Integrating learnable expert knowledge into deep learning-based multi-label ECG classification

OPEN ACCESS
IOP Publishing
Measurement Science and Technology
Meas. Sci. Technol. 36 (2025) 056113 (15pp)
https://doi.org/10.1088/1361-6501/adcce7

Integrating learnable expert knowledge into deep learning-based multi-label ECG classification

Qiao Xiao[1,2], **Yue Li**[1], **Khuan Lee**[3], **Siti Aisah Mokhtar**[2], **Iskasymar Ismail**[4], **Ahmad Luqman bin Md Pauzi**[4] and **Poh Ying Lim**[2,*]

[1] School of Computer Science, University of South China, Hengyang, Hunan, People's Republic of China
[2] Department of Community Health, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
[3] Department of Nursing and Rehabilitation, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
[4] Department of Medicine, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia

E-mail: pohying_my@upm.edu.my

Received 17 December 2024, revised 4 April 2025
Accepted for publication 15 April 2025
Published 24 April 2025

CrossMark

Abstract
Deep learning (DL) has shown great promise in electrocardiogram (ECG) analysis, revolutionizing cardiovascular medicine by enabling precise and efficient diagnosis. This study explores the integration of expert knowledge into DL models, creating a flexible structure that adapts during training to refine this knowledge and guides feature separation in higher-dimensional space, thereby improving multi-label ECG classification. By leveraging domain expertise on the relationships between specific ECG abnormalities and their corresponding changes across lead dimensions, a lead-wise prior knowledge framework (LPKF) was introduced to enhance the learning efficiency of DL models. The effectiveness of this framework was validated by the higher classification performance of LPKF-enhanced models compared to their original versions. In addition, the LPKF-enhanced Inception model outperforms recent state-of-the-art DL methods, underscoring the benefits of integrating learnable expert knowledge. The study also demonstrated the interpretability of the LPKF-enhanced Inception model using gradient-weighted class activation mapping, revealing its capability to identify crucial diagnostic symptoms from ECG signals that align with clinical criteria.

Keywords: ECG diagnosis, multi-label classification, deep learning, learnable expert knowledge

* Author to whom any correspondence should be addressed.

1. Introduction

Cardiovascular diseases, including coronary heart disease and cerebrovascular disease, are the leading cause of death worldwide, posing significant public health challenges. In the United States alone, heart disease claimed approximately 695 000 lives in 2021, accounting for one in every five deaths [1]. The electrocardiogram (ECG), a non-invasive diagnostic tool, records the electrical activity of heart and provides critical insights into cardiac function [2]. It plays a vital role in helping physicians diagnose conditions and develop effective treatment plans.

As heart disease prevalence rises, so does the demand for accurate and efficient ECG classification, driving the need for automated ECG analysis to alleviate the workload of medical professionals. Advances in artificial intelligence (AI) have led to the widespread adoption of deep learning (DL) techniques for ECG classification [3]. As most studies focus on binary or multi-class ECG classification, where each ECG record is assigned to a single and mutually exclusive category [8], these approaches have achieved classification accuracy rates consistently above 90% [4–6]. Certain DL models have even outperformed human cardiologists in ECG classification tasks [7]. However, clinical practice often presents a more complex scenario in which ECG patterns frequently correspond to multiple coexisting diagnoses, requiring additional effort for accurate diagnosis.

For multi-label classification, overlapping features and subtle differences among diagnostic categories pose significant challenges for feature separation in the feature domain [10]. Ensuring well-separated features for one category may inadvertently lead to considerable overlap from the perspective of another category due to differences in classification boundaries [9]. Unlike binary and multi-class classification, where features are assigned to mutually exclusive categories, DL models for multi-label classification must maintain feature separability across multiple categories simultaneously.

To address this, existing approaches primarily focus on enhancing classification performance by employing deeper and more complex DL architectures with extensive learnable parameters [17, 18] to improve feature separability. While increased model complexity can refine feature extraction, it does not fundamentally resolve the challenge of simultaneously maintaining clear classification boundaries across multiple overlapping categories. Consequently, multi-label ECG classification performance remains significantly lower than that of binary and multi-class tasks. For instance, reported F1 scores for binary and multi-class classification reach as high as 0.992 for atrial fibrillation detection [11] and 0.958 for arrhythmia classification [12]. In contrast, multi-label ECG classification studies [13–16] report F1 scores ranging from 0.651 to 0.887, underscoring the inherent difficulty in achieving high classification accuracy in multi-label scenarios. This performance gap further complicates the development of reliable treatment plans for patients with coexisting cardiac conditions.

Incorporating expert knowledge from medical professionals presents an alternative and potentially more effective strategy for enhancing feature extraction [19]. In [20], cardiologists assigned attention weights to ECG leads based on their diagnostic relevance, allowing the DL model to focus on the most informative leads, thereby improving tachyarrhythmia recognition. By integrating expert knowledge into feature extraction along the lead dimension, an additional degree of freedom is introduced, potentially enhancing feature separation in the feature domain. However, while this approach acknowledges the value of expert knowledge, it relies on a fixed attention structure during training, leading to rigid category separation. As a result, its impact on final classification boundaries may be limited compared to other learnable parameters.

This study investigates the integration of expert knowledge into DL models as an intrinsic structure to enhance learning capability. This expert knowledge, regarded as prior information, is derived from ECG diagnostic principles used by cardiologists, defining correlations between ECG leads and specific diagnostic categories. It could help guide the generation of more distinct features for multiple categories simultaneously. To the best of our knowledge, this is the first attempt to incorporate expert knowledge as a learnable component within DL networks to improve multi-label ECG classification performance.

The contributions of this study can be summarized as follows.

- This study proposes a lead-wise prior knowledge framework (LPKF) that integrates expert insights to improve the ability of DL models to focus on ECG leads associated with specific diagnostic categories. The proposed framework incorporates trainable expert knowledge, allowing dynamic refinement of attention weights for ECG leads and guiding the transformation of features across diagnostic categories into a higher-dimensional space for improved separability.
- The proposed LPKF consists of one branch that integrates expert knowledge to derive logits for each category separately, while another maps features to logits for all categories simultaneously. The logits from both branches are merged to facilitate mutual learning between them, thereby further distinguishing features across different categories.
- The effectiveness of the proposed LPKF is validated through extensive experiments on the publicly available ECG databases, which includes comparisons to state-of-the-art (SOTA) models, thorough ablation studies, and model interpretability analysis through gradient-weighted class activation mapping (GradCAM).

2. Related works

In the early stages of multi-label ECG classification, DL models such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their combinations are widely employed to leverage their feature extraction capabilities [21, 22]. For instance, in [23], multiple SOTA DL-based time series classification models utilizing CNNs and RNNs are evaluated on the PTB-XL database [24], establishing a benchmark for performance assessment. These models achieve areas under the receiver operating characteristic curves (AUCs) exceeding 0.900, demonstrating strong inference capability regardless of classification thresholds. However, when evaluated using specific thresholds, their average F1 scores remain around 0.7 across different multi-label classification tasks, indicating the need for further improvement. These studies primarily focus on leveraging the feature extraction capabilities of established DL architectures for ECG classification.

To enhance DL model performance, many approaches incorporate specifically designed structures to form hybrid models with diverse learning mechanisms, such as embedding layers, attention mechanisms, and graph convolutional networks (GCNs) [25, 26]. In [27], 12-lead ECG signals are first transformed into 3D features using recurrence plot transformation to better capture their recurrent behavior. Subsequently, 3DECG-Net, a model combining CNNs, LSTMs, and attention mechanisms, is designed to process these 3D features, achieving an F1-score of 0.803 on the Georgia database. In [28], a channel-wise recurrent fusion module is proposed to exploit the correlation between ECG leads by learning dependencies across the lead dimension, mimicking the temporal dependency modeling capability of LSTM. Additionally, CNNs and an attention mechanism are employed to extract features across temporal and channel domains. However, DL models with hybrid learning structures often contain a large number of learnable parameters, making them computationally expensive. Furthermore, their specialized architectures may limit adaptability, restricting integration with new learning structures as AI techniques continue to evolve.

Some studies explore pre-trained learning structures to provide additional prior knowledge for improving DL model training. In [17], an encoder consisting of CNNs and a decoder incorporating bidirectional LSTMs, both pre-trained, are used to extract semantic information from different categories. This is followed by a GCN, which captures label dependencies and exploits category correlations to enhance classification performance. In [29], twelve parallel networks are pre-trained to extract features from individual ECG leads and generate soft labels for each lead. The soft labels from the most informative ECG leads serve as additional information, which is combined with those from another model to improve classification performance. However, pre-trained models rely on the assumption that the datasets used for pre-training and fine-tuning share similar data distributions and identical classification categories, which may not always hold in real clinical practice.

In this study, the proposed DL structure is designed for adaptability, allowing the integration of various DL models as its backbone, thereby facilitating direct incorporation of newly developed architectures. Moreover, the proposed model is guided by expert knowledge, enabling adaptation to different scenarios without suffering from performance degradation due to data discrepancies encountered during pre-training.

3. Materials and methods

The overall workflow of this study is illustrated in figure 1. First, the raw ECG signals undergo preprocessing to standardize the data by scaling inputs and segment the signals for label aggregation. Next, the LPKF is developed, which includes devising lead-wise attention weights based on diagnostic knowledge, proposing two learning branches, and feature integration within the classifier. Following this, DL models are trained on the preprocessed training data. Finally, the trained models are evaluated on test data to measure how effectively they integrate domain knowledge into the DL models.

The architecture of the proposed lead-wise prior knowledge framework (LPKF) framework is illustrated in figure 2. We introduce the lead-wise prior knowledge module (LPKM), which incorporates expert knowledge to extract features from input signals within the label-level branch. Additionally, the original ECG signals are processed through a temporal attention mechanism in the global-level branch to enhance the model's focus on time-domain features. The classification results from both branches, after feature extraction by a backbone network, are subsequently merged using a specifically designed classifier to generate the final predictions. Detailed descriptions of the LPKM, temporal attention mechanism, and classifier are provided in the following subsections.

3.1. Encoding expert knowledge

Signals from different ECG leads could contribute to a specific diagnostic category with varying degrees. For instance, anteroseptal myocardial infarction (AMI) could be identified by specific changes in the heart's electrical activity patterns in leads corresponding to the anteroseptal region, often seen in leads V1–V3 [30]. The profound diagnostic insights provided by cardiology professionals shed light on the intricate relationships between various diagnostic statements and specific ECG leads.

This study first proposes a lead-wise prior knowledge attention (LPKA) which is denoted by $\mathbf{P}$, as presented in figure 3. It has the size of $C \times L$ where C and L corresponds to the numbers of categories and ECG leads in the classification task, respectively. The (c, l)th element in the attention matrix $\mathbf{P}$, i.e. $P_{c,l}$, is assigned a value of 1 to indicate the presence of an association between the cth diagnostic category and the lth lead, whereas a value of 0 denotes the absence of such an association.

The value of $P_{c,l}$ in the proposed LPKA is initially derived from established diagnostic criteria [31–33], which clearly indicates whether a diagnostic category is associated with a specific ECG lead. For instance, according to [31], incomplete left bundle branch block (ILBBB) is diagnosed based on the following criteria:

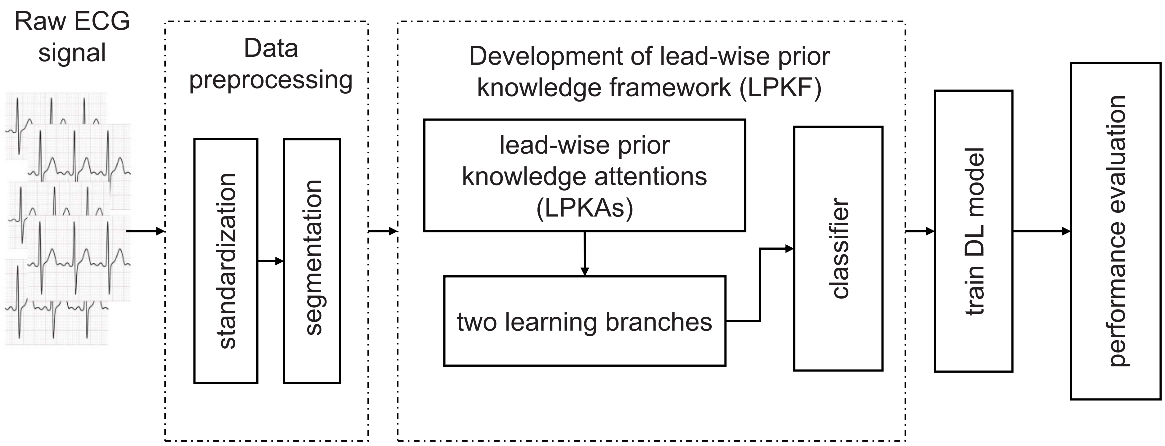

Figure 1. The overall workflow of this study.

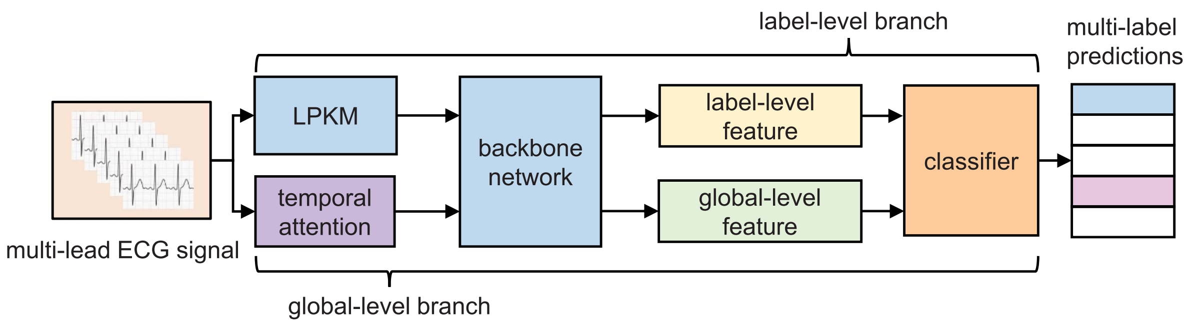

Figure 2. The overview of the proposed LPKF architecture.

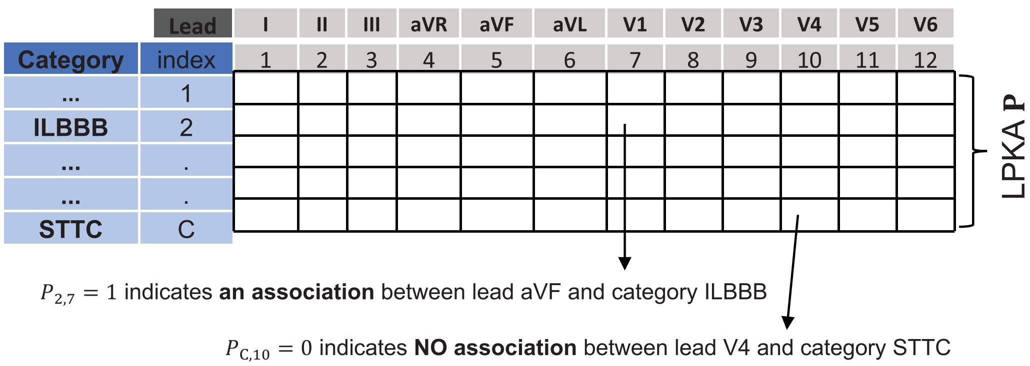

Figure 3. An illustration of the LPKA matrix $\mathbf{P}$ and the association between the leads and categories indicated by $P_{c,l}$.

(i) QRS duration between 110 and 119 ms in adults, between 90 and 100 ms in children aged 8 to 16 years, and between 80 and 90 ms in children younger than 8 years.
(ii) Presence of left ventricular hypertrophy pattern.
(iii) R peak time greater than 60 ms in leads *V4, V5, and V6*.
(iv) Absence of the Q wave in leads *I, V5, and V6*.

Based on these criteria, it is evident that ILBBB is associated with leads I, V1, V4, V5, and V6. Therefore, the attention weights of these leads for this category can be set to 1. That is, according to figure 3, the expert knowledge from the criteria can be encoded into the LPKA matrix $\mathbf{P}$ as $P_{2,1} = 1$, $P_{2,7} = 1$, $P_{2,10} = 1$, $P_{2,11} = 1$, and $P_{2,12} = 1$.

Furthermore, certain diagnostic categories comprise multiple finer-grained subcategories. For example, conduction disturbance (CD) includes eight subclasses (e.g. ILBBB, IRBBB, etc). Denote $\mathcal{D}_i = \{1, \cdots, N_{\mathcal{D}_i}\}$ as the ith higher-level hierarchical category containing $N_{\mathcal{D}_i}$ sub-categories. To determine the attention weight for the lth lead in category $\mathcal{D}_i$, it is first computed as

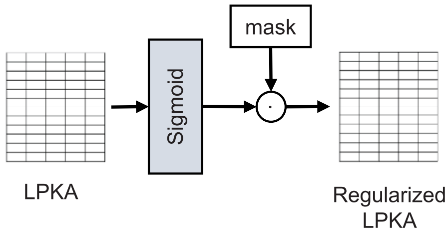

Figure 4. Process to obtain regularized LPKA.

$$\tilde{P}_{il} = \sum_{c \in \mathcal{D}_i} P_{c,l} \tag{1}$$

where $\tilde{P}_{il}$ is the attention weight of the lth lead for ith higher-level hierarchical category. To ensure proper weighting, the attention values are then so that their sum across all the leads equals unity by

$$\tilde{P}_{il} \leftarrow \frac{\tilde{P}_{il}}{\sum_l \tilde{P}_{il}}. \tag{2}$$

In this way, the attention weights for higher-level hierarchical categories are appropriately derived from their respective subcategories.

Following this approach, the attention weights for all leads corresponding to specific diagnostic categories can be systematically determined. Subsequently, these associations represented by the LPKA are verified by professional cardiologists, thereby integrating expert knowledge into the model via the LPKA matrix. For generality, in the following discussion, irrespective of hierarchical categories, we use $\mathbf{P}$ and P_{ij} to denote the LPKA matrix and its (i,j)th element, respectively.

As shown figure 4, the LPKA matrix $\mathbf{P}$ then undergoes a process to obtain the regularized LPKA, which can be formulated as

$$\mathbf{P}_{\mathrm{R}} = \mathrm{sigmoid}\left(\mathbf{P}\right) \odot \mathbf{M}, \tag{3}$$

where sigmoid is the sigmoid activation function, $\odot$ represents the element-wise multiplication, and $\mathbf{M}$ is a mask matrix with its the (i,j)th element defined as

$$M_{ij} = \begin{cases} 1 & P_{ij} \neq 0, \\ 0 & P_{ij} = 0. \end{cases} \tag{4}$$

In the following sections, the regularized LPKA will be integrated into the DL model to guide the model's focus toward ECG leads associated with specific diagnostic categories. In this study, as we incorporate the attention matrix $\mathbf{P}$ as a learnable part of the DL network, the elements P_{ij} would be dynamically updated during training. This regularization process enables non-zero weights for relevant ECG leads to be refined within the range of 0 to 1, helping further reflect the intensity of associations rather than a binary indication from the original expert knowledge.

Additionally, due to the masking process, attention weights for ECG leads unrelated to specific labels will remain at 0. Thus, during training, ECG leads with zero attention weights will not influence the classification loss, allowing the model to focus exclusively on leads with meaningful attention weights.

3.2. Feature extraction with expert knowledge

The feature extraction process based on the expert knowledge is illustrated in figure 5. Denote $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_L]$ as preprocessd ECG signals where $\mathbf{x}_\ell$ corresponds to the ℓth-lead signal. Denote M as the length of signal $\mathbf{x}_\ell$. The LPKM helps extract features from ECG signals for each statement label by exploiting the regularized LPKA. Specifically, the outer production is first performed between the preprocessd ECG signals and the regularized LPKA, which can be described by

$$\mathbf{out} = \mathbf{X} \otimes \mathbf{P}_{\mathrm{R}}. \tag{5}$$

It can be presented in details as

$$\mathbf{out} = \begin{pmatrix} \mathbf{x}_1 P_{11}^{(\mathrm{R})} & \mathbf{x}_2 P_{12}^{(\mathrm{R})} & \cdots & \mathbf{x}_L P_{1L}^{(\mathrm{R})} \\ \mathbf{x}_1 P_{21}^{(\mathrm{R})} & \mathbf{x}_2 P_{22}^{(\mathrm{R})} & \cdots & \mathbf{x}_L P_{2L}^{(\mathrm{R})} \\ \cdots & \cdots & \ddots & \cdots \\ \mathbf{x}_1 P_{C1}^{(\mathrm{R})} & \mathbf{x}_2 P_{C2}^{(\mathrm{R})} & \cdots & \mathbf{x}_L P_{CL}^{(\mathrm{R})} \end{pmatrix} \tag{6}$$

where $P_{ij}^{(R)}$ is the (i,j)th element in $\mathbf{P}_{\mathrm{R}}$. Each row in $\mathbf{out}$ essentially corresponds to the feature map that with lead-wise attention focused for each diagnostic category. In this way, for each feature map, the ECG leads correlated with a specific diagnostic category are retained effectively as input sources for the DL model, while the remaining leads are ignored during the learning process due to their attention weights being zero. The output matrix $\mathbf{out}$ is then reshaped to the size of $M \times L \times C$ which consists of C feature maps where each corresponds to one category. In this way, by integrating expert knowledge with the outer product process, the dimensionality of the features is increased while being guided by expert knowledge, which potentially aids the separation of features along category dimensions.

Denote $\mathbf{out}_c$ as the slice of the cth channel (along the lead dimension) in $\mathbf{out}$ which corresponds to the cth feature map containing L ECG leads and M signal point samples. Each feature map $\mathbf{out}_c$ essentially belongs to the cth ECG label with the lead-wise attention applied to the input multi-channel ECG signal accordingly. Then, each feature map undergoes through a channel-wise temporal attention (TA) process where each channel in $\mathbf{out}_c$ can be obtained as

$$\mathrm{TA}\left(\mathbf{out}_c\right) = \mathrm{sigmoid}(\mathrm{conv1d}([\mathrm{AvgPool}\left(\mathbf{out}_c\right); \mathrm{MaxPool}\left(\mathbf{out}_c\right)])) \odot \mathbf{out}_c \tag{7}$$

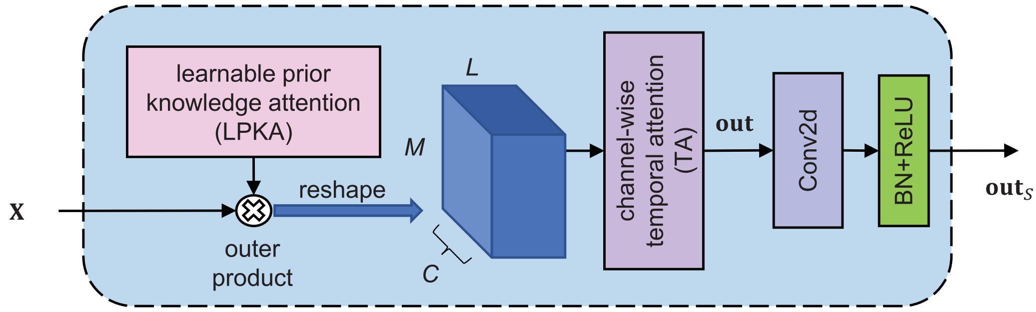

Figure 5. The structure of LPKM.

where AvgPool$(\cdot)$ and MaxPool$(\cdot)$ corresponds to the average and max pooling operations, respectively, applied to the lead dimension while preserving the temporal dimension; conv1d represents a 1D convolutional layer with a kernel size of 7; sigmoid transforms the temporal attention values to the range from 0 to 1. The resulting outcome of TA$(\mathbf{out}_c)$ has the size of $M \times L$ and c can be from 1 to C. The temporal attention process could adaptively capture temporal patterns and highlight relevant temporal features for further processing.

The features obtained from (7) for all C labels with temporal attention focused on the lead dimension are then concatenated together as

$$\mathbf{out} \leftarrow \mathrm{concat}\left(\mathrm{TA}\left(\mathbf{out}_1\right), \cdots, \mathrm{TA}\left(\mathbf{out}_C\right)\right) \tag{8}$$

where the concatenation is performed into the second dimension to form a new feature array with the size of $M \times C \times L$.

However, they potentially contained redundant information as specific ECG leads may contribute to the diagnosis of multiple labels. Hence, to improve learning efficiency by reducing the redundancy in these features, a 2D convolutional filter with a kernel size of $(1, L)$ is applied to fuse the label dimension. This operation is followed by batch normalization (BN) layer and a rectified linear unit (ReLU) activation function. The whole process can be described by

$$\mathbf{out}_S = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{conv2d}\left(\mathbf{out}\right)\right)\right) \tag{9}$$

where $\mathbf{out}_S$ is the label-level feature with a size of $M \times L$ which contains information with greater attention allocated to ECG leads according to specific labels based on the LPKA.

3.3. Integration of label- and global-level branches

Besides the label-level feature, a global-level feature, acquired by passing the original input ECG signals through a TA directly, is obtained by

$$\mathbf{out}_G = \mathrm{TA}\left(\mathbf{X}\right). \tag{10}$$

It extracts feature generally without considering label-dependent diagnostic criterion.

Furthermore, in the proposed framework, a backbone model f is utilized for further feature extraction and performs on both of the label- and global-level features individually as

$$\mathbf{out}'_S = f(\mathbf{out}_S) \tag{11}$$

$$\mathbf{out}'_G = f(\mathbf{out}_G). \tag{12}$$

The backbone model could be selected from popular classification models, with its classification layer omitted. In other words, only the feature extraction capability of the backbone model is utilized for this study.

The output label-level and global-level features from the backbone model, i.e. $\mathbf{out}'_S$ and $\mathbf{out}'_G$, are passed into a classifier, as shown in figure 6. In the global-level branch, the feature undergoes a linear layer followed by a BN layer and a dropout function, is mapped into a logit vector with the size of $C \times 1$ by another linear layer. The process to obtain the global-level logits can be described as follows

$$\mathbf{L}_G = \mathrm{linear}\left(\mathrm{dropout}\left(\mathrm{BN}\left(\mathrm{linear}\left(\mathbf{out}'_G\right)\right)\right)\right) \tag{13}$$

where the linear layer has the size of output sample C and $\mathbf{L}_G$ contains C logits, each corresponding to one label. In this way, the global-level feature is utilized to obtain the logits for all the labels simultaneously. Inspired by the label decoupling proposed in [25], in the label-level branch, unlike the global-level branch, the mapping from the final feature to one logit for each label is established individually. The logit for the ith label can be obtained by

$$l_i = \mathrm{linear}\left(\mathrm{dropout}\left(\mathrm{BN}\left(\mathrm{linear}\left(\mathbf{out}'_S\right)\right)\right)\right) \tag{14}$$

where each linear layer has the size of output sample 1. Then, the C logits are concatenated to for the label-level logits as

$$\mathbf{L}_S = \left[l_1, l_2, l_C\right]. \tag{15}$$

Finally, the two logits are fused by two learnable weights to obtain the final logits for prediction as

$$\mathbf{L}_F = \mathrm{sigmoid}\left(W_1\right) * \mathbf{L}_S + \mathrm{sigmoid}\left(W_2\right) * \mathbf{L}_G \tag{16}$$

where W_1 and W_2 are learnable weights that are utilized to merge the label- and global-level features. The ultimate logits $\mathbf{L}_F$ are leveraged for training of the DL network and final prediction.

Finally, the weights of the entire network are updated using backpropagation with gradient descent based on the loss function. Notably, since the LPKA matrix is a learnable component of the network, it can be also updated as follows:

$$\mathbf{P} \leftarrow \mathbf{P} + \lambda \frac{\partial \mathbf{L}_F}{\partial \mathbf{P}} \tag{17}$$

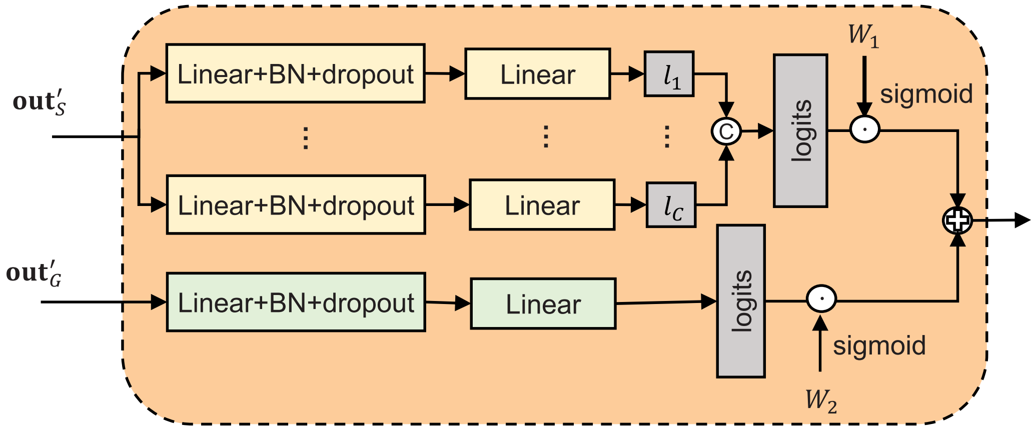

Figure 6. The structure of the classifier which integrates the label- and global-level branches.

where λ represents the learning rate. This approach allows the expert knowledge encoded in $\mathbf{P}$ to be adjusted iteratively during training.

4. Results and discussions

This section first presents the performance improvements achieved by the proposed LPKF compared to the baseline models. Next, we compare the proposed LPKF with existing SOTA models. Following that, an ablation study is conducted to justify the structural design of the proposed method. Finally, the model interpretability is assessed by visualizing the GradCAM results.

4.1. Data

In this study, we utilize two ECG databases: the PTB-XL database [24] and the CPSC2018 database [34], to evaluate the effectiveness of the proposed method.

The PTB-XL database, collected using Schiller AG equipment between October 1989 and June 1996 and released in 2020, contains 21 799 clinical 12-lead ECG recordings from 18 869 patients, each lasting 10 s. It provides hierarchical diagnostic statements, with superclass labels categorizing ECGs into five groups: normal (NORM), CD, hypertrophy (HYP), MI, and ST/T change (STTC). These superclass statements are widely used for multi-label ECG classification [35], with their distribution illustrated in figure 7(a). Additionally, the dataset includes subclass statements covering 23 diagnostic categories, as shown in figure 7(b), which provide finer-grained annotations than the superclass statements.

The CPSC2018 database, collected from 11 hospitals, contains ECG recordings classified into nine categories, with their distribution presented in figure 7(c). While most recordings have a single annotation, some include two or three labels, making this dataset well-suited for multi-label ECG classification.

It is worth noting that, in this study, we strictly adhere to the data preprocessing, train-test splits, and validation mechanisms defined in [24, 34] for the PTB-XL and CPSC2018 databases, ensuring consistency and comparability with prior works.

4.2. Performance metrics

The classification performance of the proposed LPKF is evaluated using AUC, precision, recall, F1 score, and Matthews correlation coefficient (MCC), as follows

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{18}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{19}$$

$$\text{F1} = 2\times\frac{\text{precision}\times\text{recall}}{\text{precision}+\text{recall}} \tag{20}$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}, \tag{21}$$

where TP, TN, FP and FN represent the true positive , true negative, false positive, and false negative predicted values, respectively.

Unless otherwise stated, the averaged metrics are calculated by first computing the individual metric for each category and then averaging these values, thereby assigning equal weight to each category regardless of the number of ECG samples it contains.

4.3. Experimental setup

All experiments in this study are performed on NVIDIA 3080TI GPU. All the DL models are implemented by using the PyTorch framework. During the training stage, the loss is optimized by AdamW method, with a weight decay of 1 $\times 10^{-4}$. The network is trained for 100 epochs with an initial learning rate of 0.1. To facilitate better convergence, the learning rate is decreased by a factor of 0.1 at the 50th and 75th epochs. The batch size for all experiments is set to 128.

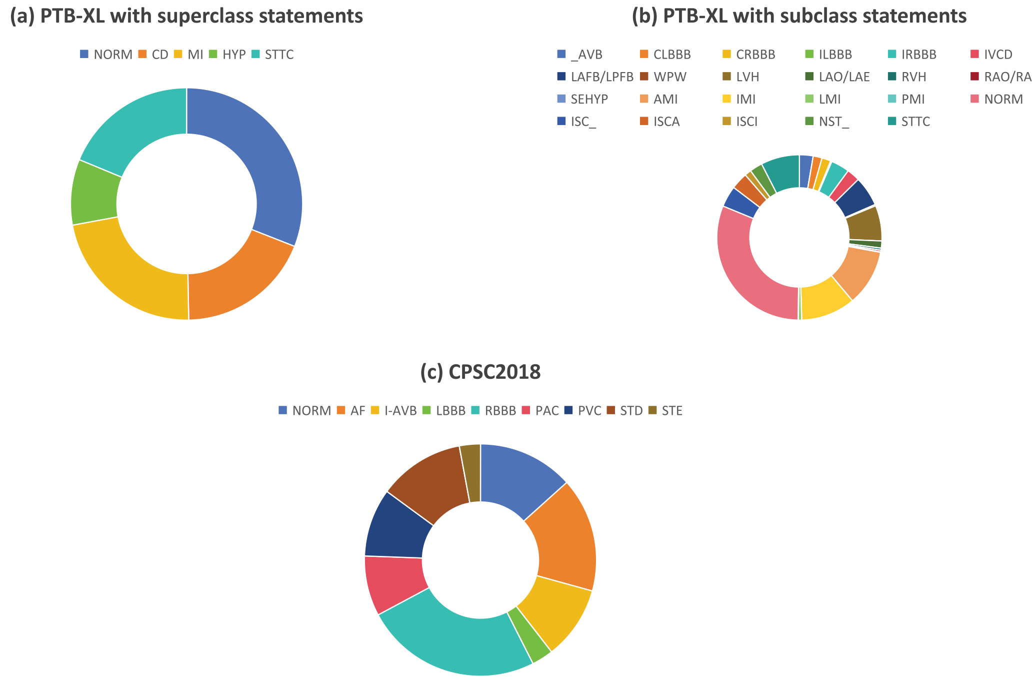

Figure 7. Distributions of categories in different databases.

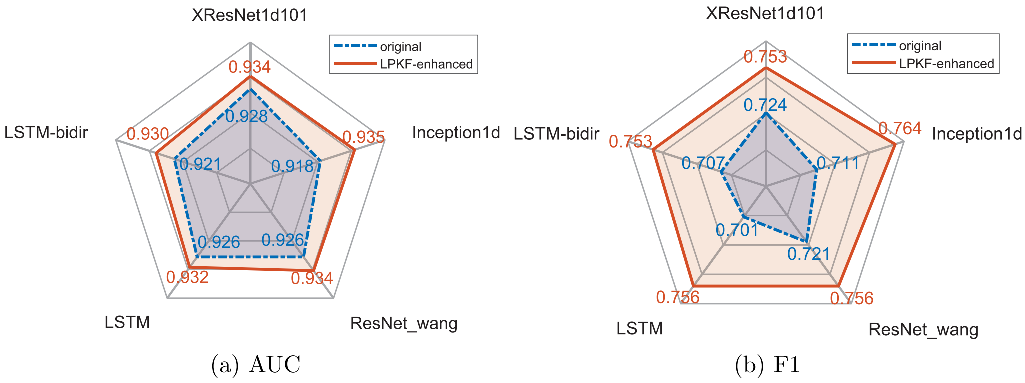

Figure 8. Performance improvement comparison with LPKF integration with superclass of PTB-XL.

4.4. Performance improvement with LPKF

The AUC and F1 scores obtained by five popular benchmark DL models on the PTB-XL database with spuderclass statements, including XResNet [36], Inception [37], ResNet_wang [38], LSTM [39] and LSTM-bidir [39], along with their LPKF-enhanced versions, are presented in figure 8. It is evident that the integration of LPKF assists all the these backbone models in achieving better classification performance. By exploiting the proposed LPKF, By exploiting the proposed LPKF, decent classification performance improvements can be observed compared to the five original DL models. The performance enhancement underscores the effectiveness and robustness of integrating LPKF into different types of backbone models can be demonstrated. In sequel, since the LPKF-enhanced Inception (LPKF-Inception) achieved the best classification performance, it will be used for further analysis.

The proposed method also achieved a higher AUC performance compared to existing relevant studies [23, 25, 40, 41] under similar experimental conditions. The existing models process raw or preprocessed ECG signals directly, while the proposed LPKF leverages features from both the LPKM and a temporal attention mechanism, providing more informative inputs to the backbone models. In addition, within the proposed LPKF, label-level and global-level features were extracted separately from the backbone models before reaching the classification layer. This actually facilitated auxiliary learning between the two branches and provided feature diversity to the classifier [42]. Consequently, the final information fusion within the classification layer led to higher classification performance. The benefits of

Table 1. Performance comparison on PTB-XL database with superclass statements.

	Precision	Recall	F1	MCC
XResNet [36]	0.721	0.741	0.724	0.638
IMLE-Net [43]	0.675	0.738	0.703	0.609
S-LSTM [44]	0.651	0.733	0.688	0.594
LDM-VGG [25]	0.712	0.719	0.715	0.621
LDM-XResNet [25]	0.762	0.734	0.741	0.661
LPKA-Inception (ours)	0.754	0.778	0.764	0.690

Table 2. Category-wise comparison between LDM-XResNet and LPKF-Inception on PTB-XL database with superclass statements.

Precision	CD	HYP	MI	NORM	STTC	Average
LDM-XResNet	0.804	0.679	0.765	0.794	0.768	0.762
LPKF-Inception	0.791	0.655	0.734	0.811	0.777	0.754
Recall						
LDM-XResNet	0.731	0.433	0.795	0.948	0.760	0.734
LPKF-Inception	0.765	0.563	0.805	0.933	0.824	0.778
F1						
LDM-XResNet	0.765	0.529	0.780	0.864	0.764	0.741
LPKF-Inception	0.778	0.605	0.768	0.868	0.800	0.764
MCC						
LDM-XResNet	0.668	0.540	0.686	0.735	0.673	0.661
LPKF-Inception	0.722	0.553	0.715	0.742	0.673	0.690

these two branches justified the design rationale behind the proposed LPKF.

4.5. Comparison to SOTA models

Table 1 presents precision, recall, F1 score, and MCC of the proposed LPKF-Inception model and SOTA models on the PTB-XL database using superclass statements. Five top-performing models with the highest F1 scores from recent literature on the same classification task are selected for comparison. While the proposed method achieves the second-highest precision, it outperforms all SOTA models in recall, F1, and MCC. Most SOTA models [6, 45, 46] focus on designing complex architectures for automatic feature extraction from ECG signals. In contrast, this study integrates expert knowledge into DL models through the proposed LPKF framework. This integration enhances feature representation in the backbone DL models, improving their learning efficiency. The results demonstrate the effectiveness of the LPKF structure in handling the complexities of multi-label ECG classification.

A detailed category-wise comparison between LDM-XResNet, which achieved the second-highest F1 and MCC scores in table 2, and the proposed LPKF-Inception is presented in table 1. While LPKF-Inception obtains a 0.8% lower precision than LDM-XResNet, it achieves a significantly 3.6% higher recall, demonstrating superior capability in identifying positive cases and reducing missed detections. This improved recall enables LPKF-Inception to achieve a better balance between false positives and false negatives, leading to higher F1 scores. Furthermore, LPKF-Inception surpasses LDM-XResNet in MCC by 2.9%, indicating a stronger correlation between predictions and ground truth.

Figure 9 presents the confusion matrix for LPKF-Inception on the PTB-XL database with superclass statements. The miss detection rates for the CD, HYP, MI, and STTC categories are 23.50%, 43.70%, 19.50%, and 17.60%, respectively-significantly higher than their corresponding false alarm rates of 6.10%, 4.10%, 10.00%, and 7.60%. In contrast, the miss detection rate for the NORM category is 6.70%, lower than its false alarm rate of 17.40%. Similar observations have been reported in previous studies [47], primarily due to dataset imbalance. Since approximately 43.63% of the ECG signals belong to the NORM category, the model learns normal ECG patterns more effectively than abnormal ones. As a result, it exhibits a higher false alarm rate for NORM while struggling to capture distinctive features of abnormal ECG signals, leading to higher miss detection rates for abnormal categories.

Tables 3 and 4 present performance comparisons on the PTB-XL database with subclass annotations and the CPSC2018 database, respectively. Due to the severe data imbalance in the subclass tasks, where 11 out of 23 categories account for less than 2% of the dataset, sample-centric metrics are used for evaluation in table 3. The results highlight two key findings: (1) Compared to the original Inception model, LPKF-Inception achieves superior classification performance, demonstrating the effectiveness of the proposed

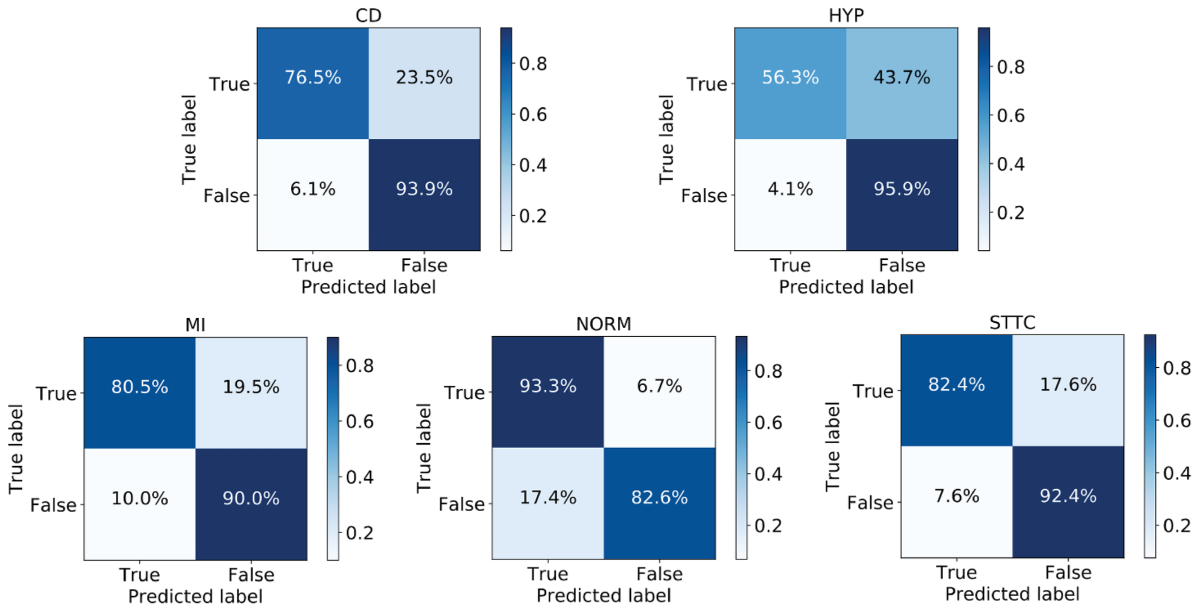

Figure 9. Confusion matrix of the LPKF-Inception on the PTB-XL database.

Table 3. Comparison of the LPKA-Inception model on PTB-XL database with subclass statements.

	Precision	Recall	F1	MCC
XResNet [36]	0.754	0.671	0.710	0.694
Inception [37]	0.747	0.686	0.714	0.698
LDM-XResNet [25]	0.704	0.737	0.720	0.702
LDM-VGG [25]	0.745	0.697	0.720	0.703
LPKA-Inception (ours)	0.722	0.742	0.732	0.715

Table 4. Comparison of the LPKA-Inception model on CPSC2018 database.

	Precision	Recall	F1	MCC
XResNet [36]	0.820	0.802	0.811	0.771
Inception [37]	0.823	0.783	0.802	0.756
LDM-XResNet [25]	0.828	0.814	0.821	0.793
LDM-VGG [25]	0.835	0.814	0.824	0.787
LPKA-Inception (ours)	0.852	0.844	0.848	0.812

LPKF framework when integrated with backbone models; (2) The proposed method consistently outperforms all other models across both classification tasks, further validating its efficacy.

The detailed category-wise comparisons between LDM-VGG and the proposed LPKF-Inception on PTB-XL database with subclass statements and CPSC2018 database are presented in figure 10 and table 5, respectively. For PTB-XL database subclass, severe data imbalance results in some categories having performance metrics of zero due to limited data availability. However, the proposed LPKF-Inception generally achieves better performance and has a greater number of categories with nonzero performance metrics compared to LDM-VGG. Similarly, for the CPSC2018 database, LPKF-Inception demonstrates superior category-wise performance. These comparisons validate the efficacy of the proposed approach.

4.6. Ablation studies

This study conducted two ablation experiments using the PTB-XL database with superclass statements. First, to evaluate the significance of the proposed learnable LPKA, we introduce two scenarios:

- L-Inception-n: The LPKA is initially determined based on expert knowledge but remains non-learnable, keeping its initial attention weights unchanged throughout the training process.
- L-Inception-p: The attention weights in the LPKA are initially assigned using the conventional initialization method [48] and remain learnable throughout training. Subsequently, fine-tuning is performed on the trained model based on strategies of gradual unfreezing and discriminative learning rates [23].

Second, a detailed investigation is conducted to compare the performance of using label-level features in isolation versus their fusion with global-level features.

Table 6 shows that L-Inception-n achieves higher F1 and MCC scores than L-Inception-p and the other two models, though slightly lower than those of LPKA-Inception. This result suggests that the initial attention weights derived solely from expert knowledge effectively guide DL model training. Even without fine-tuning, the lead-wise attention mechanism enhances the extraction of critical features. Moreover,

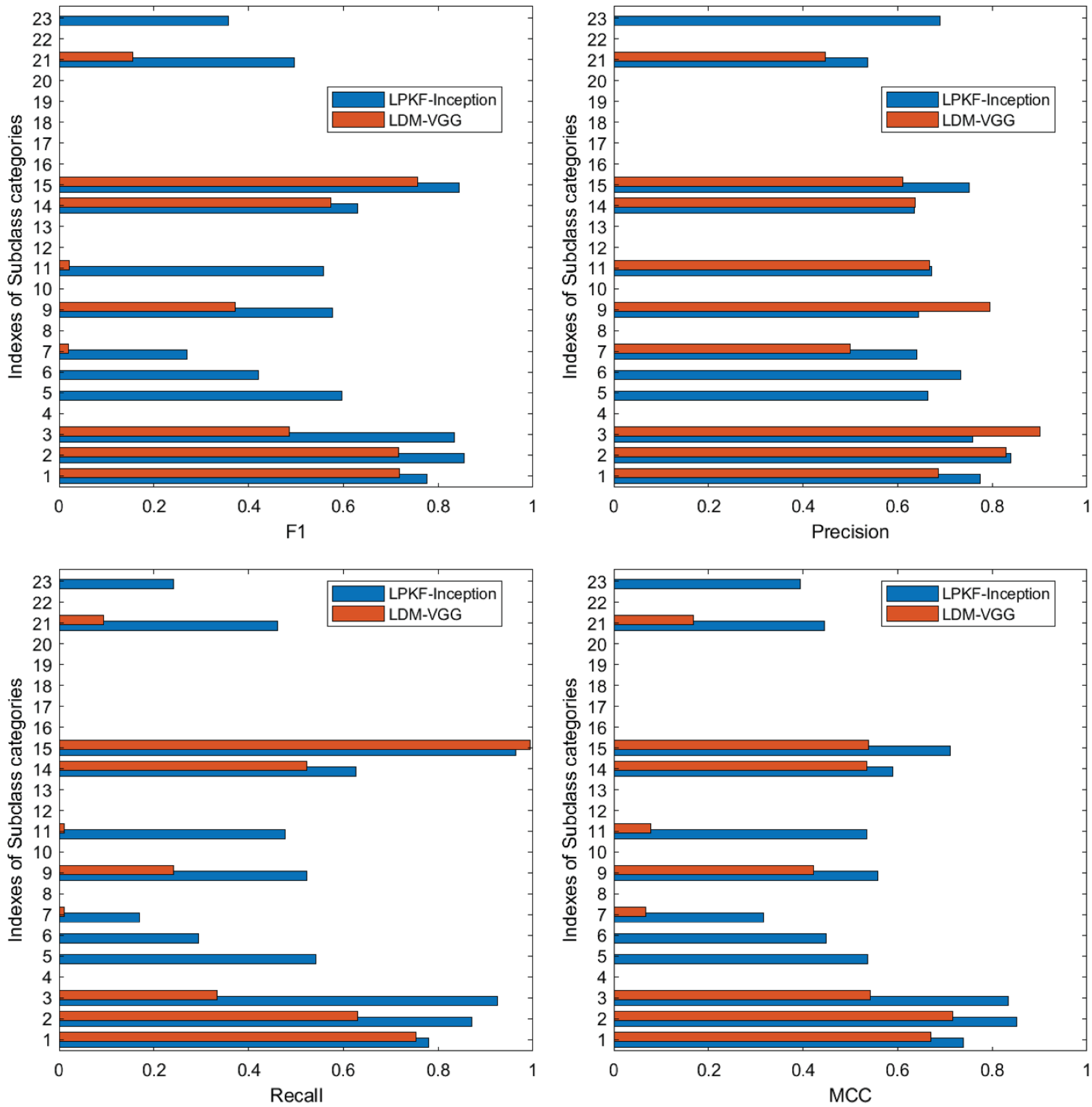

Figure 10. Category-wise comparison between LDM-VGG and LPKA-Inception on PTB-XL database with subclass statements.

updating these expert-driven attention weights during training further improves classification performance by dynamically redistributing attention, reinforcing the significance of learnable LPKAs. Additionally, L-Inception-p, which obtains LPKA through pre-training and fine-tuning, underperforms compared to all other methods except Inception, highlighting the necessity of embedding expert knowledge within the LPKF framework.

Figure 11 illustrates the initial LPKA, the corresponding final LPKA obtained at the end of the training process, and the difference between them. The results show that the distribution of attention values across ECG leads was refined through learning, enabling the model to focus more effectively on leads relevant to specific diagnostic labels. Notably, significant adjustments were observed in leads I and V6 for the CD category and lead V1 for the HYP category. These changes stemmed from the initial LPKAs, which assigned substantial attention to these leads. During training, the model slightly reduced these values, redistributing focus across other leads to optimize learning. This dynamic adjustment of attention distribution highlights the model's adaptability and effectiveness in improving feature extraction and classification performance.

To gain a better understanding of the rationale behind integrating label- and global-level features, four different strategies are evaluated and described as follows, each differing in how the two levels of features are exploited in the training and prediction stages.

- *Strategy 1*: Train the proposed Inception with LPKF model but predict labels using only the logits obtained from the label-level features

Table 5. Category-wise comparison between LDM-VGG and LPKA-Inception on CPSC2018 database.

precision	NORM	AF	IAVB	LBBB	RBBB	PAC	PVC	STD	STE	Average
LDM-VGG	0.845	0.881	0.902	0.901	0.862	0.802	0.806	0.757	0.763	0.835
LPKF-Inception	0.873	0.846	0.931	0.888	0.890	0.822	0.845	0.806	0.766	0.852
recall										
LDM-VGG	0.798	0.938	0.0809	0.842	0.926	0.788	0.836	0.792	0.617	0.814
LPKF-Inception	0.875	0.986	0.856	0.947	0.876	0.802	0.827	0.798	0.646	0.844
F1										
LDM-VGG	0.821	0.909	0.853	0.871	0.893	0.795	0.821	0.774	0.682	0.824
LPKF-Inception	0.874	0.911	0.892	0.917	0.883	0.812	0.836	0.804	0.701	0.848
MCC										
LDM-VGG	0.872	0.889	0.907	0.904	0.700	0.667	0.753	0.636	0.759	0.787
LPKF-Inception	0.884	0.889	0.913	0.912	0.746	0.712	0.787	0.689	0.775	0.812

Table 6. Performance comparison of L-Inception-n and L-Inception-p to different DL models.

	Precision	Recall	F1	MCC
Inception	0.728	0.693	0.710	0.626
LDM-XResNet	0.762	0.734	0.741	0.661
L-Inception-p	0.738	0.724	0.731	0.649
L-Inception-n	0.747	0.759	0.753	0.677
LPKA-Inception	0.754	0.778	0.764	0.690

- *Strategy 2*: Train the proposed Inception with LPKF model but predict labels using only the logits obtained from the global-level features
- *Strategy 3*: Only train the label-level branch in the proposed LPKF, i.e. only label-level features are extracted to classify all the labels
- *Strategy 4*: Only train the global-level branch in the proposed LPKF, i.e. only global-level features are extracted to classify all the labels

The classification performance of the four strategy, along with the proposed Inception with LPKF model, is depicted in table 7. It is evident that Strategies 1 and 3, which considers the label-level features, are very competitive to LPKF-Inception and relatively than Strategies 2 and 4, indicating that the label-level features provide greater benefit than leveraging the global-level features alone. This suggests that features exhibit greater separability with LPKA at the label-level branch. The lowest F1 score of 0.732 is obtained by Strategy 4, which is relatively lower than the scores achieved by Strategy 2. The learning processes of the label- and global-level branches in Strategy 2 are mutually beneficial, resulting in improved inference capability for both branches. However, in Strategy 4, the learning process is exclusively performed in the global-level branch, leading to performance degradation compared to Strategy 2. Additionally, both Strategies 1 and 2 achieves lower F1 scores compared to the LPKF-enhanced Inception model, where the logits obtain from the label- and global-level features are integrated. This aligns with the intuition that the joint consideration of both label-level and global-level features contributes to a higher inference capability of the model. It is noteworthy that the satisfactory performance achieved by Strategy 3, which outperforms strategies 2 and 4, validating the effectiveness of solely integrating expert knowledge with DL models.

Compared to [25], whose model also integrates binary and multi-label classification branches simultaneously and achieves the F1 scores of 0.741, Strategies 1 and 3 in this study outperforms their approach. Their model applies the classification layer to the same extracted features from the backbone DL models before label prediction. In contrast, the proposed LPKF extracts label-level and global-level features separately from the backbone DL models before reaching the classification layer. This approach initiates auxiliary learning between the two branches and provides feature diversity to the classifier [42], resulting in higher performance. The benefits of these two branches justify the specific design of the proposed LPKF.

4.7. Model interpretability analysis

Figure 12 depicts the LPKF-enhanced Inception model's focus on abnormal ECG signals through GradCAM visualization, highlighting regions of the input ECG signal crucial for specific disease classification. The records are randomly selected, and the two highest-weighted leads based on the diagnostic LPKA are reported. Elevated attention is evident in critical ECG regions, such as significant Q waves, ST segment elevation, large S waves, and M-shaped QRS complexes in ECG signals associated with categories of MI, STTC, HYP, and CD, respectively, which is consistent with clinical diagnostic criteria. This observation indicates the LPKF-Inception model's ability to capture distinctive features relevant to cardiac diseases, aligning with clinical considerations for healthcare practitioners.

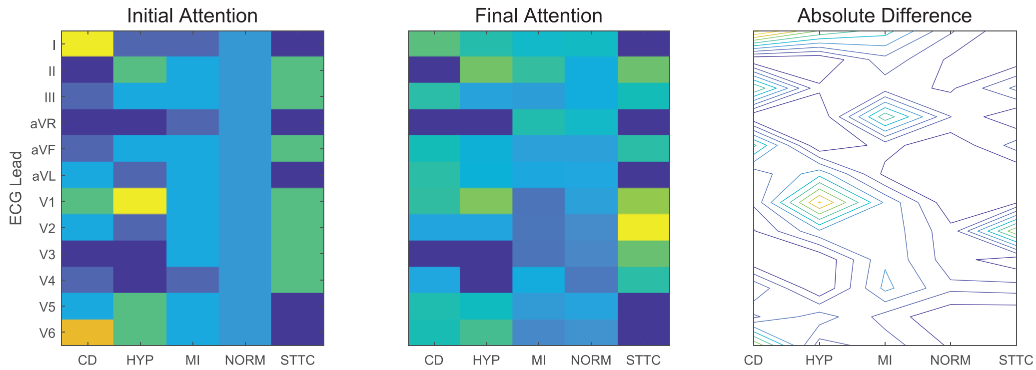

Figure 11. Comparison of LPKAs before and after training process.

Table 7. F1 comparisons in different scenarios.

	CD	HYP	MI	NORM	STTC	Average
LPKF-Inception	**0.778**	**0.605**	**0.768**	**0.868**	**0.800**	**0.764**
Strategy 1	0.768	0.611	0.781	0.867	0.758	0.757
Strategy 2	0.762	0.577	0.758	0.861	0.742	0.740
Strategy 3	0.776	0.605	0.768	0.865	0.757	0.754
Strategy 4	0.711	0.559	0.756	0.858	0.776	0.732

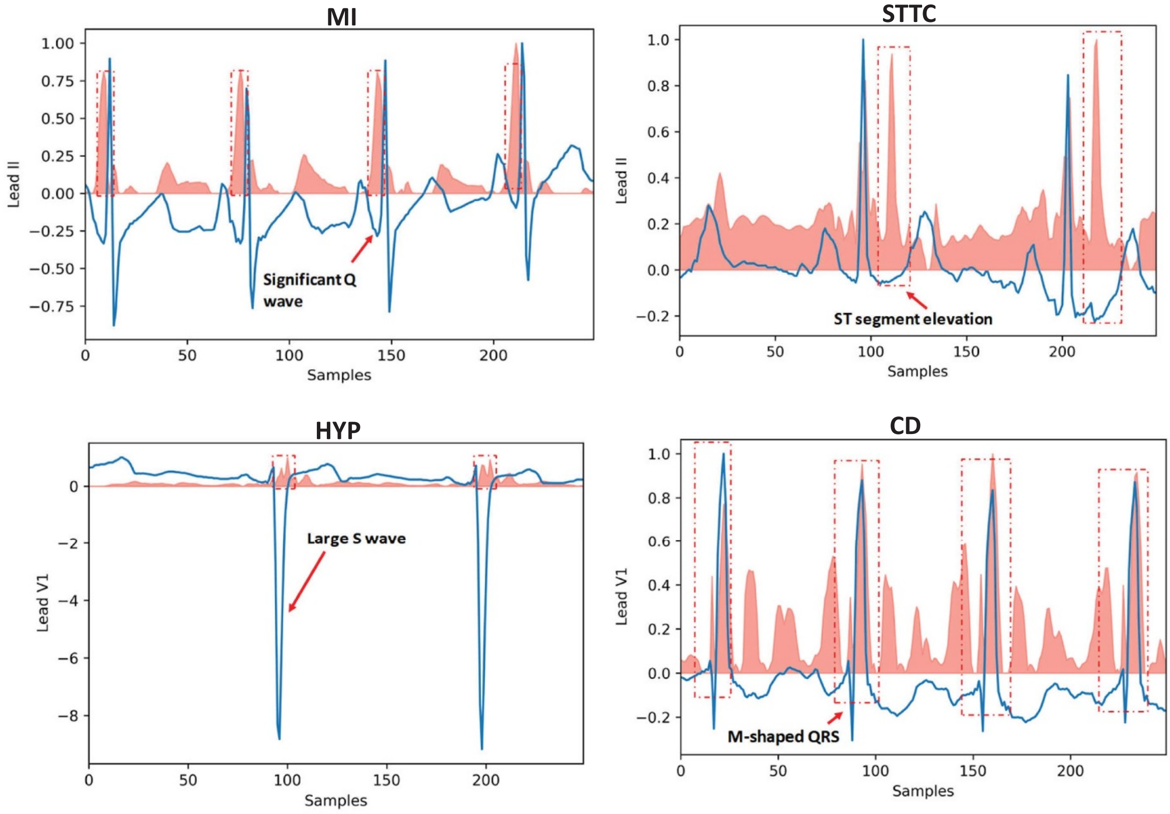

Figure 12. GradCAM for the four abnormal categories.

4.8. Limitations of the study

This study proposed a framework for integrating expert knowledge into DL models for multi-label ECG classification using publicly available ECG databases. However, its generalization to clinical practice remains to be validated. Additionally, the proposed method models the relationship between diagnostic categories and ECG leads, necessitating multi-lead input signals. As a result, it is not applicable to single-lead ECG analysis, such as those used in smartwatches. Finally, the limited number of ECG recordings in certain diagnostic categories poses a challenge for DL models in learning representative features from these sparse subsets, an issue not addressed in this study.

5. Conclusions

In this study, we proposed LPKF, which integrated expert knowledge into DL models to allocate learning efforts more effectively, improving feature separability across categories for multi-label ECG classification. Specifically, a LPKA was first developed to assign attention weights to ECG leads based on relevant heart abnormalities. Importantly, the proposed LPKA was designed to be learnable, allowing for adjustments to refine prior knowledge and better adapt to ECG data, thereby enhancing classification performance. Additionally, two learning branches were designed to combine predictions from features obtained using expert knowledge and those extracted directly from input ECG signals. The effectiveness of the proposed LPKF was validated by improvements in AUC and F1 scores achieved by LPKF-enhanced models compared to backbone models without LPKF and SOTA models, respectively. Through ablation studies and model interpretability analysis, the components within the LPKF were justified, and the LPKF-enhanced model's ability to capture important diagnostic symptoms from ECG signals was demonstrated. In summary, experimental results validated the efficacy of the proposed LPKF, highlighting its effectiveness, adaptability, and interpretability in complex multi-label ECG classification tasks.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

ORCID iDs

Qiao Xiao https://orcid.org/0000-0001-9300-688X
Poh Ying Lim https://orcid.org/0000-0002-7765-4630

References

[1] Tsao C W *et al* 2023 Heart disease and stroke statistics-2023 update: a report from the American Heart Association *Circulation* **147** e93–e621
[2] Garvey J L 2006 ECG techniques and technologies *Emerg. Med. Clin.* **24** 209–25
[3] Pyakillya B, Kazachenko N and Mikhailovsky N 2017 Deep learning for ECG classification *J. Phys.: Conf. Ser.* **913** 012004
[4] Ahmed A A, Ali W, Abdullah T A A and Malebary S J 2023 Classifying cardiac arrhythmia from ECG signal using 1D CNN deep learning model *Mathematics* **11** 562
[5] Aldughayfiq B, Ashfaq F, Jhanjhi N and Humayun M 2023 A deep learning approach for atrial fibrillation classification using multi-feature time series data from ECG and PPG *Diagnostics* **13** 2442
[6] Li Y, Luo J-H, Dai Q-Y, Eshraghian J K, Ling B W-K, Zheng C-Y and Wang X-L 2023 A deep learning approach to cardiovascular disease classification using empirical mode decomposition for ECG feature extraction *Biomed. Signal Process. Control* **79** 104188
[7] Jin Y, Li Z, Wang M, Liu J, Tian Y, Liu Y, Wei X, Zhao L and Liu C 2024 Cardiologist-level interpretable knowledge-fused deep neural network for automatic arrhythmia diagnosis *Commun. Med.* **4** 31
[8] Degirmenci M, Ozdemir M A, Izci E and Akan A 2022 Arrhythmic heartbeat classification using 2D convolutional neural networks *Irbm* **43** 422–33
[9] Liu Y, Wang K, Yuan Y, Li Q, Li Y, Xu Y and Zhang H 2020 Multi-label classification of 12-lead ECGs by using residual CNN and class-wise attention *2020 Computing in Cardiology* (IEEE) pp 1–4
[10] Li Y, Zhang Z, Zhou F, Xing Y, Li J and Liu C 2021 Multi-label classification of arrhythmia for long-term electrocardiogram signals with feature learning *IEEE Trans. Instrum. Meas.* **70** 1–11
[11] Radhakrishnan T, Karhade J, Ghosh S K, Muduli P R, Tripathy R and Acharya U R 2021 AFCNNet: automated detection of AF using chirplet transform and deep convolutional bidirectional long short term memory network with ECG signals *Comput. Biol. Med.* **137** 104783
[12] Ganguly B, Ghosal A, Das A, Das D, Chatterjee D and Rakshit D 2020 Automated detection and classification of arrhythmia from ECG signals using feature-induced long short-term memory network *IEEE Sens. Lett.* **4** 1–4
[13] Cai J, Sun W, Guan J and You I 2020 Multi-ECGnet for ECG arrythmia multi-label classification *IEEE Access* **8** 110848–58
[14] Ge Z, Jiang X, Tong Z, Feng P, Zhou B, Xu M, Wang Z and Pang Y 2021 Multi-label correlation guided feature fusion network for abnormal ECG diagnosis *Knowl.-Based Syst.* **233** 107508
[15] Krzysztof P, Sandra S, Damian L and Sławomir B 2022 Study of the few-shot learning for ECG classification based on the PTB-XL dataset *MDPI Sens.* **22** 1–25
[16] Xie X, Liu H, Chen D, Shu M and Wang Y 2022 Multilabel 12-lead ECG classification based on leadwise grouping multibranch network *IEEE Trans. Instrum. Meas.* **71** 1–11
[17] Cheng Y, Zhu W, Li D and Wang L 2024 Multi-label classification of arrhythmia using dynamic graph convolutional network based on encoder-decoder framework *Biomed. Signal Process. Control* **95** 106348
[18] Yang S, Lian C, Zeng Z, Xu B, Zang J and Zhang Z 2023 A multi-view multi-scale neural network for multi-label ECG classification *IEEE Trans. Emerg. Top. Comput. Intell.* **7** 648–60
[19] Hong S, Wu M, Zhou Y, Wang Q, Shang J, Li H and Xie J 2017 Encase: an ensemble classifier for ECG classification using expert features and deep neural networks *2017 Computing in Cardiology (Cinc)* (IEEE) pp 1–4
[20] Tao Y, Li Z, Gu C, Jiang B and Zhang Y 2022 ECG-based expert-knowledge attention network to tachyarrhythmia recognition *Biomed. Signal Process. Control* **76** 103649
[21] Ganeshkumar M, Ravi V, Sowmya V, Gopalakrishnan E and Soman K 2021 Explainable deep learning-based approach for multilabel classification of electrocardiogram *IEEE Trans. Eng. Manage.* **70** 2787–99
[22] Luo C, Jiang H, Li Q and Rao N 2019 Multi-label classification of abnormalities in 12-lead ECG using 1D CNN and LSTM *Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting: First Int. Workshop, MLMECH 2019 and 8th Joint Int. Workshop, CVII-STENT 2019, Held in Conjunction With MICCAI*

2019, (Shenzhen, China, 13 October 2019, Proc.) vol 1 (Springer) pp 55–63

[23] Strodthoff N, Wagner P, Schaeffter T and Samek W 2020 Deep learning for ECG analysis: benchmarks and insights from PTB-XL *IEEE J. Biomed. Health Inform.* **25** 1519–28

[24] Wagner P, Strodthoff N, Bousseljot R-D, Kreiseler D, Lunze F I, Samek W and Schaeffter T 2020 PTB-XL, a large publicly available electrocardiography dataset *Sci. Data* **7** 1–15

[25] Zhang S, Li Y, Wang X, Gao H, Li J and Liu C 2023 Label decoupling strategy for 12-lead ECG classification *Knowl.-Based Syst.* **263** 110298

[26] Zhou F and Chen L 2024 Leadwise clustering multi-branch network for multi-label ECG classification *Med. Eng. Phys.* **130** 104196

[27] Sadeghi A, Hajati F, Rezaee A, Sadeghi M, Argha A and Alinejad-Rokny H 2024 3DECG-Net: ECG fusion network for multi-label cardiac arrhythmia detection *Comput. Biol. Med.* **182** 109126

[28] Wen W, Zhang H, Wang Z, Gao X, Wu P, Lin J and Zeng N 2024 Enhanced multi-label cardiology diagnosis with channel-wise recurrent fusion *Comput. Biol. Med.* **171** 108210

[29] Cheng Y, Li D, Wang D, Chen Y and Wang L 2024 Multi-label arrhythmia classification using 12-lead ECG based on lead feature guide network *Eng. Appl. Artif. Intell.* **129** 107599

[30] Oldroyd C and Kucia A M 2022 Electrocardiogram (ECG) interpretation *Cardiac Care: A Practical Guide for Nurses* 103–27

[31] Surawicz B, Childers R, Deal B J and Gettes L S 2009 AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: part III: intraventricular conduction disturbances: a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society: endorsed by the International Society for Computerized Electrocardiology *Circulation* **119** e235–40

[32] Kashou A H, Goyal A, Nguyen T, Ahmed I, Chhabra L and Kukuc L G 2023 Atrioventricular block (Nursing) *Statpearls [Internet]* (StatPearls Publishing)

[33] Murphy J G 2019 *Mayo Clinic Cardiology: Concise Textbook* (CRC Press)

[34] Liu F *et al* 2018 An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection *J. Med. Imaging Health Inform.* **8** 1368–73

[35] Xiao Q, Lee K, Mokhtar S A, Ismail I, Pauzi A L bin Md, Zhang Q and Lim P Y 2023 Deep learning-based ECG arrhythmia classification: a systematic review *Appl. Sci.* **13** 4964

[36] He T, Zhang Z, Zhang H, Zhang Z, Xie J and Li M 2019 Bag of tricks for image classification with convolutional neural networks *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 558–67

[37] Ismail Fawaz H, Lucas B, Forestier G, Pelletier C, Schmidt D F, Weber J, Webb G I, Idoumghar L, Muller P-A and Petitjean F 2020 InceptionTime: finding AlexNet for time series classification *Data Mining Knowl. Discovery* **34** 1936–62

[38] Wang Z, Yan W and Oates T 2017 Time series classification from scratch with deep neural networks: a strong baseline *2017 Int. Joint Conf. on Neural Networks (IJCNN)* (IEEE) pp 1578–85

[39] Hochreiter S and Schmidhuber J 1997 Long short-term memory *Neural Comput.* **9** 1735–80

[40] Śmigiel S, Pałczyński K and Ledziński D 2021 ECG signal classification using deep learning techniques based on the PTB-XL dataset *Entropy* **23** 1121

[41] Tao R, Wang L, Xiong Y and Zeng Y-R 2024 IM-ECG: an interpretable framework for arrhythmia detection using multi-lead ECG *Expert Syst. Appl.* **237** 121497

[42] Nguyen T-N and Nguyen T-H 2021 Deep learning framework with ECG feature-based kernels for heart disease classification *Elektron. Elektrotech.* **27** 48–59

[43] Reddy L, Talwar V, Alle S, Bapi R S and Priyakumar U D 2021 IMLE-NET: an interpretable multi-level multi-channel model for ECG classification *2021 IEEE Int. Conf. on Systems, Man and Cybernetics (SMC)* (IEEE) pp 1068–74

[44] He T, Chen Y, Chen J, Wang W, and Zhou Y 2022 SEVGGNet-LSTM: a fused deep learning model for ECG classification (arXiv:2210.17111)

[45] Alkhodari M, Apostolidis G, Zisou C, Hadjileontiadis L J and Khandoker A H 2021 Swarm decomposition enhances the discrimination of cardiac arrhythmias in varied-lead ECG using ResNet-BiLSTM network activations *2021 Computing in Cardiology (CinC)* vol 48 pp 1–4

[46] Ramkumar M, Lakshmi A, Rajasekaran M P and Manjunathan A 2022 Multiscale Laplacian graph kernel features combined with tree deep convolutional neural network for the detection of ECG arrhythmia *Biomed. Signal Process. Control* **76** 103639

[47] Ran S, Li X, Zhao B, Jiang Y, Yang X and Cheng C 2023 Label correlation embedding guided network for multi-label ECG arrhythmia diagnosis *Knowl.-Based Syst.* **270** 110545

[48] He K, Zhang X, Ren S and Sun J 2015 Delving deep into rectifiers: surpassing human-level performance on ImageNet classification *Proc. IEEE Int. Conf. on Computer Vision* pp 1026–34