Learning EKG Diagnostic Models with Hierarchical Class Label Dependencies

Junheng Wang[1][0000−0003−2331−5446] and Milos Hauskrecht[1][0000−0002−7818−0633]

Department of Computer Science, University of Pittsburgh, Pittsburgh, PA, USA
{juw100,milos}@pitt.edu

Abstract. Electrocardiogram (EKG/ECG) is a key diagnostic tool to assess patient's cardiac condition and is widely used in clinical applications such as patient monitoring, surgery support, and heart medicine research. With recent advances in machine learning (ML) technology there has been a growing interest in the development of models supporting automatic EKG interpretation and diagnosis based on past EKG data. The problem can be modeled as multi-label classification (MLC), where the objective is to learn a function that maps each EKG reading to a vector of diagnostic class labels reflecting the underlying patient condition at different levels of abstraction. In this paper, we propose and investigate an ML model that considers class-label dependency embedded in the hierarchical organization of EKG diagnoses to improve the EKG classification performance. Our model first transforms the EKG signals into a low-dimensional vector, and after that uses the vector to predict different class labels with the help of the conditional tree structured Bayesian network (CTBN) that is able to capture hierarchical dependencies among class variables. We evaluate our model on the publicly available PTB-XL dataset. Our experiments demonstrate that modeling of hierarchical dependencies among class variables improves the diagnostic model performance under multiple classification performance metrics as compared to classification models that predict each class label independently.

Keywords: Electrocardiogram · Machine Learning · Bayesian Network.

1 Introduction

Electrocardiogram (EKG/ECG) is a key diagnostic tool to assess patient's cardiac condition and is widely used in patient monitoring, surgery support, and heart medicine research. Until recent years, most EKG processing depends largely on domain knowledge from experts and requires signal filtering and enhancing. Thanks to advances in machine learning (ML) methodologies and the increasing quantity and quality of EKG data, recent years has seen increased interest in the development of data driven solutions that can automatically interpret EKG signals and use it for the diagnosis of the underlying patient conditions. Such conditions are often labelled using standardized EKG vocabulary that is organized into a diagnostic class hierarchy [20]. For example, the ILMI (inferolateral myocardial infarction) and IPMI (inferoposterior myocardial infarction) labels

are aggregated in the IMI (inferior myocardial infarction) class at a lower level and the MI (myocardial infarction) class at a higher level of the hierarchy. On the other hand, the ASMI (anteroseptal myocardial infarction) is a member of AMI (anterior myocardial infarction) class which in turns also belongs to the same MI (myocardial infarction) class.

The problem of assigning class labels to EKG signals can be cast as a multi-label classification (MLC) problem. Unlike the multi-class classification problem where each instance belongs to exactly one class, in MLC each instance can have multiple class labels. In general, class labels that are assigned to individual EKGs can be at the same or the different level of abstraction. For example, a specific EKG can be assigned ILMI (inferolateral myocardial infarction), IMI (inferior myocardial infarction) as well MI (myocardial infarction) labels. The majority of the previous works on EKG classification does not consider hierarchical dependencies among class labels [14, 1, 19] and hence may result in inconsistent predictions at different levels. For example, the model may predict a class to be true while its parent class being false, or a class to be true while all its children classes are false.

In this paper, we propose and investigate an ML model that can perform MLC of EKG signals based on the hierarchical organization of EKG diagnoses and their corresponding class label dependencies. The proposed model starts from EKG signals that are transformed via ML architectures into a lower dimensional vector representation of EKG, and after that, it relies on a hierarchical organization of classes to make class label predictions. The hierarchical class dependencies our model relies on are implemented in a conditional tree structured Bayesian network (CTBN) [2] where the tree structure encodes the class hierarchy. We use multiple logistic regression models as classifiers of the CTBN, where each logistic regression model comes with its own set of trainable parameters. The trained CTBN model can make multi-label predictions in time linear in the number of classes, by computing the most probable assignment of all class variables using the tree structure of the CTBN.

We evaluate our CTBN model on the PTB-XL [20] EKG dataset that annotates each EKG using a mix of class labels at the different levels of abstraction. We show that by explicitly including the label dependencies we can improve the EKG classification performance in terms of both the exact match accuracy (EMA) and conditional log likelihood loss (CLL-loss), two criteria commonly to evaluate the MLC predictions. To prove the robustness of the CTBN model for the MLC problem we test its benefits by combining it with (five) different EKG signal transformations solutions: multilayer perceptron (MLP), recurrent neural network (RNN), long short-term memory (LSTM), gated recurrent unit (GRU), and fully convolutional network (FCN).

2 Related Work

In this section we briefly review the work related to our methodology, in particular, the work on ML models for EKG annotation, and multi-label and hierarchical classification tasks.

EKG labeling. Many efforts have been made to organize an extensive amount of EKG data into datasets [20, 10, 11] that consist of EKG waveform and diagnostic labels. Early research on EKG labeling required domain knowledge from experts, complex preprocessing on EKG signals [5, 23, 3] and hand-crafted features. More recent research efforts have explored modern ML solutions based on neural networks [1, 22] and deep learning [14, 4, 15] models to implement EKG processing and suitable EKG low-dimensional representations. Although this research does not share a common set of prediction targets, it leads to many promising results that demonstrate the potential of ML in automatic EKG labeling and diagnosis.

Multi-label classification. Specifically to the PTB-XL dataset, [17] examines several ML models including a five-layer one-dimensional convolutional neural network, a channel-wise SincNet-based [12] network architecture, and a convolutional neural network with hand picked entropy features calculated for every channel, and reports results on pair-wise classification tasks and 5- and 20-class MLC. [18] uses both the entropy features from [17] and features generated from R-wave detection methods, and performs MLC using aggregated models with different combinations of features. [19] provides a comprehensive collection of ML methods including Wavelet with shallow neural network [16], LSTM [7], FCN [21], ResNet [6], Inception [8] models and their variants. These models are benchmarked in various MLC tasks on different targets such as diagnosis, forms, and rhythms. Note that all the works mentioned above perform MLC without class-label hierarchy, and no dependency is considered when training and evaluating the models.

Hierarchical classification. The ML approach of hierarchical classification can be applied to the medical field by learning a collection of diagnostic classification models explicitly related via hierarchy. Malakouti and Hauskrecht [9] proposes a set of predictive models for multiple diagnostic categories organized in a hierarchy, and uses the hierarchy to guide the transfer of model parameters. The algorithm uses a two passes approach: the first pass follows the hierarchy in top-down fashion where the model parameters are transferred from higher-level diagnostic categories to lower-level ones; the second pass transfers the information bottom-up by adapting model parameters from lower level to their immediate parents. Their results shows improved performance when compared to independently learned models, especially for diagnosis with low priors and well-defined parent categories. [2] introduces conditional tree-structured Bayesian network (CTBN), a probabilistic approach that models conditional dependencies between classes in an effective yet computationally efficient way. Parameters of the CTBN model are captured in probabilistic prediction functions, and the structure is learned automatically from the data by maximizing the conditional log likelihood. The model makes predictions by finding the MAP assignment of class variables, with complexity linear to the number of class variables due to its tree structure. This dependency structure of class variables produces reliable probabilistic estimates and allows better performance of CTBN when comparing to other probabilistic methods.

3 Methodology

Our proposed model consists of: i) a model for generating low-dimensional summaries of the EKG inputs, and ii) an MLC model based on the CTBN supporting MLC tasks with label dependencies. The dependencies reflect the hierarchical organization of EKG classes at different levels of abstraction.

3.1 Low-dimensional EKG representation

The EKG is defined by a complex high-frequency time-series signal. Hence, the key challenge is to summarize this signal more compactly so that it can be linked to different MLC frameworks working with vector-based inputs. In this work, we consider modern neural network architectures to perform this step and generate low-dimensional representation of the EKG signals. Briefly, given an input instance X, formed by time-series of measurements, the model defines a function g that maps X to a lower-dimensional vector space $X' = g(X)$. In the case of EKG the input signal X is a tensor of shape (c, l), where c is the number of EKG channels and l is the length of the signal. The output is a k-dimensional real-valued vector $X' = (x_1, ..., x_k)$ that reduces the temporal dimension of the original EKG signal and aims to capture the key information needed to support the classification task.

The transformation of the EKG signal to a specific low-dimensional representation can be defined and learned with the help of different ML architectures. Here we consider: multilayer perceptron (MLP), fully convolutional network (FCN), and recurrent neural network (RNN). All of them have been applied to time series classification and prediction tasks, some specifically to EKG classification [15, 19] and have shown decent performance.

3.2 Multi-label classification model

The low-dimensional vector-based representations X' of EKG support MLC tasks. Briefly, we are interested in learning a model $f : X' \rightarrow Y$ where X' is a low-dimensional representation of EKG, and Y is a binary vector of m class labels. One way to define and train an MLC model $f : X' \rightarrow Y$ is to express it in terms of conditional probability $P(Y|X')$. The best assignment of labels to the input vector X' is then obtained by calculating $Y^* = \arg\max_Y P(Y|X')$.

There are different ways to represent and train $P(Y|X')$. One solution is to rely on a set of independent classification models, one model per class variable, to define $P(Y|X') = \prod_i P(Y_i|X')$. However, using an independent set of classifiers to support the MLC may fail to represent the dependencies among classes. This is important especially for EKG classification where typical multi-label annotation combines classes at the different level of abstraction. To alleviate the problem, one can resort to different MLC methods such as those defined by classifier chains [13]. Briefly, a classifier chain model decomposes the conditional probability $P(Y|X')$ over class labels into a product of conditional probabilities over components of Y using the chain rule: $P(Y|X') = \prod_i P(Y_i|pa(Y_i), X')$ where class label Y_i depends on a subset of so called parent class variables $pa(Y_i)$ that Y_i depends on.

In general, selecting and optimizing the parent subsets defining the chain classifier is a hard task. In this work we resort to **tree structured label dependency model** where each individual label may depend on at most one other class label. The model we adopt is the Conditional Tree Structure Bayesian Network (CTBN) model proposed by Batal et al [2]. It consists of a set of binary probabilistic classifiers for the transformed EKG vector X' and the value of at most one other class label. The classifiers are organized in a tree structure where the parent class denotes the class the variable depends on. Figure 1 illustrates a CTBN model with three binary class variables.

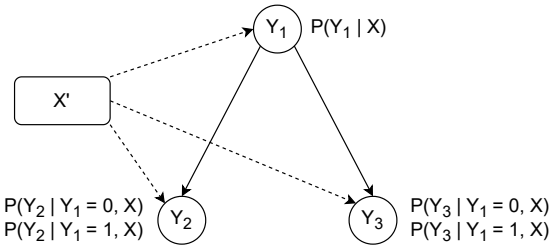

Fig. 1. An example of a CTBN with three binary class variables. The dash lines indicate the input and the solid lines model the hierarchy. Class variable Y_1 can be modeled with only one classifier, whereas Y_2 and Y_3 each requires two classifiers.

The advantages of the CTBN model are: (i) the tree structure can fit the hierarchy of class labels often used to annotate EKG, see section 4.1, and (ii) the optimal assignment of the classes for the transformed input X' can be found efficiently in the time linear in the number of classes [2]. Following Batal et al [2] we use logistic regression classifiers to define the individual classification models in the CTBN model, that is: $P(Y_i = 1|pa(Y_i) = j, X' = \mathbf{x}') = \sigma(\mathbf{w}_{\mathbf{i,j}}^{\mathbf{T}}\mathbf{x}')$ where σ is the logistic function and $\mathbf{w_{i,j}}$ are learnable weight vectors parameterizing individual models.

3.3 Training Models

The model that consists of the low-dimensional transformation g of the EKG signal and the CTBN implementing the MLC with dependencies among labels can be trained jointly using standard neural network optimization frameworks. Briefly, the CTBN part concurrently optimizes the weights of multiple logistic regression models implementing the different classifiers, where tree-based dependencies among the class variables are used to automatically select (via masking) all logistic regression models responsible for the specific training instance. We use average binary cross-entropy (BCE) loss and the AdamW optimizer with weight decay to optimize the models.

3.4 Making Predictions

The trained model allows us to transfer the original EKG signal x to its low-dimensional vector representation x', which in turn let us estimate $P(Y_i =$

$1|pa(Y_i), x')$ for all class labels. To make a prediction we want to identify the best possible assignment of class labels to all class variable, or in other words, the assignment that maximizes $y* = \arg\max_y P(Y = y|X' = x')$.

For the CTBN models, this optimization can be carried out efficiently across the tree-structure [2]. More specifically, we use a variant of the max-sum/max-product algorithm that runs in two passes with complexity linear to the number of class variables. In the first pass, information is sent upward from the leaves to the root, where each node compute the product of its local conditional probability and all probabilities sent from its children, and maximize and send the result over its value to its parent node. In the second pass, information is sent downward from the root to the leaves, where each node find its optimal assignment base on the assignment of its parent (if any) and its local conditional probability, and propagates the optimal assignment to its leaves.

4 Experiments

4.1 Data

We evaluate the proposed model on the EKG diagnostic task with 2 hierarchy levels using the publicly available PTB-XL dataset. The dataset consists of 21837 10-second long 12-lead EKGs from 18885 patients. This data is evenly distributed in terms of gender, with age covering a wide range of 0 to 95 years old. The EKG waveform is collected at a 500Hz sampling rate with 16 bit precision, and is down-sampled to 100Hz frequency. Each EKG instance is annotated by up to two cardiologists and labelled with 71 different EKG statements, using the SCP-ECG standard that covers diagnostic, form, and rhythm statements. We use the 44 diagnostic interpretations, and map them into 5 superclasses and 23 subclasses, as shown in Figure 2. Each EKG signal can be assigned multiple class labels, even at the same hierarchy level. In addition to EKG readings, PTB-XL dataset provides extensive metadata on demographics, but we do not use them in our classification models.

4.2 Methods

Our experiments compare the performance of the CTBN model to the baseline multi-label classifier that relies on a set of independent classification models where each class is predicted from a low-dimensional EKG summary vector x'. We compare these two models on five different EKG transformation approaches: MLP, RNN, LSTM, GRU, and FCN. For the MLP approach, we flatten the EKG signal input over the temporal dimension, and apply 5 fully connected layers with activation, using a hidden size of 4096 perceptrons per layer. For the recurrent approaches (RNN, LSTM, GRU), we use one unidirectional recurrent layer with hidden dimension 128. For the FCN approach, we use 5 convolution layers with activation and max pooling, followed by an adaptive average pooling layer to reduce the temporal dimension.

All models are built to make joint predictions on 28 class labels that consist of all 5 superclass and 23 subclass labels. Since superclass labels are binary and each subclass variable has exactly one superclass, two logistic regressions are needed to

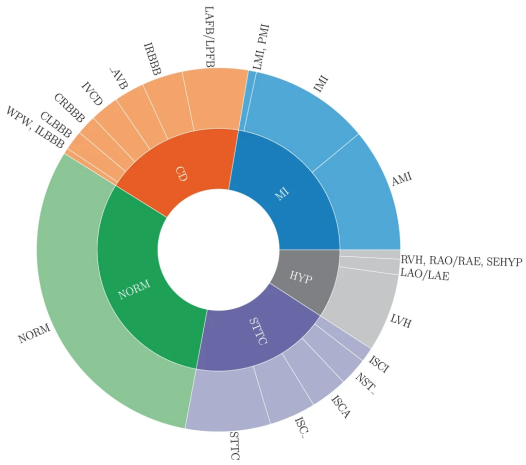

Fig. 2. A hierarchy of diagnostic classes describing the EKG readings in the PTB-XL dataset [20]. Diagnostic superclasses correspond to labels in the inner circle and subclasses to labels in the outer circle.

cover the predictions of each subclass in CTBN. However, since no EKG instance can belong to a subclass without belonging to its parent superclass label we can simplify the CTBN by considering just one logistic regression model for each subclass variable, that is, predicting probability conditioned on its parent class being true. This reduces the total number of logistic regression classifiers used in our CTBN model to 28, which is the same number of as used for the baseline MLC model that predicts each class independently.

4.3 Metrics

We consider three different metrics to evaluate the models: exact match accuracy (EMA), conditional log likelihood loss (CLL-loss), and macro F1. Briefly, the exact match accuracy (EMA) computes the percentage of the instances whose predicted class vectors are exactly the same as their true class vector, i.e. all class variables are correctly predicted. The conditional log likelihood loss (CLL-loss) is defined for probabilistic methods as: $\mathit{CLL\text{-}loss} = -\sum_{k=1}^{n} \log P(\mathbf{y}^{(k)}|\mathbf{x}^{(k)})$. The CLL-loss for a test instance $\mathbf{x}^{(k)}$ is small if the predicted probability of the true class vector $\mathbf{y}^{(k)}$ is close to 1, and the CLL-loss is large if the predicted probability of $\mathbf{y}^{(k)}$ is close to 0. Finally, the macro F1 score is the unweighted arithmetic mean of all the per-class F1 scores, which for each individual class variable is calculated as the harmonic mean of its precision and recall.

EMA evaluates the success of the models in finding the conditional joint distribution for all class variables $P(\mathbf{Y}|\mathbf{X})$ and thus is appropriate in our MLC setting. CLL-loss evaluates how much probability mass the model assigns to the true class vector and is a useful measurement for probabilistic methods. For two models that both misclassify an instance according to EMA, CLL-loss will still favor the one that assigns higher probability to the correct output. We report

the F1 score of the models, but note that it is not a very suitable metric for the MLC problem, since it is calculated separately for each class variable and then aggregated, and thus does not consider the dependencies between classes.

4.4 Results

All experimental results are obtained using 10-fold cross validation that is kept the same for all evaluated models. All the splits are obtained via stratified sampling while respecting patient assignments such that all records of the same patient are assigned to the same fold. As recommended in the PTB-XL dataset, we use folds 1-8 as training set, and folds 9 and 10 that underwent at least one human evaluation with high label quality as validation set and test set.

	Baseline (non-CTBN)			CTBN		
Transformation	EMA	CLL	macro F1	EMA	CLL	macro F1
MLP	0.368	13.0	0.168	0.385	10.4	0.153
RNN	0.237	6.7	0.045	0.332	4.2	0.043
LSTM	0.367	5.1	0.188	0.419	3.3	0.248
GRU	0.393	4.6	0.248	0.473	3.0	0.301
FCN	0.448	4.0	0.337	0.498	2.6	0.329

Table 1. Performance of the CTBN model vs non-CTBN baseline for five different EKG transformations on the PTB-XL dataset. The MLC problem is defined on the superclass and subclass labels.

Table 1 compares the performance of our CTBN model to the non-CTBN baseline defined by a set independent logistic regression models, one model per class variable. For all five EKG transformations representing three types of neural networks (fully connected, recurrent, convolutional), the CTBN model outperform its non-CTBN counterpart in terms of EMA and CLL-loss. We observe bigger improvements in EMA when using more complex EKG transformations, like the recurrent and convolutional neural networks. The results on F1 score are not consistent, as we see improvements when using LSTM and GRU, but not on MLP or FCN. This can be explained by the property of F1 score that it does not capture label dependencies and thus shows no advantage of CTBN.

	95% CI of $\Delta metric$ (lower, mean, upper)	
Transformation	ΔEMA (higher the better)	ΔCLL (lower the better)
MLP	(0.002, 0.016, 0.031)	(-3.497, -2.834, -2.154)
RNN	(0.070, 0.094, 0.117)	(-2.584, -2.488, -2.395)
LSTM	(0.043, 0.056, 0.075)	(-1.884, -1.761, -1.640)
GRU	(0.061, 0.080, 0.102)	(-1.801, -1.672, -1.540)
FCN	(0.031, 0.050, 0.067)	(-1.504, -1.402, -1.306)

Table 2. Performance of the CTBN model vs non-CTBN baseline evaluated with pair-wise statistical significance testing using 95% confidence interval.

Table 2 evaluates the performance of our CTBN model and the non-CTBN baseline using pair-wise statistical significance testing. We generate (with re-

placement) 1000 random bootstrap samples of sample size 1024 from the test set, and for each sample we evaluate our CTBN models and non-CTBN baselines using the desired metrics. We define $\Delta metric = metric_{ctbn} - metric_{baseline}$ and report the mean, upper bound, and lower bound of the 95% confidence interval of all 1000 samples. We conjecture that our CTBN models are statistically significantly better than corresponding baselines since the model consistently outperforms the baseline within an acceptable confidence interval, i.e. when ΔEMA is positive and ΔCLL is negative.

5 Conclusion

In this paper, we propose an ML model that improves EKG classification by leveraging the hierarchical class label dependencies. The model generates low-dimensional summaries of EKG instances with ML methods, and performs MLC using CTBN [2] that captures the dependencies between class variables. Our model uses logistic regression as probabilistic classifiers for the CTBN model, and can perform exact inference with complexity linear in the number of class variables. Our experimental evaluation on the PTB-XL [20] dataset shows that our approach outperforms the same ML architectures that do not incorporate class dependencies, and produces more reliable probabilistic estimates.

Acknowledgement The work presented in this paper was supported in part by NIH grants R01EB032752 and R01DK131586. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

References

1. Aziz, S., Ahmed, S. & Alouini, MS. ECG-based machine-learning algorithms for heartbeat classification. Sci Rep 11, 18738 (2021).
2. Batal, I., Hong, C., & Hauskrecht, M. (2013). An efficient probabilistic framework for multi-dimensional classification. Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management - CIKM '13.
3. Bulusu, S. C., Faezipour, M., Ng, V., Nourani, M., Tamil, L. S., & Banerjee, S. (2011). Transient st-segment episode detection for ECG Beat Classification. 2011 IEEE/NIH Life Science Systems and Applications Workshop (LiSSA).
4. Darmawahyuni, A., Nurmaini, S., Rachmatullah, M. N., Tutuko, B., Sapitri, A. I., Firdaus, F., Fansyuri, A., & Predyansyah, A. (2022). Deep learning-based electrocardiogram rhythm and beat features for heart abnormality classification. PeerJ. Computer science, 8, e825.
5. Di Marco, L. Y., & Chiari, L. (2011). A wavelet-based ECG delineation algorithm for 32-bit integer online processing. Biomedical engineering online, 10, 23.
6. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
7. Hochreiter, S., & Schmidhuber, J"urgen. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

8. Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., & Petitjean, F. (2020). InceptionTime: Finding alexnet for Time Series classification. Data Mining and Knowledge Discovery, 34(6), 1936–1962.
9. Malakouti, S., & Hauskrecht, M. (2019). Hierarchical Adaptive Multi-task Learning Framework for Patient Diagnoses and Diagnostic Category Classification. Proceedings. IEEE International Conference on Bioinformatics and Biomedicine, 2019, 19323030.
10. Moody, B., Moody, G., Villarroel, M., Clifford, G. D., & Silva, I. (2020). MIMIC-III Waveform Database (version 1.0). PhysioNet.
11. Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BiH Arrhythmia Database. IEEE Engineering in Medicine and Biology Magazine, 20(3), 45–50.
12. Ravanelli, M., & Bengio, Y. (2018). Speaker recognition from raw waveform with SincNet. 2018 IEEE Spoken Language Technology Workshop (SLT).
13. Read, J., Pfahringer, B., Holmes, G. et al. (2011). Classifier chains for multi-label classification. Mach Learn 85, 333–359.
14. Ribeiro, A.H., Ribeiro, M.H., Paixão, G.M.M. et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun 11, 1760 (2020).
15. Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., & Faubert, J. (2019). Deep learning-based Electroencephalography Analysis: A systematic review. Journal of Neural Engineering, 16(5), 051001.
16. Sharma, L. D., & Sunkaria, R. K. (2017). Inferior myocardial infarction detection using stationary wavelet transform and machine learning approach. Signal, Image and Video Processing, 12(2), 199–206.
17. Śmigiel, S., Pałczyński, K., & Ledziński, D. (2021). ECG signal classification using Deep Learning techniques based on the PTB-XL dataset. Entropy, 23(9), 1121.
18. Śmigiel, S., Pałczyński, K., & Ledziński, D. (2021). Deep learning techniques in the classification of ECG signals using R-peak detection based on the PTB-XL dataset. Sensors, 21(24), 8174.
19. Strodthoff, N., Wagner, P., Schaeffter, T., & Samek, W. (2021). Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. IEEE Journal of Biomedical and Health Informatics, 25(5), 1519–1528.
20. Wagner, P., Strodthoff, N., Bousseljot, R., Samek, W., & Schaeffter, T. (2020). PTB-XL, a large publicly available electrocardiography dataset (version 1.0.1). PhysioNet.
21. Wang, Z., Yan, W., & Oates, T. (2017). Time Series classification from scratch with Deep Neural Networks: A strong baseline. 2017 International Joint Conference on Neural Networks (IJCNN).
22. Westhuizen, J.V., & Lasenby, J. (2017). Techniques for visualizing LSTMs applied to electrocardiograms. arXiv:1705.08153: Machine Learning.
23. Zidelmal, Z., Amirou, A., Adnane, M., & Belouchrani, A. (2012). QRS detection based on wavelet coefficients. Computer methods and programs in biomedicine, 107(3), 490–496.