Supplementary Materials

APPENDIX A
RE-ANNOTATION OF THE PUBLIC DATASETS

As different public datasets have different kinds of abnormal ECG signals and adopt various standards for CVDs diagnosis, there is a strong label distribution shift among them [1]–[4]. We download the original annotations of the recordings from the Physionet [1] and visualize the class distribution in Fig. S1. It can be observed that some classes only exist in certain databases, such as AF (atrial fibrillation) and CRBBB (complete right bundle branch block). This phenomenon demonstrates that the original annotation scheme is not suitable for multi-dataset evaluation, where the test recordings come from different datasets. When employing the original annotation scheme, the test set might contain the classes that are absent in the training set. To address this issue while preserving the classes exclusive to specific databases, we re-annotate the ECG recordings into five super-classes. As shown in Table S1, a comparison between the original and our annotations is presented. Specifically, we assign five new labels to the recording from the databases (Abnormal Rhythms, ST/T Abnormalities, Conduction Disturbance, Other Abnormalities, Normal Signals). Note that each recording can belong to two or more categories simultaneously. The definition of the first three abnormalities originates from the ECG statements of the PTB-XL database [2]. To categorize the remaining CVDs that do not fit into the three abnormalities (Abnormal Rhythms, ST/T Abnormalities, Conduction Disturbance), we classify them as Other Abnormalities. If there are no potential CVDs from a given ECG segment, we regard it as Normal Signals. In practice, the 'Normal Signals' class should not co-occur with other categories, which has been considered in label correlation alignment to help our model avoid confusing predictions in which CVDs are detected in normal signals. After re-annotation, the class distributions of the four databases are shown in Table S2. Despite the presence of all CVD classes across different databases, the class distribution discrepancy among the databases still challenges the robustness of the models on unseen datasets [5].

It is important to acknowledge that our annotation scheme might not be optimal as diagnosing some CVDs can be complex and uniform definitions are challenging to establish. For example, as the 'prolonged pr interval' is one of the criteria to diagnose the '1st degree av block', it is difficult to separate them into two categories [6], [7]. However, one might argue that their association is unclear if the pr interval is prolonged but does not cross the 1st av block threshold [8]. At the same time, the prolongation is also associated with other diseases, including acute rhetheumatic fever, carditis associated with Lyme disease, and second-degree av block [9]–[11]. In order to establish a suitable annotation guideline for the two CVDs, we review the annotation schemes implemented in the two datasets, Physionet and PTB-XL. Notably, these datasets treat the 'prolonged pr interval' and '1st degree av block' as two different categories. Consequently, we annotate the segments with '1st degree av block' and 'prolonged pr interval' as Conduction Disturbance and Other Abnormalities simultaneously. Segments solely exhibiting 'prolonged pr interval' or '1st degree av block' are respectively labeled as "Other Abnormalities" or "Conduction Disturbance." On the other hand, some labels provided by the Physionet are inaccurate [1], making it difficult to design an optimal re-annotated strategy.

APPENDIX B
EFFECT OF DIFFERENT SIMILARITY METRICS

Recall that the estimated label correlation $\hat{r}_{c_1,c_2} \in [0,1]$ between class c_1 and class c_2 can be estimated by the cosine similarity between the binary label sequences (y_{c_1},y_{c_2}) on the two classes. In this section, the effect of different similarity metrics on model performance is also examined. Specifically, we use other metrics to compute the label correlation $\hat{r}_{c_1,c_2}$, such as the Pearson coefficient (Eq. S1) and the Euclidean distance (Eq. S2), given as,

$$\hat{r}_{c_1,c_2} = \hat{\rho}^2, \hat{\rho} = \frac{Cov(y_{c_1}, y_{c_2})}{\sigma_{y_{c_1}} \sigma_{y_{c_2}}}, \tag{S1}$$

$$\hat{r}_{c_1,c_2} = \frac{1}{1+d}, d = \left\| y_{c_1} - y_{c_2} \right\|, \tag{S2}$$

where $\sigma_{y_{c_1}}$ and $\sigma_{y_{c_2}}$ are the standard deviations of the elements in vector y_{c_1} and y_{c_2}. Another approach to measuring the label correlation is by computing the co-occurrence frequency of two diseases [12]. However, we argue that it is not applicable for unlabeled samples lacking binary ground truth. While it is possible to binarize the generated pseudo-labels using predetermined thresholds, this will introduce additional costs for threshold selection.

As shown in Table S3, Table S4 and Table S5, the model performance under different similarity metrics is presented. Experiment results on three protocols indicate that cosine similarity outperforms the other metrics, providing

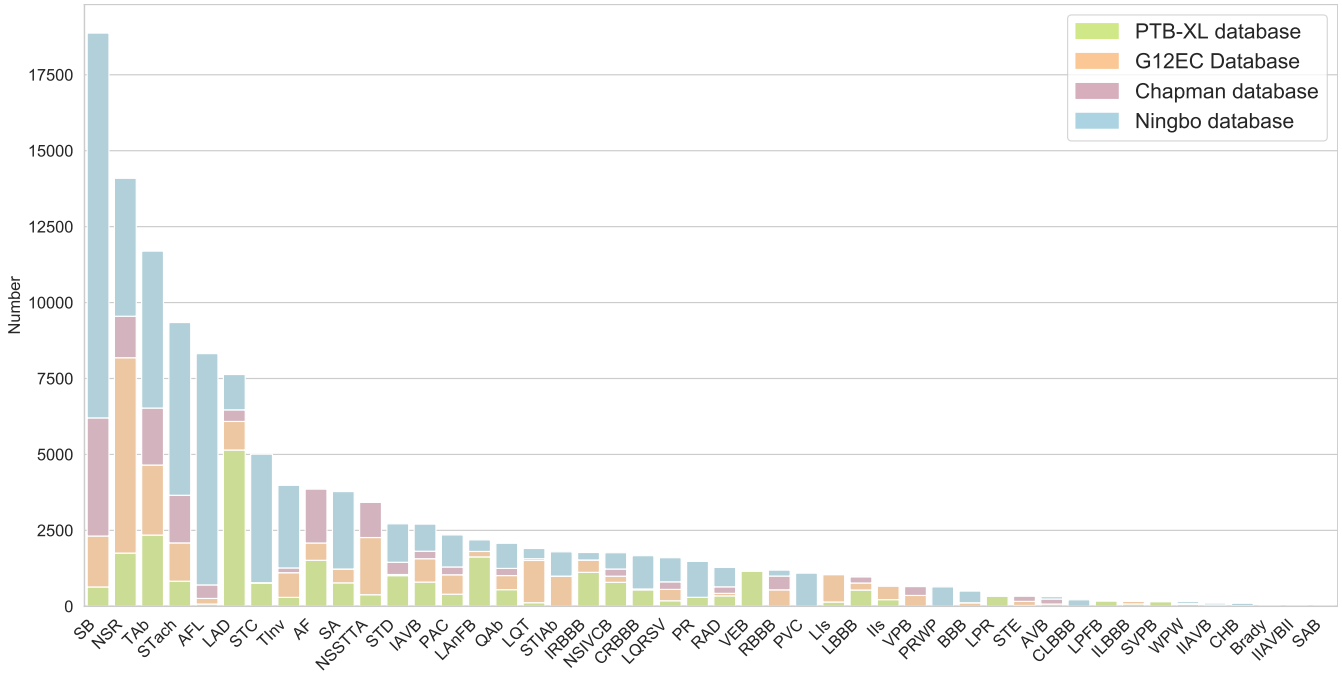


Fig. S1: Class distribution of the ECG recordings with original annotations employed in the four public databases. For simplicity, we present the abbreviation of the CVDs in the databases.

an empirical validation of its superiority. Moreover, we provide a theoretical analysis to support the conclusion further. The co-occurrence between two CVDs classes c_1 and c_2 can be represented by conditional probabilities $P(c_1 = 1|c_2 = 1)$ and $P(c_2 = 1|c_1 = 1)$, where "$c_1 = 1$" indicates the existence of CVDs c_1. Then we can drive the connection between the cosine similarity and the conditional probabilities, as

$$\begin{aligned}\hat{r}_{c_1,c_2} &= \frac{y_{c_1}^T y_{c_2}}{\|y_{c_1}\| \, \|y_{c_2}\|} \approx \frac{NP(c_1=1,c_2=1)}{\sqrt{NP(c_1=1)}\sqrt{NP(c_2=1)}} \\ &= \frac{\sqrt{P(c_1=1,c_2=1)}\sqrt{P(c_1=1,c_2=1)}}{\sqrt{P(c_1=1)}\sqrt{P(c_2=1)}} \\ &= \sqrt{\frac{P(c_1=1,c_2=1)P(c_1=1,c_2=1)}{P(c_1=1)P(c_2=1)}} \\ &= \sqrt{P(c_1=1|c_2=1)P(c_2=1|c_1=1)},\end{aligned} \tag{S3}$$

where N is the number of samples in the label sequences y_{c_1} and y_{c_2}. Eq. S3 shows that the cosine similarity can eliminate the marginal probabilities $P(c_1 = 1)$ and $P(c_2 = 1)$, which represent the class distributions of different classes and should not be considered in the computation of $\hat{r}_{c_1,c_2}$. However, other metrics fail to eliminate them, resulting in additional errors and degraded model performance when the class distributions vary across different datasets. Specifically, as demonstrated in Eq. S5 and Eq. S4, the label correlation $\hat{r}_{c_1,c_2}$ computed based on the Euclidean distance or the Pearson coefficient is influenced by the class distributions ($P(c_1 = 1)$ and $P(c_2 = 1)$), which vary across different datasets (Table S2). This limitation explains their poor performance on the Ningbo and Chapman databases, which exhibit a higher degree of class imbalance than other databases. In summary, experimental and theoretical results demonstrate that ECGMatch is compatible with different similarity metrics and achieves better performance when using cosine similarity.

$$\begin{aligned}\hat{r}_{c_1,c_2} &= \hat{\rho}^2, \hat{\rho} = \frac{(y_{c_1}-\mu_{c_1})^T(y_{c_2}-\mu_{c_2})}{\|y_{c_1}-\mu_{c_1}\| \, \|y_{c_2}-\mu_{c_2}\|} \\ &\approx \frac{(y_{c_1}-\mathbf{1}P(c_1=1))^T(y_{c_2}-\mathbf{1}P(c_2=1))}{\|y_{c_1}-\mathbf{1}P(c_1=1)\| \, \|y_{c_2}-\mathbf{1}P(c_2=1)\|} \\ &= \frac{y_{c_1}^T y_{c_2} - y_{c_1}^T \mathbf{1}P(c_2=1) - \mathbf{1}^T P(c_1=1) y_{c_2}}{\|y_{c_1}-\mathbf{1}P(c_1=1)\| \, \|y_{c_2}-\mathbf{1}P(c_2=1)\|} \\ &\quad + \frac{NP(c_1=1)P(c_2=1)}{\|y_{c_1}-\mathbf{1}P(c_1=1)\| \, \|y_{c_2}-\mathbf{1}P(c_2=1)\|} \\ &= \frac{P(c_1=1,c_2=1)-P(c_1=1)P(c_2=1)}{\sqrt{P(c_1=1)-P(c_1=1)^2}\sqrt{P(c_2=1)-P(c_2=1)^2}}.\end{aligned} \tag{S4}$$

where $\mu_{c_1} = \frac{1}{N_B}\sum_{i=1}^{N_B} y_{c_1}^i$ and $\mu_{c_2} = \frac{1}{N_B}\sum_{i=1}^{N_B} y_{c_2}^i$ are the mean values of the elements in vector y_{c_1} and y_{c_2}.

$$\begin{aligned}\hat{r}_{c_1,c_2} &= \frac{1}{1+d} = \frac{1}{1+\sqrt{(y_{c_1}-y_{c_2})^T(y_{c_1}-y_{c_2})}} \\ &= \frac{1}{1+\sqrt{y_{c_1}^T y_{c_1} - y_{c_2}^T y_{c_1} - y_{c_1}^T y_{c_2} + y_{c_2}^T y_{c_2}}} \\ &\approx \frac{1}{1+\sqrt{N(P(c_1=1)+P(c_2=1)-2P(c_1=1,c_2=1))}}.\end{aligned} \tag{S5}$$

TABLE S1: A comparison between the original and our annotation

Original annotation	Our annotation	Original annotation	Our annotation
atrial fibrillation	Abnormal Rhythms	left bundle branch block	Conduction Disturbance
atrial flutter	Abnormal Rhythms	non-specific intraventricular conduction disorder	Conduction Disturbance
bradycardia	Abnormal Rhythms	right bundle branch block	Conduction Disturbance
pacing rhythm	Abnormal Rhythms	av block	Conduction Disturbance
sinus arrhythmia	Abnormal Rhythms	complete heart block	Conduction Disturbance
sinus bradycardia	Abnormal Rhythms	2nd degree av block	Conduction Disturbance
sinus tachycardia	Abnormal Rhythms	mobitz type II atrioventricular block	Conduction Disturbance
prolonged qt interval	ST/T Abnormalities	incomplete left bundle branch block	Conduction Disturbance
t wave abnormal	ST/T Abnormalities	left posterior fascicular block	Conduction Disturbance
t wave inversion	ST/T Abnormalities	sinoatrial block	Conduction Disturbance
inferior ischaemia	ST/T Abnormalities	wolff parkinson white pattern	Conduction Disturbance
lateral ischaemia	ST/T Abnormalities	left axis deviation	Other Abnormalities
nonspecific st abnormality	ST/T Abnormalities	low qrs voltages	Other Abnormalities
st changes	ST/T Abnormalities	premature atrial contraction	Other Abnormalities
st depression	ST/T Abnormalities	poor R wave progression	Other Abnormalities
st elevation	ST/T Abnormalities	premature ventricular contractions	Other Abnormalities
st interval abnormal	ST/T Abnormalities	qwave abnormal	Other Abnormalities
bundle branch block	Conduction Disturbance	right axis deviation	Other Abnormalities
complete left bundle branch block	Conduction Disturbance	supraventricular premature beats	Other Abnormalities
complete right bundle branch block	Conduction Disturbance	ventricular premature beats	Other Abnormalities
1st degree av block	Conduction Disturbance	ventricular ectopics	Other Abnormalities
incomplete right bundle branch block	Conduction Disturbance	prolonged pr interval	Other Abnormalities
left anterior fascicular block	Conduction Disturbance	sinus rhythm	Normal Signals

TABLE S2: Class distribution of each dataset after re-annotation.

Datasets	Conduction Disturbance	Abnormal Rhythms	ST/T Abnormalities	Other Abnormalities	Normal Signals
G12EC Database [1]	2236	3977	4991	2627	1752
PTB-XL database [2]	4907	4087	4299	7296	6432
Chapman database [3]	1198	7682	2951	1445	1366
Ningbo database [4]	3843	28217	10407	5339	4542

APPENDIX C
VISUALIZATION ANALYSIS

To verify the robustness of the ECGMatch from a more intuitive perspective, we visualize its performance on the five CVD super-classes. Fig. S2 illustrates five examples of ECG recordings that were accurately classified by the ECGMatch model, while being misclassified by other models. Specifically, ECGMatch successfully detected potential CVDs in these recordings, while other models provided negative predictions. Moreover, we visualize the performance of different models in the five super-classes to provide a more comprehensive analysis. For the sake of simplicity, we present the results obtained from the mix-dataset protocol. Based on the results, it is evident that ECGMatch outperforms its competitors in the prediction of Conduction Disturbance and ST/T Abnormalities. For instance, ECGMatch achieves superior average precision in detecting Conduction Disturbance, ST/T Abnormalities and Normal Signals. The improvements compared with its best-performing competitors are 2.95%, 1.05% and 0.8%, respectively. In addition, ECGMatch improves the G-beta scores in detecting Conduction Disturbance and ST/T Abnormalities, with enhancements of 1.67% and 1.12% respectively. Regarding Abnormal Rhythms and Other Abnormalities, ECGMatch exhibits comparable performance to other models.

APPENDIX D
EXTEND RESULTS ABOUT THE EFFECT OF THE RATIO OF LABELED SAMPLES FOR MODEL TRAINING

We adjust the ratio during model training and evaluate the model performance using the mix-dataset and within-dataset protocols. The results are presented in Fig. S4 and Fig. S5. In the within-dataset protocol, ECGMatch achieves similar performance to other models across all six metrics, while reducing the required number of labeled samples by 5%. Additionally, in the mix-dataset protocol, ECGMatch demonstrates comparable ranking loss and coverage to other models, while reducing the number of labeled samples by 5%. These results provide supplementary evidence supporting the superior performance of the proposed ECGMatch in reducing the need for human annotations during model training.

APPENDIX E
EXTEND RESULTS ABOUT THE EFFECT OF DIFFERENT ANNOTATION SCHEMES

In this section, we compare the performance of different models using the annotation scheme of the Cinc2020/2021 challenges [13], [14] to further validate the superior performance of the ECGMatch. Following their pipelines, we use 25 categories from Fig. S1 for model training and

TABLE S3: Comparisons of different similarity metrics (within-dataset protocol).

Methods	G12EC	PTB	Ningbo	Chapman
Ranking loss (The smaller, the better)				
Cosine similarity	0.140±0.006	0.134±0.003	0.045±0.002	0.052±0.002
Pearson coefficient	0.150±0.010	0.134±0.003	0.051±0.003	0.059±0.002
Euclidean distance	0.143±0.004	0.133±0.002	0.054±0.002	0.057±0.003
Hamming loss (The smaller, the better)				
Cosine similarity	0.278±0.008	0.233±0.009	0.122±0.001	0.139±0.002
Pearson coefficient	0.285±0.010	0.240±0.008	0.129±0.002	0.146±0.007
Euclidean distance	0.283±0.004	0.239±0.006	0.131±0.003	0.146±0.007
Coverage (The smaller, the better)				
Cosine similarity	2.173±0.027	1.922±0.015	1.724±0.010	1.761±0.021
Pearson coefficient	2.209±0.034	1.920±0.012	1.751±0.014	1.794±0.012
Euclidean distance	2.187±0.013	1.917±0.011	1.768±0.008	1.784±0.023
MAP (The greater, the better)				
Cosine similarity	0.742±0.005	0.748±0.009	0.808±0.001	0.775±0.014
Pearson coefficient	0.733±0.007	0.744±0.009	0.797±0.004	0.759±0.010
Euclidean distance	0.737±0.004	0.746±0.011	0.794±0.006	0.755±0.003
Marco AUC (The greater, the better)				
Cosine similarity	0.854±0.003	0.880±0.005	0.925±0.001	0.912±0.002
Pearson coefficient	0.851±0.005	0.880±0.004	0.915±0.004	0.907±0.002
Euclidean distance	0.855±0.004	0.882±0.005	0.915±0.004	0.907±0.002
Marco G_{beta} score (The greater, the better)				
Cosine similarity	0.477±0.003	0.467±0.009	0.563±0.001	0.554±0.009
Pearson coefficient	0.469±0.008	0.465±0.006	0.547±0.001	0.545±0.016
Euclidean distance	0.474±0.003	0.466±0.008	0.544±0.002	0.543±0.017

evaluation. Because some of the selected CVDs only exist on certain datasets, we only conduct the mix-dataset protocol, where we combine the four databases for analysis. Compared to our and PTB-XL's scheme [2], the annotation scheme of Cinc2020/2021 presents greater challenges due to the inclusion of a larger number of classes and the presence of serious class imbalance problems. Two evaluation metrics (average precision and AUC) are used for model evaluation and the performance across three seeds is reported in Table S6 and Table S7. By averaging the performance on different classes, ECGMatch demonstrates the highest average precision (0.273±0.016) and AUC (0.803±0.008) compared to other models. Furthermore, it is observed that ECGMatch achieved the highest average precision on six CVDs and the highest AUC on ten CVDs. The results demonstrate the superior performance of the ECGMatch in challenging CVDs prediction tasks.

APPENDIX F
EXTEND RESULTS ABOUT THE COMPARISONS WITH STATE-OF-THE-ART METHODS

To highlight the contribution of the ECGMatch in ECG-based CVDs prediction, we reproduce additional state-of-the-art models for comparisons. The first one is the Mix Mean Teacher [22], a state-of-the-art model in single-label semi-supervised CVDs prediction. We reformulate its softmax loss as the multi-label binary cross entropy loss [23] to enable it for multi-label prediction. Note that the model trained with the multi-label loss can be considered as an ensemble of C single-label models, where C represents the number of categories [24]. The second one is the MSDNN [25], which achieved leading performance in multi-label CVDs prediction. To ensure a fair comparison, we use the same backbones and data augmentation techniques as our ECGMatch to implement the above two models. Table S8, Table S9, Table S10 present the results on four databases and three experiment protocols. The results demonstrate the superior performance of the ECGMatch across different experiment settings. In contrast to the ECGMatch, the above models both utilize unlabeled samples for training to alleviate the label scarcity problem. However, they ignore the label dependency between different CVDs, thus limiting their performance in multi-label classification. In conclusion, these results emphasize the importance and advantages of developing a semi-supervised multi-label model for CVDs prediction.

APPENDIX G
EXTEND RESULTS ON THE SENSITIVITY ANALYSIS

In this section, we conduct a sensitivity analysis to investigate the impact of the hyperparameter K, which denotes the number of nearest neighbors for pseudo-label generation and refinement. To be specific, we adjust the value of K from 2 to 10 and keep the other hyperparameters fixed. The average model performance across four databases is reported in Fig. S6 using the cross-dataset protocol, where the training and test sets could have

TABLE S4: Comparisons of different similarity metrics (cross-dataset protocol)

Methods	G12EC	PTB	Ningbo	Chapman
Ranking loss (The smaller, the better)				
Cosine similarity	0.203±0.004	0.248±0.005	0.102±0.006	0.068±0.002
Pearson coefficient	0.202±0.003	0.234±0.007	0.115±0.013	0.078±0.004
Euclidean distance	0.203±0.007	0.243±0.005	0.098±0.003	0.088±0.003
Hamming loss (The smaller, the better)				
Cosine similarity	0.331±0.007	0.310±0.001	0.253±0.008	0.219±0.003
Pearson coefficient	0.335±0.002	0.320±0.012	0.281±0.010	0.249±0.005
Euclidean distance	0.328±0.007	0.331±0.014	0.276±0.010	0.262±0.003
Coverage (The smaller, the better)				
Cosine similarity	2.415±0.016	2.379±0.023	1.971±0.025	1.803±0.008
Pearson coefficient	2.415±0.010	2.324±0.029	2.027±0.046	1.852±0.014
Euclidean distance	2.422±0.025	2.364±0.018	1.963±0.017	1.898±0.016
MAP (The greater, the better)				
Cosine similarity	0.657±0.009	0.591±0.012	0.689±0.002	0.748±0.004
Pearson coefficient	0.650±0.002	0.586±0.006	0.670±0.003	0.742±0.005
Euclidean distance	0.651±0.007	0.594±0.009	0.671±0.004	0.729±0.001
Marco AUC (The greater, the better)				
Cosine similarity	0.805±0.004	0.800±0.010	0.874±0.002	0.900±0.002
Pearson coefficient	0.799±0.003	0.801±0.006	0.869±0.001	0.893±0.005
Euclidean distance	0.801±0.002	0.802±0.006	0.869±0.001	0.887±0.001
Marco G_{beta} score (The greater, the better)				
Cosine similarity	0.403±0.002	0.369±0.001	0.442±0.003	0.516±0.006
Pearson coefficient	0.397±0.002	0.368±0.003	0.430±0.005	0.490±0.006
Euclidean distance	0.399±0.004	0.368±0.005	0.432±0.005	0.481±0.005

TABLE S5: Comparisons of different similarity metrics (mix-dataset protocol).

Methods	Ranking loss	Hamming loss	Coverage	MAP	Marco AUC	Marco G_{beta} score
Cosine similarity	0.150±0.001	0.270±0.001	2.101±0.009	0.658±0.006	0.838±0.003	0.442±0.002
Pearson coefficient	0.148±0.007	0.269±0.007	2.098±0.037	0.659±0.003	0.840±0.002	0.444±0.003
Euclidean distance	0.157±0.004	0.282±0.009	2.137±0.015	0.647±0.004	0.834±0.003	0.435±0.004

different class distributions. The results demonstrate that the optimal setting for all CVDs is $K = 5$. In addition, it is observed that the model performance shows relatively low sensitivity to the changes of K, thereby providing additional evidence to support the stability of ECGMatch in multi-label CVDs prediction.

REFERENCES

[1] E. A. P. Alday, A. Gu, A. J. Shah, C. Robichaux, A.-K. I. Wong, C. Liu, F. Liu, A. B. Rad, A. Elola, S. Seyedi *et al.*, "Classification of 12-lead ECGs: the physionet/computing in cardiology challenge 2020," *Physiological measurement*, vol. 41, no. 12, p. 124003, 2020.

[2] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," *Scientific data*, vol. 7, no. 1, p. 154, 2020.

[3] 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," *Scientific data*, vol. 7, no. 1, p. 48, 2020.

[4] J. Zheng, H. Chu, D. Struppa, J. Zhang, S. M. Yacoub, H. El-Askary, A. Chang, L. Ehwerhemuepha, I. Abudayyeh, A. Barrett *et al.*, "Optimal multi-stage arrhythmia classification approach," *Scientific reports*, vol. 10, no. 1, p. 2898, 2020.

[5] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[6] S. H. Oldroyd, B. S. Q. Rodriguez, and A. N. Makaryus, "First degree heart block," in *StatPearls [Internet]*. StatPearls Publishing, 2022.

[7] D. T. Huang and T. Prinzi, *Clinical Cardiac Electrophysiology in Clinical Practice*. Springer, 2014.

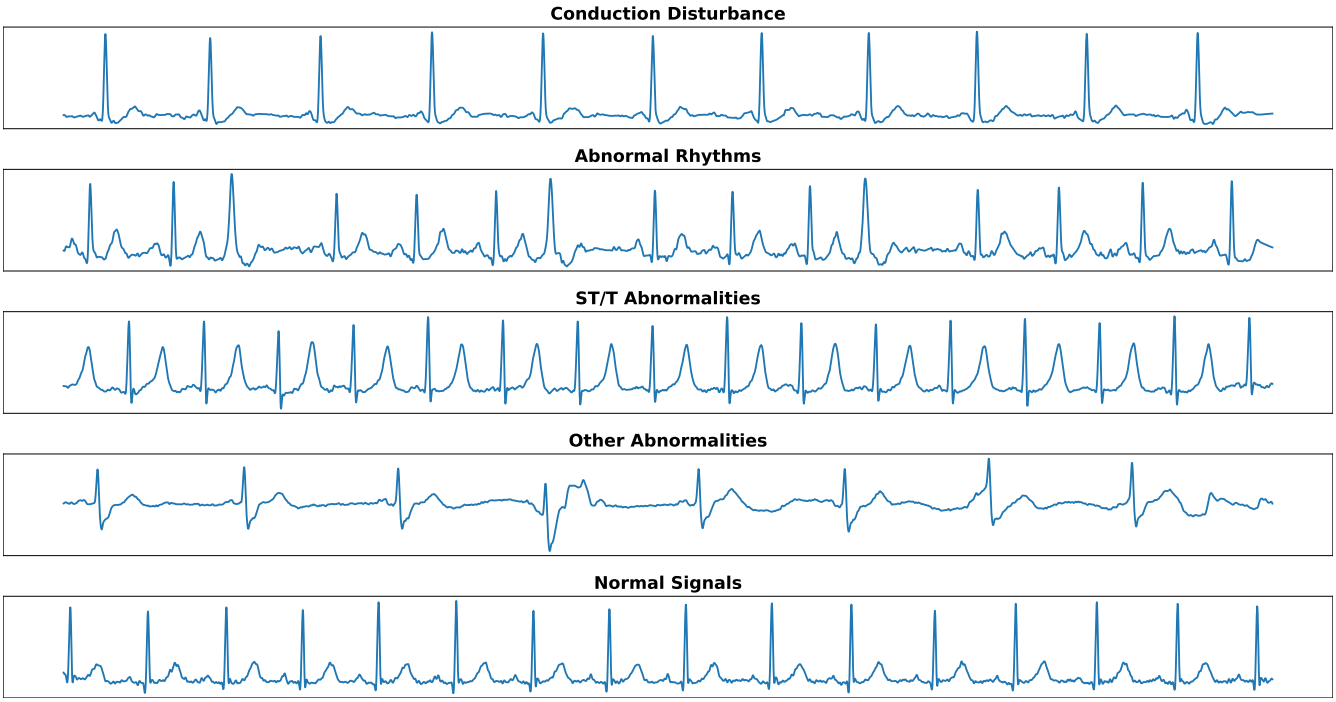


Fig. S2: Visualizations of five cases where ECGMatch predicted the right labels, while other models failed to do so.

[8] R. Pranata, E. Yonas, V. Chintya, H. Deka, and S. B. Raharjo, "Association between pr interval, first-degree atrioventricular block and major arrhythmic events in patients with brugada syndrome–systematic review and meta-analysis," *Journal of Arrhythmia*, vol. 35, no. 4, pp. 584–590, 2019.

[9] W. KF, "Beitrage zur kenntnis der menschlichen herztatigkeit," *Arch Anat Physiol*, vol. 1, pp. 1–2, 1907.

[10] J. M. Costello, M. E. Alexander, K. M. Greco, A. R. Perez-Atayde, and P. C. Laussen, "Lyme carditis in children: presentation, predictive factors, and clinical course," *Pediatrics*, vol. 123, no. 5, pp. e835–e841, 2009.

[11] M. Karacan, N. Ceviz, and H. Olgun, "Heart rate variability in children with acute rheumatic fever," *Cardiology in the Young*, vol. 22, no. 3, pp. 285–292, 2012.

[12] Z. Ge, X. Jiang, Z. Tong, P. Feng, B. Zhou, M. Xu, Z. Wang, and Y. Pang, "Multi-label correlation guided feature fusion network for abnormal ECG diagnosis," *Knowledge-Based Systems*, vol. 233, p. 107508, 2021.

[13] M. Reyna, N. Sadr, A. Gu, E. A. Perez Alday, C. Liu, S. Seyedi, A. Shah, and G. Clifford, "Will two do? varying dimensions in electrocardiography: The physionet/computing in cardiology challenge 2021," *PhysioNet*, 2022.

[14] E. Perez Alday, A. Gu, A. Shah, C. Robichaux, A.-K. I. Wong, C. Liu, F. Liu, B. Rad, A. Elola, S. Seyedi *et al.*, "Classification of 12-lead ECGs: the physionet," *Computing in Cardiology challenge*, 2020.

[15] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[16] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 596–608.

[17] B. Zhang, Y. Wang, W. Hou, H. WU, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[18] B. Chen, J. Jiang, X. Wang, P. Wan, J. Wang, and M. Long, "Debiased self-training for semi-supervised learning," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022.

[19] J. Huang, A. Huang, B. C. Guerra, and Y.-Y. Yu, "PercentMatch: Percentile-based dynamic thresholding for multi-label semi-supervised classification," *arXiv preprint arXiv:2208.13946*, 2022.

[20] H. Chen, R. Tao, Y. Fan, Y. Wang, J. Wang, B. Schiele, X. Xie, B. Raj, and M. Savvides, "Softmatch: Addressing the quantity-quality tradeoff in semi-supervised learning," in *The Eleventh International Conference on Learning Representations*, 2023.

[21] 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 *International Conference on Learning Representations*, 2021.

[22] 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 Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3791–3801, 2022.

[23] T. Durand, N. Mehrasa, and G. Mori, "Learning a deep convnet for multi-label classification with partial labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[24] T. Kobayashi, "Two-way multi-label loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7476–7485.

[25] J. Lai, H. Tan, J. Wang, L. Ji, J. Guo, B. Han, Y. Shi, Q. Feng, and W. Yang, "Practical intelligent diagnostic algorithm for wearable 12-lead ECG via self-supervised learning on large-scale dataset," *Nature Communications*, vol. 14, no. 1, p. 3741, 2023.

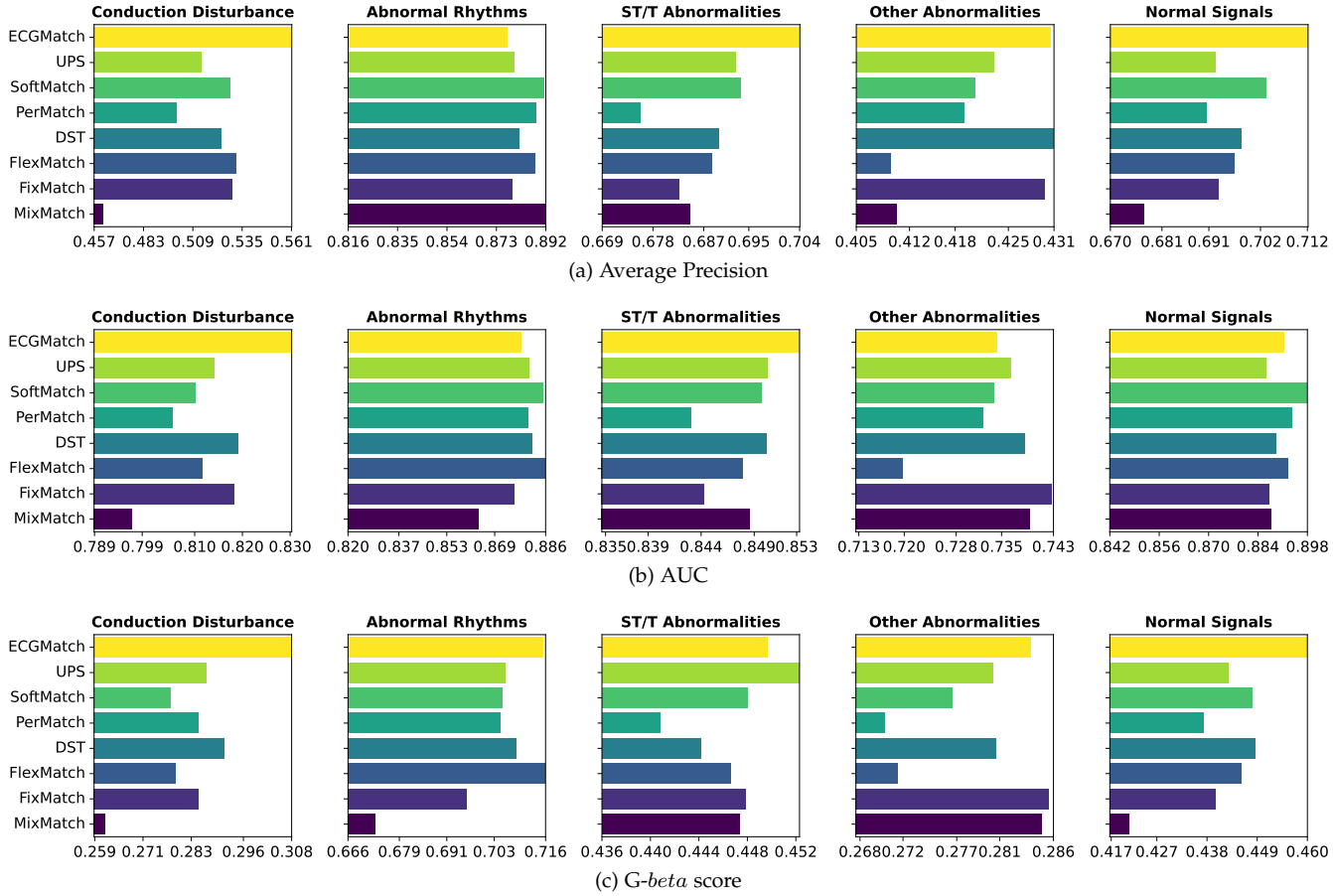


Fig. S3: Performance comparison of different models in different CVD classes (mix-dataset protocol).

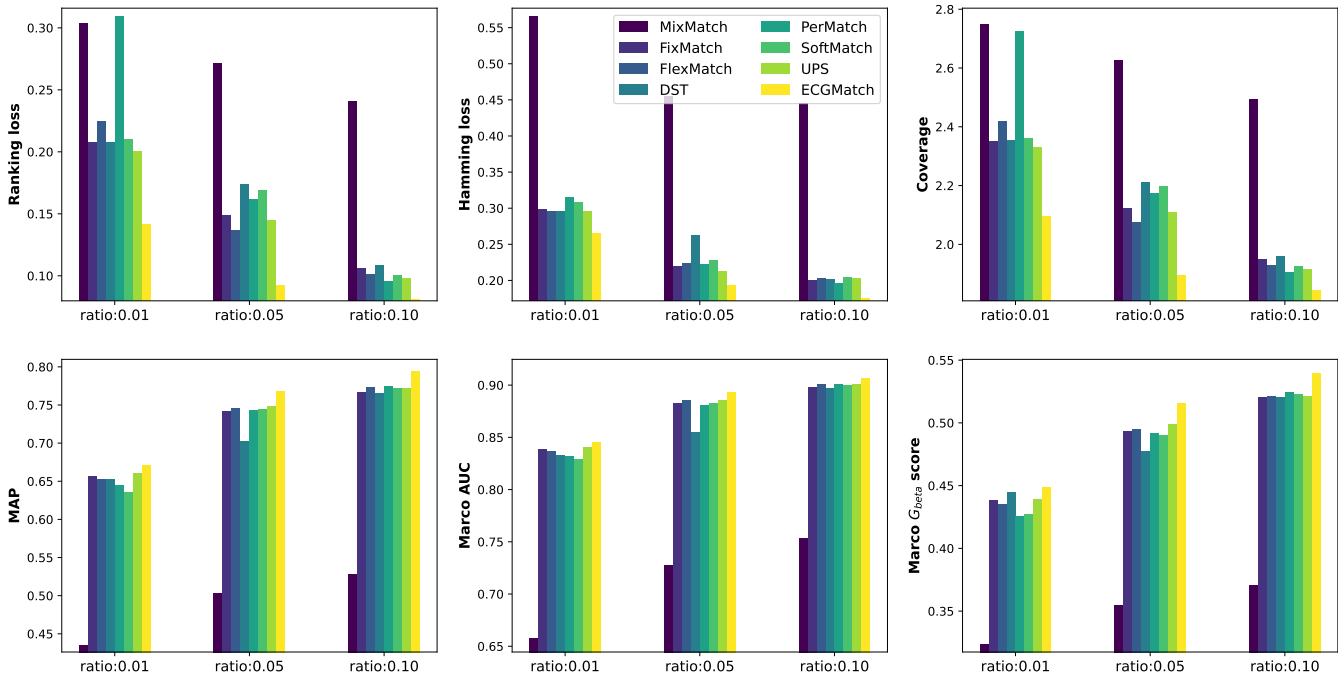


Fig. S4: Performance comparison of different models under various labeled sample ratios (within-dataset protocol).

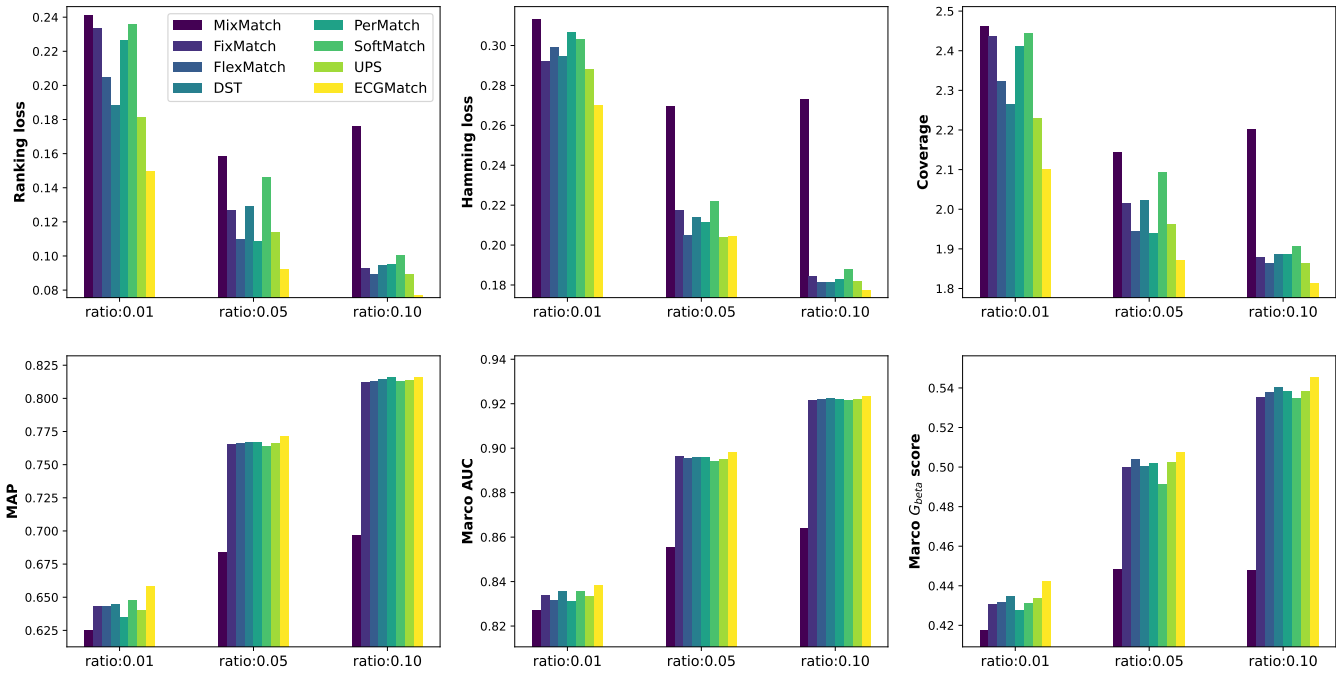


Fig. S5: Performance comparison of different models under various labeled sample ratios (mix-dataset protocol).

TABLE S6: Comparison results between ECGMatch and the state-of-the-art models using the mix-dataset protocol, under the annotation scheme of Physionet. The mean average precision and standard deviations on four databases are shown across three seeds. For each CVD, the model with the best performance is highlighted in bold.

Methods	MixMatch [15]	FixMatch [16]	FlexMatch [17]	DST [18]	PerMatch [19]	SoftMatch [20]	UPS [21]	**ECGMatch**
AF	0.082±0.021	0.150±0.009	**0.194±0.053**	0.142±0.023	0.147±0.009	0.186±0.025	0.130±0.016	0.164±0.022
AFL	0.202±0.032	**0.390±0.029**	0.316±0.007	0.355±0.054	0.307±0.062	0.303±0.054	0.377±0.052	0.389±0.071
BBB	**0.067±0.014**	0.054±0.033	0.045±0.025	0.061±0.037	0.064±0.026	0.041±0.007	0.046±0.018	0.050±0.007
CLBBB/LBBB	0.337±0.083	0.381±0.174	0.342±0.074	0.489±0.068	**0.529±0.057**	0.459±0.035	0.485±0.033	0.484±0.148
CRBBB/RBBB	0.249±0.119	0.465±0.161	0.601±0.051	0.562±0.057	0.543±0.021	0.611±0.052	0.590±0.069	**0.677±0.059**
IAVB	0.071±0.015	0.205±0.092	0.127±0.058	0.177±0.022	0.172±0.030	0.129±0.055	0.120±0.044	**0.235±0.057**
IRBBB	0.043±0.014	0.090±0.026	0.108±0.013	0.107±0.005	0.105±0.014	**0.116±0.037**	0.100±0.017	0.112±0.018
LAD	0.308±0.066	0.293±0.009	0.293±0.013	**0.345±0.073**	0.283±0.027	0.338±0.020	0.304±0.020	0.330±0.028
LAnFB	0.156±0.028	0.210±0.016	0.225±0.086	**0.331±0.094**	0.237±0.040	0.238±0.065	0.261±0.070	0.291±0.051
LQRSV	0.037±0.005	0.042±0.005	0.044±0.011	0.048±0.004	**0.051±0.003**	0.048±0.008	0.043±0.002	0.042±0.003
NSIVCB	0.038±0.006	0.041±0.006	0.048±0.011	0.045±0.006	0.044±0.001	0.048±0.009	0.044±0.005	**0.054±0.008**
NSR	0.553±0.039	0.738±0.052	0.712±0.037	0.728±0.020	0.711±0.066	0.721±0.061	0.679±0.027	**0.752±0.018**
PAC/SVPB	0.041±0.003	0.045±0.009	0.063±0.017	0.049±0.006	**0.074±0.020**	0.057±0.011	0.049±0.005	0.057±0.013
PR	0.230±0.068	0.604±0.073	0.559±0.059	0.481±0.088	0.423±0.131	0.428±0.140	0.551±0.069	**0.670±0.050**
PRWP	0.031±0.004	0.031±0.000	0.038±0.012	**0.047±0.006**	0.033±0.005	0.028±0.000	0.038±0.006	0.030±0.010
VPB/PVC	0.060±0.016	0.075±0.029	**0.124±0.028**	0.067±0.007	0.084±0.030	0.116±0.006	0.072±0.010	0.088±0.021
LPR	0.023±0.012	0.070±0.038	0.047±0.030	0.043±0.011	**0.080±0.005**	0.049±0.026	0.042±0.012	0.046±0.004
LQT	0.049±0.019	0.104±0.019	0.132±0.049	0.116±0.004	0.121±0.024	0.117±0.032	**0.151±0.008**	0.109±0.019
QAb	0.038±0.005	0.042±0.008	0.042±0.004	0.048±0.005	0.050±0.005	0.050±0.012	0.053±0.005	**0.055±0.003**
RAD	0.060±0.020	0.066±0.020	0.077±0.020	0.072±0.013	0.064±0.015	**0.087±0.018**	0.073±0.012	0.052±0.006
SA	0.053±0.003	**0.091±0.009**	0.075±0.015	0.077±0.024	0.062±0.004	0.067±0.010	0.064±0.005	0.077±0.008
SB	0.626±0.131	0.778±0.026	**0.787±0.025**	0.763±0.009	0.726±0.058	0.783±0.022	0.741±0.033	0.761±0.048
STach	0.630±0.032	0.863±0.002	0.841±0.030	0.863±0.029	0.866±0.028	0.863±0.017	0.881±0.023	**0.896±0.034**
TAb	0.185±0.038	0.291±0.039	0.291±0.018	0.293±0.005	0.295±0.009	0.243±0.036	**0.305±0.018**	0.281±0.016
TInv	0.081±0.008	0.128±0.039	0.135±0.008	0.131±0.005	0.122±0.015	0.114±0.032	**0.139±0.022**	0.129±0.017
Average	0.170±0.025	0.250±0.022	0.251±0.008	0.258±0.013	0.248±0.004	0.250±0.020	0.254±0.012	**0.273±0.016**

TABLE S7: Comparison results between ECGMatch and the state-of-the-art models using the mix-dataset protocol, under the annotation scheme of Physionet. The mean AUC and standard deviations on four databases are shown across three seeds. For each CVD, the model with the best performance is highlighted in bold.

Methods	MixMatch [15]	FixMatch [16]	FlexMatch [17]	DST [18]	PerMatch [19]	SoftMatch [20]	UPS [21]	**ECGMatch**
AF	0.628±0.083	0.798±0.026	0.796±0.032	0.806±0.031	0.809±0.014	**0.825±0.015**	0.780±0.030	0.817±0.026
AFL	0.722±0.065	0.844±0.024	0.808±0.003	0.837±0.025	0.803±0.032	0.802±0.023	0.838±0.024	**0.851±0.028**
BBB	**0.879±0.027**	0.813±0.088	0.774±0.077	0.833±0.049	0.878±0.029	0.822±0.022	0.772±0.096	0.867±0.047
CLBBB/LBBB	0.957±0.007	0.943±0.031	0.941±0.024	**0.964±0.009**	0.961±0.012	0.963±0.012	0.960±0.006	**0.964±0.005**
CRBBB/RBBB	0.867±0.112	0.935±0.032	0.964±0.010	0.962±0.009	0.966±0.003	0.964±0.005	0.963±0.007	**0.979±0.005**
IAVB	0.668±0.053	0.775±0.021	0.682±0.022	0.804±0.028	0.792±0.029	0.723±0.016	0.710±0.066	**0.841±0.033**
IRBBB	0.656±0.101	0.760±0.021	0.794±0.013	0.802±0.033	0.795±0.009	**0.812±0.026**	0.774±0.026	0.800±0.018
LAD	0.770±0.054	0.757±0.018	0.762±0.023	**0.794±0.043**	0.751±0.016	0.785±0.029	0.755±0.004	0.773±0.022
LAnFB	0.847±0.037	0.832±0.032	0.855±0.048	**0.902±0.037**	0.865±0.018	0.854±0.015	0.834±0.024	0.880±0.037
LQRSV	0.674±0.027	0.650±0.018	0.649±0.058	0.677±0.001	**0.682±0.018**	0.669±0.034	0.668±0.009	0.681±0.008
NSIVCB	0.623±0.060	0.649±0.064	0.654±0.064	0.689±0.015	0.671±0.023	0.675±0.029	0.648±0.030	**0.693±0.039**
NSR	0.710±0.034	**0.813±0.031**	0.801±0.021	0.802±0.017	0.793±0.046	0.809±0.035	0.789±0.010	0.810±0.017
PAC/SVPB	0.586±0.030	0.575±0.051	0.642±0.070	0.618±0.027	**0.692±0.023**	0.625±0.047	0.603±0.029	0.631±0.047
PR	0.898±0.032	0.927±0.028	0.940±0.015	0.951±0.006	0.933±0.023	0.937±0.027	0.929±0.018	**0.955±0.021**
PRWP	0.763±0.062	0.712±0.042	0.738±0.035	**0.787±0.025**	0.725±0.014	0.719±0.008	0.750±0.018	0.785±0.008
VPB/PVC	0.755±0.052	0.767±0.087	0.804±0.049	0.790±0.017	0.787±0.051	0.807±0.015	0.764±0.032	**0.816±0.024**
LPR	0.655±0.033	0.708±0.020	0.722±0.091	**0.786±0.041**	0.768±0.029	0.783±0.064	0.728±0.095	0.763±0.015
LQT	0.637±0.093	0.705±0.045	0.730±0.023	0.754±0.025	**0.763±0.014**	0.716±0.027	0.744±0.010	0.751±0.013
QAb	0.585±0.041	0.578±0.074	0.576±0.024	0.639±0.026	0.634±0.010	0.622±0.064	0.626±0.015	**0.662±0.044**
RAD	0.802±0.056	0.768±0.058	0.802±0.040	0.778±0.014	0.766±0.030	**0.820±0.016**	0.770±0.016	0.770±0.079
SA	0.538±0.024	**0.665±0.024**	0.632±0.032	0.616±0.062	0.599±0.016	0.603±0.042	0.607±0.023	0.621±0.040
SB	0.856±0.049	0.921±0.025	**0.935±0.018**	0.916±0.008	0.909±0.038	0.934±0.011	0.895±0.015	0.913±0.024
STach	0.936±0.013	0.969±0.006	0.973±0.005	0.965±0.012	0.976±0.006	0.974±0.003	0.972±0.006	**0.978±0.005**
TAb	0.557±0.075	0.696±0.054	0.680±0.027	0.702±0.018	**0.725±0.014**	0.626±0.045	0.704±0.003	0.718±0.017
TInv	0.632±0.009	0.698±0.082	0.695±0.033	0.739±0.019	0.730±0.051	0.685±0.045	0.730±0.015	**0.747±0.002**
Average	0.728±0.040	0.770±0.032	0.774±0.019	0.797±0.004	0.791±0.010	0.782±0.011	0.773±0.013	**0.803±0.008**

TABLE S8: Comparison results between ECGMatch and the state-of-the-art models using the within-dataset protocol. The mean performance and standard deviations on four databases are shown across three seeds.

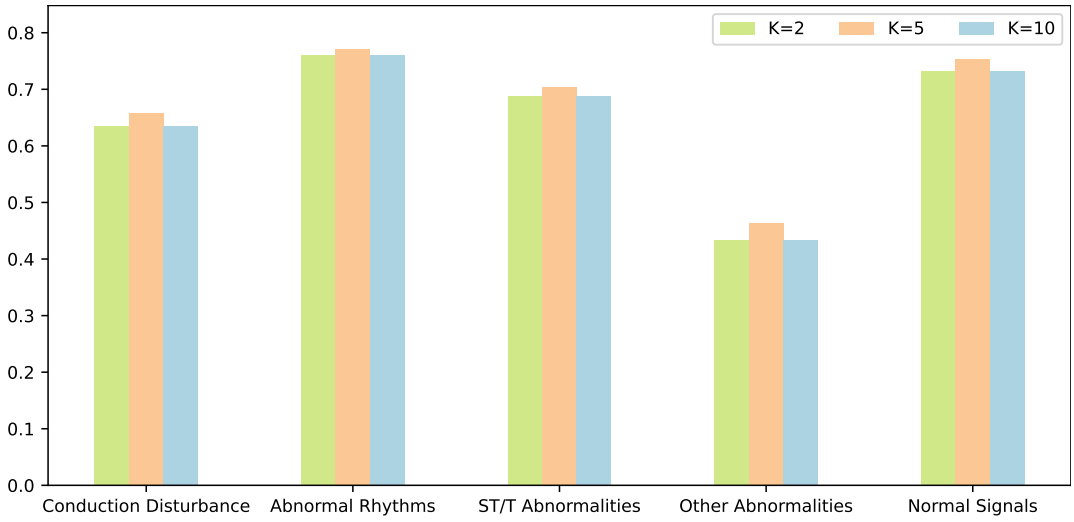
Methods	G12EC	PTB	Ningbo	Chapman
Ranking loss (The smaller, the better)				
Mixed Mean Teacher	0.214±0.010	0.161±0.013	0.184±0.030	0.261±0.019
MSDNN	0.168±0.001	0.232±0.114	0.211±0.045	0.107±0.030
ECGMatch	**0.140±0.006**	**0.134±0.003**	**0.045±0.002**	**0.052±0.002**
Hamming loss (The smaller, the better)				
Mixed Mean Teacher	0.342±0.006	0.266±0.012	0.136±0.006	0.207±0.022
MSDNN	0.270±0.043	0.301±0.107	0.123±0.011	0.221±0.091
ECGMatch	**0.278±0.008**	**0.233±0.009**	**0.122±0.001**	**0.139±0.002**
Coverage (The smaller, the better)				
Mixed Mean Teacher	2.462±0.039	2.020±0.049	2.279±0.116	2.583±0.076
MSDNN	2.292±0.003	2.306±0.444	2.361±0.168	1.978±0.113
ECGMatch	**2.173±0.027**	**1.922±0.015**	**1.724±0.010**	**1.761±0.021**
MAP (The greater, the better)				
Mixed Mean Teacher	0.684±0.004	0.723±0.012	0.771±0.007	0.715±0.004
MSDNN	0.728±0.007	0.726±0.020	0.801±0.001	0.744±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)				
Mixed Mean Teacher	0.824±0.004	0.871±0.004	0.891±0.003	0.888±0.004
MSDNN	0.846±0.006	0.873±0.007	0.915±0.003	0.896±0.005
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)				
Mixed Mean Teacher	0.430±0.008	0.447±0.008	0.523±0.003	0.483±0.015
MSDNN	0.454±0.009	0.401±0.078	0.522±0.032	0.491±0.046
ECGMatch	**0.477±0.003**	**0.467±0.009**	**0.563±0.001**	**0.554±0.009**

TABLE S9: Comparison results between ECGMatch and the state-of-the-art models using the cross-dataset protocol. The mean performance and standard deviations on four databases are shown across three seeds.

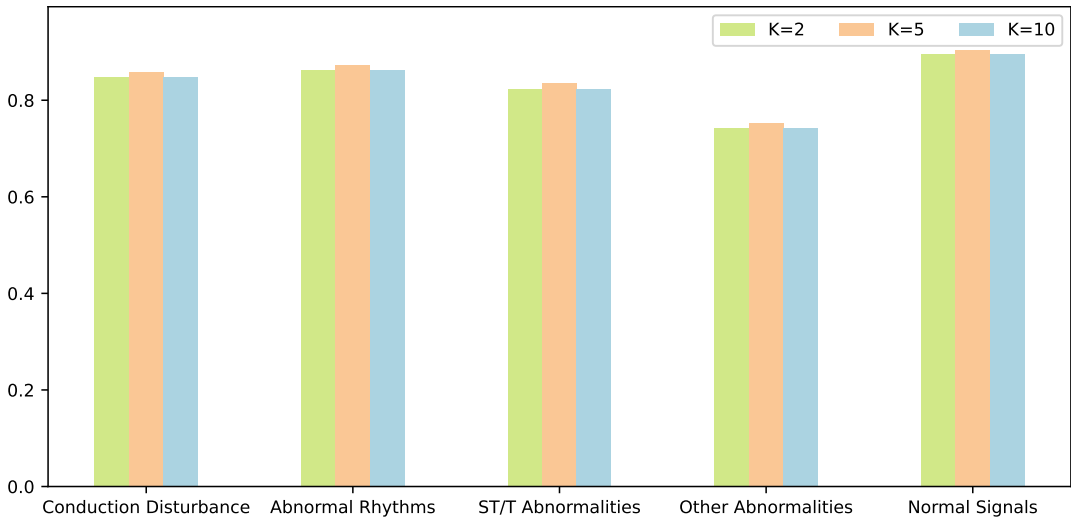
Methods	G12EC	PTB	Ningbo	Chapman
Ranking loss (The smaller, the better)				
Mixed Mean Teacher	0.265±0.048	0.275±0.028	0.191±0.013	0.183±0.017
MSDNN	0.263±0.027	0.298±0.014	0.139±0.027	0.092±0.006
ECGMatch	**0.203±0.004**	**0.248±0.005**	**0.102±0.006**	**0.068±0.002**
Hamming loss (The smaller, the better)				
Mixed Mean Teacher	0.368±0.003	0.354±0.008	0.332±0.021	0.285±0.016
MSDNN	0.305±0.029	0.293±0.010	0.232±0.026	0.205±0.042
ECGMatch	**0.331±0.007**	**0.310±0.001**	**0.253±0.008**	**0.219±0.003**
Coverage (The smaller, the better)				
Mixed Mean Teacher	2.653±0.174	2.470±0.098	2.322±0.046	2.281±0.070
MSDNN	2.648±0.100	2.561±0.051	2.117±0.103	1.912±0.024
ECGMatch	**2.415±0.016**	**2.379±0.023**	**1.971±0.025**	**1.803±0.008**
MAP (The greater, the better)				
Mixed Mean Teacher	0.610±0.010	0.545±0.010	0.631±0.005	0.707±0.003
MSDNN	0.619±0.009	0.529±0.006	0.631±0.003	0.716±0.007
ECGMatch	**0.657±0.009**	**0.591±0.012**	**0.689±0.002**	**0.748±0.004**
Marco AUC (The greater, the better)				
Mixed Mean Teacher	0.779±0.009	0.773±0.009	0.854±0.001	0.876±0.002
MSDNN	0.781±0.007	0.764±0.008	0.858±0.003	0.878±0.003
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)				
Mixed Mean Teacher	0.385±0.004	0.339±0.005	0.391±0.010	0.458±0.010
MSDNN	0.338±0.026	0.281±0.010	0.415±0.022	0.482±0.009
ECGMatch	**0.403±0.002**	**0.369±0.001**	**0.442±0.003**	**0.516±0.006**

TABLE S10: Comparison results between ECGMatch and the state-of-the-art models using the mix-dataset protocol. The mean performance and standard deviations on four databases are shown across three seeds.

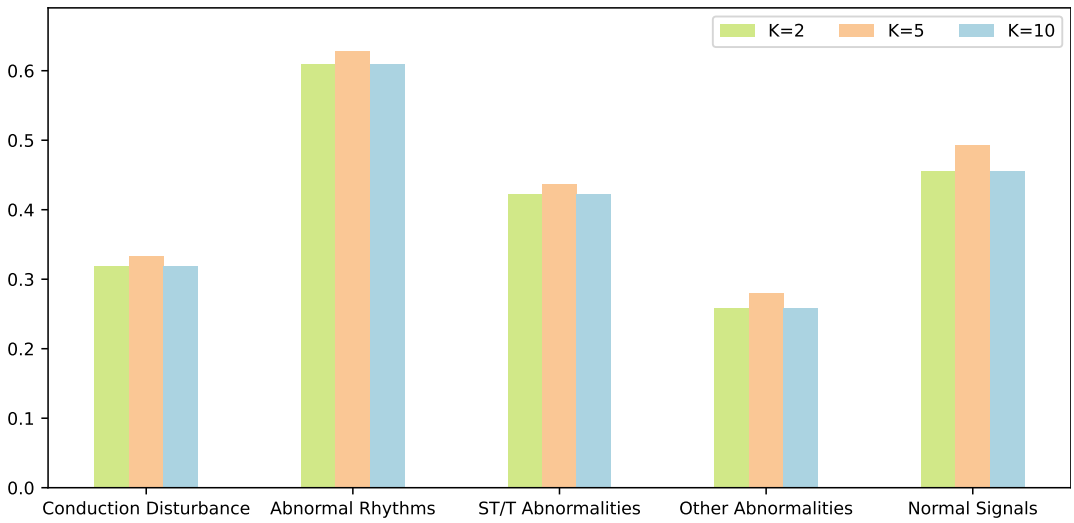
Methods	Ranking loss	Hamming loss	Coverage	MAP	Marco AUC	Marco G_{beta} score
Mixed Mean Teacher	0.213±0.010	0.314±0.008	2.357±0.047	0.633±0.006	0.827±0.003	0.424±0.004
MSDNN	0.299±0.085	0.432±0.075	2.727±0.352	0.605±0.013	0.811±0.004	0.337±0.040
ECGMatch	**0.150±0.001**	**0.270±0.001**	**2.101±0.009**	**0.658±0.006**	**0.838±0.003**	**0.442±0.002**



(a) Average Precision



(b) AUC



(c) G-$beta$ score

Fig. S6: The performance of the ECGMatch with varying hyperparameter K (cross-dataset protocol).