Towards Generalization of Cardiac Abnormality Classification Using Reduced-lead Multi-source ECG Signal

Xiaoyu Li[1],Chen Li[2], Xian Xu[4], Yuhua Wei[3], Jishang Wei[5], Yuyao Sun[4], Buyue Qian[6], Xiao Xu[4],

[1] **School of Electronic and Information Engineering, Xi'an Jiaotong University, China**

[2] **SPKLSTN Lab, Department of Computer Science and Technology, Xi'an Jiaotong University, China**

[3] **National Engineering Lab of Big Data Analytics, Xi'an Jiaotong University, China**

[4] **Ping An Health Technology, Shanghai, China**

[5] **HP Labs, Palo Alto, America**

[6] **The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China**

E-mail: `xiaoyuli@stu.xjtu.edu.cn, cli@xjtu.edu.cn`

Abstract.

Objective: ECG is a common non-invasive tool to measure the status of the heart. Although 12-lead ECG is a common non-invasive tool to identify cardiac abnormalities, fewer-lead ECG signal is promising for smaller, lower-cost and easier-to-use device. Besides, ECG signals may be collected from different hospital, cities and countries, with different data and label distribution shift. Our objective is to develop a generalizable classifier to identifies 26 cardiac abnormalities based on reduced-lead multi-source ECG signals.

Approach: Firstly, a series of pre-processing methods were proposed and applied on various data sources in order to mitigate the problem of data divergence. Secondly, we ensembled two SE_ResNet models and one rule-based model to enhance the performance of various ECG abnormalities' classification. Thirdly, we introduce a Sign Loss to tackle the problem of class imbalance, and thus improve the model's generalisability.

Main Results: In the PhysioNet/Computing in Cardiology Challenge 2020, our proposed approach achieved a challenge validation score of 0.682, and a full test score of 0.514, placed us 3rd out of 40 in the official ranking.

Significance: We proposed an accurate and robust predictive framework that combines deep neural networks and clinical knowledge to automatically classify multiple ECG abnormalities. Our framework is able to identify 27 ECG abnormalities from multi-lead ECG signals regardless of discrepancies in data sources and the imbalance of data labelling. We trained our framework on five datasets and validated it on six datasets from various countries. The outstanding performance demonstrate the effectiveness of our proposed framework.

Model Update: We add this part in the preprint to describe the main difference

between current solution and the codebase during the challenge. 1. We apply a source-based mask to ASL loss to alleviate the problem of label distribution shift. This change is supposed to significantly improve the generalizability of our model. 2. We choose not to finetune the model on CPSC and Georgia datasets anymore. The finetuning helps achieve better intra-source performance on hidden CPSC and Georgia datasets, but not on hidden unseen dataset. 3. We consider to extend the self-supervised learning part to try to publish another paper. Related code is not included in the submission. There is a lot work to do for the self-supervised learning and contrastive learning. we add related content here for now and might remove these for the final version. We will also rewrite some details, add more result analysis and comparison later.

1. Introduction

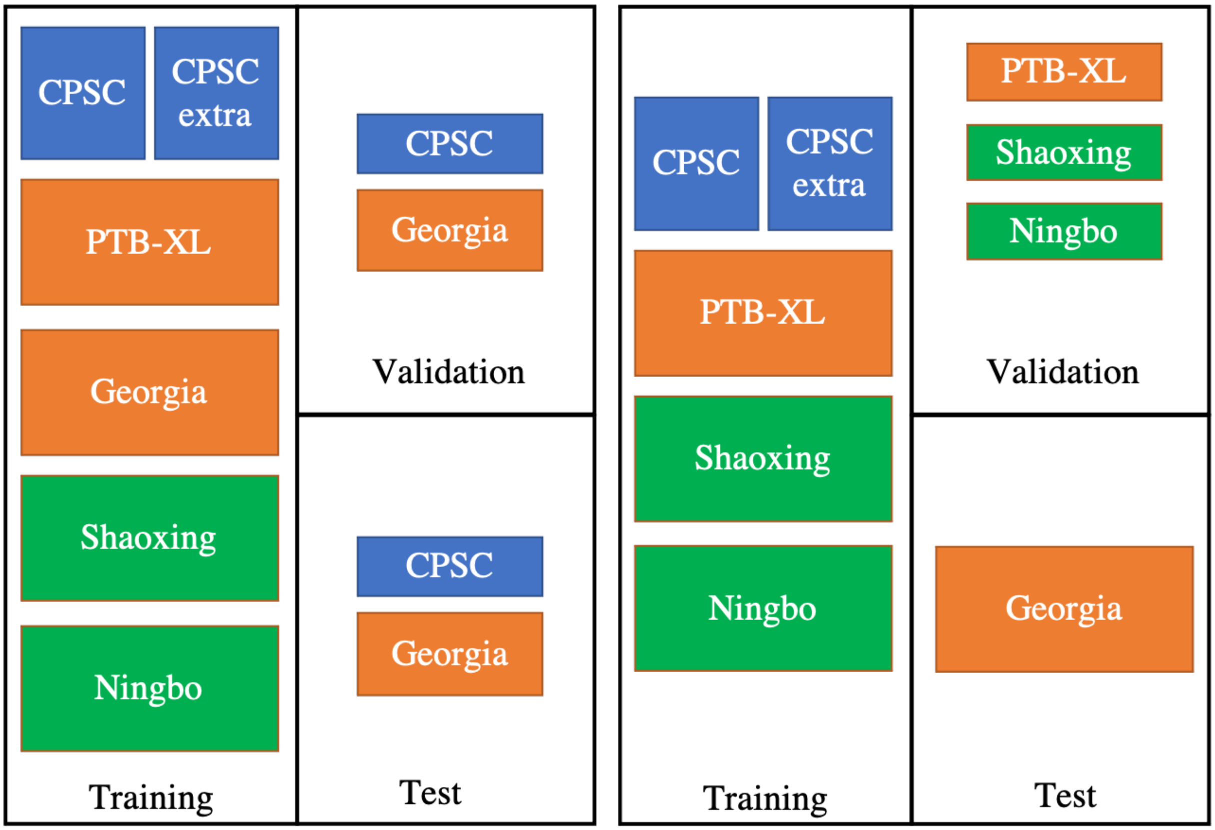
Electrocardiogram (ECG) is the most common non-invasive tool to screen and diagnose cardiac arrhythmias. However, such diagnoses are labor-intensive and require years of training. With the advancing machine learning and deep learning techniques, computer-aid methods are promising to detect cardiac abnormalities. The PhysioNet/Computing in Cardiology Challenge 2021 provides a platform to develop automatic models for classifying cardiac abnormalities from reduced-lead ECG recordings Perez Alday et al. (2020); Reyna et al. (2021).

We start Challenge 2021 by analysing the Challenge 2020. Last year, most of the top teams utilize deep convolutional neural network (CNN) and attention mechanism (Transformer or Squeeze-And-Excitation) Natarajan et al. (2020); Zhao et al. (2020); Zhu et al. (2020); Jia et al. (2020), which means CNN is the winner and attention mechanism counts. From the results on hidden test sets of Challenge 2020, we observe that models perform well on the data source which also appears in the training sets, but generalize poorly on the unseen hidden undisclosed dataset. To better understand and compare the performance of the models, we calculate the challenge score of random output on all training sets and on the Georgia training set as two weak baselines. The two baselines achieve the scores as 0.2832 and 0.3263 respectively. What surprizes us is that many models can not compete against the two random baselines, especially on the unseen dataset. We also observe that the final scores on the whole test set are basically aligned with the scores on the hidden undisclosed set, which means the performance on the unseen data source is vital, and generalization is the key. Consequently, when investigating various settings and techniques, we simultaneously consider intra-source performance and inter-source generalization to select the best ones.

Challenge 2021 provides multi-source datasets from different countries Liu et al. (2018), INCART Tihonenko et al. (2008), PTB (-XL) Bousseljot et al. (1995); Wagner et al. (2020), Chapman Zheng et al. (2020) and Ningbo .

To compare various techniques and settings in terms of intra-source performance and inter-source generalization using these datasets, we design the two data split settings, as shown in Fig.1. The data split in Fig.1.(a) is for intra-source performance. CPSC and Georgia datasets are randomly split into a training set, a validation set, and a test set, which matches the online validation set. The data split in Fig.1.(b) is for inter-source generalization. The Georgia dataset is split solely as the test set.

Under the two data split settings, we investigate different techniques, including domain alignment methods, different network architectures, domain knowledge aid, semi-supervised learning, and ensemble learning. Our solution mainly consists of SE-ResNet, peak detection as a self-supervised auxiliary task, and ensemble learning.



(a) Data split for intra-source performance

(b) Data split for inter-source generalization

Figure 1: Two data split settings for intra-source performance and inter-source generalization respectively. (a). CPSC and Georgia datasets are randomly split into a training set, a validation set, and a test set. (b). Georgia dataset is solely split as the test set.

2. Methods

2.1. Datasets

In PhysioNet/Computing in Cardiology Challenge 2021, the challenge data were partitioned into three parts, i.e., training dataset, official validation set, and official test dataset. Organizers made the training data and their corresponding diagnosis publicly available, whereas the official validation dataset and the official test dataset were kept hidden. The official validation dataset helps contestants validate their models. However, the source and size of which have not been published by the organizer. The official test dataset was used to calculate the final score for ranking. The details of training datasets and scored 26 ECG abnormalities are listed in the appendix.

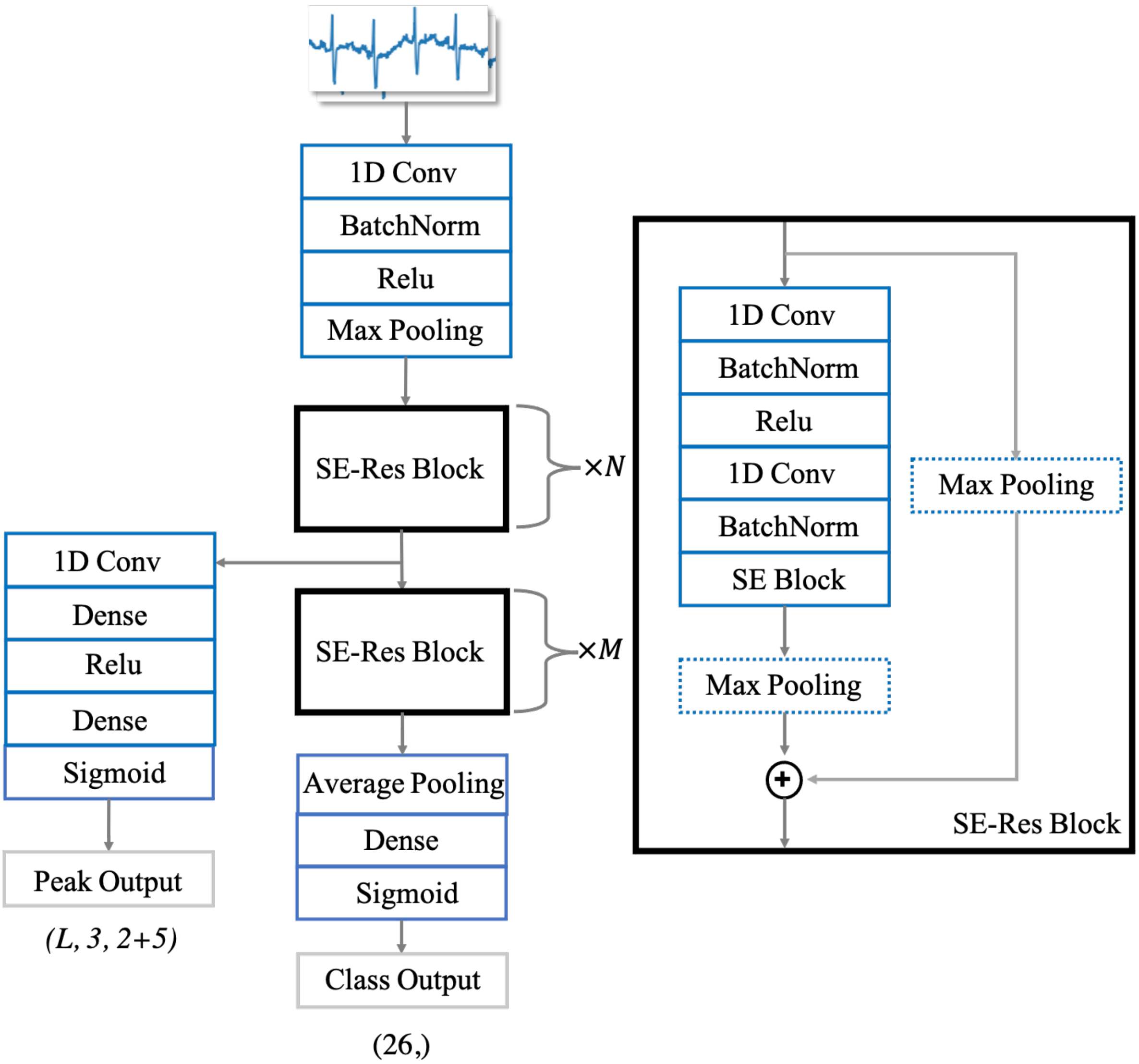


Figure 2: Network structure

2.2. Preprocessing

We remove INCART and PTB datasets. All recordings are resampled to 500 Hz or 250Hz. We randomly cut or zero-pad the recordings to the length of 4992 (about 10 seconds or 20 seconds). We add the label ”BBB” to the samples with the label “LBBB” or ”RBBB”

2.3. Abnormality classification task

The backbone of our SE-ResNet is the same as Jia et al. (2020), and the whole network structure is as Fig.2. We add a branch on the middle of the backbone, which outputs the results for peak detection. There is a hyperparameter array $[t_1, t_2, t_3, t_4]$. For each t_i, i denotes current stage number, t_i denotes the number of the SE-Res Block in the

Lead number	Officially provided leads	Selected leads
12	I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, V6	I, II, V1, V2, V3, V4, V5, V6
6	I, II, III, aVR, aVL, aVF	I, II
4	I, II, III, V2	I, II, V2
3	I, II, V2	I, II, V2
2	I, II	I, II

Table 1: Lead number and selected leads.

stage. There are 4 stages. At the first SE-Res Block in each stage, a Max-Pooling layer with a stride of 2 is applied to downsample the recording. We keep all of the settings the same for different-lead recordings, and this array is set as $[3, 4, 6, 3]$

Inspired by Petr et al. (2021), we also randomly choose the different lead configuration as data augmentation. Considering the spatial relationship among the leads, we step further to only select certain leads for different lead configuration as in Table.1.

For the abnormality classification task, the main difference between our solution and Jia et al. (2020) is that we utilize an asymmetric loss (ASL) Ben-Baruch et al. (2020) and a label mask for the multi-label classification problem. The ASL loss is a strong multi-label version of focal loss. There are 26 labels for each recording, but usually, only a small number of the labels are positive.

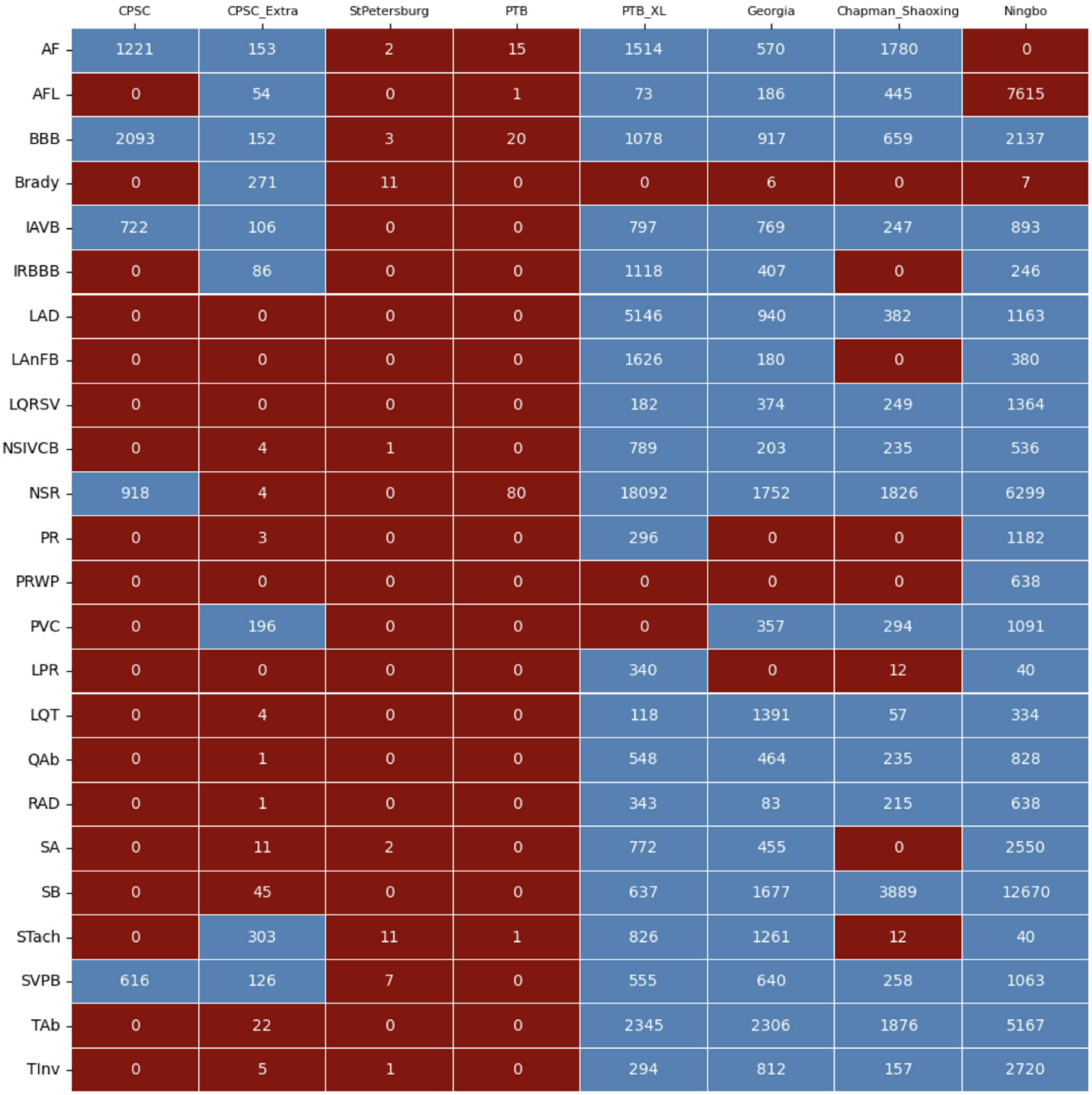
$$L_{classification} = \frac{1}{K}\sum_{k=0}^{K} ASL(p_k, y_k) \tag{1}$$

We utilize ASL to alleviate such a label imbalance problem. Three hyperparameters $(\lambda_+, \lambda_-, m)$ need to be set. We set λ_+ as 1, λ_- as 4 and m as 0.05 to reduce the contribution of negative labels. For details of ASL loss, please refer to Ben-Baruch et al. (2020).

Besides imbalance, the label quality can vary from different classes and sources. We propose to evaluate the labels according to different database details. For example, there are only six classes are labeled on the CPSC dataset. This means that labels from other 20 classes should not contribute to the loss. Similarly, considering the label number the some background knowledge, we manully set the label mask for each class and each database as in Figure.3.

2.4. Peak detection as an auxiliary task

We extend Li et al. (2019) to detect all 5 kinds of peaks. We add a branch before the last stage of SE-ResNet. The branch consists of a multi-layer-perception and a convolutional layer with a filter size of $3 * (2 + 5)$. The output of the branch is then reshaped to $(L, 3, 2 + 5)$. L is the length of the feature map before the last stage. The number L means, we treat the recording as L grids and detect peaks in each grid, shown



	CPSC	CPSC_Extra	StPetersburg	PTB	PTB_XL	Georgia	Chapman_Shaoxing	Ningbo
AF	1221	153	2	15	1514	570	1780	0
AFL	0	54	0	1	73	186	445	7615
BBB	2093	152	3	20	1078	917	659	2137
Brady	0	271	11	0	0	6	0	7
IAVB	722	106	0	0	797	769	247	893
IRBBB	0	86	0	0	1118	407	0	246
LAD	0	0	0	0	5146	940	382	1163
LAnFB	0	0	0	0	1626	180	0	380
LQRSV	0	0	0	0	182	374	249	1364
NSIVCB	0	4	1	0	789	203	235	536
NSR	918	4	0	80	18092	1752	1826	6299
PR	0	3	0	0	296	0	0	1182
PRWP	0	0	0	0	0	0	0	638
PVC	0	196	0	0	0	357	294	1091
LPR	0	0	0	0	340	0	12	40
LQT	0	4	0	0	118	1391	57	334
QAb	0	1	0	0	548	464	235	828
RAD	0	1	0	0	343	83	215	638
SA	0	11	2	0	772	455	0	2550
SB	0	45	0	0	637	1677	3889	12670
STach	0	303	11	1	826	1261	12	40
SVPB	616	126	7	0	555	640	258	1063
TAb	0	22	0	0	2345	2306	1876	5167
TInv	0	5	1	0	294	812	157	2720

Figure 3: Label mask.

as Fig.4. The number 3 means we detect 3 peaks in a grid at most. The 2 + 5 means, for each grid, we predict peak detection confidence C, peak relative position x, and 5 peak classes (PQRST). The ground truth of the peaks is calculated by Makowski et al. (2021). The loss of peak detection is as the formula (2), where $\mathbb{1}$ denotes whether to calculate the corresponding loss item when there is a peak inside the grid or not, λ_{coord} and λ_{noobj} are set as 5 and 0.2 to balance positive grids and negative grids.

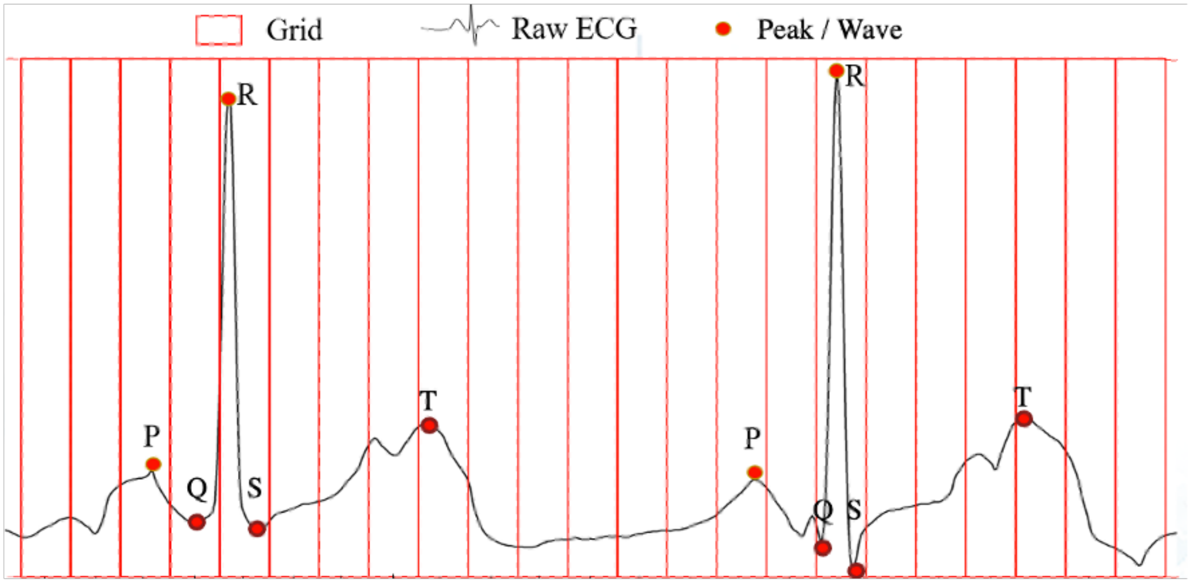


Figure 4: One recording as multiple grids.

$$\begin{aligned} L_{detection} = \lambda_{coord} & \sum_{i=0}^{L} \boldsymbol{1}_i^{obj} (x_i - \hat{x}_i)^2 \\ & + \sum_{i=0}^{L} \boldsymbol{1}_i^{obj} (C_i - \hat{C}_i)^2 \\ & + \lambda_{noobj} \sum_{i=0}^{L} \boldsymbol{1}_i^{noobj} (C_i - \hat{C}_i)^2 \\ & + \sum_{i=0}^{L} \boldsymbol{1}_i^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2 \end{aligned} \tag{2}$$

Then we combine the classification loss and peak detection loss with a hyperparameter α. We set the α as 0.1. During training, we pre-train the SE-ResNet only with the peak detection task first. Then we combine both tasks to finetune the network.

$$L_{final} = L_{classification} + \alpha * L_{detection} \tag{3}$$

2.5. Ensemble

We also integrate different models for final classification. An ideal generalizable classifier learns the features which truly capture the patterns for 26 classes instead of the features dependent on the data source. Considering there are only 6 classes for CPSC dataset, an ideal classifier should generalize poorly on CPSC data. Thus, a source classifier with the same backbone as in SE-ResNet is trained to predict whether the recording is from CPSC datasets. If true, the recording is sent to the classifier, which only outputs

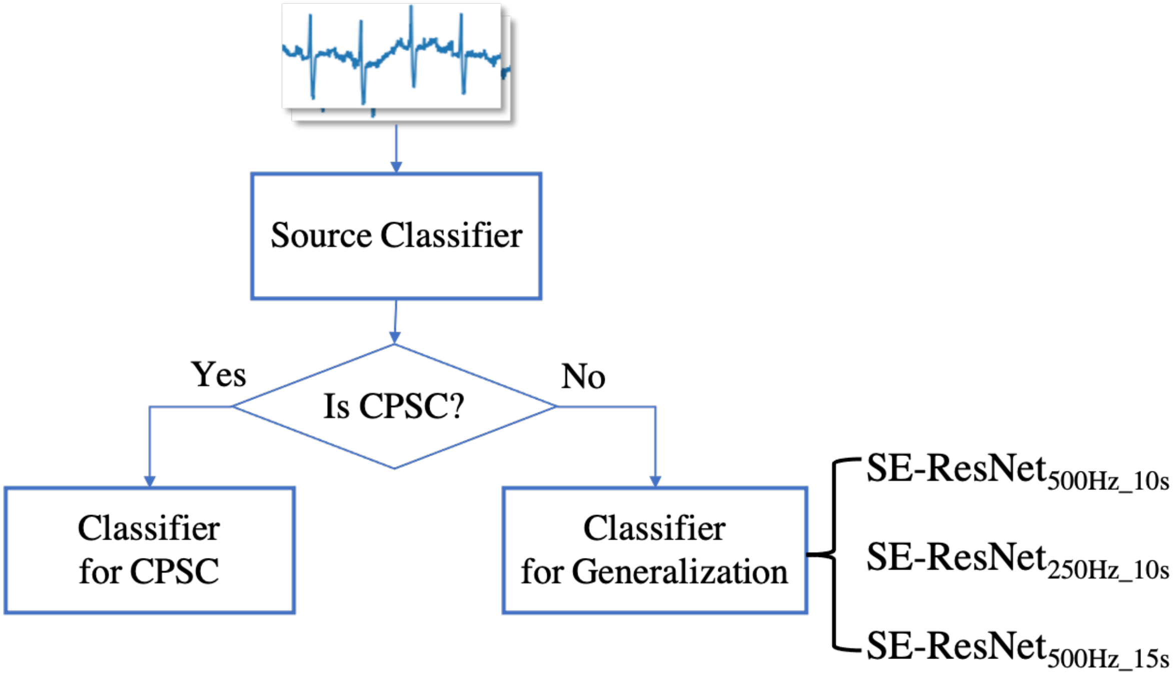


Figure 5: Ensemble structure

6 classes for CPSC. If not, we utilize the classifier for generalization. The pipeline is shown as in Fig.5. Each of the classifier for CPSC and genralization is integrated with three SE-ResNet models, $SE\text{-}ResNet_{500Hz\_10s}$, $SE\text{-}ResNet_{250Hz\_10s}$, $SE\text{-}ResNet_{500Hz\_15s}$. These SE-ResNet models are trained with different input lengths and sampling rates.

2.6. Evaluation Metrics

We adopted the official evaluation metrics from PhysioNet/Computing in Cardiology Challenge 2021 Reyna et al. (2021). To conform to real-world clinical practice, where some misdiagnoses are less harmful than others, the misdiagnoses that end in similar outcomes or treatments as the ground truth diagnoses will still be awarded partial credit.

To be more specific, $C = [c_i]$ defined as the collection of our predictions. The multiclass confusion matrix is $A = [a_{ij}]$, where a_{ij} is the normalized number of recordings in a database that were classified as belonging to class c_i but actually belong to class c_j. The a_{ij} is calculated by $a_{ij} = \sum_{k=1}^{n} a_{ijk}$, where a_{ijk} is defined by

$$a_{ijk} = \begin{cases} \frac{1}{|x_k \cup y_k|}, & \text{if } c_i \in x_k \text{ and } c_j \in y_k \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

For n ECG signal recordings, x_k is the set of positive labels and y_k is the collection of positive predictions for the k-th recording. Hence, the $|x_k \cup y_k|$ is the number of unique classes with a positive label and predictions for recording k. A reward metrics $W = [w_{ij}]$ is defined by the challenge organizer, where w_{ij} denotes the reward for a

Models	SE-ResNet	Masked CE	Masked ASL	+warmup&earlyStop	+postprocess
12lead	0.432	0.482	0.528	0.542	0.551
6lead, 2lead	0.419	0.462	0.509	0.518	0.526
4lead, 3lead	0.425	0.471	0.516	0.530	0.535

Table 2: The results of challenge score for inter-source generalization.

positive classifier output for class c_i with a positive label c_j. Then, the unnormalized score will be calculated by equation 5.

$$Unnormalized\_Score = \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}a_{ij} \tag{5}$$

$$Normalized\_Score = \frac{Unnormalized\_Score - Inactive\_Score}{Correct\_Score - Inactive\_Score} \tag{6}$$

After normalization by equation 6, a score of 1 will be assigned to the classifier that always predicts the true label, while a score of 0 will be assigned to an inactive classifier. The Inactive Score is the score for an inactive classifier that always outputs a normal class, while the Correct Score is the score for the model that always predicts the true class. The detailed calculation of Normalized_Score is shown in Reyna et al. (2021). The Normalized Score will be in the range between 0 and 1, and the higher the score indicates the better performance of the model.

2.7. Training details and post-processing

For training settings, we set batch size as 256, optimizer as *AdamW* with a learning rate of 0.0006. We also linearly warm up the learning rate for the first 3 epochs, and then adopt a cosine learning schedule. We also add early stopping with a patience number of 5..

As for post-processing, if all of 26 labels are predicted negative for a recording, we label the *NSR* class as positive. If the label of *TInv* is positive, we also label the *TAb* class as positive.

3. Results

The offline results for inter-source generalization are in the Table.2. The results prove the effectiveness of the components in our model and also show that reduced-lead ECG also contains rich information for abnormality classification.

The online test results are in the Table.3...

Leads	12	6	4	3	2
Online Validation	0.??	0.??	0.??	0.??	0.??
Ranking	?th	?th	?th	?th	?th

Table 3: The results of online test.

4. Conclusions

In this paper, we propose a generalizable deep learning model to classify cardiac abnormalities for reduced-lead multi-source ECG signal. We compare various methods and techniques in the term of inter-source performance. We treat different leads as augmentation and simplify the lead configuration. We also propose a masked asymmetric loss to deal with label imbalance and label distribution shift. Extensive experiments and Challenge2021 test results demonstrate the effectiveness of our proposed models.

Acknowledgments

This work has been supported by National Natural Science Foundation of China (61772409); The consulting research project of the Chinese Academy of Engineering (The Online and Offline Mixed Educational Service System for "The Belt and Road" Training in MOOC China); Project of China Knowledge Centre for Engineering Science and Technology; The innovation team from the Ministry of Education (IRT_17R86); and the Innovative Research Group of the National Natural Science Foundation of China (61721002).

References

Ben-Baruch, E., Ridnik, T., Zamir, N., Noy, A., Friedman, I., Protter, M. and Zelnik-Manor, L. (2020). Asymmetric Loss For Multi-Label Classification.

Bousseljot, R., Kreiseler, D. and Schnabel, A. (1995). Nutzung der EKG-Signaldatenbank CARDIODAT der ptb über das internet, *Biomedizinische TechnikBiomedical Engineering* **40**(s1): 317–318.

Jia, W., Xu, X., Xu, X., Sun, Y. and Liu, X. (2020). Automatic Detection and Classification of 12-lead ECGs Using a Deep Neural Network, *2020 Computing in Cardiology*, IEEE, pp. 1–4.

Li, X., Qian, B., Wei, J., Zhang, X., Chen, S., Zheng, Q. and none, n. (2019). Domain Knowledge Guided Deep Atrial Fibrillation Classification and Its Visual Interpretation, *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 129–138.

Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., Liu, Y., Ma, C., Wei, S., He, Z., Li, J. and Kwee, E. N. Y. (2018). An Open Access Database for Evaluating the

Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection, *Journal of Medical Imaging and Health Informatics* **8**(7): 1368—1373.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C. and Chen, S. H. A. (2021). NeuroKit2: A Python Toolbox for Neurophysiological Signal Processing, *Behavior Research Methods* **53**(4): 1689–1696.

Natarajan, A., Chang, Y., Mariani, S., Rahman, A., Boverman, G., Vij, S. and Rubin, J. (2020). A Wide and Deep Transformer Neural Network for 12-Lead ECG Classification, *2020 Computing in Cardiology*, IEEE, pp. 1–4.

Perez Alday, E. A., Gu, A., Shah, A., Robichaux, C., Wong, A.-K. I., Liu, C., Liu, F., Rad, B. A., Elola, A., Seyedi, S., Li, Q., Sharma, A., Clifford, G. D. and Reyna, M. A. (2020). Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020, *Physiological Measurement* **41**.

Petr, N., Adam, I., Radovan, S., Ivo, V., Zuzana, K., Pavel, J. and Filip, P. (2021). Classification of ECG using Ensemble of Residual CNNs with Attention Mechanism, *Computing in Cardiology* **48**.

Reyna, M. A., Sadr, N., Perez Alday, E. A., Gu, A., Shah, A., Robichaux, C., Rad, B. A., Elola, A., Seyedi, S., Ansari, S., Li, Q., Sharma, A. and Clifford, G. D. (2021). Will Two Do? Varying Dimensions in Electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021, *Computing in Cardiology* **48**: 1–4.

Tihonenko, V., Khaustov, A., Ivanov, S., Rivin, A. and Yakushenko, E. (2008). St Petersburg INCART 12-lead arrhythmia database, *PhysioBank, PhysioToolkit, and PhysioNet* . doi: 10.13026/C2V88N.

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W. and Schaeffter, T. (2020). PTB-XL, a Large Publicly Available Electrocardiography Dataset, *Scientific Data* **7**(1): 1–15.

Zhao, Z., Fang, H., Relton, S. D., Yan, R., Liu, Y., Li, Z., Qin, J. and Wong, D. C. (2020). Adaptive Lead Weighted ResNet Trained with Different Duration Signals for Classifying 12-lead ECGs, *2020 Computing in Cardiology*, IEEE, pp. 1–4.

Zheng, J., Zhang, J., Danioko, S., Yao, H., Guo, H. and Rakovski, C. (2020). A 12-lead Electrocardiogram Database for Arrhythmia Research Covering More Than 10,000 Patients, *Scientific Data* **7**(48): 1–8.

Zhu, Z., Wang, H., Zhao, T., Guo, Y., Xu, Z., Liu, Z., Liu, S., Lan, X., Sun, X. and Feng, M. (2020). Classification of Cardiac Abnormalities From ECG Signals Using SE-ResNet, *2020 Computing in Cardiology*, IEEE, pp. 1–4.

Appendix

Table 4: Basic information of six different datasets that forms the challenge training dataset.

Databases	ECG Recordings	ECG Abnormality Categories
CPSC	6,877	9
CPSC2	3,453	72
INCART	74	37
PTB	516	17
PTB-XL	21,837	50
Georgia	10,344	67
Chapman, Ningbo	45,152	67

Table 5: 26 scored ECG abnormalities and their corresponding abbreviations.

ECG Abnormality	Abbreviation
Atrial fibrillation	AF
Atrial flutter	AFL
Bundle Branch Block	BBB
Bradycardia	Brady
Complete left bundle branch block	LRBBB
Complete right bundle branch block	CRBBB
1st degree AV block	IAVB
Incomplete right bundle branch block	IRBBB
Left axis deviation	LAD
Left anterior fascicular block	LAnFB
Prolonged PR interval	LPR
Low QRS voltages	LQRSV
Prolonged QT interval	LQT
Nonspecific intraventricular conduction disorder	NSIVCB
Sinus rhythm	NSR
Supraventricular premature beats	SVPB
Pacing rhythm	PR
Poor R wave Progression	PRWP
Premature ventricular contractions	PVC
Q wave abnormal	QAb
Right axis deviation	RAD
Sinus arrhythmia	SA
Sinus bradycardia	SB
Sinus tachycardia	STach
T wave abnormal	TAb
T wave inversion	TInv