LightX3ECG: A Lightweight and eXplainable Deep Learning System for 3-lead Electrocardiogram Classification

Khiem H. Le [a,b,*], Hieu H. Pham [a,b], Thao BT. Nguyen [a,b], Tu A. Nguyen [a,b], Tien N. Thanh [c] and Cuong D. Do [a,b]

[a] VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam
[b] College of Engineering and Computer Science, VinUniversity, Hanoi, Vietnam
[c] College of Health Sciences, VinUniversity, Hanoi, Vietnam

ARTICLE INFO

Keywords:
Reduced-lead ECG Classification
1D-CNNs, Attention
Explainable AI (XAI)
Model Compression

ABSTRACT

Cardiovascular diseases (CVDs) are a group of heart and blood vessel disorders that is one of the most serious dangers to human health, and the number of such patients is still growing. Early and accurate detection plays a key role in successful treatment and intervention. Electrocardiogram (ECG) is the gold standard for identifying a variety of cardiovascular abnormalities. In clinical practices and most of the current research, standard 12-lead ECG is mainly used. However, using a lower number of leads can make ECG more prevalent as it can be conveniently recorded by portable or wearable devices. In this research, we develop a novel deep learning system to accurately identify multiple cardiovascular abnormalities by using only three ECG leads which are I, II, and V1. Specifically, we use three separate One-dimensional Convolutional Neural Networks (1D-CNNs) as backbones to extract features from three input ECG leads separately. The architecture of 1D-CNNs is redesigned for high performance and low computational cost. A novel Lead-wise Attention module is then introduced to aggregate outputs from these three backbones, resulting in a more robust representation which is then passed through a Fully-Connected (FC) layer to perform classification. Moreover, to make the system's prediction clinically explainable, the Grad-CAM technique is modified to produce a highly meaningful lead-wise explanation. Finally, we employ a pruning technique to reduce system size, forcing it suitable for deployment on hardware-constrained platforms. The proposed lightweight, explainable system is named LightX3ECG. We get classification performance in terms of F1 scores of 0.9718 and 0.8004 on two large-scale ECG datasets, i.e., Chapman and CPSC-2018, respectively, which surpass current state-of-the-art methods while achieving higher computational and storage efficiency. Visual examinations and a sanity check are also performed to strictly demonstrate the strength of our system's interpretability.

1. Introduction

Cardiovascular diseases (CVDs) are one of the primary sources of death globally, accounting for 17.9 million deaths in 2019, representing 32% of all deaths worldwide. Also, three-quarters of these deaths take place in low- and middle-income countries, according to World Health Organization [1]. Therefore, it is critical to detect these heart problems as soon as possible so that treatment may begin with counseling and medications. Electrocardiogram (ECG) is a waveform representation of the electrical activity of the heart obtained by placing electrodes on the body surface. The usual structure of an ECG beat [1], as illustrated in Figure 1, consists of three main components: P wave, which represents depolarization of atria; QRS complex, which represents depolarization of ventricles; and T wave, which represents repolarization of ventricles. Other parts of the signal include PR, QT intervals, or PR, ST segments. This electrical signal is a widely used, non-invasive tool for identifying cardiovascular abnormalities in patients. However, ECG analysis is a professional and time-consuming task, it requires cardiologists with a high degree of training to carefully examine and recognize pathological patterns in ECG recordings. This challenge, coupled with the rapid increase in ECG data, makes computer-aided, automatic ECG analysis more and more essential, especially in low- and middle-income countries, where high-quality and experienced cardiologists are extremely scarce.

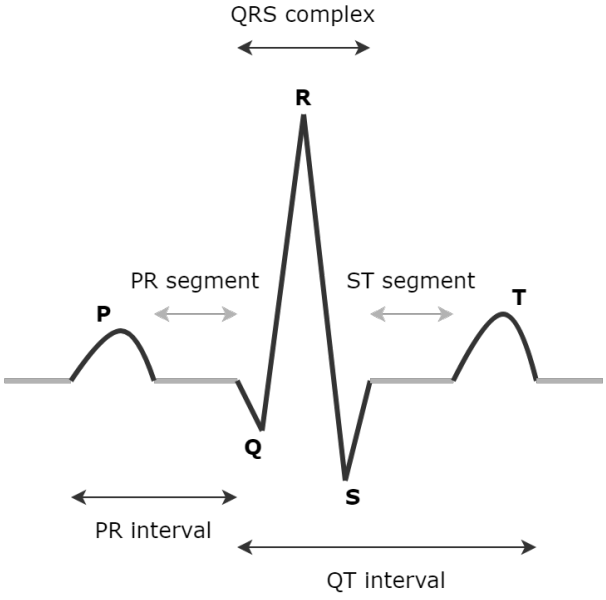


Figure 1: The usual structure of an ECG beat

*Corresponding author: cuong.dd@vinuni.edu.vn
[1] www.who.int/health-topics/cardiovascular-diseases

The 12-lead ECG, which is standard for hospital and clinic usage, is typically recorded from electrodes placed on the patient's limbs and on the surface of the chest. Thus, twelve ECG leads can be broken down into two main types: six limb leads (I, II, III, aVR, aVL, aVF) and six chest leads (V1, V2, V3, V4, V5, V6). Conventional 12-lead ECG has also been demonstrated to be effective for various ECG analysis tasks by many previous efforts [2, 3, 4]. Acquiring 12-lead ECG, on the other hand, is heavily relied on clinical equipment with limited accessibility, particularly for medical institutes in remote areas. Over the past decade, especially in recent years, breakthroughs in ECG technologies have led to the development of smaller, lower-cost, and easier-to-use ECG-enabled devices [5, 6, 7, 8]. These advancements have paved the way for point-of-care screening and continuous monitoring using signals recorded by these devices [9, 10]. However, these devices only produce a subset of standard twelve leads, sometimes even just one lead. This raises an urgent need for building ECG analysis methods that only rely on this subset of leads rather than the entire set. Beyond monitoring, developing and integrating analysis tools into these devices can also aid in the early detection of CVDs, as well as support and save time for cardiologists in their manual analysis process.

In this study, we use a combination of only three ECG leads (I, II, and V1) as input for the proposed system to strike a balance between high classification performance and ease of signal acquisition. Leads I and II are used because they are easy to acquire and favored by cardiologists for quick review. They also represent relatively enough information for six limb leads, according to some laws and equations [11]. Lead V1 is used to incorporate information about chest leads into the input. Importantly, the combination of three leads I, II, and V1 resemble an orthogonal set of leads, which can constitute all ECG leads by a good linear approximation, therefore can possibly perform similarly for the diagnosis of many cardiovascular abnormalities [12, 13].

Existing approaches for automatic ECG analysis can be divided into two categories: traditional methods and deep learning-based methods. In traditional methods, which are also known as two-stage methods, human experts hand-craft meaningful features from raw ECG signals such as statistical features (e.g., mean, standard deviation, variance, and percentile) or time- and frequency-domain features, referred to as expert features [14, 15]. Then, these features are concatenated and fed into some kinds of machine learning algorithms. The performance of these methods significantly depends on the capability of the machine learning algorithms applied and the hand-crafted feature extraction stage, which requires expertise to select optimal features. The second approach is to use end-to-end deep learning models that offer a high model capability without the need for domain knowledge and an explicit feature extraction stage [16]. These types of models have gained significant improvements compared to the former approach [4]. Deep learning models have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection, and many other areas such as drug discovery and genomics [17]. Despite their superior performance, deep learning models are plagued by two well-known drawbacks: their black-box nature and increasingly large model size which limit their applicability in real-world scenarios. In this study, we aim to design an accurate ECG classification system that also overcomes these two issues.

In almost all previous works on deep learning-based 12-lead ECG classification, all twelve leads are standardized to the same length, then vertically stacked together to form a unified input and fed into a followed deep learning model [4, 3]. This strategy works well when dealing with 12-lead ECG. However, when dealing with a smaller number of leads, such as three, we propose to use three distinct models as separate backbones to handle three input ECG leads separately, which will be demonstrated in this study to give us better performance. This multi-input strategy is reasonable since these kinds of signals usually require separate treatment. In more detail, we employ three distinct redesigned One-dimensional Squeeze-and-Excitation Residual Networks (1D-SEResNets) [18], an improved version of ResNet architecture with the Squeeze-and-Excitation modules, which are highly effective for dealing with ECG data, to extract features from three input signals. Then, inspired by the attention mechanism [19, 20], we design a novel Lead-wise Attention module as our aggregation technique to explore the most essential input lead and merge outputs of these backbones, resulting in a more robust representation that is then sent through an FC layer to perform classification.

Although deep learning models can achieve state-of-the-art performance in a range of predictive tasks, they are often viewed as black boxes. In many applications, especially in the medical domain, understanding the model's behavior is as important as the accuracy of its predictions since it is difficult for cardiologists or pathologists to accept unexplainable decisions [21]. This makes Explainable AI (XAI) become a highly active research topic in the past few years [22]. In this study, we also construct an XAI framework for our 3-lead ECG classification task using class activation maps. Our XAI technique called Lead-wise Grad-CAM provides three different class activation maps for three input ECG leads, giving more clinical interpretability to our system. Another disadvantage of deep learning models, as previously discussed, is the expansion in model size. The majority of existing ECG classification research is primarily concerned with enhancing classification performance while paying little attention to model size, leading to memory-intensive models that are impractical for hardware-constrained platforms deployment [23]. To improve the proposed system's suitability for point-of-care screening and remote monitoring deployment on these platforms, we apply a pruning technique to make the system lightweight and easy to distribute while just slightly sacrificing its performance.

To summarize, our main contributions are as follows:

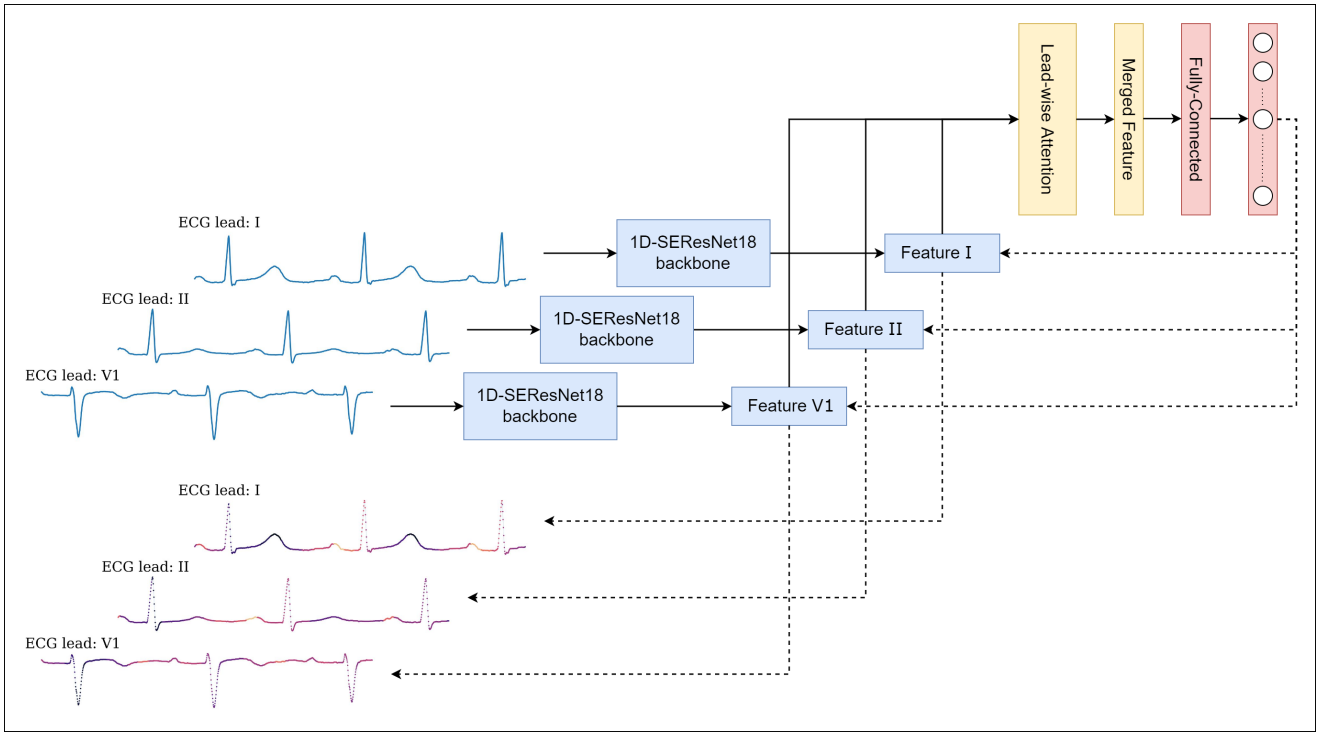


Figure 2: An overview of the proposed system. Dashed arrows indicate the interpreting stage

- We propose an accurate deep learning system for 3-lead ECG classification which consists of three redesigned 1D-SEResNet backbones followed by a novel Lead-wise Attention module and an FC layer, as shown in Figure 2.
- A novel XAI technique named Lead-wise Grad-CAM is introduced, which is adapted from the common Grad-CAM technique on the system's architecture, giving a better explanation for the made prediction.
- We further employ a pruning technique to reduce the system's space on memory while mostly preserving its classification performance.
- Extensive experiments are conducted on two large-scale multi-lead ECG datasets, i.e., Chapman and CPSC-2018 where our system shows superior performance in both multi-class and multi-label classification manners while enhancing compactness and clinical interpretability.

The rest of this article is organized as follows. Section 2 provides a survey of literature related to our work. In Section 3, we present the components of the proposed system in detail. Section 4 describes the experimental setup and our results. An ablation study is performed in Section 5. Finally, we give further discussions and conclude this work in Section 6.

2. Related Works

In this section, we discuss some research directions and existing works that are highly related to our work, including deep learning-based ECG analysis, reduced-lead ECG classification, and explainable AI for ECG classification.

Deep learning-based ECG Analysis. In the research community, deep learning-based methods have been the preferred approach for ECG analysis over the last few years [16]. Specifically, 1D-CNNs have become popular when dealing with ECG data because of their one-dimension structure. Acharya et al. [24] early developed a 9-layer 1D-CNN to identify 5 different types of cardiovascular abnormalities. Recently, researchers have begun to use more sophisticated 1D-CNN architectures, particularly ones whose 2D version achieves high image classification accuracy. For instance, Zhang et al. [25] proposed using 1D-ResNet34, Zhu et al. [26] ensembled two 1D-SEResNet34s and one set of expert rules to respectively identify 9 and 27 types of abnormalities. Furthermore, in order to capture both spatial and temporal patterns in ECG signals, Yao et al. [27] constructed Time-Incremental ResNet18 (TI-ResNet18), a combination of a 1D-ResNet18 and an LSTM network, Murugesan et al. [28] combined an Inception and an LSTM network to constitute ECGNet, identifying 9 and 3 types of abnormalities, respectively. Other than CVD detection, deep learning models have been employed on ECG data for other variety of tasks. Li et al. [29] combined a sparse Autoencoder and Hidden Markov Model for diagnosing obstructive sleep apnea. Moreover, Santamaria-Granados et al. [30] focused on emotion, classifying the affective state of a person. Attia et al. [31] performed a proof of concept study on non-invasive drug assessment based on ECG signals. Rahman et al. [32] and Özdemir et al. [33] tried to early diagnose COVID-19 using ECG trace images. Deshmane et al. [34] designed an ECG-based biometric human identification using machine learning and deep learning techniques in smart health applications.

Reduced-lead ECG Classification. In recent years, some small, low-cost, and easy-to-use ECG-enable devices with different advantages have been introduced in the market [6, 7, 8]. These devices are different from clinical equipment in that they only provide a subset of standard twelve ECG leads, sometimes just one. Thus, in most cases, newer methods are being developed to do ECG classification based on single- or reduced-lead data rather than standard 12-lead data. While single-lead ECG is currently limiting in performance, early studies have suggested that reduced-lead ECG may hold potential. Hannun et al. [35] used single-lead ECG data from Zio Patch devices to identify atrial fibrillation. Drew et al. [36] demonstrated that interpolated 12-lead ECG, which is derived from a reduced-lead set (limb leads plus V1 and V5), is comparable to standard 12-lead ECG for diagnosing wide-QRS-complex tachycardias and acute myocardial ischemia. Xue [37] used the same set of leads to evaluate the changes in morphology due to a matrix-based 12-lead conversion and the possibility of adapting the changes with the new criteria trained with a large ischemia ECG database. Green et al. [38] also found that the leads III, aVL, and V2 together yielded a similar performance as the full 12-lead ECG for diagnosing acute coronary syndrome. Cho et al. [39] claimed that myocardial infarction could be detected not only with a conventional 12-lead ECG but also with a limb 6-lead ECG. Although the potential of reduced-lead ECG was verified, there has not been much research done in this area yet. Our work provides further support to demonstrate the ability of reduced-lead ECG for identifying a wide range of cardiovascular abnormalities, not just a few.

Explainable AI for ECG Classification. While the black-box nature of deep learning models may be ignorable in many contexts, it leads to a lack of responsibility and trusts in decisions made in sensitive areas like medicine and healthcare. Hence, researchers have started to bring popular XAI techniques applied to image data into ECG data. Hughes et al. [40] proposed to use of Linear Interpretable Model-Agnostic Explanations (LIME). Zhang et al. [25], Anand et al. [41] applied SHapley Additive exPlanations (SHAP) analysis to test the interpretability of an ECG classification model. LIME and SHAP are both perturbation-based techniques that provide explanations based on the variation of output after applying perturbations to input. Some disadvantages of these techniques are combinatorial complexity explosion and producing explanations by very concrete class activation maps [42]. Due to inherent smoothing in provided explanations, some XAI techniques such as Grad-CAM [43] and its variants are recently more preferred. Vijayarangan et al. [44], Raza et al. [45] employed Grad-CAM on 1D-CNN for single-lead ECG classification. Ganeshkumar et al. [46] further applied Grad-CAM on a multi-lead circumstance but generated the same class activation map for multiple input signals. In this work, we leverage the system's architecture with a multi-input strategy and our Lead-wise Attention module to adapt Grad-CAM and provide one different informative class activation map for each of the three input leads.

3. Proposed System

In this section, we present the whole proposed system in detail. Firstly, the architecture of 1D-SEResNet backbones is described. Next, we sequentially introduce our novel Lead-wise Attention module and XAI technique, Lead-wise Grad-CAM. The pruning technique, which is used to establish LightX3ECG, is briefly discussed last. An overview of our LightX3ECG is shown in Figure 2.

3.1. 1D-SEResNet Backbones

To achieve high performance and low computational cost backbones, we redesign 1D-SEResNet18 [18], which consists of 18 main layers, in two steps as follows.

First, Convolution (Conv) layers are modified with a much larger kernel size to expand those receptive fields in order to capture longer patterns in ECG signals. This strategy has been suggested as more effective for ECG data in specific [47], and time-series data, in general, [48]. Second, we replace all of the standard Conv layers with Depth-wise Separable Conv (DSConv) layers for reducing the number of parameters of the model. Introduced in MobileNets [49, 50], DSConv splits the computation of standard Conv into two parts. The first part is depth-wise, in which each filter only convolutes each input channel. Another part is point-wise, using a 1x1 filter to combine multi-channel outputs of depth-wise layers. This design reduces the total number of parameters of our system by 80%. This architecture is used for all three backbones and is illustrated in Figure 3.

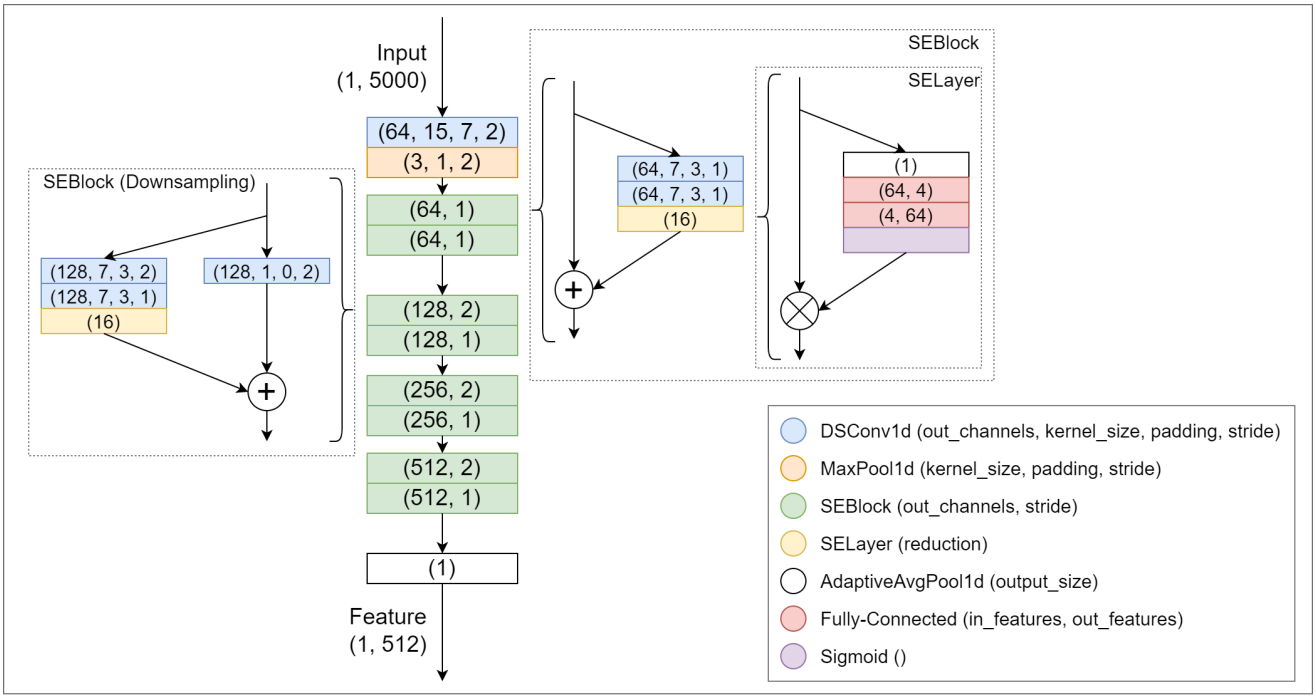
3.2. Lead-wise Attention

To achieve an end-to-end classification system, the outputs, also known as features or embeddings, extracted from backbones, must be combined. Typically, one can combine these features by simply applying a summation or concatenation operation to them, but this is usually ineffective due to their simplicity. Inspired by the success of the attention mechanism in many areas [20], we propose a Lead-wise Attention module to more effectively ensemble these features together and acquire a final robust feature which is then routed to the last FC layer, the classifier, to perform classification. Our Lead-wise Attention module is described in Figure 4.

Firstly, features from backbones are concatenated and sent through a sequential list of layers including an FC, a BatchNorm, a Dropout, followed by another FC layer and a Sigmoid function to determine the attention score, or importance score for each feature. Subsequently, the final feature is obtained by taking a weighted sum over these features by corresponding generated scores. This module can be formulated:

$$\mathrm{f}_{\text{merged}} = \sum_{i=1}^{3} \boldsymbol{\alpha}_i \mathrm{f}_i, \tag{1}$$

$$\boldsymbol{\alpha} = \texttt{Sigmoid}(\texttt{FC}(\texttt{FC}(\texttt{Concat}[\mathrm{f}_i | i = \overline{1,3}]))). \tag{2}$$



* The BatchNorm, Dropout layer, and ReLU function are ignored in the above figure for simplicity.

Figure 3: The architecture of 1D-SEResNet backbones

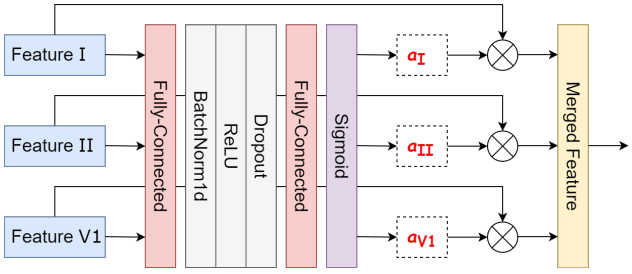


Figure 4: The proposed Lead-wise Attention module

3.3. Lead-wise Grad-CAM

A class activation map, or CAM, is a heatmap that highlights class-specific regions of an image that the model looked at to classify that image. In the domain of imaging, Grad-CAM [43] is one of the most famous techniques to provide interpretability to 2D-CNNs which uses values of gradients flowing into the final Conv layer to produce a CAM [51, 52]. In this work, we subtly adapt Grad-CAM to our system for the same aim, which we refer to as Lead-wise Grad-CAM, in the following steps.

First, similar to standard Grad-CAM, we employ values of gradients flowing into the final Conv layers of three backbones to gather three distinct CAMs $C_{i,i=\overline{1,3}}$ corresponding to three input ECG leads. In addition to CAMs provided by Grad-CAM, the proposed system has an additional source of interpretability, the importance scores $\boldsymbol{\alpha}_{i,i=\overline{1,3}}$ gathered from the Lead-wise Attention module that show the contribution of each backbone's feature to the prediction of the system, therefore, show the contribution of each input signal. To take advantage of this insight from our Lead-wise Attention module, we multiply three CAMs by corresponding importance scores to get more informative heatmaps. Finally, for visualization, these heatmaps are normalized and overlaid on corresponding input ECG lead:

$$\boldsymbol{M}_i = \texttt{normalize}(\boldsymbol{\alpha}_i \boldsymbol{C}_i) \tag{3}$$

3.4. Pruning

A deep learning-based method often involves a large model and massive computation. Hence, when operating the proposed system on portable or wearable devices, issues such as insufficient memory or computational resources are noticeable. As a direct solution, we apply the weights pruning [53, 54] technique to compress the system and make it can be executed completely on these devices.

Weights pruning is a post-training model compression technique to make a trained model more sparse. This is accomplished by increasing the number of zero-valued elements present in the model's weights. In this work, we prune 80% weights of the system with the lowest L_1-norm in order to reduce the system's space on memory 3 times while mostly maintaining its classification performance, and finally establishing LightX3ECG as a result. The idea is that weights with small L_1-norm, or absolute value, contribute little to the prediction of the system, so they are less important and can be zeroed out.

4. Experiments and Results

In this section, we comprehensively describe our study design and all experimental results. Two datasets and implementation details are introduced first, then we report the performance of LightX3ECG and its interpretability.

4.1. Datasets

To benchmark the performance of the proposed system, we conduct experiments on two of the largest public real-world datasets for ECG classification, i.e., Chapman and CPSC-2018. Diagnosis class frequency and patient characteristics of these two datasets are shown in Table 1.

Chapman [55]. Chapman University and Shaoxing People's Hospital collaborated to establish this large-scale *multi-class* ECG dataset which consists of 10.646 12-lead ECG recordings. Each recording is taken over 10 seconds with a sampling rate of 500 Hz and labeled with 11 common diagnostic classes. The amplitude unit is a microvolt. These 11 classes are grouped into 4 categories including AFIB, GSVT, SB, and SR. AFIB consists of atrial fibrillation and atrial flutter, GSVT contains supraventricular tachycardia, atrial tachycardia, atrioventricular node reentrant tachycardia, atrioventricular reentrant tachycardia, and sinus atrium to atrial wandering rhythm, SB only includes sinus bradycardia, and SR includes sinus rhythm and sinus irregularity.
CPSC-2018 [56]. In 2018, the first China Physiological Signal Challenge organized during the 7th International Conference on Biomedical Engineering and Biotechnology released a publicly available large-scale *multi-label* ECG dataset. This dataset contains 6.877 12-lead ECG recordings with a sampling rate of 500 Hz and durations ranging from 6 to 60 seconds. Millivolt is the amplitude unit. These ECG recordings are labeled with 9 diagnostic classes including NSR (normal sinus rhythm), AF (atrial fibrillation), IAVB (first-degree atrioventricular block), LBBB (left bundle branch block), RBBB (right bundle branch block), PAC (premature atrial contraction), PVC (premature ventricular contraction), STD (ST-segment depression), STE (ST-segment elevation).

Table 1: Description of two datasets
Mean and standard deviation are reported for age

Chapman			
Class	Frequency (%)	Male (%)	Age
AFIB	2225 (20.90)	1298 (58.34)	72.90 ± 11.68
GSVT	2307 (21.67)	1152 (49.93)	55.44 ± 20.49
SB	3889 (36.53)	2481 (63.80)	58.34 ± 13.95
SR	2225 (20.90)	1025 (46.07)	50.84 ± 19.25
CPSC-2018			
Class	Frequency (%)	Male (%)	Age
NSR	918 (13.35)	363 (39.54)	41.56 ± 18.45
AF	1221 (17.75)	692 (56.67)	71.47 ± 12.53
IAVB	722 (10.50)	490 (67.87)	66.97 ± 15.67
LBBB	236 (03.43)	117 (49.58)	70.48 ± 12.55
RBBB	1857 (27.00)	1203 (64.78)	62.84 ± 17.07
PAC	616 (08.96)	328 (53.25)	66.56 ± 17.71
PVC	700 (10.18)	357 (51.00)	58.37 ± 17.90
STD	869 (12.64)	252 (29.00)	54.61 ± 17.49
STE	220 (03.20)	180 (81.82)	52.32 ± 19.77

4.2. Implementation Details

To ensure the reproducibility of our results, the experimental setup is described in detail below.

Data Preprocessing: As a deep learning system requires inputs to be of the same length, all ECG recordings are fixed at 10 seconds in length in both datasets. This is done by truncating the part exceeding the first 10 seconds for longer recordings and padding shorter ones with zero. We take leads I, II, and V1 from each ECG recording to construct the input with the shape of 3x5000 and feed it into our system.
Data Augmentation: To reach a better generalization, we additionally propose the DropLead augmentation technique which randomly drops one of three input signals with a probability of 50% during training. This is accomplished by masking the selected signal with all of zero. DropLead is not applied during the inference stage.

Training and Evaluation: For evaluation, we apply a 10-fold cross-validation strategy following some previous works [55, 25]. We stratify and divide each of the two datasets into 10 folds and perform 10 rounds of training and evaluation. At each round, 8 folds; 1 fold; and 1 remaining fold are used as training, validation, and test set, respectively. In the multi-label classification manner, the optimal threshold of each class is searched in a range (0.05, 0.95) with a step of 0.05 to achieve the best F1 score on the validation set. We report the average performance of 10 rounds on the test set in terms of precision, recall, F1 score, and accuracy. For training, the proposed system is optimized from scratch by Adam optimizer [57] with an initial learning rate of 1e-3 and a weight decay of 5e-5 for 70 epochs. We use the Cosine Annealing scheduler [58] in the first 40 epochs to reschedule the learning rate to 1e-4 and then keep it constant in the last 30 epochs. Cross-entropy and binary cross-entropy are utilized as loss functions in multi-class and multi-label manners, respectively. Finally, after weights pruning is applied, our system is fine-tuned for 5 epochs with the same setting except the learning rate is held constant at 1e-4. All experiments are run on a machine with an NVIDIA GeForce RTX 3090 TURBO 24G.

4.3. System Performance

We get F1 scores of 0.9718 and 0.8004 on two datasets, i.e., Chapman and CPSC-2018, respectively. Overall, accuracy for each class exceeds 0.92 and the average exceeds 0.96 in both. However, we also observe that F1 scores of PAC and STE classes are limited, which could be due to the insufficiency of these diagnosis classes in the CPSC-2018 dataset. Detailed performance is presented in Table 2.

Table 2: Performance detail of LightX3ECG on two datasets

Chapman				
Class	Precision	Recall	F1 score	Accuracy
AFIB	0.9750	0.9662	0.9706	0.9878
GSVT	0.9510	0.9612	0.9561	0.9807
SB	0.9823	0.9987	0.9904	0.9930
SR	0.9860	0.9550	0.9703	0.9878
Average	0.9736	0.9703	0.9718	0.9873
CPSC-2018				
Class	Precision	Recall	F1 score	Accuracy
NSR	0.6903	0.8342	0.7554	0.9266
AF	0.9344	0.9461	0.9402	0.9789
IAVB	0.9014	0.8828	0.8920	0.9775
LBBB	0.9038	0.8704	0.8868	0.9913
RBBB	0.9454	0.9428	0.9441	0.9702
PAC	0.6972	0.5758	0.6307	0.9353
PVC	0.8796	0.7197	0.7917	0.9637
STD	0.7870	0.7824	0.7847	0.9469
STE	0.6486	0.5217	0.5783	0.9746
Average	0.8209	0.7862	0.8004	0.9628

For benchmarking, we compare LightX3ECG with some popular ECG classification methods, which can be considered state-of-the-art including 1D-ResNet34 [25], 1D-SEResNet34 [26], TI-ResNet18 [27], InceptionTime [48], and ECGNet [28]. For fair comparisons, all of these methods are implemented and trained using 3-lead ECG as input and settings similar to our system. Comparisons of F1 scores, complexity, and compactness are shown in Table 3. LightX3ECG outperforms other methods in both datasets while achieving the lowest computational cost with FLOPs at 1.34B. In terms of storage, our system only takes up 6.52MB on disk, which is much less than the other three methods. Additionally, the performance of the system without applying weights pruning shows that effectively using this technique helps reduce the system's space significantly with a negligible side-effect.

4.4. System Interpretability

A comprehensive validation is conducted to demonstrate LightX3ECG's interpretability, including a visual check and a methodical check.

4.4.1. Visual examinations

For visual check, we carefully review the explanation from the system for a sample ECG recording, drawn from the CPSC-2018 dataset, belonging to each of the diagnosis classes and compare it with some cardiological evidence collected from a variety of sources [59, 60, 61, 62, 63, 64] and the LITFL ECG Library [65].

1) NSR (normal sinus rhythm). An NSR ECG recording has a normal P wave preceding each QRS complex, which is also standard, as seen in Figure 1. Also, P waves upright in leads I and II. From activation maps in Figure 5, we can see that system strongly focuses on regions of P waves in leads I and II. Thus, the explanation is consistent with the diagnostic criteria of NSR. The importance scores indicate that lead I contributed more to the system's prediction than others.

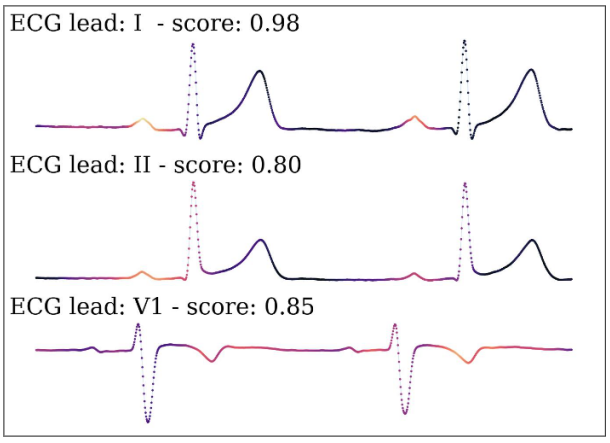


Figure 5: The explanation for a sample NSR ECG recording

2) AF (atrial fibrillation). An AF ECG recording has irregular QRS complexes with the lack of P waves. Also, fibrillatory waves are usually visible in lead V1. From activation maps in Figure 6, we can see that system recognizes the lack of P waves in leads I and II, and fibrillatory waves in lead V1. Thus, the explanation is consistent with the diagnostic criteria of AF. The importance scores indicate that three leads contributed roughly equally to the system's prediction.

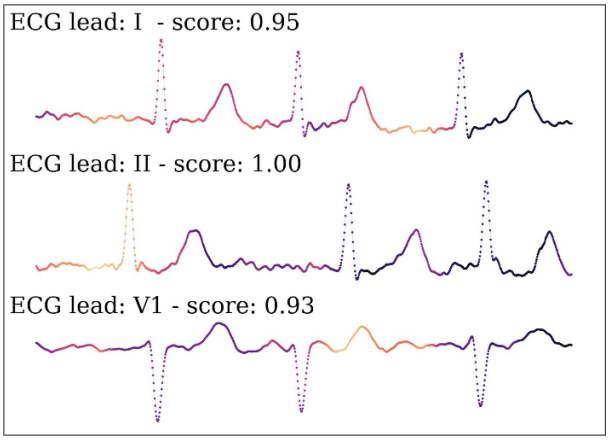


Figure 6: The explanation for a sample AF ECG recording

3) IAVB (first-degree atrioventricular block). An IAVB ECG recording has prolonged PR intervals. Also, P waves are buried in the preceding T wave. From activation maps in Figure 7, we can see that system recognizes the prolonged PR intervals in leads I and II. Thus, the explanation is consistent with the diagnostic criteria of IAVB. The importance scores indicate that three leads contributed roughly equally to the system's prediction.

Table 3: Comparison of the proposed system to other methods

Method	F1 on Chapman	F1 on CPSC-2018	No. Params (M)	No. FLOPs (B)	Size (MB)
1D-ResNet34 [25]	0.9624	0.7684	16.61	5.91	58.18
1D-SEResNet34 [26]	0.9659	0.7845	16.76	5.91	58.75
TI-ResNet18 [27]	0.9647	0.7872	11.39	1.42	40.51
InceptionTime [48]	0.9417	0.7352	**0.45**	2.29	**1.63**
ECGNet [28]	0.9652	0.7880	1.03	1.97	3.75
LightX3ECG (Ours)	**0.9718**	**0.8004**	5.31	**1.34**	6.52
LightX3ECG (w/o pruning)	0.9722	0.8010	5.31	1.34	19.28

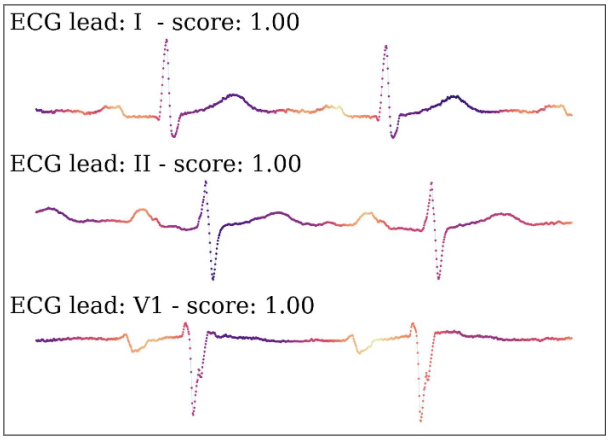


Figure 7: The explanation for a sample IAVB ECG recording

4) LBBB (left bundle branch block). An LBBB ECG recording has broad QRS complexes. Also, S waves are fairly deep in lead V1. From activation maps in Figure 8, we can see that system recognizes broad QRS complexes in lead I, and deep S waves in lead V1. Thus, the explanation is consistent with the diagnostic criteria of LBBB. The importance scores indicate that leads I and V1 mostly contributed to the system's prediction.

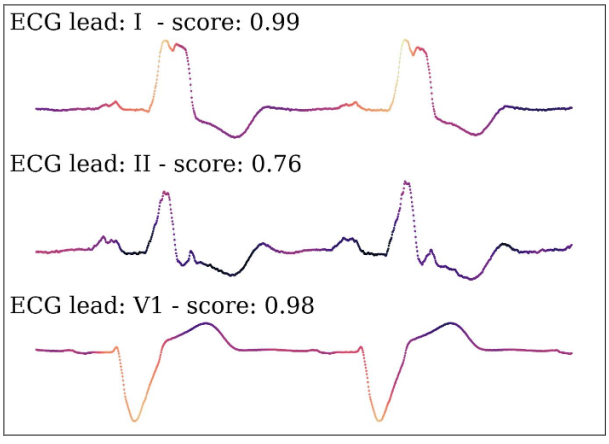


Figure 8: The explanation for a sample LBBB ECG recording

5) RBBB (right bundle branch block). An RBBB ECG recording has wide slur S waves in lead I. Also, "M-shaped" QRS complexes are visible in lead V1. From activation maps in Figure 9, we can see that system recognizes wide slur S waves in lead I, and "M-shaped" QRS complexes in lead V1. Thus, the explanation is consistent with the diagnostic criteria of RBBB. The importance scores indicate that leads I and V1 mostly contributed to the system's prediction.

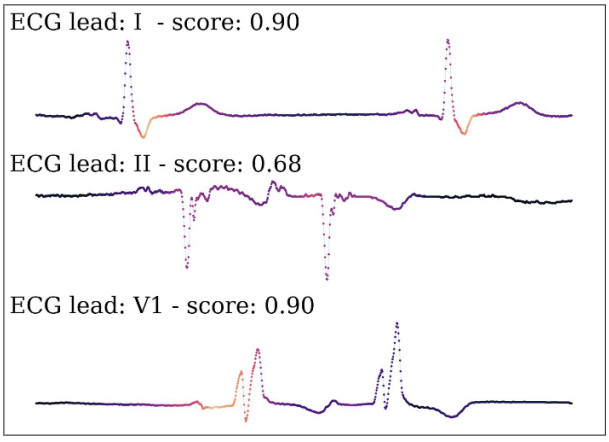


Figure 9: The explanation for a sample RBBB ECG recording

6) PAC (premature atrial contraction). A PAC ECG recording has abnormal (non-sinus) P waves followed by a normal QRS complex. Also, P waves are usually negative in lead II. From activation maps in Figure 10, we can see that system recognizes non-sinus P waves in leads II and V1, specifically negative P waves in lead II. Thus, the explanation is consistent with the diagnostic criteria of PAC. The importance scores indicate that leads II and V1 mostly contributed to the system's prediction.

7) PVC (premature ventricular contraction). A PVC ECG recording has some sporadic periods that are abnormal compared to surrounding periods. Also, QRS complexes in these periods are irregular too. From activation maps in Figure 11, we can see that system recognizes abnormal periods compared to surrounding periods in lead II and irregular QRS complexes in these periods. Thus, the explanation is consistent with the diagnostic criteria of PVC. The importance scores indicate that lead II contributed more to the system's prediction than others.

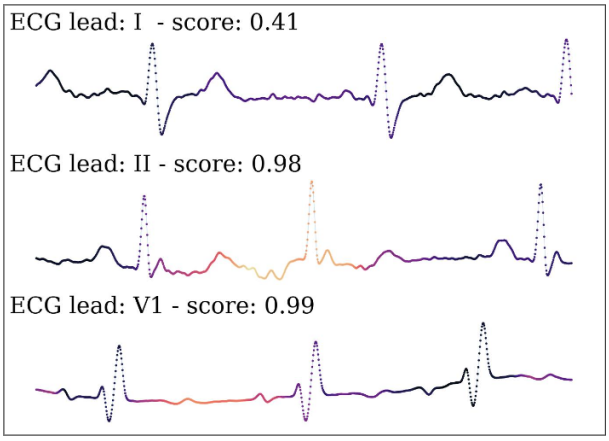


Figure 10: The explanation for a sample PAC ECG recording

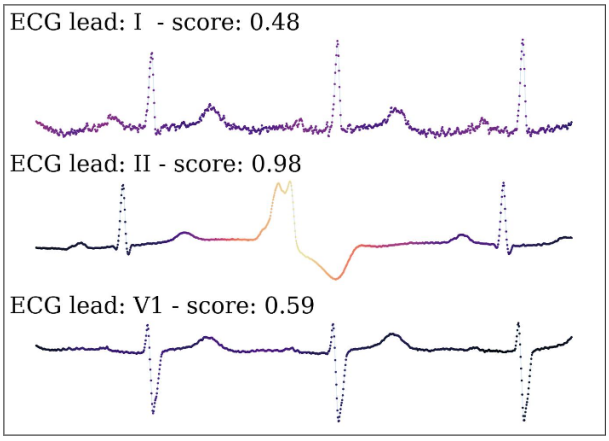


Figure 11: The explanation for a sample PVC ECG recording

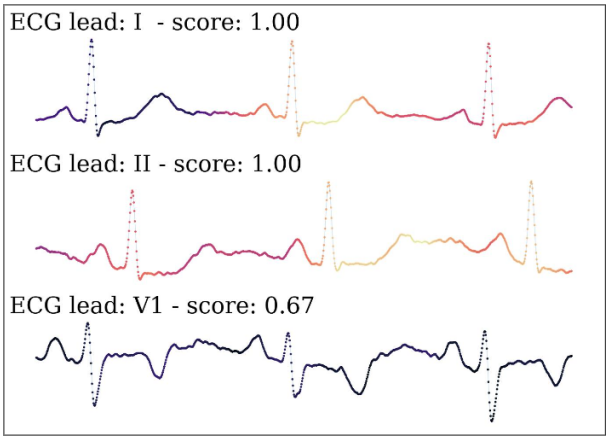


Figure 12: The explanation for a sample STD ECG recording

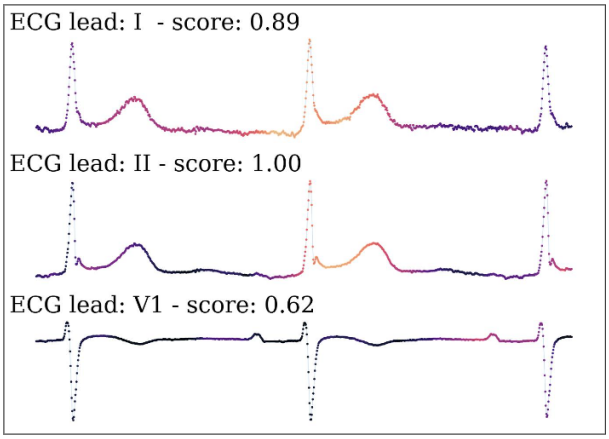


Figure 13: The explanation for a sample STE ECG recording

8) STD (ST-segment depression). As its name, an STD ECG recording has depressed ST segments. From activation maps in Figure 12, we can see that system recognizes depressed ST segments in leads I and II. Thus, the explanation is consistent with the diagnostic criteria of STD. The importance scores indicate that leads I and II mostly contributed to the system's prediction.

9) STE (ST-segment elevation). As its name, an STE ECG recording has elevated ST segments. From activation maps in Figure 13, we can see that system recognizes elevated ST segments in leads I and II. Thus, the explanation is consistent with the diagnostic criteria of STE. The importance scores indicate that leads I and II mostly contributed to the system's prediction.

4.4.2. Sanity check

Recent works in the literature on XAI research have strongly emphasized the importance of implementing sanity checks [66] in order to assess the quality of XAI techniques methodically [67, 68]. These types of checks verify whether or not the provided explanation is related to the model's parameters or the data used for training, hence, evaluating whether an XAI technique is suitable to deploy or not.

For this purpose, we perform a simple parameter randomization test, which is one of two forms of sanity checks, to assess our Lead-wise GradCAM technique. In particular, by using Lead-wise GradCAM, we compare explanations for a hundred ECG recordings from our trained system (original system) with those from the randomized system. the randomized system is accomplished by randomly reinitializing the final FC layer, classifier, of the original system. Figure 14 shows an example of this comparison, as we expect, explanations differ. We also report the average Spearman's rank correlation of these explanations in Table 4. Lead-wise GradCAM and SHAP analysis [25] both pass this sanity check, but our technique gives a lower correlation score.

Table 4: Spearman's rank correlation of explanations between the original system and randomized system

Method	Chapman	CPSC-2018
SHAP [25]	0.16	0.18
Lead-wise Grad-CAM (Ours)	**0.10**	**0.11**

Table 5: Comparison with different chest leads on the performance of LightX3ECG in terms of F1 scores

Dataset	(I, II, and V1)	(I, II, and V2)	(I, II, and V3)	(I, II, and V4)	(I, II, and V5)	(I, II, and V6)
Chapman	**0.9718**	0.9702	0.9705	0.9702	0.9714	0.9711
CPSC-2018	**0.8004**	0.8002	0.7997	0.7959	0.8001	0.7992

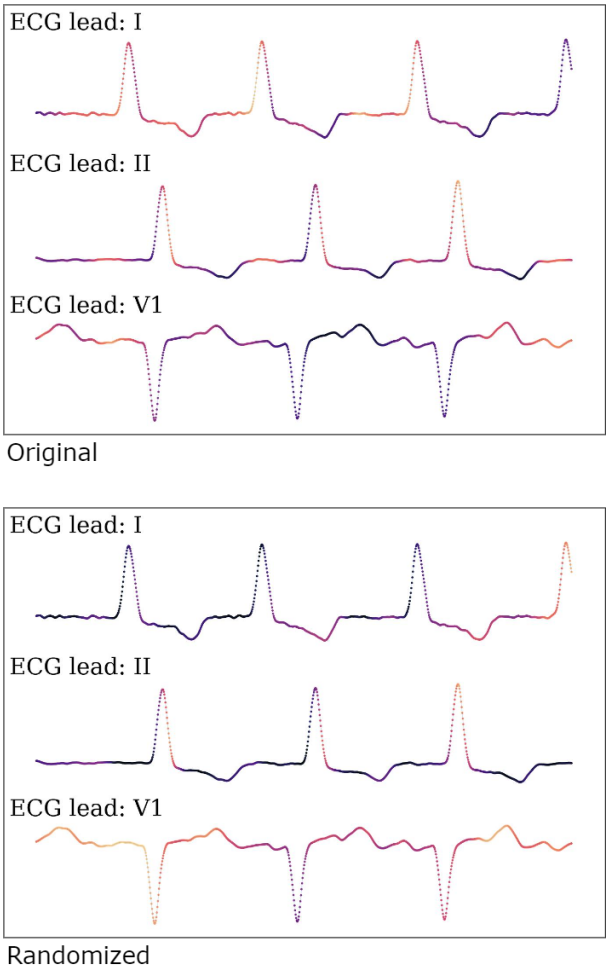


Figure 14: An example of a comparison between explanations from the original system and the randomized system

5. Ablation Studies

In this section, we conduct two types of ablation studies to validate the effect of three selected input leads and explore the contribution of the Lead-wise Attention module on the performance of LightX3ECG.

5.1. Chest Lead Substitution

In particular, we fix leads I and II as our limb leads while substituting lead V1 with another chest lead to create a new orthogonal combination of three leads that we use as input for the system. Table 5 shows that the performance is fairly consistent among combinations and the combination of leads (I, II, and V1) produces the best performance in both datasets by a slight margin.

5.2. Lead-wise Attention Analysis

We conduct a thorough investigation of the proposed Lead-wise Attention (Att) module in order to verify two crucial questions. First, how much does the module improve the overall performance of the system? Second, what is the effect of the module when it is integrated into other networks?

To address the first question, we respectively replace the Lead-wise Attention module with two simple, common operators that are feature averaging and concatenation to show the change in the system's performance. We can observe from Table 6 that the proposed module significantly contributes to the overall system compared to the two mentioned operators, making it surpass other networks.

Table 6: Contribution of the Lead-wise Attention module to the whole system

Operator/Module	F1 on Chapman	F1 on CPSC-2018
Averaging	0.9683	0.7846
Concatenation	0.9694	0.7881
Lead-wise Attention	**0.9722**	**0.8010**

For the second question, we conduct experiments by integrating the Lead-wise Attention module into other networks described in Subsection 4.3 including TI-ResNet18 [27], InceptionTime [48], and ECGNet [28], then compare the results. From Table 7, we can observe that the module consistently gives boosts to the performance of these networks.

Table 7: Integration of Lead-wise Attention module into other networks

Network	F1 on Chapman		F1 on CPSC-2018	
	w/o Att	w/ Att	w/o Att	w/ Att
TI-ResNet18	0.9647	**0.9698**	0.7872	**0.7902**
InceptionTime	0.9417	**0.9438**	0.7352	**0.7412**
ECGNet	0.9652	**0.9703**	0.7880	**0.7916**

6. Discussions

After illustrating the proposed system, experimental results, and ablation analysis, we further provide comprehensive discussions to explain the strengths and weaknesses of our system below.

When compared to previous works on the same topic, our system performs competitively on diagnosis and interpretation. The following reasons may contribute to improvements: (i) We propose to process three input ECG leads separately by three distinct backbones. Also, the backbone architecture is designed to be efficient for ECG signals. (ii) The Lead-wise Attention module is the most important component of the system, which significantly contributes to the system's better performance. (ii) Based on the attention module, the XAI technique can provide a lead-wise explanation and explore the most important lead contributing to the system's prediction. However, there are still some drawbacks to the proposed system: (i) The system might miss important information when using reduced-lead ECG data, resulting in the impossibility to detect certain types of cardiovascular abnormalities. (ii) The multi-input architecture of LightX3ECG is not suitable for small-scale datasets and leads to difficulty in training, as well as a high storage cost which needs a practical technique like weights pruning to compensate.

7. Conclusion

In this article, we introduce an efficient and accurate deep learning system that uses an orthogonal set of three 10-second ECG leads (I, II, and V1) to identify cardiovascular abnormalities. We pose a new state-of-the-art for the 3-lead ECG classification task, where the proposed system outperforms most of the existing methods available for ECG classification in terms of F1 scores, complexity, and compactness. Additionally, we focus heavily on the XAI framework, which can give a more meaningful and clinical explanation for the system's prediction, making it more valuable in medical contexts. Our system is also compressed to be ready for the production stage. Moreover, our source code is made available to the public to encourage further development [2]. In the future, LightX3ECG will be improved to identify wider varieties of cardiovascular abnormalities, as well as be generalized on different sources of data. Demographic data such as age and gender will be incorporated to boost current performance. And a novel XAI framework for this multi-modal input will be also developed.

Acknowledgment

The authors would like to thank the VinUni-Illinois Smart Health Center (VISHC) and the source of funding: VinUni Seed Grant 2020. Besides, this work was also funded by Vingroup Joint Stock Company (Vingroup JSC), and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2021.DA00128.

References

[1] Wikipedia. "Electrocardiography". https://en.wikipedia.org/wiki/Electrocardiography, 2020.

[2] Erick A. Perez Alday, Annie Gu, Amit J. Shah, Chad Robichaux, An Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Qiao Li, Ashish Sharma, Gari D. Clifford, and Matthew A. Reyna. Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020. *Physiological Measurement*, 41, 2020.

[3] Junsang Park, Junho An, Jinkook Kim, Sunghoon Jung, Yeongjoon Gil, Yoojin Jang, Kwanglo Lee, and Il young Oh. Study on the Use of Standard 12-Lead ECG Data for Rhythm-Type ECG Classification Problems. *Computer Methods and Programs in Biomedicine*, 214, 2022.

[4] Shenda Hong, Yuxi Zhou, Junyuan Shang, Cao Xiao, and Jimeng Sun. Opportunities and Challenges of Deep Learning Methods for Electrocardiogram Data: A Systematic Review. *Computers in Biology and Medicine*, 122, 2020.

[5] Abigail C. Teron, Pedro A. Rivera, and Miguel A. Goenaga. ECG Holter Monitor With Alert System and Mobile Application. In *Signal Processing, Sensor/Information Fusion, and Target Recognition XXV*, volume 9842, 2016.

[6] Mintu P. Turakhia, Donald D. Hoang, Peter Zimetbaum, Jared D. Miller, Victor F. Froelicher, Uday N. Kumar, Xiangyan Xu, Felix Yang, and Paul A. Heidenreich. Diagnostic Utility of a Novel Leadless Arrhythmia Monitoring Device. *American Journal of Cardiology*, 112, 2013.

[7] John S. Chorba, Avi M. Shapiro, Le Le, John Maidens, John Prince, and Steve Pham et al. Deep Learning Algorithm for Automated Cardiac Murmur Detection via a Digital Stethoscope Platform. *Journal of the American Heart Association*, 10, 2021.

[8] Gregory M. Marcus. The Apple Watch Can Detect Atrial Fibrillation: So What Now? *Nature Reviews Cardiology*, 17, 2020.

[9] Patrik Bachtiger, Camille F. Petri, Francesca E. Scott, Se Ri Park, Mihir A. Kelshiker, and Harpreet K. Sahemey et al. Point-Of-Care Screening for Heart Failure With Reduced Ejection Fraction Using Artificial Intelligence During ECG-Enabled Stethoscope Examination in London, UK: A Prospective, Observational, Multicentre Study. *The Lancet Digital Health*, 4, 2022.

[10] Shu Li Guo, Li Na Han, Hong Wei Liu, Quan Jin Si, De Feng Kong, and Fu Su Guo. The Future of Remote ECG Monitoring Systems. *Journal of Geriatric Cardiology*, 13, 2016.

[11] John G. Webster. *The Physiological Measurement Handbook*. CRC Press, 2014.

[12] Ch L. Levkov. Orthogonal Electrocardiogram Derived From the Limb and Chest Electrodes of the Conventional 12-Lead System. *Medical & Biological Engineering & Computing*, 25, 1987.

[13] Sidharth Maheshwari, Amit Acharyya, Michele Schiariti, and Paolo Emilio Puddu. Frank Vectorcardiographic System From Standard 12 Lead ECG: An Effort to Enhance Cardiovascular Diagnosis. *Journal of Electrocardiology*, 49, 2016.

[14] S. Karpagachelvi, M. Arthanari, and M. Sivakumar. ECG Feature Extraction Techniques - A Survey Approach. *International Journal of Computer Science and Information Security*, 8, 2010.

[15] Shanti Chandra, Ambalika Sharma, and Girish Kumar Singh. Feature Extraction of ECG Signal. In *Journal of Medical Engineering and Technology*, volume 42, 2018.

[16] Sulaiman Somani, Adam J. Russak, Felix Richter, Shan Zhao, Akhil Vaid, Fayzan Chaudhry, Jessica K. De Freitas, Nidhi Naik, Riccardo Miotto, Girish N. Nadkarni, Jagat Narula, Edgar Argulian, and Benjamin S. Glicksberg. Deep Learning and the Electrocardiogram: Review of the Current State-Of-The-Art. *Europace*, 23, 2021.

[17] Y. G. LeCun, Y. Bengio, and Hinton. Deep Learning. Nature 521 (7553): 436. *Nature*, 521, 2015.

[18] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-Excitation Networks. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 42, 2020.

[19] Minh Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-Based Neural Machine Translation. In *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 2015.

[20] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. An Attentive Survey of Attention Models. *ACM Transactions*

[2] https://github.com/lhkhiem28/LightX3ECG

on Intelligent Systems and Technology, 12, 2021.

[21] Shinjini Kundu. AI in Medicine Must Be Explainable. *Nature Medicine*, 27, 2021.

[22] Filip Karlo Dosilovic, Mario Brcic, and Nikica Hlupic. Explainable Artificial Intelligence: A Survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings*, 2018.

[23] Yingchun Wang, Jingyi Wang, Weizhan Zhang, Yufeng Zhan, Song Guo, Qinghua Zheng, and Xuanyu Wang. A Survey on Deploying Mobile Deep Learning Applications: A Systemic and Technical Perspective. *Digital Communications and Networks*, 8, 2022.

[24] U. Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Muhammad Adam, Arkadiusz Gertych, and Ru San Tan. A Deep Convolutional Neural Network Model to Classify Heartbeats. *Computers in Biology and Medicine*, 89, 2017.

[25] Dongdong Zhang, Samuel Yang, Xiaohui Yuan, and Ping Zhang. Interpretable Deep Learning for Automatic Diagnosis of 12-Lead Electrocardiogram. *iScience*, 24, 2021.

[26] Zhaowei Zhu, Xiang Lan, Tingting Zhao, Yangming Guo, Pipin Kojodjojo, Zhuoyang Xu, Zhuo Liu, Siqi Liu, Han Wang, Xingzhi Sun, and Mengling Feng. Identification of 27 Abnormalities From Multi-Lead ECG Signals: An Ensembled Se-Resnet Framework With Sign Loss Function. *Physiological Measurement*, 42, 2021.

[27] Qihang Yao, Ruxin Wang, Xiaomao Fan, Jikui Liu, and Ye Li. Multi-class Arrhythmia detection from 12-lead varied-length ECG using Attention-based Time-Incremental Convolutional Neural Network. In *Information Fusion*, volume 53, 2020.

[28] Balamurali Murugesan, Vignesh Ravichandran, Keerthi Ram, S. P. Preejith, Jayaraj Joseph, Sharath M. Shankaranarayana, and Mohanasankar Sivaprakasam. ECGNet: Deep Network for Arrhythmia Classification. In *MeMeA 2018 - 2018 IEEE International Symposium on Medical Measurements and Applications, Proceedings*, 2018.

[29] Kunyang Li, Weifeng Pan, Yifan Li, Qing Jiang, and Guanzheng Liu. A Method to Detect Sleep Apnea Based on Deep Neural Network and Hidden Markov Model Using Single-Lead ECG Signal. *Neurocomputing*, 294, 2018.

[30] Luz Santamaria-Granados, Mario Munoz-Organero, Gustavo Ramirez-Gonzalez, Enas Abdulhay, and N. Arunkumar. Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS). *IEEE Access*, 7, 2019.

[31] Zachi I. Attia, Alan Sugrue, Samuel J. Asirvatham, Michael J. Ackerman, Suraj Kapa, Paul A. Friedman, and Peter A. Noseworthy. Noninvasive Assessment of Dofetilide Plasma Concentration Using a Deep Learning (Neural Network) Analysis of the Surface Electrocardiogram: A Proof of Concept Study. *PLoS ONE*, 13, 2018.

[32] Tawsifur Rahman, Alex Akinbi, Muhammad E.H. Chowdhury, Tarik A. Rashid, Abdulkadir Şengür, Amith Khandakar, Khandaker Reajul Islam, and Aras M. Ismael. Cov-ECGnet: COVID-19 Detection Using ECG Trace Images With Deep Convolutional Neural Network. *Health Information Science and Systems*, 10, 2022.

[33] Mehmet Akif Ozdemir, Gizem Dilara Ozdemir, and Onan Guren. Classification of COVID-19 Electrocardiograms by Using Hexaxial Feature Mapping and Deep Learning. *BMC Medical Informatics and Decision Making*, 21, 2021.

[34] Madhura Deshmane and Swati Madhe. ECG Based Biometric Human Identification Using Convolutional Neural Network in Smart Health Applications. In *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*, 2018.

[35] Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network. *Nature Medicine*, 25, 2019.

[36] Barbara J. Drew, Michele M. Pelter, Donald E. Brodnick, Anil V. Yadav, Debbie Dempel, and Mary G. Adams. Comparison of a New Reduced Lead Set ECG With the Standard ECG for Diagnosing Cardiac Arrhythmias and Myocardial Ischemia. *Journal of Electrocardiology*, 35, 2002.

[37] J. Q. Xue. Adapting ECG Morphology Changes From Reduced-Lead Set by Specifically Trained Algorithms for Acute Ischemia Detection. In *Computers in Cardiology*, volume 34, 2007.

[38] Michael Green, Mattias Ohlsson, Jakob Lundager Forberg, Jonas Björk, Lars Edenbrandt, and Ulf Ekelund. Best Leads in the Standard Electrocardiogram for the Emergency Detection of Acute Coronary Syndrome. *Journal of Electrocardiology*, 40, 2007.

[39] Younghoon Cho, Joon myoung Kwon, Kyung Hee Kim, Jose R. Medina-Inojosa, Ki Hyun Jeon, Soohyun Cho, Soo Youn Lee, Jinsik Park, and Byung Hee Oh. Artificial Intelligence Algorithm for Detecting Myocardial Infarction Using Six-Lead Electrocardiography. *Scientific Reports*, 10, 2020.

[40] J. Weston Hughes, Jeffrey E. Olgin, Robert Avram, Sean A. Abreau, Taylor Sittler, Kaahan Radia, Henry Hsia, Tomos Walters, Byron Lee, Joseph E. Gonzalez, and Geoffrey H. Tison. Performance of a Convolutional Neural Network and Explainability Technique for 12-Lead Electrocardiogram Interpretation. *JAMA Cardiology*, 6, 2021.

[41] Atul Anand, Tushar Kadian, Manu Kumar Shetty, and Anubha Gupta. Explainable AI Decision Model for ECG Data of Cardiac Disorders. *Biomedical Signal Processing and Control*, 75, 2022.

[42] Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-Based Methods for Explaining Deep Neural Networks: A Survey. *Pattern Recognition Letters*, 150, 2021.

[43] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *International Journal of Computer Vision*, volume 128, 2020.

[44] Sricharan Vijayarangan, Balamurali Murugesan, R. Vignesh, S. P. Preejith, Jayaraj Joseph, and Mohansankar Sivaprakasam. Interpreting Deep Neural Networks for Single-Lead ECG Arrhythmia Classification. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, volume 2020-July, 2020.

[45] Ali Raza, Kim Phuc Tran, Ludovic Koehl, and Shujun Li. Designing ECG Monitoring Healthcare System With Federated Transfer Learning and Explainable AI. *Knowledge-Based Systems*, 236, 2022.

[46] Ganeshkumar M., Vinayakumar Ravi, Sowmya V, Gopalakrishnan E.A, and Soman K.P. Explainable Deep Learning-Based Approach for Multilabel Classification of Electrocardiogram. *IEEE Transactions on Engineering Management*, 2021.

[47] Jinyong Cheng, Qingxu Zou, and Yunxiang Zhao. ECG Signal Classification Based on Deep CNN and Bilstm. *BMC Medical Informatics and Decision Making*, 21, 2021.

[48] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre Alain Muller, and François Petitjean. Inceptiontime: Finding Alexnet for Time Series Classification. *Data Mining and Knowledge Discovery*, 34, 2020.

[49] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[50] Andrew Howard, Mark Sandler, Bo Chen, Weijun Wang, Liang Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Quoc Le, and Hartwig Adam. Searching for MobileNetV3. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-October, 2019.

[51] Harsh Panwar, P. K. Gupta, Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, Prakhar Bhardwaj, and Vaishnavi Singh. A Deep Learning and Grad-Cam Based Color Visualization Approach for Fast Detection of COVID-19 Cases Using Chest X-Ray and Ct-Scan Images. *Chaos, Solitons and Fractals*, 140, 2020.

[52] Sivaramakrishnan Rajaraman, Kamolrat Silamut, Md. A. Hossain, I. Ersoy, Richard J. Maude, Stefan Jaeger, George R. Thoma, and Sameer K. Antani. Understanding the Learned Behavior of Customized Convolutional Neural Networks Toward Malaria Parasite Detection in Thin Blood Smear Images. *Journal of Medical Imaging*, 5, 2018.

[53] Miguel A. Carreira-Perpiñán and Yerlan Idelbayev. 'Learning-Compression' Algorithms for Neural Net Pruning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.
[54] Song Han, Jeff Pool, John Tran, and William J. Dally. Learning Both Weights and Connections for Efficient Neural Networks. In *Advances in Neural Information Processing Systems*, volume 2015-January, 2015.
[55] Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-Lead Electrocardiogram Database for Arrhythmia Research Covering More Than 10,000 Patients. *Scientific Data*, 7, 2020.
[56] Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, Jianqing Li, and Eddie Ng Yin Kwee. An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection. *Journal of Medical Imaging and Health Informatics*, 8, 2018.
[57] Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.
[58] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic Gradient Descent With Warm Restarts. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2017.
[59] Cecilia Gutierrez and Daniel G. Blanchard. Atrial Fibrillation: Diagnosis and Treatment. *American Family Physician*, 83, 2011.
[60] Borys Surawicz and Timothy K. Knilans. *Chou's Electrocardiography in Clinical Practice: Adult and Pediatric*. W.B. Saunders, 2008.
[61] S. Serge Barold, Arzu Ilercil, Fabio Leonelli, and Bengt Herweg. First-Degree Atrioventricular Block: Clinical Manifestations, Indications for Pacing, Pacemaker Management & Consequences During Cardiac Resynchronization. *Journal of Interventional Cardiac Electrophysiology*, 17, 2006.
[62] Ary L. Goldberger, Zachary D. Goldberger, and Alexei Shvilkin. *Goldberger's Clinical Electrocardiography: A Simplified Approach: Ninth Edition*. Elsevier, 2017.
[63] M. Alventosa-Zaidin, L. Guix Font, M. Benitez Camps, C. Roca Saumell, G. Pera, M. Teresa Alzamora Sas, R. Forés Raurell, O. Rebagliato Nadal, A. Dalfó-Baqué, and J. Brugada Terradellas. Right Bundle Branch Block: Prevalence, Incidence, and Cardiovascular Morbidity and Mortality in the General Population. *European Journal of General Practice*, 25, 2019.
[64] Borys Surawicz, Rory Childers, Barbara J. Deal, and Leonard S. Gettes. AHA/ACCF/HRS Recommendations for the Standardization and Interpretation of the Electrocardiogram. *Circulation*, 119, 2009.
[65] LITFL. "ECG Library". https://litfl.com, 2022.
[66] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems*, volume 2018-December, 2018.
[67] Finale Doshi-Velez and Been Kim. A Roadmap for a Rigorous Science of Interpretability. *arXiv preprint arXiv:1702.08608v1*, 2017.
[68] Chih Kuan Yeh, Cheng Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (In)Fidelity and Sensitivity of Explanations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A. Appendix

In this section, we provide supplementary figures for Subsection 4.4, including more examples for visual examinations, related to Figures 5-13, and more examples for the sanity check, related to Figure 14.

A.1. Visual examinations

1) NSR (normal sinus rhythm)

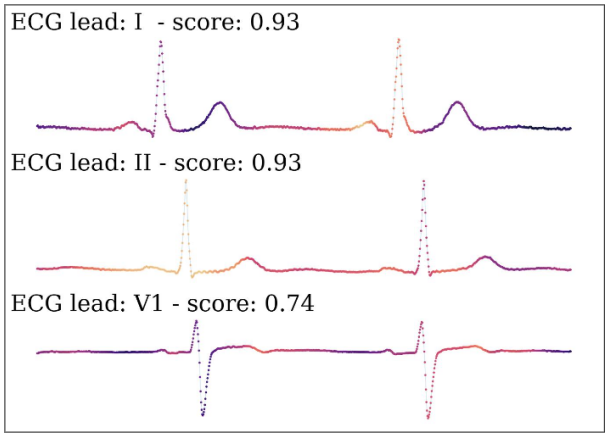


Figure 15: The explanation for another NSR ECG recording

2) AF (atrial fibrillation)

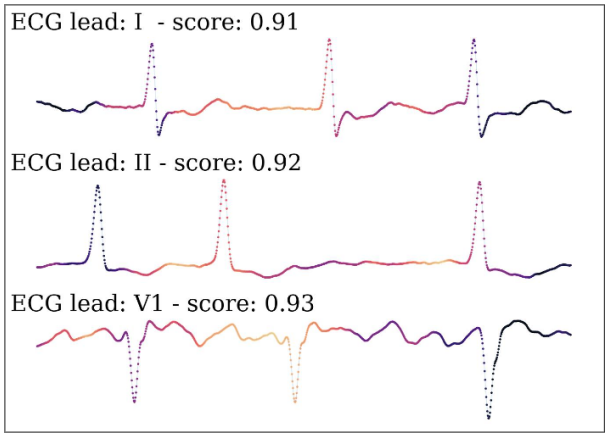


Figure 16: The explanation for another AF ECG recording

3) IAVB (first-degree atrioventricular block)

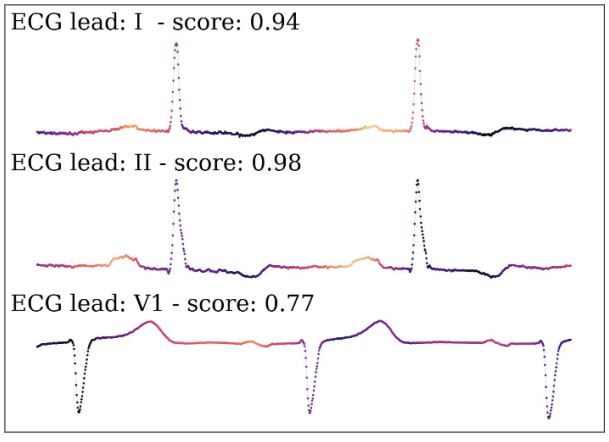


Figure 17: The explanation for another IAVB ECG recording

4) LBBB (left bundle branch block)

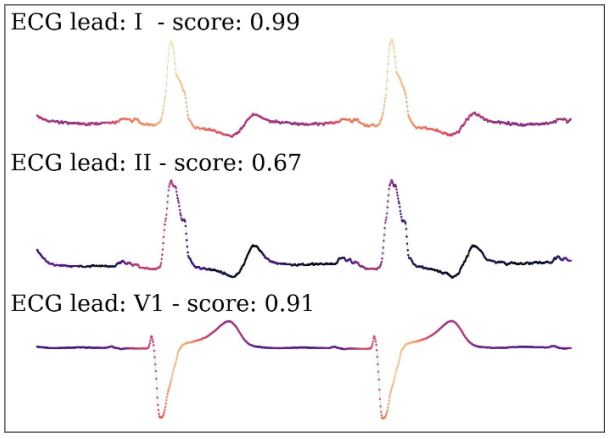


Figure 18: The explanation for another LBBB ECG recording

5) RBBB (right bundle branch block)

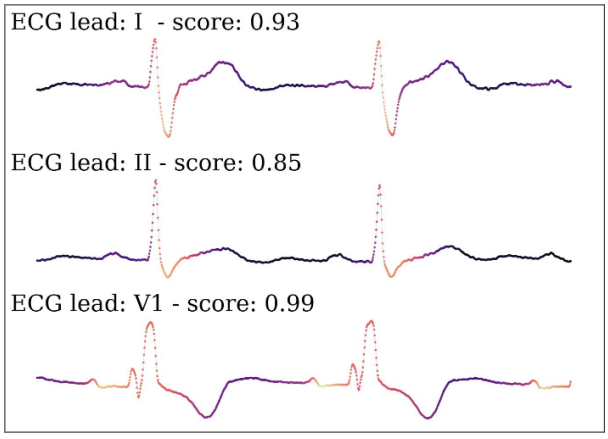


Figure 19: The explanation for another RBBB ECG recording

6) PAC (premature atrial contraction)

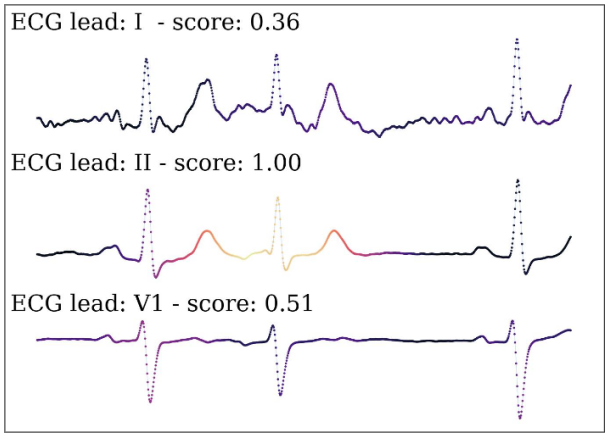


Figure 20: The explanation for another PAC ECG recording

7) PVC (premature ventricular contraction)

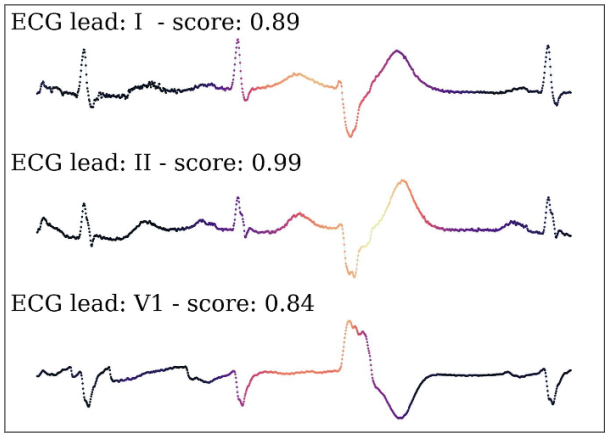


Figure 21: The explanation for another PVC ECG recording

8) STD (ST-segment depression)

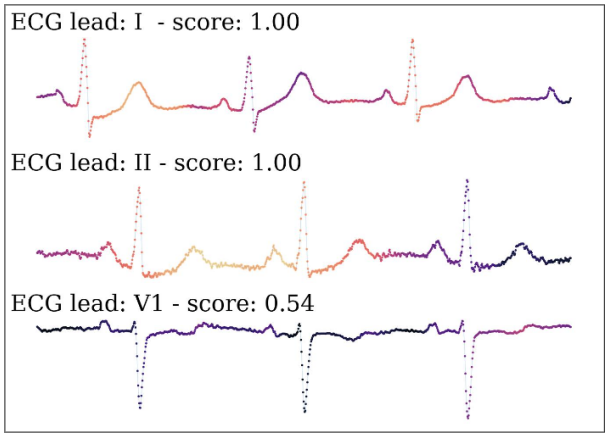


Figure 22: The explanation for another STD ECG recording

9) STE (ST-segment elevation)

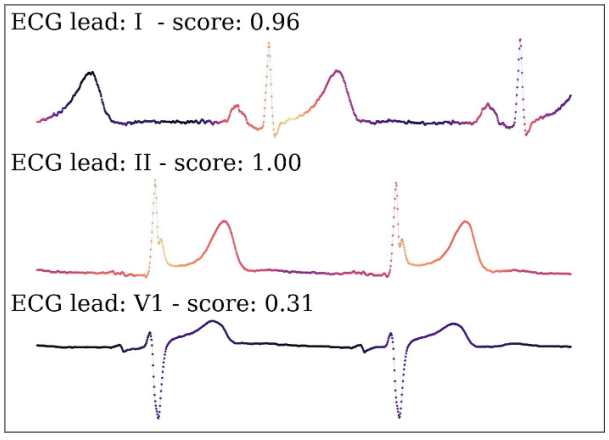


Figure 23: The explanation for another STE ECG recording

A.2. Sanity check

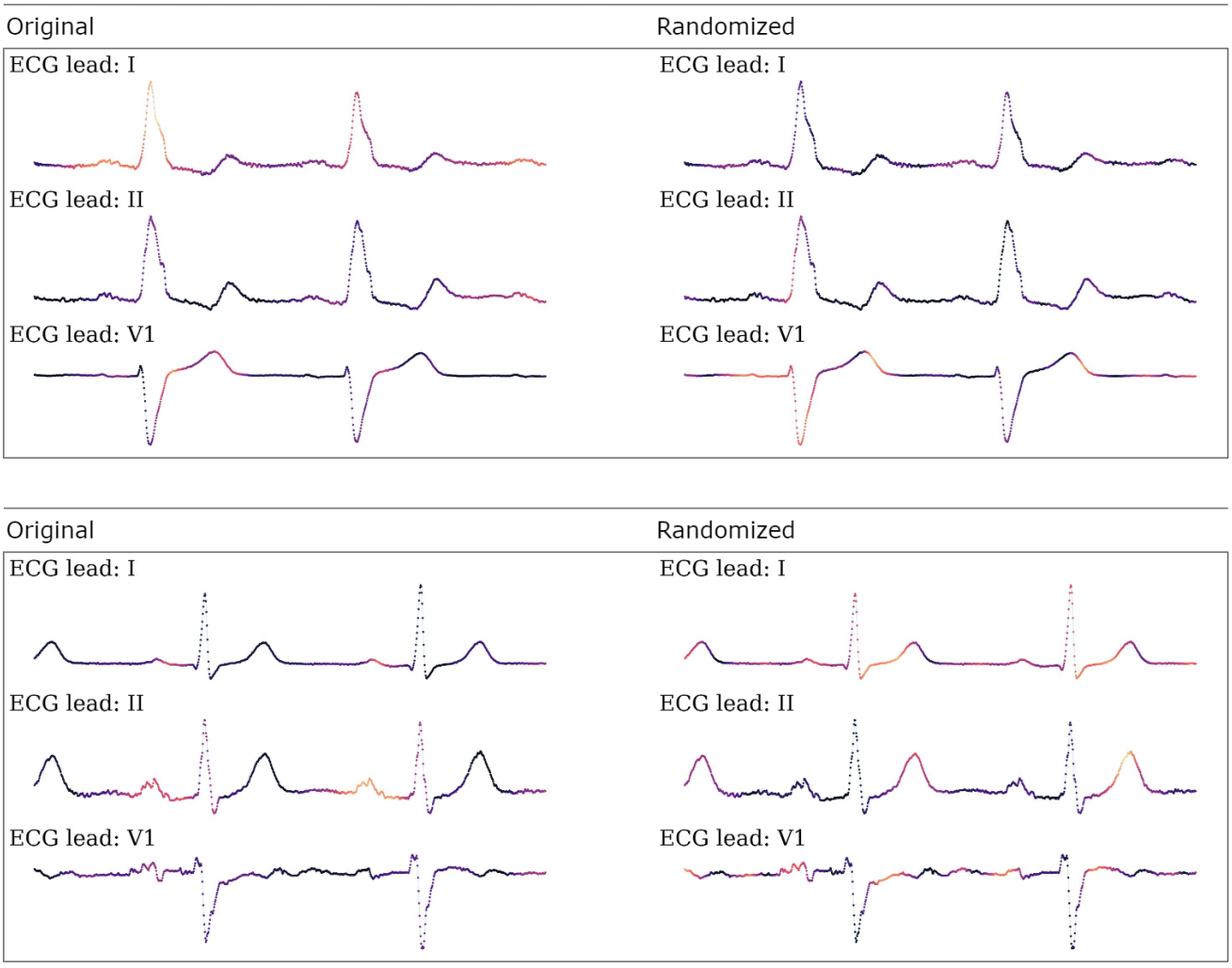


Figure 24: Some examples of comparison between explanations from the original system and the randomized system