Semi-Supervised Learning for Multi-Label Cardiovascular Diseases Prediction: A Multi-Dataset Study

Rushuang Zhou, *Graduate Student Member, IEEE*, Lei Lu, Zijun Liu, *Graduate Student Member, IEEE*, Ting Xiang, *Graduate Student Member, IEEE*, Zhen Liang, *Member, IEEE*, David A. Clifton, Yining Dong, and Yuan-Ting Zhang, *Fellow, IEEE*

***Abstract*—Electrocardiography (ECG) is a non-invasive tool for predicting cardiovascular diseases (CVDs). Current ECG-based diagnosis systems show promising performance owing to the rapid development of deep learning techniques. However, the label scarcity problem, the co-occurrence of multiple CVDs and the poor performance on unseen datasets greatly hinder the widespread application of deep learning-based models. Addressing them in a unified framework remains a significant challenge. To this end, we propose a multi-label semi-supervised model (ECGMatch) to recognize multiple CVDs simultaneously with limited supervision. In the ECGMatch, an ECGAugment module is developed for weak and strong ECG data augmentation, which generates diverse samples for model training. Subsequently, a hyperparameter-efficient framework with neighbor agreement modeling and knowledge distillation is designed for pseudo-label generation and refinement, which mitigates the label scarcity problem. Finally, a label correlation alignment module is proposed to capture the co-occurrence information of different CVDs within labeled samples and propagate this information to unlabeled samples. Extensive experiments on four datasets and three protocols demonstrate the effectiveness and stability of the proposed model, especially on unseen datasets. As such, this model can pave the way for diagnostic systems that achieve robust performance on multi-label CVDs prediction with limited supervision.**

***Index Terms*—Cardiovascular diseases, electrocardiograph, multi-label learning, semi-supervised learning.**

Manuscript received 16 June 2023; revised 1 November 2023; accepted 7 December 2023. Date of publication 14 December 2023; date of current version 3 April 2024. This work was supported by an NIHR Research Professorship, an RAEng Research Chair, the InnoHK Hong Kong projects under the Hong Kong Center for Cerebro-cardiovascular Health Engineering (COCHE), in part by the NIHR Oxford Biomedical Research Centre (BRC), in part by the Pandemic Sciences Institute at the University of Oxford, in part by the National Natural Science Foundation of China under Grant 22322816, and in part by the City University of Hong Kong Project under Grant 9610640. Recommended for acceptance by A.N. Veeraraghavan. *(Corresponding authors: Yining Dong; Yuan-Ting Zhang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Emory University, Physikalisch-Technische Bundesanstalt, Chapman University, Ningbo First Hospital of Zhejiang University and Shaoxing People's Hospital.

Rushuang Zhou, Zijun Liu, and Ting Xiang are with the Department of Biomedical Engineering, City University of Hong Kong, Hong Kong, SAR, China, and also with the Hong Kong Center for Cerebro-Cardiovascular Health Engineering (COCHE), Hong Kong, SAR, China (e-mail: rrushuang2-c@my.cityu.edu.hk; zijunliu4-c@my.cityu.edu.hk; txiang7-c@my.cityu.edu.hk).

Lei Lu is with the Department of Engineering Science, University of Oxford, OX1 2JD Oxford, U.K. (e-mail: lei.lu@eng.ox.ac.uk).

David A. Clifton is with the Department of Engineering Science, University of Oxford, OX1 2JD Oxford, U.K., and also with the Oxford-Suzhou Institute of Advanced Research (OSCAR), Suzhou 215123, China (e-mail: davidc@robots.ox.ac.uk).

Zhen Liang is with the School of Biomedical Engineering, Medical School, Shenzhen University, Shenzhen 518060, China (e-mail: zhenliang.szu@gmail.com).

Yining Dong is with the School of Data Science, City University of Hong Kong, Hong Kong, SAR, China, and also with the Hong Kong Center for Cerebro-Cardiovascular Health Engineering (COCHE), Hong Kong, SAR, China (e-mail: yinidong@cityu.edu.hk).

Yuan-Ting Zhang is with Micro Sensing and Imaging Technologies Limited, Hong Kong, SAR, China, and also with Wearable Intelligent Sensing Technologies Limited, Hong Kong, SAR, China (e-mail: ytzhanghicas@gmail.com).

Code is available at https://github.com/KAZABANA/ECGMatch.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2023.3342828, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2023.3342828

I. Introduction

CARDIOVASCULAR diseases (CVDs) have become the world's leading cause of morbidity and mortality in recent years [1]. As a non-invasive test, the 12-lead electrocardiography (ECG) is widely used for diagnosing CVDs. With the rapid development of deep learning and artificial intelligence, AI-aided automatic diagnosis systems have attracted considerable interest in clinical practice. Most of these systems are designed for a well-defined setting where the annotated samples are sufficient and identically distributed, with each sample only belonging to one CVDs class. Unfortunately, the complex real-world setting differs from this ideal setting, where annotated ECG segments are tough to collect, and multiple CVDs can be identified from each segment. Furthermore, the real-world training and test data may not be sampled from the same distribution, which greatly hurts the model performance. The difference between the real-world setting and the ideal setting restricts the clinical applications of current systems. In a nutshell, there are three challenges in the clinical applications of automatic diagnosis systems: 1) Label scarcity problem. 2) Poor performance on unseen datasets. 3) Co-occurrence of multiple CVDs.

In recent years, semi-supervised learning (SSL) has shown great potential in addressing the label scarcity problem in clinical applications. The main idea of the SSL models is to utilize unlabeled samples for model training, which are easier to collect compared with labeled samples [2], [3]. By leveraging the abundant information within the unlabeled samples, SSL models

often outperform fully-supervised models when the number of labeled samples is limited [4], [5], [6]. Consequently, numerous studies were proposed to extend the success of SSL to ECG-based CVDs prediction. For example, Oliveira et al. applied existing SSL models for ECG signal classification. Experiments on the MIT-BIH database [7] demonstrated the superiority of the SSL models compared with fully-supervised models [8]. To improve the model performance on unseen datasets, Feng et al. proposed a transfer learning framework to transfer the model trained on a label-sufficient dataset to a label-scarce target dataset. Comprehensive results on four benchmarks demonstrated the robustness of the proposed framework. At the same time, multi-label learning sheds new light on how to detect multiple CVDs from one ECG recording simultaneously. In contrast to single-label learning, multi-label learning generates multiple predictions for a given sample, with each prediction indicating whether the sample belongs to a specific category [9]. Multi-label learning models have the capability to detect multiple diseases from ECG signals, while single-label models are limited to recognizing only one disease at a time. As a result, numerous models have been proposed to leverage multi-label learning for ECG-based CVDs prediction. For example, Strodthoff et al. evaluated the performance of existing models on the PTB-XL database [10], [11] and demonstrated the feasibility of using multi-label learning models for CVDs prediction. Subsequently, Ge et al. and Ran et al. proposed utilizing the relationship between different cardiac diseases to enhance the model performance [12], [13]. Lai et al. realized efficient prediction of CVDs by developing a multi-scale deep neural network that was initialized through self-supervised pretraining [14]. More details about the existing models for ECG-based CVDs prediction are presented in Section II.

To the best of our knowledge, no prior study has proposed and validated a unified framework to tackle the aforementioned three challenges simultaneously. Specifically, most previous studies only alleviated one of the aforementioned problems without comprehensively considering the other two challenges in CVDs prediction. For example, many SSL-based models ignored the co-occurrence of multiple CVDs and their performance on unseen datasets was not satisfying [6], [8], [15], [16], [17]. Previous multi-label learning models could detect multiple CVDs from ECG signals, but their effectiveness relied heavily on sufficient labeled data and handcrafted prior knowledge [11], [12], [13]. These significant deficiencies imply that these models are still far from being applicable in real-world scenarios. Therefore, in this study, we propose a multi-label semi-supervised framework (ECGMatch) that can use only 1% of the annotated samples to achieve good results in cross-dataset multi-label CVDs prediction. Here, we introduce how the proposed framework addresses the aforementioned problem simultaneously.

First, a novel ECGAugment module is developed to alleviate the label scarcity problem by generating diverse samples. It exploits the intrinsic characteristics of ECG signals and dramatically outperforms traditional methods [8], [18]. Moreover, we design a pseudo-label generation module that utilizes the interaction between the student and teacher networks to generate pseudo-labels for the unlabeled samples. Specifically, we formulate the generation task as a knowledge distillation process. During training, the teacher stores the learned knowledge in two memory banks, and the student visits the banks to assign pseudo-labels for the unlabeled samples using a K-Nearest voting strategy. To mitigate the negative impact of inaccurate pseudo-labels, we propose a neighbor agreement modeling method and develop a hyperparameter-efficient module for refining these labels. During the K-Nearest voting process, the degree of agreement among neighbors can be utilized to estimate the pseudo-label confidence, which is an important indicator for discovering trustworthy pseudo-labels. In multi-label learning, the advantage of the proposed hyperparameter-efficient refinement module is more significant as it only relies on the number of neighbors K rather than numerous thresholds and complex control strategies [5], [6], [15], [19].

To capture the co-occurrence of different CVDs, we introduce a label correlation alignment module. It quantitatively estimates the co-occurrence information using limited labeled data and transfers this knowledge to unlabeled data. In practice, we compute a correlation matrix to represent the co-occurrence information, and align the matrices computed on labeled and unlabeled data to complete a knowledge transfer process. Finally, we conduct extensive experiments on four public datasets across three protocols. The results comprehensively validate the superiority of the ECGMatch, especially on unseen datasets. In summary, the main contributions and novelties are listed below.

- We proposed a robust pipeline for ECG signal augmentation, which shows remarkable improvements compared with previous methods.
- An efficient method for pseudo-label refinement is developed for multi-label learning with limited supervision. It has fewer parameters than threshold-based methods but shows better performance.
- A novel approach is proposed to align the label correlation computed on labeled and unlabeled data, which provides a reliable solution to capture the co-occurrence of multiple CVDs.
- A unified semi-supervised framework for multi-label CVDs prediction is proposed, which is the first one to address three critical challenges in this area.

II. Related Work

A. ECG-Based CVDs Prediction Using Deep Learning

Over the past decade, the potential and feasibility of utilizing ECG signals to diagnose a wide spectrum of CVDs have been demonstrated by numerous previous studies [11], [18], [20], [21], [22], [23], [24], [25], [26], [27], [28]. With the rapid development of deep learning techniques, many studies used end-to-end deep learning models to achieve accurate predictions of the CVDs. For example, Kiranyaz et al. designed a real-time one-dimensional convolutional neural network (CNN) that achieved superior performance in ECG-based CVDs prediction compared with traditional models [20]. Hannun et al. conducted a comprehensive evaluation of a deep neural network (DNN) for ECG signal classification. The extensive results showed that the DNN model achieved a similar diagnosis performance to cardiologists, thus demonstrating its enormous potential in clinical applications [22]. Subsequently, several methods were

proposed to enhance the accuracy of the DNN model. For example, Ribeiro et al. proposed a unidimensional residual neural network architecture that outperformed cardiology resident medical doctors in recognizing six kinds of CVDs [23]. Huang et al. introduced a novel deep reinforcement learning framework called snippet policy network V2 (SPN-V2) for the early prediction of CVDs based on ECG signals. Using a novel keen-guided neuroevolution algorithm, the SPN-V2 network achieved a stable balance between recognition accuracy and earliness [27]. However, despite the significant advancements in ECG-based CVDs prediction using deep learning methods in recent years, such methods may experience a performance drop when the number of labeled samples is limited [17].

B. Semi-Supervised Learning for ECG-Based CVDs Prediction

Semi-supervised learning has achieved great success in reducing the requirements on laborious annotations for model training [4], [5], [6], [15], [29]. As a result, an increasing number of studies have proposed using SSL to develop robust models for ECG-based CVDs prediction with limited supervision [8], [16], [17]. For instance, Zhai et al. proposed a semi-supervised model to transfer knowledge learned from large datasets to small datasets. Extensive experiments demonstrated that the performance of the proposed model was comparable to other methods which required numerous annotated samples [16]. Oliveira et al. applied different SSL models for ECG-based CVDs prediction, such as MixMatch [4] and FixMatch [5]. When only 15% of the ECG data was labeled, the SSL models achieved comparable prediction performance obtained by fully supervised models [8]. Motivated by the mean teacher algorithm [30], Zhang et al. proposed the mixed mean teacher model for automatic atrial fibrillation detection using ECG, which significantly reduced the workload of data annotation by 98% while achieving comparable performance as fully supervised models [17]. To address the distribution shifts across different datasets, Feng et al. proposed a SSL framework based on two complementary modules: semantic-aware feature alignment (SAFA) and prototype-based label propagation (PBLP) [31]. Comprehensive experiments verified that the proposed model achieved promising performance on target datasets using limited labeled target samples.

However, previous SSL studies for ECG-based CVDs prediction have two main limitations. *1) Previous SSL studies developed single-label classification models for CVDs prediction, which were greatly limited in clinical applications.* Specifically, they simply formulated the CVDs prediction task as a single-label problem, where each ECG signal can only belong to one category. However, multiple CVDs, such as atrial fibrillation and right bundle branch block, usually co-occur in one ECG segment [13]. This phenomenon suggests that the CVDs prediction task should be formulated as a multi-label problem, where each ECG signal belongs to multiple categories. *2) Previous studies did not consider CVDs prediction on unseen datasets.* The training and test data in previous studies were from the same dataset, which is often unrealistic in real-world applications. While some studies applied transfer learning to transfer knowledge from the training datasets to unseen datasets [11], [31], their methods still needed labeled samples from the test dataset, which led to information leaking.

TABLE I
FREQUENTLY USED NOTATIONS AND DESCRIPTIONS

Notation	Description
B	batch size
C	the number of CVDs category
R	label correlation matrix
K	the number of nearest neighbor
α	neighbor agreement
$D_B \backslash D_U$	labeled\unlabeled datasets
$Y_b \backslash \hat{Y}_u$	true\pseudo labels
$X_b \backslash X_u$	labeled\unlabeled ECG samples
$Z \backslash P$	feature\prediction banks
$M_s \backslash M_t$	student\teacher networks
$w(\cdot) \backslash g(\cdot)$	weak\strong augmentations
$f(\cdot) \backslash h(\cdot)$	feature extractor\multi-label classifier

C. Multi-Label Model for ECG-Based CVDs Prediction

Many studies have investigated the feasibility of using multi-label learning to simultaneously detect multiple arrhythmia types from ECG signals. Strodthoff et al. proposed a pilot study that evaluated the performance of different models for ECG-based multi-label CVDs classification and found that ResNet and Inception-based CNN architectures achieved the best performance [11]. Ge et al. utilized Bayesian conditional probability to capture the association between ECG abnormalities and used it to guide the feature fusion of ECG-based models. Promising experimental results showed that the multi-label correlation guided feature fusion network outperformed other competitors [12]. Ran et al. proposed a label correlation embedding guided network (LCEGNet) to capture the relationship between different ECG abnormalities and improve the model performance by learning arrhythmia-specific features [13]. However, annotating multi-label ECG data is prohibitively expensive, leading to a critical bottleneck in real-world applications. This problem highlights the need for semi-supervised learning in ECG-based multi-label CVDs prediction, which will be investigated in our study.

III. METHODOLOGY

A. Overview

In semi-supervised learning for multi-label CVDs prediction, the training ECG data is divided into the labeled and unlabeled sets, given as $D_B = \{X_b, Y_b\} = \{x_b^i, y_b^i\}_{i=1}^{N_B}$ and $D_U = \{X_u, -\} = \{x_u^i, -\}_{i=1}^{N_U}$. N_B and N_U are the number of samples in D_B and D_U. X_b contains the labeled 12-leads ECG recordings, and Y_b represents the corresponding multi-label ground truth. Specifically, the c-th dimension in y_b^i contains the ground truth of category c, $y_b^{i,c} \in \{0, 1\}$. Hence, a given ECG recording might belong to multiple categories simultaneously. To clarify the narrative, the frequently used notations are summarized in Table I. As shown in Fig. 1, the proposed ECGMatch includes three modules: ECGAugment module, pseudo-label generation and refinement module, and label correlation alignment module. **In the ECGAugment module**, motivated by the weak-strong

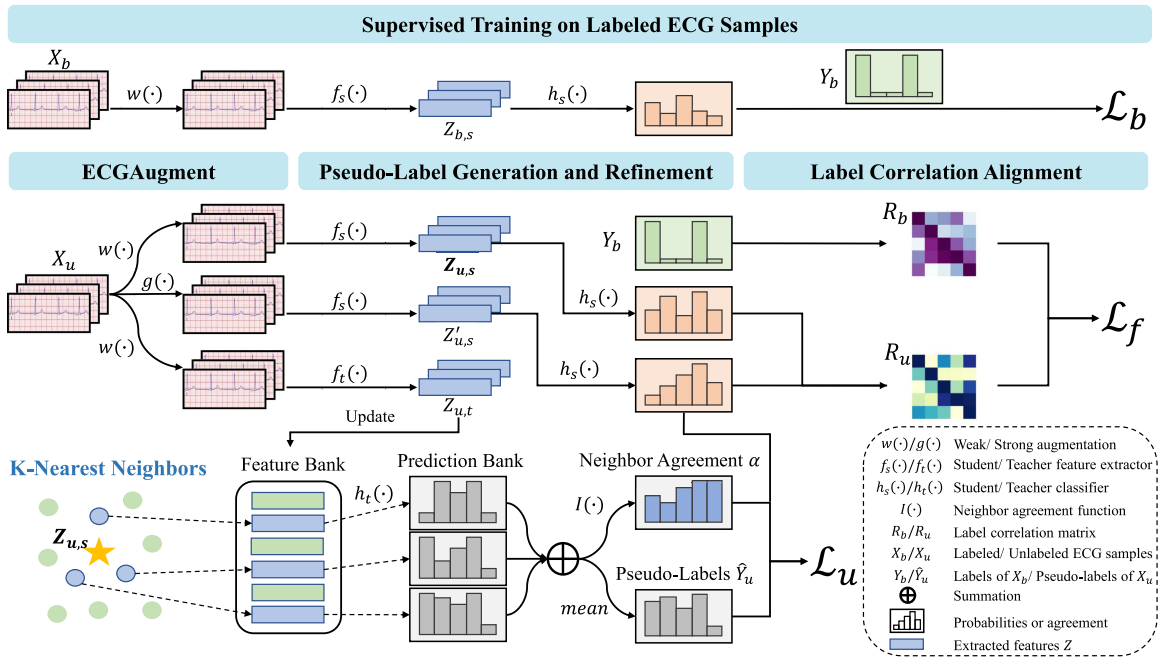


Fig. 1. Overall schematics of the ECGMatch. It consists of three losses and four parts. *1) Supervised training on the labeled ECG samples:* the student network $M_s = \{f_s(\cdot), h_s(\cdot)\}$ outputs CVDs predictions for the labeled samples X_b and computes the supervised loss $\mathcal{L}_b$ in (3). *2) ECGAugment for the unlabeled samples:* apply weak and strong augmentations to the unlabeled samples X_u. *3) Pseudo-label generation and refinement:* generate pseudo-labels for the unlabeled samples using two memory banks maintained by the teacher network $M_t = \{f_t(\cdot), h_t(\cdot)\}$; refine the raw pseudo-labels based on a neighbor agreement function $I(\cdot)$; computes the unsupervised loss $\mathcal{L}_u$ defined in (7). *4) Label correlation alignment:* estimate the label correlation matrices for the labeled and unlabeled samples and compute the loss $\mathcal{L}_f$ in (11).

augmentation method [5], we design a novel augmentation pipeline by investigating the intrinsic characteristics of the ECG signals, named as ECGAugment. **In the pseudo-label generation and refinement module**, we introduce a knowledge distillation method for pseudo-label generation. Then, we propose a neighbor agreement modeling method to compute the importance score for the pseudo labels, which can alleviate the negative effect of the inaccurate pseudo-labels. **In the label correlation alignment module**, we propose to align the label correlation matrices computed on the labeled data and unlabeled data by Frobenius norm regularization, which enables the model to capture the label dependency between different CVDs. More details about the proposed ECGMatch are presented below.

B. ECGAugment

One critical method to tackle the label scarcity problem is efficient data augmentation [29]. Although ECG augmentation methods have been well investigated in previous studies [17], [32], how to properly define a weak and strong augmentation pipeline for ECG-based semi-supervised learning is still challenging. Hence, we propose a novel augmentation pipeline for the ECG signals by leveraging their characteristic, termed as ECGAugment. Specifically, 'weak' augmentation $w(\cdot)$ is defined by randomly choosing one transformation to augment a 12-lead ECG signal $x \in \mathbb{R}^{12\times L}$, where L is the length of x. 1) **Signal Dropout**: we randomly set the ECG signal values within a random time window to zero. Its length and location are randomly generated from uniform distributions. This transformation enables the model to handle weak signals caused by bad contact of ECG electrodes [33]. 2) **Temporal Flipping**: motivated by previous studies [18], [34], we flip the original ECG signal along the temporal axis, which means the signal is read in reverse. 3) **Channel Reorganization**: Each row of x represents the ECG signal recorded at one lead (channel). Hence, we randomly change the order of the row vectors in the signal matrix x to shuffle its channel organization. 4) **Random Noise**: inspired by the noise contamination technique in ECG-based contrastive learning and adversarial learning [18], [35], we add a Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma)$ to the original signal x.

Motivated by the RandAugment technique for image augmentation [36], we define the 'strong' augmentation $g(\cdot)$ by randomly selecting $T \leq 4$ transformations to perturb the input signal x. Specifically, a transformation queue is randomly generated and transformations within the queue are applied one after another. For a random queue $\{2, 1, 3\}$, we successively apply Temporal Flipping, Signal Dropout, and Channel Reorganization to the input signal. Compared with traditional sequential perturbations which fix the number and the order of transformations [8], [18], the proposed method dramatically increases the diversity of the augmented samples by introducing extra randomness, which greatly increases the model performance [29], [36].

C. Pseudo-Label Generation for Multi-Label Learning

The key to robust semi-supervised learning is accurate pseudo-label generation, which has been demonstrated by many previous studies [4], [5], [6], [37], [38], [39]. However, previous studies mainly consider a single-label condition, where each

sample belongs to one category only. In contrast, we focus on a multi-label condition in this study, where each sample belongs to multiple categories simultaneously. Here, we generate the pseudo-labels using a knowledge distillation method. Specifically, we introduce a teacher model $M_t = \{f_t(\cdot), h_t(\cdot)\}$ and a student model $M_s = \{f_s(\cdot), h_s(\cdot)\}$, where $f(\cdot)$ is a feature extractor and $h(\cdot)$ is a multi-label classifier. As shown in Fig. 1, we first apply the weak augmentation $w(\cdot)$ and the strong augmentation $g(\cdot)$ to the unlabeled ECG recordings x_u, respectively. The teacher model extracts deep features $z_{u,t} = f_t(w(x_u))$ from the weak-augmented signals $w(x_u)$ and outputs the corresponding predictions $p_{u,t} = sigmoid(h_t(z_{u,t}))$. Then we store them in two memory banks (feature bank $Z = \{z^n_{u,t}\}^{N_U}_{n=1}$ and prediction bank $P = \{p^n_{u,t}\}^{N_U}_{n=1}$), N_U is the number of samples in D_U. Note that Z and P are updated on the fly with the current mini-batch. In this study, $p^n_{u,t} = [p^{n,1}_{u,t}, \ldots, p^{n,c}_{u,t}]$ is a C dimensional vector where the c-th element represents the prediction of class c, $p^{n,c}_{u,t} \in [0, 1]$. During training, the student model extracts a feature vector $z^i_{u,s} = f_s(w(x^i_u))$ from a given unlabeled sample x^i_u and assigns a pseudo-label ($\hat{y}^i_u$) for it using a widely used soft voting method [38]. Specifically, $\hat{y}^i_u$ is computed by integrating the predictions of its K-Nearest neighbors $\{z^k_{u,t}\}^K_{k=1}$ in the feature bank Z, given as

$$\hat{y}^i_u = \frac{1}{K}\sum_{k=1}^{K} p^k_{u,t}, \tag{1}$$

where $p^k_{u,t}$ is the prediction of $z^k_{u,t}$, which is the k-th nearest neighbor of the feature vector $z^i_{u,s}$ in the feature bank Z, K is the number of neighbors. $\{p^k_{u,t}\}^K_{k=1}$ is acquired by visiting the prediction bank P. To conduct the knowledge distillation process, we minimized the binary cross entropy loss between the prediction of the student model and the pseudo-label $\hat{y}^i_u$ given by the prediction bank. Motivated by the weak-strong consistency regularization method [5], we apply a strong augmentation $g(\cdot)$ to the unlabeled sample x^i_u and compute the corresponding student prediction by $q^i_{u,s} = sigmoid(h_s(z'_{u,s}))$, $z'_{u,s} = f_s(g(x^i_u))$. Then we compute the binary cross entropy loss between the pseudo-labels and the student predictions of the unlabeled samples, defined as

$$\mathcal{L}_u = -\frac{1}{B_u C}\sum_{i=1}^{B_u}\sum_{c=1}^{C}(1-\hat{y}^{i,c}_u)\log(1-q^{i,c}_{u,s}) + \hat{y}^{i,c}_u \log q^{i,c}_{u,s}, \tag{2}$$

where C is the number of categories in the dataset, and B_u is the number of unlabeled samples in the current mini-batch. Using the ground truth of the labeled samples in the mini-batch, we compute the supervised binary cross-entropy loss, defined as

$$\mathcal{L}_b = -\frac{1}{BC}\sum_{i=1}^{B}\sum_{c=1}^{C}(1-y^{i,c}_b)\log(1-p^{i,c}_{b,s}) + y^{i,c}_b \log p^{i,c}_{b,s}, \tag{3}$$

where B is the number of labeled samples in the current mini-batch, $p^{i,c}_{b,s} = sigmoid(h_s(f_s(w(x^i_b))))$ is the prediction given by the student model and $y^{i,c}_b \in \{0, 1\}$ is the corresponding ground truth. Combing (2) and (3), we compute the overall loss for semi-supervised multi-label classification, defined as

$$\mathcal{L} = \mathcal{L}_b + \lambda\mathcal{L}_u, \tag{4}$$

where λ is a hyperparameter controlling the weight of $\mathcal{L}_u$. Before pseudo-label generation, the teacher model M_t is pre-trained on the labeled dataset D_B using the (3). Then in the knowledge distillation process, the student model M_s is updated by stochastic gradient descent to minimize (4). To stabilize the maintained feature bank Z and the prediction bank P, the parameters θ_t of the teacher model M_t are updated by the momentum moving average of the parameters θ_s of the student model [38], [40], defined as

$$\theta_t = m\theta_t + (1-m)\theta_s, \tag{5}$$

where m is a momentum hyperparameter.

D. Pseudo-Label Refinement Based on Neighbor Agreement Modeling

Inaccurate pseudo-labels can hurt the model performance in semi-supervised learning. Consequently, the generated pseudo-labels $\hat{y}_u$ should be further refined to avoid this problem. Previous studies [5], [6] utilized fixed or dynamic thresholds to remove the pseudo-labels with low confidence. However, it is difficult to set up separate optimized thresholds for different categories in multi-label classification. Moreover, designing update strategies for dynamic thresholds needs numerous hyperparameters [15], [37]. Hence, we proposed a novel pseudo-label refinement method based on *neighbors agreement modeling* (NAM). It refines the raw pseudo-labels $\hat{y}_u$ by computing their neighbor agreement based on a neighbor agreement function $I(\cdot)$ and then adjusts their importance in the loss propagation. Compared with traditional threshold-based refinement method [5], [6], [41], NAM replaces the threshold control process with an importance weighting process, which is more hyperparameter-efficient in semi-supervised multi-label classification.

Recall that we have generated the raw pseudo-label $\hat{y}^i_u$ for the unlabeled sample x^i_u by averaging the prediction $p^k_{u,t}$ of its K-Nearest neighbors ((1)). Here, we sum up the neighbors' predictions $p^{k,c}_{u,t} \in [0, 1]$ and apply a neighbor agreement function $I(\cdot)$ to compute the neighbor agreement of the pseudo-label $\hat{y}^i_u$ on the c-th category.

$$\alpha^{i,c}_u = I\left(\sum_{k=1}^{K} p^{k,c}_{u,t}\right) = \left|\frac{2}{K}\sum_{k=1}^{K} p^{k,c}_{u,t} - 1\right|, \tag{6}$$

where K is the number of nearest neighbors. $\alpha^{i,c}_u \in [0, 1]$ is the neighbor agreement which can also be regarded as the model confidence on the pseudo-label $\hat{y}^i_u$. Combing the (6) and (2), we can rewrite the unsupervised binary cross entropy loss as

$$\mathcal{L}_u = -\frac{1}{B_u C}\sum_{i=1}^{B_u}\sum_{c=1}^{C}\alpha^{i,c}_u[(1-\hat{y}^{i,c}_u)\log(1-q^{i,c}_{u,s}) + \hat{y}^{i,c}_u \log q^{i,c}_{u,s}], \tag{7}$$

where $\alpha_u^{i,c}$ controls the weight of the $\hat{y}_u^{i,c}$ in loss computation. Specifically, the (6) allocates high weights ($\alpha_u^{i,c} \approx 1$) to the pseudo-labels with high agreement on neighbors' prediction ($\sum_{k=1}^{K} p_{u,t}^{k,c} \approx K$ or 0). Note that the label refinement process of the NAM module is only related to one hyperparameter K, which is the number of nearest neighbors. On the contrary, previous threshold-based methods need to set up fixed or dynamic thresholds for C independent categories in multi-label classification, which is less efficient in hyperparameters grid searching than the proposed NAM module. In addition, threshold-based methods typically discard pseudo-labels whose confidences are lower than the pre-defined thresholds. The selection of the thresholds is sensitive to the class distribution in training datasets [5], [42], which can result in suboptimal generalization performance when applied to the unseen dataset with a different class distribution. In contrast, the proposed NAM employs a soft method that adjusts the importance of the generated pseudo-labels rather than directly rejecting them. It is less sensitive to the class distribution in the training data compared with the threshold-based methods [15] and can enhance the model performance on the unseen dataset.

E. Label Correlation Alignment

The co-occurrence of CVDs leads to a strong relationship between different categories, which should be considered to achieve better prediction performance in multi-label classification [43], [44]. Previous studies focused on the label dependency within the labeled samples and utilized the semantic relationship between different categories to guide the model training [45], [46], [47]. However, it is hard to define the relationship among various CVDs without sufficient prior knowledge. On the other hand, ignoring the label dependency within the large-scale unlabeled samples results in information waste [46], [48]. Hence, we propose jointly capturing the label dependency within the labeled and unlabeled samples by computing two label correlation matrices (R_b and R_u). In practice, R_b is calculated using the labeled samples while R_u is estimated using the unlabeled samples. The computation process does not need extra prior information like the word-embedding correlation between different labels [46]. Then we minimize the discrepancy between the R_b and R_u to align the label dependencies computed by the labeled and unlabeled samples, which enhances the model performance in multi-label classification.

Firstly, we introduce how to compute the label correlation matrix R_b based on the labeled sample set $D_B = \{X_b, Y_b\}$. $Y_b = [y_b^1; y_b^2; \ldots; y_b^{N_B}]$ is a $N_B \times C$ label matrix, where C is the number of categories. The label correlation $\hat{r}_{c_1,c_2} \in [0,1]$ between classes c_1 and c_2 can be estimated by the similarity between the label sequences (y_{c_1}, y_{c_2}) on the two classes, where $y_{c_1} = [y_b^{1,c_1}; y_b^{2,c_1}, \ldots, y_b^{N_B,c_1}]$ and $y_{c_2} = [y_b^{1,c_2}; y_b^{2,c_2}, \ldots, y_b^{N_B,c_2}]$. We find that cosine similarity is more efficient for label correlation analysis than other metrics such as the Pearson coefficient. As shown in (8), with binarized labels y_{c_1} and y_{c_2}, it estimates the conditional probabilities between the classes c_1 and c_2 without being influenced by the class distributions of different datasets. The proof of (8) and detailed analysis of different similarity metrics can be found in Appendix B, available online. Based on the cosine similarity, the correlation $\hat{r}_{c_1,c_2}$ is computed as

$$\hat{r}_{c_1,c_2} = \frac{y_{c_1}^T y_{c_2}}{\|y_{c_1}\| \, \|y_{c_2}\|} \approx \sqrt{P(c_1=1|c_2=1)P(c_2=1|c_1=1)}, \tag{8}$$

And the label correlation matrix R_b can be computed by

$$R_b = \begin{bmatrix} \hat{r}_{1,1} & \hat{r}_{1,2} & \cdots & \hat{r}_{1,C} \\ \hat{r}_{2,1} & \hat{r}_{2,2} & \cdots & \hat{r}_{2,C} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{r}_{C,1} & \hat{r}_{C,2} & \cdots & \hat{r}_{C,C} \end{bmatrix} = N(Y)^T N(Y), \tag{9}$$

where $N(Y)$ is a normalization function which normalizes the column vectors of Y to unit vectors. For the unlabeled sample, we estimate the label correlation matrix R_u using the model prediction P_u output by the student network M_s. To improve the robustness of the estimated R_u, we simultaneously use the strongly-augmented and weakly-augmented samples to increase the sample size for computation. Hence, R_u is estimated as

$$R_u = N(P_u)^T N(P_u), P_u = \left[q_{u,s}^1; p_{u,s}^1; \ldots; q_{u,s}^{B_u}; p_{u,s}^{B_u}\right], \tag{10}$$

where P_u is a $2B_u \times C$ matrix containing the student predictions of the strongly and weakly augmented unlabeled samples ($g(x_u)$ and $w(x_u)$), and B_u is the number of unlabeled samples in the current mini-batch. Specifically, $q_{u,s}^i = sigmoid(h_s(f_s(g(x_u^i))))$ and $p_{u,s}^i = sigmoid(h_s(f_s(w(x_u^i))))$. The label correlation matrices represent the dependency and the semantic relationship between different CVDs, which should be consistent across the labeled and unlabeled data. Consequently, we minimize the discrepancy between the R_u and R_b using Frobenius norm regularization, defined as

$$\mathcal{L}_f = \|R_b - R_u\|_F, \tag{11}$$

where $\|\cdot\|_F$ represents the Frobenius norm of a given matrix. Finally, we formulate the final loss of the proposed ECGMatch by combing the supervised multi-label classification loss (3), the importance weighted unsupervised multi-label classification loss (7) and the label correlation alignment loss (11)

$$\mathcal{L} = \mathcal{L}_b + \lambda_u \mathcal{L}_u + \lambda_f \mathcal{L}_f, \tag{12}$$

where λ_u and λ_f are two hyperparameters controlling the importance of different objective functions. We present the complete algorithm for ECGMatch in Algorithm 1.

IV. Experiments and Datasets

A. Public ECG Databases

To evaluate the performance of the proposed ECGMatch model, we conduct experiments on four well-known public databases released on the PhysioNet website: The Georgia 12-lead ECG Challenge (G12EC) Database [49], the

Algorithm 1: ECGMatch Algorithm.

Input:
- Label dataset $D_B = \{X_b, Y_b\} = \{x_b^i, y_b^i\}_{i=1}^{N_B}$ and unlabeled dataset $D_U = \{X_u, -\} = \{x_u^i, -\}_{i=1}^{N_U}$;
- Student model M_s and teacher model M_t; Feature bank Z and prediction bank P; Batch size B

Output: Trained student model M_s;

pretrain the teacher model using (3) and D_B; compute the label correlation matrix R_b using (9);
for 1 to $Epoch$ **do**
  for 1 to $iteration$ **do** $//iteration = \frac{N_B}{B}$
  ## Mini-batch sampling and ECGAugment ##
    sample labeled data $\{x_b, y_b\}$ from D_B;
    sample unlabeled data $\{x_u, -\}$ from D_U;
    apply ECGAugment to x_b and x_u;
    compute the supervised loss $\mathcal{L}_b$ using $\{x_b, y_b\}$ and (3);
  ## Pseudo-label generation ##
    update the feature and prediction banks using x_u and M_t;
    generate pseudo-labels $\hat{y}_u$ for x_u using (1);
  ## Pseudo-label refinement ##
    compute the neighbor agreement α_u of $\hat{y}_u$ using (6);
    compute the unsupervised loss $\mathcal{L}_u$ using (7)
  ## Label correlation alignment ##
    compute the label correlation matrix R_u using (10);
    compute the loss $\mathcal{L}_f$ using (11); compute the final loss $\mathcal{L}$ using (12);
    update network M_s by minimizing $\mathcal{L}$ and stochastic gradient descent; update network M_t using (5).
  end for
  apply an early-stop strategy to avoid overfitting;
end for

Physikalisch-Technische Bundesanstalt (PTB-XL) database [10], the Chapman-Shaoxing databases [50] and the Ningbo databases [51]. The G12EC database contains 10,344 available ECG recordings, each lasting between 5 and 10 seconds long with a sampling frequency of 500 Hz. The PTB-XL database contains 22,353 available ECG recordings and each recording is around 10 seconds long at a sampling frequency of 500 Hz. The Chapman-Shaoxing database consists of ECG recordings from 10,646 subjects, while the Ningbo database contains 40258 ECG recordings. Each recording is sampled at 500 Hz, with a duration of 10 seconds. Unfortunately, the aforementioned databases employed different label annotation schemes and contained different kinds of CVDs, which led to a substantial category gap across databases. As a detailed discussion about the category gap problem is beyond the scope of our study, we simply addressed this issue by using a consistent label annotation scheme to re-annotate the databases. In summary, we re-annotated the ECG signals from the datasets by categorizing them into five classes (Abnormal Rhythms, ST/T Abnormalities, Conduction Disturbance, Other Abnormalities, and Normal Signals). Note that the ECG signals might belong to two or more categories simultaneously. Details about the annotation scheme can be found in Appendix A, available online. To preprocess the signals, we first normalized the length of the raw signals into 6144 samples in the time domain by zero-padding. Next, we applied a bandpass filter (1.0-47.0 Hz) to eliminate noise components within the raw ECG recordings. Finally, the signals were normalized using z-score normalization.

B. Implementation Details

In our implementation, we use the Attention-based Convolutional Neural Network [52] as the feature extractor $f(\cdot)$ in Fig. 1, where the dimension of the output feature z is 128. The classifier $h(\cdot)$ is designed as 128 neurons (input layer)-128 neurons (hidden layer 1)-5 neurons (output layer)-Sigmoid activation. The teacher network $M_t = \{f_t, h_t\}$ is pre-trained on the labeled sample set D_B, and the parameters of the student network $M_s = \{f_s, h_s\}$ are initialized with those of M_t. In the semi-supervised training process, the parameters of M_t are updated by (5), with a momentum of 0.999. We use the standard stochastic gradient descent (SGD) optimizer with a momentum of 0.9 for parameter optimization. The initial learning rate is set to 3e-2 with an exponential learning rate decay schedule as $\eta = \frac{\eta_0}{(1+\gamma e/E)^{-p}}$, where η_0 is the initial learning rate, e is the current training step and $E = 5000$ is the max training step. In each mini-batch, the number of labeled samples B is 64 and the number of unlabeled samples B_u is 448. The weights λ_u and λ_f in (12) are searched within the range of [0,1.6] with a step of 0.4.

C. Experimental Protocols for Model Evaluation

To assess the robustness of the proposed model on multi-label CVDs classification, we propose three distinct protocols for model evaluation when taking into account various clinical applications. **1) Within-dataset protocol.** For model training and evaluation, the training, validation, and testing data are

TABLE II
COMPARISON RESULTS BETWEEN ECGMATCH AND THE STATE-OF-THE-ART MODELS USING THE WITHIN-DATASET PROTOCOL

Methods	MixMatch [4]	FixMatch [5]	FlexMatch [6]	DST [54]	PerMatch [41]	SoftMatch [15]	UPS [19]	**ECGMatch**
				Ranking loss (The smaller, the better)				
G12EC	0.349±0.033	0.217±0.041	0.160±0.009	0.189±0.019	0.167±0.006	0.199±0.045	0.177±0.016	**0.140±0.006**
PTB-XL	0.345±0.004	0.170±0.010	0.146±0.004	0.279±0.148	0.200±0.065	0.158±0.016	0.156±0.001	**0.134±0.003**
Ningbo	0.178±0.014	0.106±0.048	0.153±0.024	0.082±0.020	0.200±0.020	0.197±0.047	0.172±0.082	**0.045±0.002**
Chapman	0.214±0.027	0.103±0.051	0.088±0.014	0.146±0.099	0.080±0.008	0.122±0.014	0.075±0.008	**0.052±0.002**
				Hamming loss (The smaller, the better)				
G12EC	0.538±0.032	0.306±0.004	0.303±0.004	0.330±0.013	0.321±0.009	0.311±0.009	0.294±0.008	**0.278±0.008**
PTB-XL	0.439±0.018	0.257±0.005	0.265±0.013	0.407±0.207	0.251±0.005	0.265±0.021	0.255±0.006	**0.233±0.009**
Ningbo	0.421±0.094	0.148±0.015	0.139±0.005	0.134±0.004	0.136±0.008	0.138±0.005	0.136±0.003	**0.122±0.001**
Chapman	0.423±0.070	0.168±0.009	0.189±0.008	0.180±0.009	0.182±0.006	0.196±0.003	0.167±0.010	**0.139±0.002**
				Coverage (The smaller, the better)				
G12EC	2.990±0.115	2.471±0.143	2.255±0.037	2.369±0.085	2.281±0.031	2.395±0.163	2.325±0.064	**2.173±0.027**
PTB-XL	2.728±0.007	2.065±0.028	1.971±0.015	2.481±0.555	2.187±0.261	2.007±0.061	2.016±0.009	**1.922±0.015**
Ningbo	2.317±0.054	1.978±0.200	2.164±0.094	1.880±0.080	2.346±0.086	2.347±0.186	2.241±0.329	**1.724±0.010**
Chapman	2.466±0.128	1.981±0.190	1.912±0.038	2.121±0.362	1.888±0.038	2.040±0.041	1.859±0.038	**1.761±0.021**
				MAP (The greater, the better)				
G12EC	0.479±0.012	0.703±0.008	0.719±0.008	0.690±0.016	0.711±0.007	0.717±0.010	0.719±0.011	**0.742±0.005**
PTB-XL	0.546±0.023	0.737±0.013	0.738±0.012	0.739±0.005	0.737±0.014	0.730±0.012	0.740±0.014	**0.748±0.009**
Ningbo	0.484±0.059	0.796±0.006	0.793±0.005	0.791±0.006	0.786±0.005	0.794±0.002	0.797±0.005	**0.808±0.001**
Chapman	0.500±0.073	0.730±0.005	0.732±0.005	0.721±0.015	0.736±0.005	0.736±0.004	0.737±0.012	**0.775±0.014**
				Marco AUC (The greater, the better)				
G12EC	0.668±0.023	0.841±0.004	0.850±0.004	0.833±0.014	0.843±0.005	0.846±0.004	0.848±0.005	**0.854±0.003**
PTB-XL	0.775±0.015	0.875±0.005	0.877±0.004	0.778±0.138	0.876±0.005	0.874±0.005	0.877±0.006	**0.880±0.005**
Ningbo	0.718±0.053	0.916±0.005	0.913±0.004	0.909±0.005	0.906±0.002	0.913±0.002	0.915±0.004	**0.925±0.001**
Chapman	0.749±0.052	0.897±0.003	0.900±0.003	0.898±0.008	0.900±0.004	0.899±0.002	0.901±0.007	**0.912±0.002**
				Marco G_{beta} score (The greater, the better)				
G12EC	0.339±0.004	0.452±0.006	0.460±0.004	0.448±0.011	0.450±0.009	0.447±0.005	0.465±0.007	**0.477±0.003**
PTB-XL	0.352±0.008	0.454±0.007	0.453±0.010	0.393±0.080	0.460±0.003	0.450±0.009	0.458±0.003	**0.467±0.009**
Ningbo	0.360±0.027	0.544±0.011	0.541±0.007	0.545±0.010	0.536±0.004	0.542±0.006	0.544±0.006	**0.563±0.001**
Chapman	0.368±0.053	0.523±0.012	0.526±0.015	0.523±0.024	0.521±0.012	0.523±0.013	0.530±0.017	**0.554±0.009**

The mean performance and standard deviations on four databases are shown across three seeds.

randomly sampled from one dataset in a ratio of 0.8 : 0.1 : 0.1. Then, we split the training data into labeled and unlabeled data in a ratio of 0.05 : 0.95. Finally, the average performance and standard deviations of four datasets are computed across three random seeds. **2) Mix-dataset protocols.** In this scenario, we randomly sample the training, validation, and testing data from four datasets simultaneously in a ratio of 0.8 : 0.1 : 0.1. The training data is split into labeled and unlabeled data in a ratio of 0.01 : 0.99. The average performance and standard deviations are calculated across three random seeds. This protocol considers a multi-center setting, where the training data contains samples from different datasets (centers). **3) Cross-dataset protocols.** To evaluate the model performance on the unseen testing dataset (s), we use three datasets for model training and validation and reserve the remaining one for testing. For example, we can reserve the G12EC dataset as the unseen test set and sample the training (90%) and validation data (10%) from the remaining three datasets (PTB-XL, Chapman, Ningbo). Only 1% of the training data is labeled, while the remaining 99% is unlabeled. We repeat the evaluation process until each dataset is used once as the unseen test set and report the average performance and standard deviations across three random seeds. This protocol serves as an external validation of the proposed model, which evaluates the model's generalization ability across different independent datasets.

We evaluate the performance of various models using multiple multi-label metrics including ranking loss, hamming loss, coverage, mean average precision (MAP), macro AUC and marco-G_{beta}. It is important to note that lower values of ranking loss, hamming loss and coverage indicate better performance, while lower values of MAP, macro AUC and marco-G_{beta} score mean worse performance. More details about these metrics can be found in [53]. The following section presents a comparison between the proposed ECGMatch and the existing literature based on the three experimental protocols and six evaluation metrics mentioned above. As there is limited research on the application of SSL for ECG-based multi-label classification, we replicated several state-of-the-art (SOTA) models that were originally implemented for image or text classification: MixMatch [4], FixMatch [5], FlexMatch [6], DST [54], PercentMatch [41], SoftMatch [15], UPS [19]. We ensure consistency across all compared models by employing identical backbones and augmentation strategies (ECGAugment). We also use the same set of common hyperparameters, such as learning rate and batch size. For the model-specific parameters such as the sharpen-temperature in FixMatch [5], we utilize the optimal settings recommended by the referenced studies.

V. RESULTS AND DISCUSSION

A. Comparisons With State-of-The-Art Methods

The performance of different models on different protocols is presented in Tables II, III, and IV. The results show that

TABLE III
COMPARISON RESULTS BETWEEN ECGMATCH AND THE STATE-OF-THE-ART MODELS USING THE CROSS-DATASET PROTOCOL

Methods	MixMatch [4]	FixMatch [5]	FlexMatch [6]	DST [54]	PerMatch [41]	SoftMatch [15]	UPS [19]	**ECGMatch**
Ranking loss (The smaller, the better)								
G12EC	0.333±0.026	0.290±0.020	0.243±0.023	0.251±0.009	0.255±0.018	0.281±0.033	0.248±0.011	**0.203±0.004**
PTB-XL	0.474±0.071	0.285±0.015	0.254±0.009	0.265±0.009	0.272±0.016	0.271±0.022	0.270±0.019	**0.248±0.005**
Ningbo	0.319±0.222	0.148±0.029	0.138±0.049	0.127±0.014	0.156±0.026	0.169±0.009	0.160±0.035	**0.102±0.006**
Chapman	0.212±0.016	0.147±0.040	0.126±0.017	0.126±0.027	0.169±0.053	0.169±0.042	0.156±0.043	**0.068±0.002**
Hamming loss (The smaller, the better)								
G12EC	0.380±0.003	0.343±0.007	0.337±0.006	0.357±0.007	0.350±0.009	0.352±0.015	0.349±0.006	**0.331±0.007**
PTB-XL	0.517±0.114	0.365±0.019	0.362±0.014	0.360±0.029	0.345±0.019	0.372±0.020	0.359±0.009	**0.310±0.001**
Ningbo	0.353±0.056	0.287±0.012	0.308±0.022	0.280±0.028	0.307±0.018	0.315±0.002	0.288±0.011	**0.253±0.008**
Chapman	0.273±0.013	0.259±0.008	0.276±0.011	0.267±0.014	0.277±0.014	0.275±0.003	0.262±0.015	**0.219±0.003**
Coverage (The smaller, the better)								
G12EC	2.892±0.094	2.740±0.078	2.586±0.092	2.599±0.020	2.619±0.071	2.728±0.125	2.594±0.036	**2.415±0.016**
PTB-XL	3.206±0.275	2.512±0.069	2.400±0.037	2.432±0.038	2.451±0.050	2.454±0.084	2.445±0.073	**2.379±0.023**
Ningbo	2.822±0.838	2.160±0.121	2.131±0.198	2.084±0.070	2.196±0.104	2.245±0.036	2.208±0.150	**1.971±0.025**
Chapman	2.352±0.067	2.142±0.176	2.046±0.068	2.035±0.095	2.217±0.206	2.235±0.173	2.183±0.181	**1.803±0.008**
MAP (The greater, the better)								
G12EC	0.591±0.012	0.616±0.013	0.632±0.005	0.630±0.010	0.622±0.007	0.621±0.005	0.630±0.004	**0.657±0.009**
PTB-XL	0.518±0.030	0.532±0.007	0.551±0.006	0.538±0.006	0.558±0.004	0.545±0.008	0.553±0.009	**0.591±0.012**
Ningbo	0.560±0.067	0.663±0.006	0.665±0.003	0.667±0.004	0.649±0.001	0.658±0.003	0.667±0.003	**0.689±0.002**
Chapman	0.702±0.005	0.730±0.007	0.726±0.005	0.727±0.005	0.710±0.001	0.728±0.004	0.726±0.005	**0.748±0.004**
Marco AUC (The greater, the better)								
G12EC	0.755±0.008	0.779±0.007	0.789±0.005	0.784±0.009	0.781±0.008	0.783±0.004	0.787±0.001	**0.805±0.004**
PTB-XL	0.733±0.035	0.767±0.008	0.780±0.004	0.771±0.002	0.780±0.006	0.773±0.008	0.779±0.010	**0.800±0.010**
Ningbo	0.810±0.045	0.869±0.003	0.867±0.001	0.867±0.002	0.864±0.003	0.866±0.003	0.872±0.001	**0.874±0.002**
Chapman	0.864±0.005	0.888±0.004	0.889±0.004	0.889±0.002	0.880±0.001	0.888±0.000	0.890±0.002	**0.900±0.002**
Marco G_{beta} **score** (The greater, the better)								
G12EC	0.376±0.002	0.387±0.005	0.394±0.001	0.388±0.004	0.390±0.005	0.389±0.005	0.392±0.005	**0.403±0.002**
PTB-XL	0.312±0.032	0.337±0.003	0.346±0.006	0.345±0.005	0.353±0.004	0.347±0.009	0.347±0.011	**0.369±0.001**
Ningbo	0.380±0.028	0.419±0.007	0.413±0.014	0.429±0.006	0.408±0.008	0.408±0.003	0.426±0.003	**0.442±0.003**
Chapman	0.463±0.003	0.484±0.006	0.470±0.014	0.489±0.004	0.456±0.009	0.467±0.003	0.482±0.009	**0.516±0.006**

The mean performance and standard deviations on four databases are shown across three seeds.

TABLE IV
COMPARISON RESULTS BETWEEN ECGMATCH AND THE STATE-OF-THE-ART MODELS USING THE MIX-DATASET PROTOCOL

Methods	MixMatch [4]	FixMatch [5]	FlexMatch [6]	DST [54]	PerMatch [41]	SoftMatch [15]	UPS [19]	**ECGMatch**
Ranking loss	0.241±0.057	0.233±0.026	0.205±0.014	0.189±0.023	0.227±0.031	0.236±0.033	0.181±0.012	**0.150±0.001**
Hamming loss	0.313±0.014	0.292±0.006	0.299±0.009	0.295±0.007	0.307±0.014	0.303±0.006	0.288±0.004	**0.270±0.001**
Coverage	2.462±0.218	2.437±0.102	2.322±0.061	2.265±0.098	2.411±0.125	2.445±0.114	2.230±0.057	**2.101±0.009**
MAP	0.625±0.021	0.643±0.009	0.643±0.009	0.645±0.011	0.635±0.013	0.647±0.005	0.640±0.009	**0.658±0.006**
Marco AUC	0.827±0.011	0.834±0.005	0.832±0.004	0.836±0.006	0.831±0.004	0.835±0.003	0.834±0.005	**0.838±0.003**
Marco G_{beta}	0.417±0.008	0.431±0.003	0.432±0.004	0.435±0.003	0.428±0.008	0.431±0.007	0.434±0.004	**0.442±0.002**

The mean performance and standard deviations on four databases are shown across three seeds.

ECGMatch achieves the leading performance in all experimental protocols, which demonstrates its superiority. For example, the averaged performance of the ECGMatch is better than the threshold-based SOTA models, such as FixMatch [5], FlexMatch [6], DST [54], especially when the test data comes from an unseen dataset. Specifically, the performance difference between the ECGMatch and the other models in the cross-dataset protocol is more distinct than that in the within-dataset and mix-dataset protocols. This phenomenon suggests that the NAM module is more efficient for pseudo-label refinement in multi-label classification than fixed or dynamic threshold strategies, especially on unseen datasets. Additionally, we notice that the ECGMatch achieves better performance than PercentMatch [41] and UPS [19], which are the latest models designed for semi-supervised multi-label learning. This observation indicates that capturing label relationships within the labeled and unlabeled samples benefits multi-label classification, while this property is ignored in the above two competitors. More details about the contributions of each component are listed in the next sub-section. In summary, the noteworthy improvements in different protocols demonstrate the potential of the proposed ECGMatch to be implemented in various clinical applications.

B. Ablation Study

In order to quantitatively assess the contribution of different modules in the ECGMatch, we successively remove one of them and evaluate the model performance using the three established protocols. Tables V, VI, and VII report the ablation studies on different experimental protocols. 1) When the pseudo-label generation module is removed ($\lambda_u = 0$, (12)), the performance of the proposed model decreases in all the experimental protocols, which demonstrates the advantages of introducing pseudo-labels for semi-supervised learning. For example, in the within-dataset protocols (Table V), the hamming loss on the Chapman dataset increases from 0.139 ± 0.002 to 0.163 ± 0.009 while the MAP decreases from 0.775 ± 0.014 to 0.761 ± 0.010. Notably, as the

TABLE V
ABLATION STUDY OF THE PROPOSED ECGMATCH (WITHIN-DATASET PROTOCOL)

Methods	G12EC	PTB	Ningbo	Chapman
Ranking loss (The smaller, the better)				
Without pseudo-label generation	0.145±0.008	0.140±0.004	0.053±0.004	0.060±0.003
Without pseudo-label refinement	0.164±0.012	0.140±0.005	0.046±0.002	0.066±0.013
Without label correlation alignment	0.150±0.010	0.138±0.007	0.063±0.010	0.061±0.002
ECGMatch	**0.140±0.006**	**0.134±0.003**	**0.045±0.002**	**0.052±0.002**
Hamming loss (The smaller, the better)				
Without pseudo-label generation	0.300±0.014	0.244±0.006	0.127±0.006	0.163±0.009
Without pseudo-label refinement	0.296±0.009	0.241±0.013	0.138±0.012	0.163±0.014
Without label correlation alignment	0.290±0.007	0.243±0.005	0.131±0.003	0.151±0.011
ECGMatch	**0.278±0.008**	**0.233±0.009**	**0.122±0.001**	**0.139±0.002**
Coverage (The smaller, the better)				
Without pseudo-label generation	2.192±0.031	1.946±0.019	1.758±0.018	1.795±0.010
Without pseudo-label refinement	2.229±0.047	1.946±0.021	1.722±0.011	1.822±0.055
Without label correlation alignment	2.220±0.048	1.939±0.032	1.804±0.046	1.803±0.013
ECGMatch	**2.173±0.027**	**1.922±0.015**	**1.724±0.010**	**1.761±0.021**
MAP (The greater, the better)				
Without pseudo-label generation	0.728±0.006	0.739±0.009	0.801±0.002	0.761±0.010
Without pseudo-label refinement	0.725±0.018	0.741±0.012	0.805±0.004	0.739±0.040
Without label correlation alignment	0.731±0.007	0.741±0.010	0.794±0.006	0.751±0.006
ECGMatch	**0.742±0.005**	**0.748±0.009**	**0.808±0.001**	**0.775±0.014**
Marco AUC (The greater, the better)				
Without pseudo-label generation	0.849±0.005	0.878±0.004	0.920±0.001	0.905±0.004
Without pseudo-label refinement	0.852±0.003	0.877±0.007	0.922±0.002	0.906±0.005
Without label correlation alignment	0.851±0.005	0.879±0.005	0.915±0.004	0.906±0.003
ECGMatch	**0.854±0.003**	**0.880±0.005**	**0.925±0.001**	**0.912±0.002**
Marco G_{beta} score (The greater, the better)				
Without pseudo-label generation	0.467±0.010	0.463±0.007	0.553±0.005	0.538±0.020
Without pseudo-label refinement	0.460±0.006	0.465±0.012	0.539±0.016	0.539±0.011
Without label correlation alignment	0.467±0.011	0.463±0.005	0.544±0.002	0.541±0.019
ECGMatch	**0.477±0.003**	**0.467±0.009**	**0.563±0.001**	**0.554±0.009**

parameter λ_u is set to zero, the following refinement module is also disabled. 2) The significant negative effect of removing the pseudo-label refinement module is observed in the results. In the cross-dataset protocol (Table VI), the hamming loss on the Chapman dataset increases from 0.219 ± 0.003 to 0.242 ± 0.007 while the MAP drops from 0.748 ± 0.004 to 0.732 ± 0.006. This phenomenon indicates that increasing the importance of the trust-worthy pseudo-labels in loss computation greatly enhances the model performance. 3) Comparing the results with and without the label correlation alignment module ($\lambda_f = 0$, (12)), a significant performance drop is observed when the module is removed. In the mix-dataset protocols (Table VII), the hamming loss increases from 0.270 ± 0.001 to 0.282 ± 0.010 while the MAP decreases from 0.658 ± 0.006 to 0.640 ± 0.010. This phenomenon demonstrates the benefits of capturing the correlation between different categories, which has also been reported in other multi-label learning studies [43], [44], [46].

C. Comparison of Different Augmentation Strategies

In this section, we further investigate the effectiveness of the ECGAugment module in ECG signal augmentation. Using the aforementioned protocols, we compare the performance of the ECGAugment with the fixed sequential perturbations proposed in previous studies [18], [34]. Following their setups, we apply two successive Gaussian perturbations to the ECG recordings from a mini-batch for strong augmentation and a single Gaussian perturbation for weak augmentation. The averaged performance across four datasets is shown in Fig. 2, where an obvious performance enhancement attributed to the ECGAugment could be observed from the performance of different models. For the evaluation metrics where smaller is better, the blue zones (ECGAugment) in the radar charts are surrounded by the red zones (fixed sequential perturbations). Conversely, for the metrics where greater is better, the blue zones cover the red zones. These phenomena demonstrate the superiority of the proposed ECGAugment in the downstream classification tasks. In other words, it increases the sample diversity by enhancing the randomness in data augmentation, which can improve the model performance [29], [36], [55].

D. Statistical Analysis

To statistically analyze the performance difference between the ECGMatch and other SOTA models, a commonly used *Friedman test* and the post-hoc *Bonferroni-Dunn test* are employed. Following the pipeline of the aforementioned tests [56], we use the performance of different models in the within-dataset protocol and cross-dataset protocol for comparison. Table VIII presents the Friedman statistics F_F and the associated critical value for each metric (comparing models $k = 8$, datasets $N = 4$). Based on the results (F_F>3.2590), we can reject the null hypothesis that the compared models show no significant difference in performance at a 0.05 significance level. Then

TABLE VI
ABLATION STUDY OF THE PROPOSED ECGMATCH (CROSS-DATASET PROTOCOL)

Methods	G12EC	PTB	Ningbo	Chapman
Ranking loss (The smaller, the better)				
Without pseudo-label generation	0.215±0.012	0.259±0.007	0.109±0.005	0.082±0.007
Without pseudo-label refinement	0.225±0.011	0.273±0.010	0.143±0.015	0.079±0.000
Without label correlation alignment	0.226±0.019	0.252±0.002	0.126±0.010	0.101±0.009
ECGMatch	**0.203±0.004**	**0.248±0.005**	**0.102±0.006**	**0.068±0.002**
Hamming loss (The smaller, the better)				
Without pseudo-label generation	0.338±0.002	0.355±0.010	0.271±0.006	0.242±0.007
Without pseudo-label refinement	0.358±0.008	0.350±0.014	0.302±0.009	0.266±0.030
Without label correlation alignment	0.344±0.001	0.341±0.014	0.281±0.010	0.270±0.003
ECGMatch	**0.331±0.007**	**0.310±0.001**	**0.253±0.008**	**0.219±0.003**
Coverage (The smaller, the better)				
Without pseudo-label generation	2.455±0.038	2.425±0.026	1.994±0.014	1.867±0.032
Without pseudo-label refinement	2.515±0.020	2.510±0.080	2.144±0.129	1.816±0.012
Without label correlation alignment	2.498±0.067	2.393±0.014	2.085±0.037	1.949±0.033
ECGMatch	**2.415±0.016**	**2.379±0.023**	**1.971±0.025**	**1.803±0.008**
MAP (The greater, the better)				
Without pseudo-label generation	0.643±0.006	0.568±0.006	0.665±0.007	0.732±0.006
Without pseudo-label refinement	0.635±0.013	0.571±0.013	0.664±0.018	0.739±0.008
Without label correlation alignment	0.639±0.010	0.572±0.011	0.670±0.003	0.718±0.006
ECGMatch	**0.657±0.009**	**0.591±0.012**	**0.689±0.002**	**0.748±0.004**
Marco AUC (The greater, the better)				
Without pseudo-label generation	0.796±0.002	0.782±0.007	0.866±0.002	0.890±0.005
Without pseudo-label refinement	0.788±0.007	0.790±0.003	0.865±0.004	0.889±0.005
Without label correlation alignment	0.792±0.004	0.794±0.011	0.869±0.001	0.884±0.004
ECGMatch	**0.805±0.004**	**0.800±0.010**	**0.874±0.002**	**0.900±0.002**
Marco G_{beta} score (The greater, the better)				
Without pseudo-label generation	0.397±0.003	0.357±0.007	0.441±0.005	0.497±0.005
Without pseudo-label refinement	0.390±0.003	0.355±0.001	0.421±0.001	0.506±0.017
Without label correlation alignment	0.392±0.002	0.357±0.011	0.430±0.005	0.476±0.008
ECGMatch	**0.403±0.002**	**0.369±0.001**	**0.442±0.003**	**0.516±0.006**

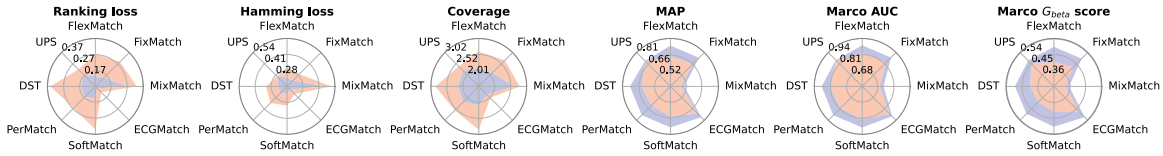
TABLE VII
ABLATION STUDY OF THE PROPOSED ECGMATCH (MIX-DATASET PROTOCOLS)

Methods	Ranking loss	Hamming loss	Coverage	MAP	Marco AUC	Marco G_{beta} score
Without pseudo-label generation	0.156±0.003	0.286±0.005	2.122±0.015	0.651±0.007	0.837±0.003	0.438±0.002
Without pseudo-label refinement	0.165±0.011	0.285±0.006	2.163±0.048	0.651±0.004	0.829±0.005	0.436±0.004
Without label correlation alignment	0.169±0.004	0.282±0.010	2.183±0.019	0.640±0.010	0.832±0.004	0.432±0.005
ECGMatch	**0.150±0.001**	**0.270±0.001**	**2.101±0.009**	**0.658±0.006**	**0.838±0.003**	**0.442±0.002**

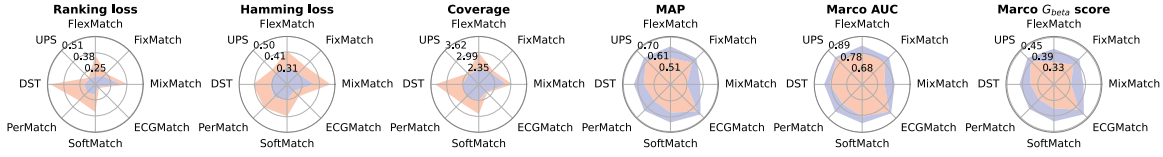
TABLE VIII
FRIEDMAN STATISTICS F_F FOR EACH METRIC AND THE CRITICAL VALUE AT 0.05 SIGNIFICANCE LEVEL (THE NUMBER OF COMPARING MODELS $k = 8$ AND DATASETS $N = 4$)

Within-dataset protocol								
Evaluation metric	F_F	critical values	Evaluation metric	F_F	critical values	Evaluation metric	F_F	critical values
Ranking loss	5.4706	3.2590	MAP	17.1600	3.2590	Coverage	5.1290	3.2590
Hamming loss	7.1818	3.2590	Marco AUC	16.0189	3.2590	Marco G_{beta}	9.0000	3.2590
Cross-dataset protocol								
Evaluation metric	F_F	critical values	Evaluation metric	F_F	critical values	Evaluation metric	F_F	critical values
Ranking loss	20.4419	3.2590	MAP	5.6154	3.2590	Coverage	22.8462	3.2590
Hamming loss	5.0640	3.2590	Marco AUC	11.0000	3.2590	Marco G_{beta}	5.0640	3.2590

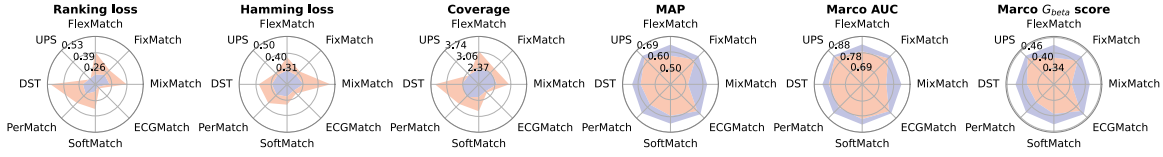
the post-hoc *Bonferroni-Dunn test* are applied to describe the performance gap between the control model (ECGMatch) and the other models. For each evaluation metric, we calculate the average rank of all the models across four datasets and determine the rank differences between the control model and the compared models. Note that the top-performing model is assigned a rank of 1, and the second-best model gets a rank of 2, and so on. The control model (ECGMatch) is significantly better than the compared model if their rank difference is larger than one *critical difference* (CD=4.6592 in our experiment). Fig. 3 presents the mean rank of different models on different evaluation metrics. It is evident that the proposed ECGMatch ranks the best in



(a) Within-dataset protocol

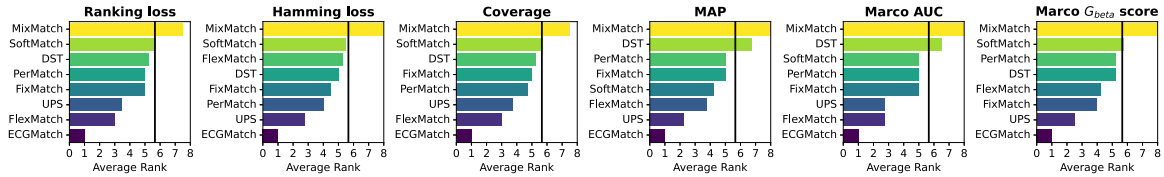


(b) Cross-dataset protocol

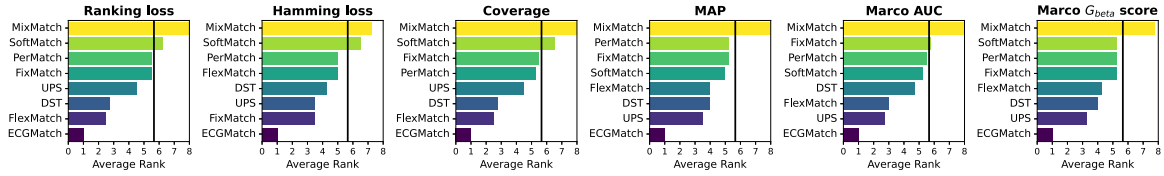


(c) Mix-dataset protocol

Fig. 2. Performance comparison of two augmentation pipelines using radar charts. The vertices of the red zone denote the performance of the model with the fixed sequential perturbations, while the vertices of the blue zone represent the performance of the model with the proposed ECGAugment.



(a) Within-dataset protocol



(b) Cross-dataset protocol

Fig. 3. Comparison of ECGMatch against other compared models based on the Bonferroni-Dunn test (cross-dataset protocol). ECGMatch is deemed to have a significantly better performance than one compared model if their average ranks differ by at least one *critical difference* $= 4.6592$, as denoted by the intersection of the bar with the black lines.

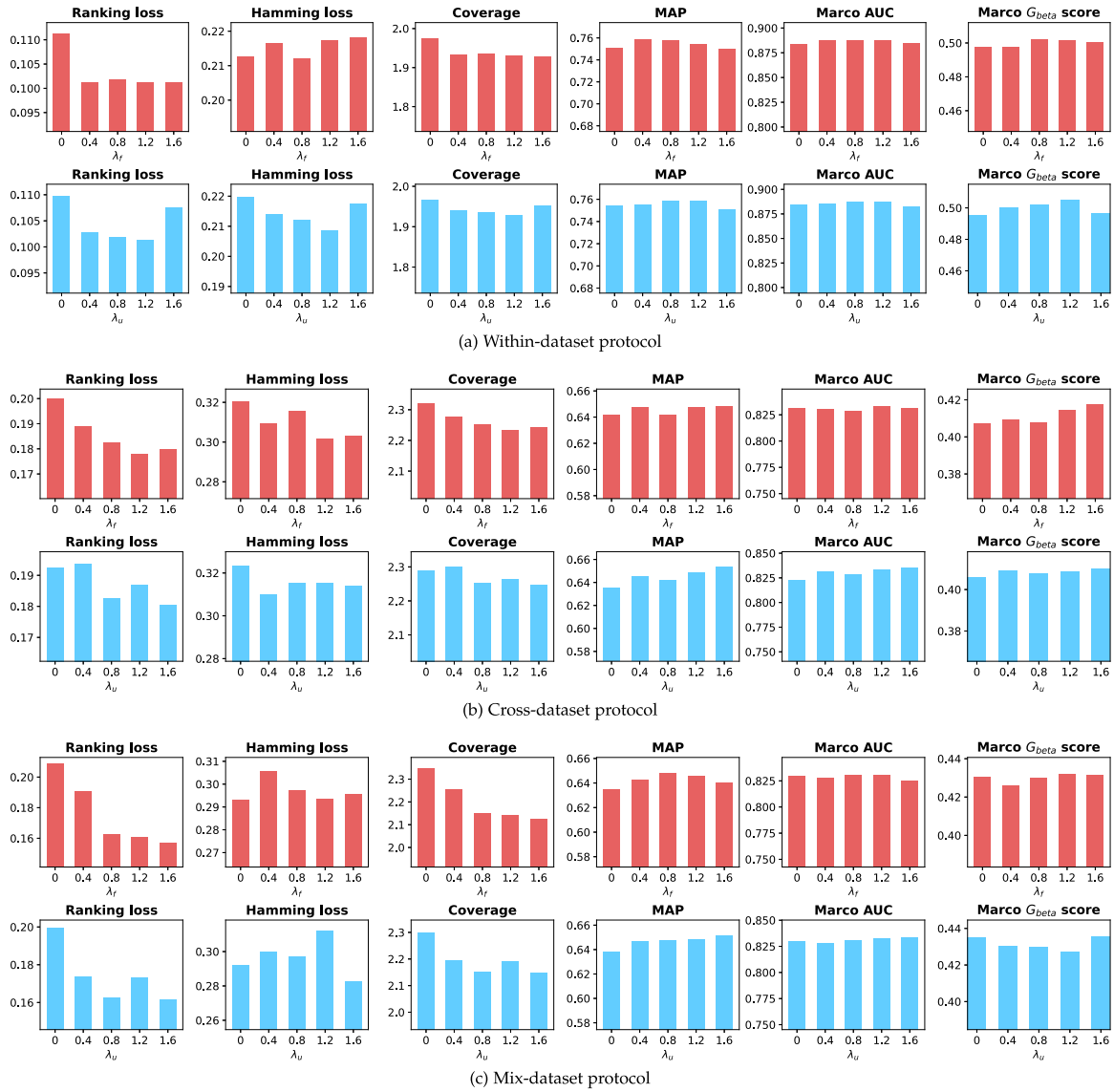
terms of all the metrics and outperforms some competitors at a 0.05 significance level, such as MixMatch [4], DST [54] and SoftMatch [15]. In summary, these statistical results demonstrate the superiority of the proposed ECGMatch.

E. Sensitivity Analysis

In this section, we use a grid-search method to investigate the impact of varying hyperparameters on the performance of the proposed model. For simplicity, we only focus on two critical hyperparameters λ_u and λ_f in (12). Specifically, λ_u controls the weight of the unsupervised binary cross entropy loss L_u, while λ_f controls the weight of the label correlation alignment loss L_f. In the grid search process, we adjust the values of the hyperparameters and use different evaluation protocols to evaluate the average model performance across four datasets. First, we fix λ_u at 0.8 and adjust λ_f from 0 to 1.6 in steps of 0.4. Then, we fix λ_f at 0.8 and adjust λ_u in the same manner. As illustrated in Fig. 4, the performance of the proposed ECGMatch in each evaluation metric is relatively insensitive to the changes of the two hyperparameters, which suggests its stability in clinical applications.

F. Effect of the Ratio of Labeled Samples

The performance of SSL models is influenced by the ratio of labeled samples to the total number of training samples [4], [5], [6]. We adjust the ratio during model training and investigate whether ECGMatch can reduce the need for labeled samples



(a) Within-dataset protocol

(b) Cross-dataset protocol

(c) Mix-dataset protocol

Fig. 4. Average model performance in different protocols under varying hyperparameters.

compared to other models. For simplicity, we present the experiment results of the cross-dataset protocol in Fig. 5. The results of the other protocols can be found in Appendix D, available online. Comparing the model performance under ratios 0.05 and 0.01, we can observe that ECGMatch achieves comparable performance to other models across all six metrics while reducing the required number of labeled samples by 5%. This phenomenon validates the efficiency of the proposed ECGMatch in reducing the requirement for human annotations during model training.

G. *Effect of Different Annotation Schemes*

In this section, we conduct experiments using another annotation scheme to further validate the superior performance of the ECGMatch. Specifically, the scheme from the PTB-XL database is employed for model training and evaluation. The scheme categorizes CVDs into four super-classes: Conduction Disturbance, ST/T Abnormalities, Myocardial Infarction and Hypertrophy. The recordings with sinus rhythm are classified as normal recordings. Given that Myocardial Infarction and Hypertrophy are predominantly present in the PTB-XL database [10], [49], we conduct the within-dataset protocol to evaluate the model performance on the database and present the results in Table. IX. It can be observed that the proposed ECGMatch achieves superior performance compared with the state-of-the-art models. For example, ECGMatch increases the MAP from 0.688 ± 0.007 to 0.705 ± 0.002 and decreases the Ranking loss from 0.144 ± 0.007 to 0.130 ± 0.003. In Appendix E, available online, we report the experiment results on the annotation scheme adopted by Cinc2020/2021 challenges [49], [57], where the proposed ECGMatch consistently performs better than other models. Based on the aforementioned experiments, we can conclude that the proposed ECGMatch achieves superior performance across various annotation schemes, which demonstrates its robustness in different CVDs prediction tasks.

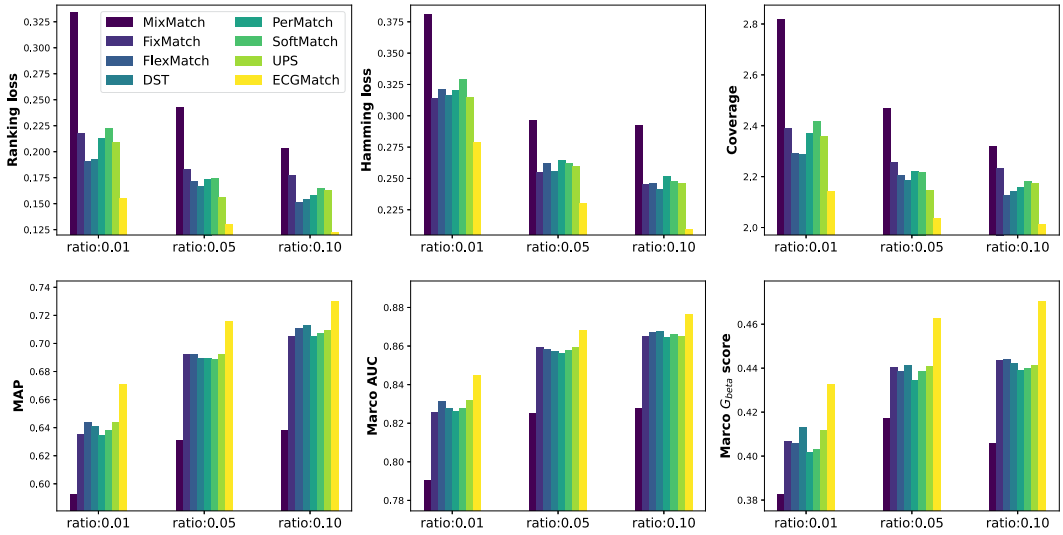


Fig. 5. Performance comparison of different models under various labeled sample ratios (cross-dataset protocol).

TABLE IX
COMPARISON RESULTS BETWEEN ECGMATCH AND THE STATE-OF-THE-ART MODELS USING THE WITHIN-DATASET PROTOCOL, UNDER THE ANNOTATION SCHEME OF PTB-XL

Methods	MixMatch [4]	FixMatch [5]	FlexMatch [6]	DST [54]	PerMatch [41]	SoftMatch [15]	UPS [19]	**ECGMatch**
Ranking loss	0.226±0.069	0.181±0.019	0.174±0.023	0.199±0.007	0.144±0.007	0.200±0.038	0.160±0.013	**0.130±0.003**
Hamming loss	0.379±0.024	0.290±0.010	0.249±0.013	0.346±0.028	0.253±0.004	0.246±0.009	0.260±0.009	**0.229±0.003**
Coverage	2.147±0.225	2.008±0.073	1.982±0.090	2.068±0.027	1.860±0.024	2.086±0.149	1.923±0.055	**1.804±0.012**
MAP	0.548±0.030	0.659±0.014	0.687±0.007	0.588±0.042	0.680±0.001	0.688±0.007	0.674±0.014	**0.705±0.002**
Marco AUC	0.793±0.012	0.852±0.005	0.863±0.002	0.821±0.020	0.861±0.001	0.864±0.005	0.858±0.009	**0.870±0.003**
Marco G_{beta}	0.346±0.010	0.402±0.005	0.417±0.006	0.373±0.016	0.414±0.003	0.422±0.005	0.416±0.006	**0.424±0.003**

The mean performance and standard deviations are shown across 3 seeds.

VI. CONCLUSION

In this study, we point out three important real-world challenges in ECG-based CVDs prediction: 1) Label scarcity problem. 2) Poor performance on unseen datasets. 3) Co-occurrence of multiple CVDs. To address the challenges simultaneously, we propose a novel framework (ECGMatch) that combines data augmentation, pseudo-label learning, and label correlation alignment modules to formulate a unified framework. Further, we re-annotate four public datasets and propose three practical experimental protocols to conduct a multi-dataset evaluation of the proposed model. Extensive experiments on three protocols and four datasets convincingly demonstrated the superiority of the proposed model against other SOTA models. We believe the proposed ECGMatch can provide a reliable baseline for future research on ECG-based CVDs prediction. However, the class imbalance problem in the ECG datasets continues to pose a significant challenge. Therefore, we advocate for future research on this ongoing issue.

REFERENCES

[1] B. B. Kelly et al., "Promoting cardiovascular health in the developing world: A critical challenge to achieve global health," 2010.
[2] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, Sep. 2006, vol. 2, doi: 10.7551/mitpress/9780262033589.001.0001.
[3] H. Lee, S. Shin, and H. Kim, "ABC: Auxiliary balanced classifier for class-imbalanced semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 7082–7094.
[4] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "MixMatch: A holistic approach to semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–11.
[5] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 596–608.
[6] B. Zhang et al., "FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 18 408–18 419.
[7] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH Arrhythmia database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, May/Jun. 2001.
[8] L. C. Oliveira, Z. Lai, H. M. Siefkes, and C.-N. Chuah, "Generalizable semi-supervised learning strategies for multiple learning tasks using 1-D biomedical signals," in *Proc. Int. Conf. Neural Inf. Process. Syst. Workshop*, 2022, pp. 1–7.
[9] W. Liu, H. Wang, X. Shen, and I. W. Tsang, "The emerging trends of multi-label learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7955–7974, Nov. 2022.
[10] P. Wagner et al., "PTB-XL: A large publicly available electrocardiography dataset," *Sci. Data*, vol. 7, no. 1, 2020, Art. no. 154.
[11] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep learning for ECG analysis: Benchmarks and insights from PTB-XL," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 5, pp. 1519–1528, May 2021.
[12] Z. Ge et al., "Multi-label correlation guided feature fusion network for abnormal ECG diagnosis," *Knowl.-Based Syst.*, vol. 233, 2021, Art. no. 107508.
[13] S. Ran, X. Li, B. Zhao, Y. Jiang, X. Yang, and C. Cheng, "Label correlation embedding guided network for multi-label ECG arrhythmia diagnosis," *Knowl.-Based Syst.*, vol. 270, 2023, Art. no. 110545.
[14] J. Lai et al., "Practical intelligent diagnostic algorithm for wearable 12-lead ECG via self-supervised learning on large-scale dataset," *Nature Commun.*, vol. 14, no. 1, 2023, Art. no. 3741.
[15] H. Chen et al., "SoftMatch: Addressing the quantity-quality tradeoff in semi-supervised learning," in *Proc. 11th Int. Conf. Learn. Representations*, 2023, pp. 1–21.

[16] X. Zhai, Z. Zhou, and C. Tin, "Semi-supervised learning for ECG classification without patient-specific labeled data," *Expert Syst. Appl.*, vol. 158, 2020, Art. no. 113411.
[17] P. Zhang, Y. Chen, F. Lin, S. Wu, X. Yang, and Q. Li, "Semi-supervised learning for automatic atrial fibrillation detection in 24-hour Holter monitoring," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 8, pp. 3791–3801, Aug. 2022.
[18] D. Kiyasseh, T. Zhu, and D. A. Clifton, "CLOCS: Contrastive learning of cardiac signals across space, time, and patients," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5606–5615.
[19] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–20.
[20] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ECG classification by 1-D convolutional neural networks," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 664–675, Mar. 2016.
[21] G. D. Clifford et al., "AF classification from a short single lead ECG recording: The physionet/computing in cardiology challenge 2017," *Comput. Cardiol.*, vol. 44, pp. 1–4, 2017.
[22] A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Med.*, vol. 25, no. 1, pp. 65–69, 2019.
[23] A. H. Ribeiro et al., "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nature Commun.*, vol. 11, no. 1, 2020, Art. no. 1760.
[24] D. Kiyasseh, T. Zhu, and D. Clifton, "A clinical deep learning framework for continually learning from cardiac signals across diseases, time, modalities, and institutions," *Nature Commun.*, vol. 12, no. 1, 2021, Art. no. 4221.
[25] Y. Jin et al., "A novel interpretable method based on dual-level attentional deep neural network for actual multilabel arrhythmia detection," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2021.
[26] H. Tesfai et al., "Lightweight shufflenet based CNN for arrhythmia classification," *IEEE Access*, vol. 10, pp. 111 842–111 854, 2022.
[27] Y. Huang, G. G. Yen, and V. S. Tseng, "Snippet policy network for multi-class varied-length ECG early classification," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 6, pp. 6349–6361, 2023, doi: 10.1109/TKDE.2022.3160706.
[28] Y. Huang, G. G. Yen, and V. S. Tseng, "A novel constraint-based knee-guided neuroevolutionary algorithm for context-specific ECG early classification," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 11, pp. 5394–5405, Nov. 2022.
[29] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6256–6268.
[30] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–10.
[31] P. Feng, J. Fu, N. Wang, Y. Zhou, B. Zhou, and Z. Wang, "Semantic-aware alignment and label propagation for cross-domain arrhythmia classification," *Knowl.-Based Syst.*, vol. 264, 2023, Art. no. 110323.
[32] A. Raghu, D. Shanmugam, E. Pomerantsev, J. Guttag, and C. M. Stultz, "Data augmentation for electrocardiograms," in *Proc. Conf. Health Inference Learn.*, 2022, pp. 282–310.
[33] M. Zihlmann, D. Perekrestenko, and M. Tschannen, "Convolutional recurrent neural networks for electrocardiogram classification," in *Proc. Comput. Cardiol.*, 2017, pp. 1–4.
[34] N. Nonaka and J. Seita, "Electrocardiogram classification by modified efficientNet with data augmentation," in *Proc. Comput. Cardiol.*, 2020, pp. 1–4.
[35] X. Han, Y. Hu, L. Foschini, L. Chinitz, L. Jankelson, and R. Ranganath, "Deep learning models for electrocardiograms are susceptible to adversarial attack," *Nature Med.*, vol. 26, no. 3, pp. 360–363, 2020.
[36] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 702–703.
[37] Y. Wang et al., "FreeMatch: Self-adaptive thresholding for semi-supervised learning," in *Proc. 11th Int. Conf. Learn. Representations*, 2023, pp. 1–20.
[38] D. Chen, D. Wang, T. Darrell, and S. Ebrahimi, "Contrastive test-time adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 295–305.
[39] Y. Gao et al., "Visual prompt tuning for test-time domain adaptation," 2022, *arXiv:2210.04831*.
[40] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.
[41] J. Huang, A. Huang, B. C. Guerra, and Y.-Y. Yu, "PercentMatch: Percentile-based dynamic thresholding for multi-label semi-supervised classification," 2022, *arXiv:2208.13946*.
[42] D. Berthelot et al., "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–13.
[43] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, 2007.
[44] W. Liu, I. W. Tsang, and K.-R. Müller, "An easy-to-hard learning paradigm for multiple classes and multiple labels," *J. Mach. Learn. Res.*, vol. 18, pp. 1–38, 2017.
[45] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. F. Wang, "Learning deep latent space for multi-label classification," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1–7.
[46] L. Wang, Y. Liu, H. Di, C. Qin, G. Sun, and Y. Fu, "Semi-supervised dual relation learning for multi-label classification," *IEEE Trans. Image Process.*, vol. 30, pp. 9125–9135, 2021.
[47] J. Liang, F. Xu, and S. Yu, "A multi-scale semantic attention representation for multi-label image recognition with graph networks," *Neurocomputing*, vol. 491, pp. 14–23, 2022.
[48] L. Wang, Y. Liu, C. Qin, G. Sun, and Y. Fu, "Dual relation semi-supervised multi-label learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 04, 2020, pp. 6227–6234.
[49] E. A. P. Alday et al., "Classification of 12-lead ECGs: The physionet/computing in cardiology challenge 2020," *Physiol. Meas.*, vol. 41, no. 12, 2020, Art. no. 124003.
[50] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski, "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients," *Sci. data*, vol. 7, no. 1, 2020, Art. no. 48.
[51] J. Zheng et al., "Optimal multi-stage arrhythmia classification approach," *Sci. Rep.*, vol. 10, no. 1, 2020, Art. no. 2898.
[52] P. Nejedly et al., "Classification of ECG using ensemble of residual CNNs with attention mechanism," in *Proc. Comput. Cardiol.*, 2021, pp. 1–4.
[53] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
[54] B. Chen, J. Jiang, X. Wang, P. Wan, J. Wang, and M. Long, "Debiased self-training for semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 32424–32437.
[55] R. Gontijo-Lopes, S. J. Smullin, E. D. Cubuk, and E. Dyer, "Affinity and diversity: Quantifying mechanisms of data augmentation," 2020, *arXiv: 2002.08973*.
[56] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.
[57] M. Reyna et al., "Will two do? Varying dimensions in electrocardiography: The physioNet/computing in cardiology challenge 2021," in *Proc. Comput. Cardiol.*, 2022.

Rushuang Zhou (Graduate Student Member, IEEE) is currently working toward the PhD degree with the Department of Biomedical Engineering, City University of Hong Kong, Hong Kong.

Lei Lu is a postdoctoral research assistant with the Department of Engineering Science. University of Oxford.

Zijun Liu (Graduate Student Member, IEEE) is currently working toward the PhD degree with the Department of Biomedical Engineering, City University of Hong Kong, Hong Kong.

Ting Xiang (Graduate Student Member, IEEE) is currently working toward the PhD degree with the Department of Biomedical Engineering, City University of Hong Kong, Hong Kong.

Zhen Liang (Member, IEEE) received the PhD degree from the Hong Kong Polytechnic University, Hong Kong, in 2013. From 2012 to 2017, she was an algorithm development scientist with NeuroSky, Inc., Hong Kong. From 2018 to 2019, she was a specially–appointed assistant professor with the Graduate School of Informatics, Kyoto University, Japan. She is currently an assistant professor with the School of Biomedical Engineering, Health Science Center, Shenzhen University, China. Her current research interests include brain encoding and decoding system, affective computing, visual attention, and neural engineering.

David A. Clifton is a professor with the Clinical Machine Learning and leads the Computational Health Informatics (CHI) Lab, Department of Engineering Science, University of Oxford.

Yining Dong received the BEng degree in electronic engineering from Tsinghua University, Beijing, China, in 2011, and the PhD degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2016, She is currently an assistant professor with the School of Data Science, City University of Hong Kong, Hong Kong. Her research interests include process data analytics, statistical machine learning, smart manufacturing, and new material design.

Yuan-Ting Zhang (Fellow, IEEE) is currently the chairman and director with the Hong Kong Center for Cerebro-cardiovascular Health Engineering and the chair professor with City University of Hong Kong. He is a LRG member of Karolinska Institutet MWLC. He was the Sensing System Architect in Health Technology and Sensing Hardware Divisions with Apple Inc., California, USA, the founding director with the Key Lab for Health Informatics of Chinese Academy of Sciences (CAS) and the founding director of CAS-SIAT Institute of Biomedical and Health Engineering. He dedicated his service to the Chinese University of Hong Kong from 1994 to 2015, where he served as the first Head of the Division of Biomedical Engineering. He has been the editor-in-chief for *IEEE Reviews in Biomedical Engineering* since 2016 and chair of the IEEE 1708 Working Group. He was the editor-in-chief for *IEEE Transactions on Information Technology in Biomedicine* and the first editor-in-chief of *IEEE Journal of Biomedical and Health Informatics*. He served as vice preside of *IEEE Engineering in Medicine and Biology Society*. He was also the chair of 2016-2018 IEEE Award Committee in Biomedical Engineering and a member of IEEE Medal Panel for Healthcare Technology Award. His research interests include unobtrusive sensing and wearable devices, and neural muscular modeling. He was selected on the lists of China's Most Cited Researchers by Elsevier and the top 2% Researcher worldwide by Stanford University. He won a number of international awards including IEEEEMBS best journal paper awards, IEEE-EMBS Outstanding Service Award, IEEE-SA 2014 Emerging Technology Award, and Earl Owen Lecture with SMITIBEC2018 in Korea. He is elected as IAMBE Fellow, AIMBE fellow and AAIA fellow for his contributions to the development of wearable and m-Health technologies.