor to detect atrial fibrillation in general practice”. In: *Scandinavian journal of primary health care* 34.3, pp. 304–308.

Küppers, F. et al. (June 2020). “Multivariate confidence calibration for object detection”. In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Lai, C., S. Zhou, and N.A. Trayanova (Oct. 2021). "Optimal ECG-lead selection increases generalizability of deep learning on ECG abnormality classification". In: *Philosophical Transactions of the Royal society A* 379.2212, p. 20200258.

Lee, Kuang-Huei et al. (June 2018). "CleanNet: Transfer learning for scalable image classifier training with label noise". In: arXiv: 1711.07131 [cs.CV].

Li, Y. and W. Cui (Jan. 2019). "Identifying the mislabeled training samples of ECG signals using machine learning". In: *Biomedical Signal Processing and Control* 47, pp. 168–176.

Liu, S.H., D.C. Cheng, and C.M. Lin (Dec. 2013). "Arrhythmia identification with two-lead electrocardiograms using artificial neural networks and support vector machines for a portable ECG monitor system". In: *Sensors* 13.1, pp. 813–828.

Mathews, S.M., C. Kambhamettu, and K.E. Barner (Aug. 2018). "A novel application of deep learning for single-lead ECG classification". In: *Computers in biology and medicine* 99, pp. 53–62.

Murphy, A.H. and E.S. Epstein (Oct. 1967). "Verification of probabilistic predictions: A brief review". In: *Journal of Applied Meteorology and Climatology* 6.5, pp. 748–755.

Naeini, Mahdi Pakdaman, Gregory Cooper, and Milos Hauskrecht (Feb. 2015). "Obtaining well calibrated probabilities using bayesian binning". In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Pasolli, E. and F. Melgani (May 2015). "Genetic algorithm-based method for mitigating label noise issue in ECG signal classification". In: *Biomedical Signal Processing and Control* 19, pp. 130–136.

Reyna, M. A. et al. (Sept. 2021). "Will Two Do? Varying Dimensions in Electrocardiography: The Physionet/Computing in Cardiology Challenge 2021". In: *Computing in Cardiology* 48, pp. 1–4.

Reyna, M.A. et al. (July 2022). "Issues in the automated classification of multilead ECGs using heterogeneous labels and populations." In: *Physiological Measurement*.

Shiyovich, A. et al. (Oct. 2010). "Accuracy of diagnosing atrial flutter and atrial fibrillation from a surface electrocardiogram by hospital physicians: analysis of data from internal medicine departments". In: *The American journal of the medical sciences* 340.4, pp. 271–275.

Sohn, J. et al. (May 2020). "Reconstruction of 12-lead electrocardiogram from a three-lead patch-type device using a LSTM network". In: *Sensors* 20.11, p. 3278.

Stepien, K. and I. Grzegorczyk (Sept. 2017). "Classification of ECG recordings with neural networks based on specific morphological features and regularity of the signal". In: *2017 Computing in Cardiology (CinC)*, pp. 1–4.

Wang, J. (Oct. 2021). "An intelligent computer-aided approach for atrial fibrillation and atrial flutter signals classification using modified bidirectional LSTM network". In: *Information Sciences* 574, pp. 320–332.

Wu, P. and S. Tian (Nov. 2020). "Using semi-supervised cluster method to correct the mislabeled training samples of ECG signals". In: *9th Data Driven Control and Learning Systems Conference (DDCLS*, pp. 260–265.

Yang, W. et al. (July 2019). “A novel approach for multi-lead ECG classification using DL-CCANet and TL-CCANet”. In: *Sensors* 19.14, p. 3214.

Zhang, Chiyuan et al. (Feb. 2021). “Understanding deep learning (still) requires rethinking generalization”. In: *Communications of the ACM* 64.3, pp. 107–115.