Injecting Domain Knowledge in Deep Learning Models for Automatic Identification of Myocardial Infarction from Electrocardiograms

Silvia Ibrahimi[1], Massimo W. Rivolta[1], Roberto Sassi[1]

[1] Dipartimento di Informatica, Università degli Studi di Milano, Milan, Italy

Abstract

Deep Learning (DL) models for automatic ECG interpretation became widely investigated in recent years. However, their performance varies highly across models and datasets. One of the main reasons is the possibility that the DL model might learn spurious correlations present in a dataset between inputs and outcomes. In this study, we proposed a novel training strategy potentially able to force the domain knowledge into a DL model, by complementing, only during training, an end-to-end approach with features known to be relevant for the outcome. We tested the approach for the creation of a DL model tuned to identify myocardial infarction (MI) from the standard 12-lead electrocardiograms (ECGs). Two models were trained: one with standard backpropagation (full model) and the second one (split model) with the proposed approach, on the PTB Diagnostic ECG Database. An explainable AI technique was then used to identify which ECG leads were considered relevant by both models for each MI site, and were compared with guidelines for MI site identification. The validation accuracy was 0.85 and 0.69 for full and split models, respectively. Despite the lower performance achieved with the proposed approach, the number of relevant leads was higher (10 vs 4), suggesting that the domain knowledge was likely percolated into the DL model, made it more robust and capable of better generalization on other dataset.

1. Introduction

Deep Learning (DL) models for automatic ECG interpretation become widely investigated in recent years. Models have been trained from few hundreds of electrocardiograms (ECGs) to a few millions (*e.g.*, [1, 2]). However, their performance highly varies across models and datasets, as demonstrated by the Physionet Challenge 2020 [3], which was focused on automatic ECG interpretation.

Validating a DL model for ECG interpretation requires extensive tests on multiple datasets with a wide range of clinical conditions. Therefore, it often becomes prohibitive to have sufficient training and validation data. For instance, Ribeiro *et al.* [2] successfully trained a DL model able to identify only 6 cardiac abnormalities with $2 \cdot 10^6$ ECGs.

Another major problem is represented by the lack of different datasets with patients having similar clinical conditions. For example, in our previous study [4], we demonstrated that a high test-set performance achieved on a single dataset does not guarantee that the model, trained for automatic identification of myocardial infarction (MI), can be considered reliable. In fact, applying an explainable AI (XAI) technique, we quantified that the model was using completely different information (spurious correlations) with respect to the clinical guidelines for the identification of MI [5]. This drawback is likely present in other state-of-the-art DL models [6], even when trained on multiple datasets.

Being able to inject domain knowledge into the training phase of DL models seem a valid approach to mitigate the risk that the model would learn spurious correlations from the data at disposal. Indeed, since ECG interpretation is a highly established and investigated domain, the clinical decision rules, born out of decades of clinical investigation, may be considered reliable enough (at least for certain conditions). However, injecting such rules into the training phase is not straightforward. For example, Shahin *et al.* [7] proposed an innovative training strategy which makes use of two loss functions to reduce the correlation with confounding factors in hearbeat classification tasks. Another strategy could be knowledge distillation from a large DL model, known to perform well, called “teacher”, to a smaller model named “student” [8]. However, the teacher-student paradigm is still fully data-driven and does not incorporate the domain knowledge.

In this study, we proposed an innovative methodology to inject the domain knowledge into the training strategy of a DL model aiming to automatically identify MI from standard 12-lead ECGs. In order to assess the performance of the proposed approach, we compared the novel methodology with a DL model having the same architecture but trained with standard backpropagation and verified whether the most relevant leads for the localization of MI were properly employed in the process.

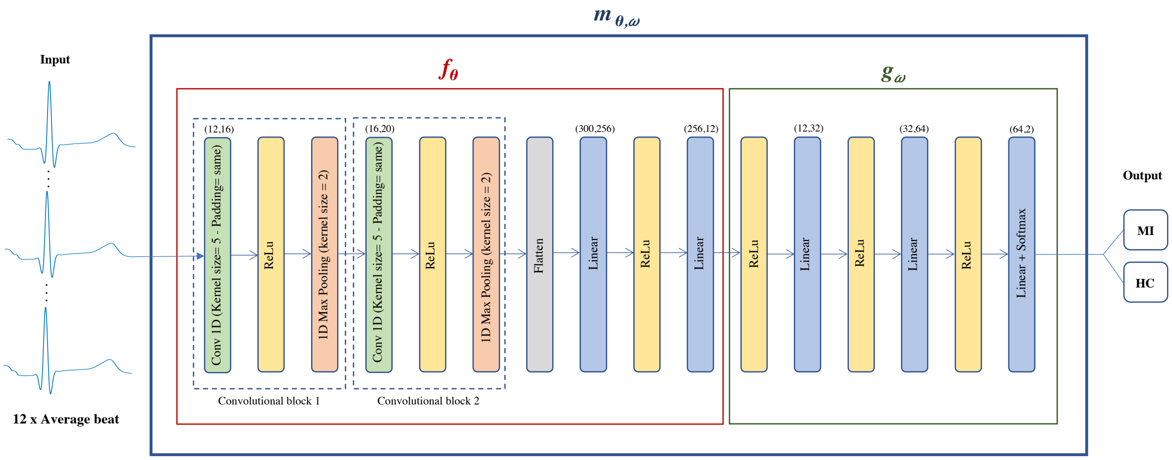

Figure 1. Diagram of the DL model and its components.

2. Methods

2.1. Dataset

ECG signals were downloaded from the open-access PTB Diagnostic ECG Database, freely available on physionet.org [9, 10]. The database provided 549 acquisitions from 290 subjects (209 men and 81 women) aged from 17 to 87 years. Each ECG signal was sampled at 1 kHz and 16 bit resolution. In this work, we considered the standard 12 leads available.

For each subject, there were one to five recordings available. In the study, we only used one trace per subject, specifically the oldest one when multiples were present. After quality assessment, we included all 52 healthy control (HC) subjects and 145 MI patients. Some of the patients were excluded from the study due to low quality recordings (for details, refer to the Sec 2.2).

2.2. Preprocessing and feature extraction

Each ECG signal was filtered using a Butterworth filter (3rd order, $0.5 - 40$ Hz, zero-phase) to reduce baseline wandering, high-frequency noise, and powerline interference. After denoising, beats were detected using the *gqrs* algorithm available on Physionet [9] and applied on the vector magnitude of the 12-lead ECG (*i.e.*, square root of the sum of squares). Beat positions were aligned on the Q point using the Woody's algorithm [11] applied to the vector magnitude. The signal quality of each lead was evaluated by computing the mean Pearson's correlation coefficient between each QRS complex (from $Q - 20$ ms to $Q + 100$ ms), and an average QRS template. An ECG trace was deemed of good quality when the average cross-correlation was higher than 0.9 for every lead. For each ECG, we computed the average beat for all leads. The average considered only heart beats whose inter-beat time interval, *i.e.*, $QQ_k = Q_k - Q_{k-1}$, with k the beat index, did not vary more than 50 ms with respect to the median QQ value. The considered QRST segments lasted 600 ms starting from 150 ms before the Q point. The amplitude of the ST segment for each lead was determined as the average of the samples between 250 ms and 300 ms on the average template (100 to 150 ms from the Q point).

2.3. Injecting domain knowledge

The core idea for injecting the domain knowledge into a DL model is to constraint a hidden layer to estimate specific properties related with the cardiac abnormality to identify. This constraint ensures that the next layers must use relevant information for the identification of the cardiac condition. In order to implement such approach (Figure 1), the dataset needs to be augmented with such specific properties. Formally, the dataset becomes the set $\mathcal{D} = \{\mathbf{x}_n, \mathbf{p}_n, y_n\}_n$ where $\mathbf{x}_n$ is the input vector of the DL model for the n-th patient, $\mathbf{p}_n$ is the vector of additional properties and y_n is the scalar containing the binary label. In this study, the vector $\mathbf{x}_n$ was the 12-lead ECG, the vector $\mathbf{p}_n$ contained 12 ST segment amplitudes (one per lead), which are relevant for the identification of MI infarction and its localization, and y_n was one hot-encoded (MI vs HC).

The DL model $m_{\theta,\omega}$ was designed as the composition of two mathematical functions:

$$m_{\theta,\omega}(\mathbf{x}) = g_\omega(f_\theta(\mathbf{x})) \qquad (1)$$

where $m_{\theta,\omega}(\mathbf{x})$ is the full DL model providing the probabilities of MI and HC, $f_\theta(\mathbf{x})$ represents all layers from the input to the constraint hidden layer, g_ω is the function describing the composition of all next layers up to the final classification one; θ and ω are vectors containing all model's parameters.

Several are the strategies to construct the proposed model with such a constraint. In this study, we split the training of the model $m_{\theta,\omega}$ into two different phases. The first one trained the model $f_\theta(\mathbf{x})$ to estimate the 12 ST amplitudes by minimizing the mean squared error between all $\mathbf{p}_n$ and the output $f_\theta(\mathbf{x}_n)$. The second one minimized the cross-entropy between $m_{\theta,\omega}(\mathbf{x}_n)$ and $\mathbf{y}_n$ while freezing all parameters θ, so that only the ω parameters could be learnt.

2.4. DL model

In order to perform binary classification (MI vs HC), we used a Convolutional Neural Network (CNN) model consisting of 2 convolutional blocks (1D convolution + ReLu + max pooling) and 5 fully connected layers. The input of the model was a matrix 12×600 representing the 600 samples of each of the 12 leads, whereas the output layer had 2 artificial neurons with a softmax function applied. Figure 1 shows the details of the proposed model.

2.5. Experiments

Two DL models with identical architectures were trained and their performance compared: the first one, called "full model", was not forced to incorporate domain knowledge and was trained using the standard backpropagation algorithm; the second one, named "split model", involved the proposed training strategy in two parts for training both f_θ and g_ω. The first part of the network was trained to predict the 12 ST amplitudes stored in the vectors $\mathbf{p}_n$ from the input ECG $\mathbf{x}_n$, while the second part dealt with binary classification of HC vs MI. The mean squared error loss was used for the training of f_θ and the cross entropy loss for the training of both g_ω and $m_{\theta,\omega}$.

The dataset was split into training set (0.7) and validation set (0.3). Since it was imbalanced, the cross-entropy loss was adjusted to weight the classes according to their number of subjects. The full model and each component of the split model were trained for 30 epochs with a batch size of 8, a learning rate of 0.002 and Adam optimizer.

In order to quantify whether the domain knowledge was properly injected into the DL model, we employed the popular occlusion method from the XAI domain [12]. Here, our interest was to determine whether the proper ECG leads were leveraged for the prediction of MI. To achieve this objective, after training both models, we considered the MI recordings within the entire dataset. For each of these recordings, we identified the three most important leads for classification by means of the occlusion method. To do so, we repeatedly set to 0 one lead at the time and quantified the absolute difference in probability between the modified ECG and its original version provided by both models. The three leads with the highest absolute difference were considered the most significant. Then, to identify the most relevant leads for each area of infarction, we examined the recordings in that specific site and selected the three most frequently occurring leads.

Once the most relevant leads for each site of infarction and model were identified, they were compared to match those reported in the guidelines for MI identification [5].

3. Results

The full and split models obtained a validation accuracy of 0.85 and 0.69, respectively. Table 1 shows the recall values for both models and each MI site (we do not report the results for other regions available in the dataset which contained only one recording). Recalls varied from 0.77 to 0.93 and from 0.40 to 1.00 for full and split models.

When using the proposed training strategy, the f_θ component of the split model obtained high performance for estimating the 12 ST amplitudes from the average beats. The R^2 between ST predicted and ST across leads was 0.82 (median=0.76 and interquartile range=0.94). In addition, despite the overall lower performance of the split model, the number of leads identified as consistent with the guidelines for MI identification was higher for the proposed approach (10 vs 4; bold text in Table 1). Specifically, for the anterior, antero-lateral, and inferior infarction areas, both models identified the same number of leads, with one common lead for inferior infarction. For the anterio-septal infarction area, the split model identified three leads that matched the guidelines, while the full model did not identify any. For infero-postero-lateral area, neither model leveraged a relevant lead.

4. Discussion and conclusions

In this study, we proposed an innovative training strategy to inject the domain knowledge into a DL model trained to identify MI from 12-lead ECGs. The strategy introduced a constraint into one of the hidden layers of the neural network to aid the next layers leveraging the proper information for MI identification.

We tested the hypothesis that the new approach would lead to a more robust DL model by comparing it with another one with the same architecture but trained with the standard back-propagation algorithm. The results showed that the full model generally performed better in terms of recall. However, the split model selected more leads in agreement with the guidelines, suggesting that domain knowledge, in the form of relevant leads, was likely incorporated into the model and thus likely to be more robust.

On the same dataset, in our previous work [4], we quantified that a model trained on average QRST segment did not leverage the proper ECG leads to identify MI. This

Table 1. Number of subjects, recall values and the three most relevant leads for both full and split models for each infarct site. The bold text indicates the leads matching the guidelines for MI identification at each infarct site.

Area	Subjects	Recall Full/Split	Full			Split		
			1st lead	2nd lead	3rd lead	1st lead	2nd lead	3rd lead
Anterior	16	0.81/0.44	II	**V5**	**V6**	I	**V3**	**V2**
Antero-lateral	15	0.93/0.40	V1	**I**	V4	V2	**V3**	V1
Anterio-septal	26	0.77/0.50	V5	V6	II	**V1**	**V2**	**V3**
Inferior	30	0.93/0.87	V5	V6	**I**	**I**	V3	V2
Infero-lateral	23	0.91/0.91	V5	V6	V4	**I**	**V2**	**V1**
Infero-postero-lateral	8	0.88/1.00	V5	V6	V4	I	V3	V2

problem was due to the fact that the model learnt a spurious correlation between the QRS complex and the outcome. This statement was confirmed by repeating the training by feeding the model using only the ST-T segment of the average beat, which resulted in a better match between the relevant leads and guidelines. In this study, we proved that it is possible to reduce such correlation by imposing a constraint into the DL model.

Despite further analyses are necessary to prove the efficacy of the new approach, we may speculate that the recalls for the full model are overestimated and likely due to the use of the same dataset in the validation scheme, which was also confirmed by our previous study [4]. It is necessary to verify this hypothesis with a different dataset.

As future works, it may be worth exploring the use of longer recordings (*e.g.*, 10 s) and different features such as T wave polarity instead of just a representative average of the ST segment. In addition, it is possible to introduce more complex constraints. For instance, a proper encoding of two or more contiguous leads in the hidden layer could be beneficial for MI identification. The approach can be adapted to address different types of cardiac abnormalities.

Acknowledgement

Part of this research was supported by the MUSA – Multilayered Urban Sustainability Action – project, funded by the European Union – NextGenerationEU, under the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strengthening of research structures and creation of R&D "innovation ecosystems", set up of "territorial leaders in R&D".

References

[1] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25(1):65–69.

[2] Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun 2020;11:1–9.

[3] Perez Alday EA, Gu AJ, Shah A, Robichaux C, Ian Wong AK, et al. Classification of 12-lead ECGs: The Physionet/Computing in Cardiology challenge 2020. Physiol Meas 2021;41:124003.

[4] Bodini M, Rivolta MW, Sassi R. Interpretability analysis of machine learning algorithms in the detection of ST-elevation myocardial infarction. Comput Cardiol 2020; 47:1–4.

[5] Thygesen K, Alpert JS, Jaffe AS, et al. Fourth universal definition of myocardial infarction (2018). J Am Coll Cardiol 2018;72(18):2231–64.

[6] Bodini M, Rivolta MW, Sassi R. Opening the black box: interpretability of machine learning algorithms in electrocardiography. Phil Trans R Soc A 2021;379:20200253.

[7] Shahin M, Oo E, Ahmed B. Adversarial multi-task learning for robust end-to-end ECG-based heartbeat classification. In Conf Proc IEEE Eng Med Biol Soc. IEEE, 2020; 341–344.

[8] Li R, Meng L, Liu Y, Hu S, Qiao G. Arrhythmia classification method based on knowledge distillation and 2d ECG images. In 2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE). IEEE, 2023; 493–497.

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

[10] Bousseljot RD, Kreiseler D, Schnabel A. The PTB diagnostic ECG database, 2004. URL `https://physionet.org/content/ptbdb/`.

[11] Woody CD. Characterization of an adaptive filter for the analysis of variable latency neuroelectric signals. Med Biol Eng 1967;5(6):539–554.

[12] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In Computer Vision – ECCV 2014. 2014; 818–833.

Address for correspondence:

Silvia Ibrahimi
Dipartimento di Informatica,
Università degli Studi di Milano,
Via Celoria 18, Milan 20133, Italy
silvia.ibrahimi@unimi.it