DATA AUGMENTATION VIA WASSERSTEIN GEODESIC PERTURBATION FOR ROBUST ELECTROCARDIOGRAM PREDICTION

Jiacheng Zhu[1,*], Jielin Qiu[1,*], Zhuolin Yang[2], Michael Rosenberg[3], Emerson Liu[4], Bo Li[2], Ding Zhao[1]
[1]Carnegie Mellon University, [2]University of Illinois at Urbana-Champaign,
[3]University of Colorado School of Medicine, [4]Allegheny General Hospital

ABSTRACT

There has been an increased interest in applying deep neural networks to automatically interpret and analyze the 12-lead electrocardiogram (ECG). However, the imbalance and heterogeneity of real-world datasets place obstacles to the efficient training of neural networks. Moreover, deep learning classifiers could be vulnerable to adversarial examples and perturbations and could lead to catastrophic outcomes for clinical trials and insurance claims. In this paper, we propose a physiologically-inspired data augmentation to improve the performance, generalization, and to increase the robustness of ECG prediction models. We obtain augmented samples by perturbing the data distribution towards other classes along the geodesic in Wasserstein space. To better utilize the domain knowledge, we design a ground metric that recognizes the difference between ECG signals based on physiological features. Learning from 12-lead ECG signals, our model is able to distinguish five categories of cardiac conditions. Our results demonstrate improvements in accuracy and robustness reflecting the effectiveness of our data augmentation method.

1 INTRODUCTION

The 12-lead Electrocardiogram (ECG) is the foundation for cardiology and electrophysiology. ECG provides unique information about the structure and electrical activity of the heart and systemic conditions, through changes in timing and morphology of the recorded waveforms. Achievement of reliable ECG reading would be a significant achievement, such that critical and timely ECG interpretations of acute cardiac conditions can lead to efficient and cost-effective intervention. With the development of machine learning and deep learning methods, it may be possible to identify additional previously unrecognized signatures of disease. Many methods have been explored for diagnosing physiological signals, i.e., EEG, ECG, EMG, etc (Liu et al., 2019; Shanmugam et al., 2019; Côté-Allard et al., 2019). Due to limited data and sensitive modeling frameworks, the diagnosis results are not always robust. Also, deep learning models for ECG data have been shown to be susceptible to adversarial attack (Han et al., 2020; Hossain et al., 2021b; Chen et al., 2020).

To tackle the problem caused by *adversarial data distributions*, people have proposed both empirical and certified robust learning approaches, such as adversarial training (Madry et al., 2017) and certified defense approaches (Cohen et al., 2019; Li et al., 2020; 2021). Despite conventional deep learning algorithms, since different categories of ML algorithms are being deployed to safety-critical domains, there is a need for provable robustness guarantees for different ML algorithms, such as ensemble learning (Yang et al., 2021b;a), reinforcement learning (Wu et al., 2021b;a), and federated learning (Xie et al., 2021; Xie et al.).

It has already been shown that *data augmentation* strategies (Rebuffi et al., 2021a;b; Gao et al., 2020; Volpi et al., 2018; Ng et al., 2020) or more training data (Carmon et al., 2019) can improve the performance and increase the robustness of deep learning models. Specifically, augmenting data with random Gaussian noise (Cohen et al., 2019) or transformations (Li et al., 2021) yields certifiable smoothed models. Mixup methods (Zhang et al., 2018; Greenewald et al., 2021), which augment data with weighted averages of training points, also promote the certifiable robustness (Jeong et al.,

[*]Equal contribution

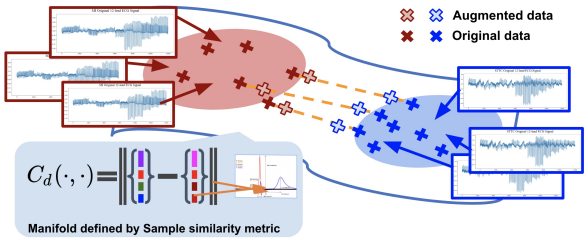

Figure 1: Our data augmentation creates perturbed samples toward the closest other-class samples. The perturbation lies on the geodesic between two distributions.

2021). However, different types of data usually contain domain-specific properties. In particular, the temporal structure in data such as audio data (Yang et al., 2018) and natural language (Wang et al., 2021) should not be ignored when performing robust training.

In this paper, we propose a new data augmentation method from a probability perspective. We perturb the data distribution towards other classes along the geodesic in a Wasserstein space. Also, the ground metric of this Wasserstein space is computed via a set of physiological features so that the perturbation lies on a manifold that exploits the physiology properties of ECG data. We employ a Multi-Feature Transformer as base classifier to evaluate the performance of our proposed method.

2 Related Work

ECG deep learning and robustness With the development in machine learning, many models have been applied to ECG disease detection (Kiranyaz et al., 2015; Nonaka & Seita, 2021; Khurshid et al., 2021; Raghunath et al., 2021; Giudicessi et al., 2021; Strodthoff et al., 2021). The transformer model has recently been adopted in several ECG applications, i.e., arrhythmia classification, abnormalities detection, stress detection, etc (Yan et al., 2019; Che et al., 2021; Natarajan et al., 2020; Behinaein et al., 2021; Song et al., 2021; Weimann & Conrad, 2021). Robustness of ECG has recently drawn more attention. Venton et al. (2021) generated clean and noisy ECG datasets to test the robustness of different models. Hossain et al. (2021a) proposed Conditional GAN, which claimed to be robust against adversarial attacked ECG signals. Venton (2021) explored the impact of different physiological noise types, and signal-to-noise ratios (SNRs) of noise.

Data augmentation for sequential data Zhang et al. (2018) proposed Mixup, an effective model regularizer for data augmentation that encourages the model to behave linearly in-between training examples, which has been applied in sequential data. It generates out-of-manifold samples through linearly interpolating inputs and their corresponding labels of random sample pairs. Zhang et al. (2020) augmented the queried samples by generating extra labeled sequences. Guo et al. (2020) created new synthetic examples by softly combining input/output sequences from the training set. Guo (2020) embraced nonlinear interpolation policy for both the input and label pairs, where the mixing policy for the labels is adaptively learned based on the mixed input. However, the data augmentation for electrocardiograms has not been well explored.

3 Methods

3.1 Robust deep learning with data augmentation

It is imperative to obtain a deep learning model that is operational in the presence of potentially adversarial shift in data distribution. Through the framework of distributional robust optimization (Weber et al., 2022), we denote P as the joint data distribution over features $X \in \mathcal{X}$ and labels $Y \in \mathcal{Y}$, and let $h_\theta : \mathcal{X} \mapsto \mathcal{Y}$ be a family of predictive function parameterized by θ. Given a loss function $l : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$, we wish to solve the following optimization problem:

$$\min_{\theta} \sup_{Q \in \mathcal{U}_P} \mathbb{E}_{(X,Y)\sim Q}[l(h_\theta(X), Y)], \tag{1}$$

where $\mathcal{U}_P \subseteq \mathcal{P}(\mathcal{Z})$ is a set of probability distribution. Intuitively, this objective finds the worst-case optimal predictor h^*_θ when the data distribution P is perturbed towards some distribution $\mathcal{U}_P$.

A promising way to enable robust learning is to provide adversarially perturbed samples with data augmentation (Volpi et al., 2018; Rebuffi et al., 2021c). Consider we can rewrite the joint data distribution $P(X, Y)$ as the product of conditional distributions $P(X, Y) = P(X|Y)P(Y)$. Also, since we focus on the k-classification problem, we denote $P_k(X) = P(X|Y_k)$ as the data distribution of one class k. When doing the data augmentation, we want to perturb the data distribution $P_i(X)$ towards another class $P_j(X), i \neq j$ as we believe it is those data samples lie on the geodesic that serve as adversarial samples (Courty et al., 2017; Moosavi-Dezfooli et al., 2017).

3.2 Data augmentation by perturbation on the geodesic

The Monge formulation of optimal transport finds a map $T : \mathbb{R}^d \mapsto \mathbb{R}^d$ than transports a distribution P towards Q:

$$T^* = \arg\inf_T \int \|x - T(x)\|^p dP(x), \tag{2}$$

where a minimizer T^* is the optimal transport map such that $T^*_{\#}P = Q$, where $T^*_{\#}P$ is the pushforward of P. Given distributions P and Q, if T^* exits, then map $T_t(x) = (1-t)x + tT^*(x)$ gives the path of a particle of mass at x and $P_t = T_{t\#}P$ is the *geodesic* connecting P to Q.

However, the minimizer might not always exit. Thus we introduce Kantorovich that finds an optimal coupling π of given measures $\mu \in \mathcal{M}(\mathcal{C}_s), \nu \in \mathcal{M}(\mathcal{C}_t)$ to minimize

$$\inf_{\pi \in \Pi} \int_{\mathcal{C}_s \times \mathcal{C}_t} d(c_s, c_t)\, \mathrm{d}P(c_s, c_t), \text{ subject to } \mathcal{P} = \left\{ P : \gamma^{\mathcal{C}_s}_{\#} P = \mu, \gamma^{\mathcal{C}_t}_{\#} P = \nu \right\} \tag{3}$$

where $\mathcal{C}_s$ and $\mathcal{C}_t$ are the source and target context space, $d(\cdot, \cdot) : \mathcal{C}_s \times \mathcal{C}_t \mapsto \mathbb{R}^+$is a distance function, $\gamma^{\mathcal{C}_s}, \gamma^{\mathcal{C}_t}$ are projections from $\mathcal{C}_s \times \mathcal{C}_t$ onto $\mathcal{C}_s$ and $\mathcal{C}_t$ respectively. In a more general case, to obtain the perturbation between P_i and P_j corresponds to the problem of Wasserstein barycenter, which interpolate between distributions along the geodesic

$$\tilde{P}_{ij} = \inf_{\tilde{P}_\alpha} (1-\alpha)W(P_i, \tilde{P}_\alpha) + \alpha W(\tilde{P}_\alpha, P_j) \text{ where } \alpha \in (0, \epsilon) \tag{4}$$

Then, the augmented samples can be obtained $(\tilde{x}_i, y_i) \sim \tilde{P}_{ij}$. We will show an algorithmic derivation of this augmentation procedure leads to a similar framework with mixup guided by batch optimal transport, but we propose to better exploit the data manifold structure with a user specified ground metric based on (time domain and frequency domain) physiological features.

3.3 Algorithm

In practice, we only observe discrete training samples that represents empirical distribution of P_i and P_j. Consider $\mathbf{X}_i = \{\mathbf{x}^i_l\}_{l=1}^{n_i}$ and $\mathbf{X}_j = \{\mathbf{x}^j_l\}_{l=1}^{n_j}$ are two set of features from class i and j respectively. The empirical distributions are written as $\hat{P}_i = \sum_{l=1}^{n_i} p^i_l \delta_{x^i_l}$ and $\hat{P}_j = \sum_{l=1}^{n_j} p^j_l \delta_{x^j_l}$ where δ_x is the Dirac function at location $x \in \Omega$, p^i_l and p^j_l are probability mass associated to the sample. Then the Wasserstein distance between empirical measures Eq.(3) becomes

$$\pi^* = \arg\min_{\pi \in \Pi} \sum_{l=1,k=1}^{n_i, n_j} C(\mathbf{x}^i_l, \mathbf{x}^j_k)\pi_{l,k}. \tag{5}$$

In the special case where the ground metric $C(\cdot, \cdot)$ is l2 norm, we can follow the barycentric mapping to obtain the pushforward $\hat{\mathbf{X}}_i = T^{ij}_{t\#}P_j = n_i \pi^* \mathbf{X}_j$. Then we can explicitly perturb the $\mathbf{X}_i$ towards $\mathbf{X}_j$ by:

$$\tilde{X}_{ij} = (1-\alpha)\mathbf{X}_i + \alpha \hat{\mathbf{X}}_i, \tag{6}$$

When selecting a batch of samples, our method interpolate the class i samples with a set of pushforward samples $\hat{\mathbf{X}}_i$, rather than $\mathbf{X}_j$, which better exploit the geometric structure of data distribution.

4 Results

We carried out experiments on the PTB-XL dataset (Wagner et al., 2020), which contains clinical 12-lead ECG of 10-second length. The augmented data generated by our proposed method is used to improve classification robustness among different categories. In specific, (1) In the augmentation

procedure, we randomly sample a batch of ECG signal from both the source and target categories and then use formulation in Equation (6) to get the barycentric mapping samples. (2) We mix the original data and augmented data and then process them for the MF-Transformer. (More details are shown in the Appendix.) Examples of augmented data are shown in Fig. 2, where we can find the augmented data preserves the semi-periodic nature, and the results of each lead fit well with the ECG pattern compared with original ECG signals by domain knowledge.

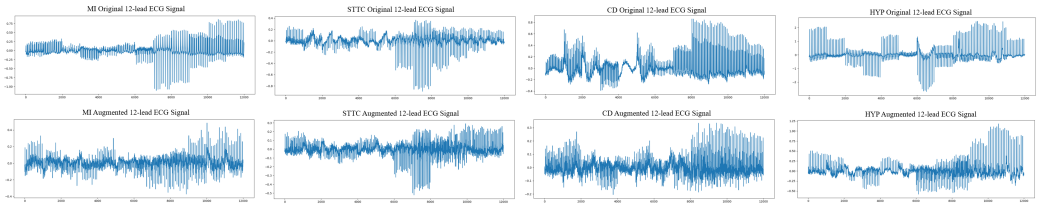

Figure 2: Examples of 10-s 12-lead original ECG signals and augmented ECG signals within different conditions. Top row: original signals; Bottom row: augmented signals.

To evaluate our method, we used a MF-Transformer model as classifier (details of the model are introduced in the Appendix). We trained the MF-Transformer model with (1) original PTB-XL data, (2) oversampling augmented data, and (3) our augmented data. Table 2 and Fig. 3 show that compared with baseline results and oversampling results, our augmented method not only improved the classification accuracy of each category but also improved the average classification result from 71.80% (original) and 72.05% (oversampling) to 75.82% (ours), which demonstrated the robustness improvement. (More details are introduced in the Appendix)

Table 1: Comparison of classification results by different data augmentation methods.

Methods	Average Accuracy	F1-score
MF-Transformer-Raw	71.80 %	0.669
MF-Transformer-Oversampling	72.05 %	0.717
MF-Transformer-Ours	**75.82 %**	**0.757**

Table 2: Comparison of the AUROC result on the clean test set and adversarial test set. The task is to diagnose Conduction Disturbance (CD). The quantitative results demonstrate our method helps train a more robust predictive model. We use PGD attack to generate adversarial samples.

Conduction Disturbance	Clean AUROC	$\epsilon = 0.001$	$\epsilon = 0.002$
MF-Transformer-Raw	0.816	0.816	0.762
MF-Transformer-Oversampling	0.879	0.822	0.763
TaskAug Raghu et al. (2022)	0.831	–	–
MF-Transformer-Ours	**0.891**	**0.838**	**0.770**

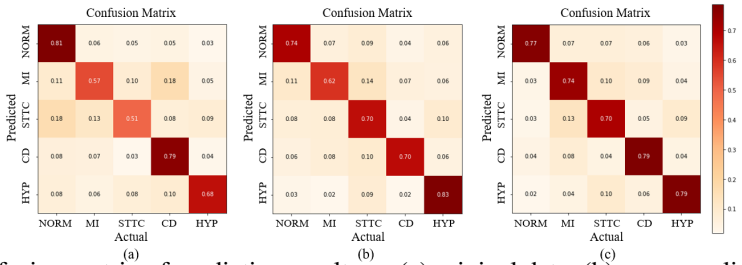

Figure 3: Confusion matrix of prediction results on (a) original data; (b) oversampling data; and (c) our augmented data.

5 Conclusions and Future Work

In this paper, we proposed a new method for electrocardiograms data augmentation. We augmented the minority category from the majority category with Wasserstein Geodesic Perturbation. We showed that after data augmentation, there are both accuracy and robustness improvements on the classification results over five ECG categories, which demonstrate the effectiveness of our method.

REFERENCES

U. Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Muhammad Adam, Arkadiusz Gertych, and Ru San Tan. A deep convolutional neural network model to classify heartbeats. *Computers in biology and medicine*, 89:389–396, 2017.

Salah Al-Zaiti, Lucas Besomi, Zeineb Bouzid, Ziad Faramand, Stephanie O. Frisch, Christian Martin-Gill, Richard E. Gregg, Samir F. Saba, Clifton Callaway, and Ervin Sejdić. Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. *Nature Communications*, 11, 2020.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015.

Behnam Behinaein, Anubha Bhatti, Dirk Rodenburg, Paul C. Hungler, and Ali Etemad. A transformer architecture for stress detection from ecg. *2021 International Symposium on Wearable Computers*, 2021.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

Chao Che, Peiliang Zhang, Min Zhu, Yue Qu, and Bo Jin. Constrained transformer network for ecg signal processing and arrhythmia classification. *BMC Medical Informatics and Decision Making*, 21, 2021.

Huangxun Chen, Chenyu Huang, Qianyi Huang, Qian Zhang, and Wei Wang. Ecgadv: Generating adversarial electrocardiogram to misguide arrhythmia classification system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3446–3453, 2020.

S. ClementVirgeniya and E. Ramaraj. A novel deep learning based gated recurrent unit with extreme learning machine for electrocardiogram (ecg) signal recognition. *Biomed. Signal Process. Control.*, 68:102779, 2021.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Ulysse Côté-Allard, Cheikh Latyr Fall, Alexandre Drouin, Alexandre Campeau-Lecours, Clément Gosselin, Kyrre Glette, François Laviolette, and Benoit Gosselin. Deep learning for electromyographic hand gesture signal classification using transfer learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27:760–771, 2019.

Nicolas Courty, Rémi Flamary, and Mélanie Ducoffe. Learning wasserstein embeddings. *arXiv preprint arXiv:1710.07457*, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

Chris Drummond and Robert C. Holte. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. 2003.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Adrien Corenflos, Nathalie T. H. Gayraud, Hicham Janati, Ievgen Redko, Antoine Rolet, Antony Schutz, Danica J. Sutherland, Romain Tavenard, Alexander Tong, Titouan Vayer, and Andreas Mueller. Pot: Python optimal transport. 2021.

Xiang Gao, Ripon K. Saha, Mukul R. Prasad, and Abhik Roychoudhury. Fuzz testing based data augmentation to improve robustness of deep neural networks. *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, pp. 1147–1158, 2020.

John R. Giudicessi, Matthew Schram, J. Martijn Bos, Conner Galloway, Jacqueline Baras Shreibati, Patrick W. Johnson, Rickey E. Carter, Levi W Disrud, Robert B Kleiman, Zachi I. Attia, Peter A. Noseworthy, Paul A. Friedman, David E. Albert, and Michael J. Ackerman. Artificial intelligence-enabled assessment of the heart rate corrected qt interval using a mobile electrocardiogram device. *Circulation*, 2021.

Kristjan Greenewald, Anming Gu, Mikhail Yurochkin, Justin Solomon, and Edward Chien. k-mixup regularization for deep learning via optimal transport. *arXiv preprint arXiv:2106.02933*, 2021.

Demi Guo, Yoon Kim, and Alexander M. Rush. Sequence-level mixed sample data augmentation. In *EMNLP*, 2020.

Hongyu Guo. Nonlinear mixup: Out-of-manifold data augmentation for text classification. In *AAAI*, 2020.

Xintian Han, Yuxuan Hu, Luca Foschini, Larry Chinitz, Lior Jankelson, and Rajesh Ranganath. Deep learning models for electrocardiograms are susceptible to adversarial attack. *Nature medicine*, 26(3):360–363, 2020.

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21:1263–1284, 2009.

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328, 2008.

Robert C. Holte, Liane Acker, and Bruce W. Porter. Concept learning and the problem of small disjuncts. In *IJCAI*, 1989.

Khondker Fariha Hossain, Sharif Amit Kamran, Xingjun Ma, and A. Tavakkoli. Ecg-atk-gan: Robustness against adversarial attacks on ecg using conditional generative adversarial networks. *ArXiv*, abs/2110.09983, 2021a.

Khondker Fariha Hossain, Sharif Amit Kamran, Alireza Tavakkoli, Lei Pan, Xingjun Ma, Sutharshan Rajasegarar, and Chandan Karmaker. Ecg-adv-gan: Detecting ecg adversarial examples with conditional generative adversarial networks. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 50–56. IEEE, 2021b.

Jongheon Jeong, Sejun Park, Minkyu Kim, Heung-Chang Lee, Do-Guk Kim, and Jinwoo Shin. Smoothmix: Training confidence-calibrated smoothed classifiers for certified robustness. *Advances in Neural Information Processing Systems*, 34, 2021.

Shaan Khurshid, Samuel N. Friedman, Christopher Reeder, Paolo Di Achille, Nathaniel Diamant, Pulkit Singh, Lia X. Harrington, Xin Wang, Mostafa A. Al-Alusi, Gopal Sarma, Andrea S. Foulkes, Patrick T. Ellinor, Christopher D Anderson, Jennifer E. Ho, Anthony A. Philippakis, Puneet Batra, and Steven A. Lubitz. Electrocardiogram-based deep learning and clinical risk factors to predict atrial fibrillation. *Circulation*, 2021.

Serkan Kiranyaz, Turker Ince, Ridha Hamila, and M. Gabbouj. Convolutional neural networks for patient-specific ecg classification. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2608–2611, 2015.

Linyi Li, Xiangyu Qi, Tao Xie, and Bo Li. Sok: Certified robustness for deep neural networks. *arXiv preprint arXiv:2009.04131*, 2020.

Linyi Li, Maurice Weber, Xiaojun Xu, Luka Rimanic, Bhavya Kailkhura, Tao Xie, Ce Zhang, and Bo Li. Tss: Transformation-specific smoothing for robustness certification. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 535–557, 2021.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Multimodal emotion recognition using deep canonical correlation analysis. *ArXiv*, abs/1908.05349, 2019.

Yamin Liu, Hanshuang Xie, Qineng Cao, Jiayi Yan, Fan Wu, Huaiyu Zhu, and Yun Pan. Multi-label classification of multi-lead ecg based on deep 1d convolutional neural networks with residual and attention mechanism. *2021 Computing in Cardiology (CinC)*, 48:1–4, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Harold Martin, Ulyana Morar, Walter Izquierdo, Mercedes Cabrerizo, Anastasio Cabrera, and Malek Adjouadi. Real-time frequency-independent single-lead and single-beat myocardial infarction detection. *Artificial intelligence in medicine*, 121:102179, 2021.

David Mease, Abraham J. Wyner, and Andreas Buja. Boosted classification trees and class probability/quantile estimation. *J. Mach. Learn. Res.*, 8:409–439, 2007.

George B. Moody and Roger G. Mark. The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20:45–50, 2001.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, and Stefano Soatto. Robustness of classifiers to universal perturbations: A geometric perspective. *arXiv preprint arXiv:1705.09554*, 2017.

Annamalai Natarajan, Yale Chang, Sara Mariani, Asif Rahman, Gregory Boverman, Shruti Gopal Vij, and Jonathan Rubin. A wide and deep transformer neural network for 12-lead ecg classification. *2020 Computing in Cardiology*, pp. 1–4, 2020.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. Ssmba: Self-supervised manifold based data augmentation for improving out-of-domain robustness. *ArXiv*, abs/2009.10195, 2020.

Naoki Nonaka and Jun Seita. In-depth benchmarking of deep neural network architectures for ecg diagnosis. In Ken Jung, Serena Yeung, Mark Sendak, Michael Sjoding, and Rajesh Ranganath (eds.), *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pp. 414–439. PMLR, 06–07 Aug 2021.

Jielin Qiu, Jiacheng Zhu, Michael Rosenberg, Emerson Liu, and D. Zhao. Optimal transport based data augmentation for heart disease diagnosis and prediction. *ArXiv*, abs/2202.00567, 2022.

Aniruddh Raghu, Divya Shanmugam, Eugene Pomerantsev, John Guttag, and Collin M Stultz. Data augmentation for electrocardiograms. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 282–310. PMLR, 07–08 Apr 2022. URL `https://proceedings.mlr.press/v174/raghu22a.html`.

Sushravya Raghunath, John M. Pfeifer, Alvaro E. Ulloa-Cerna, Arun Nemani, Tanner Carbonati, Linyuan Jing, David P. vanMaanen, Dustin N. Hartzel, Jeffery A. Ruhl, Braxton F. Lagerman, Daniel B. Rocha, Nathan J. Stoudt, Gargi Schneider, Kipp W. Johnson, Noah Zimmerman, Joseph B. Leader, H. Lester Kirchner, Christoph J. Griessenauer, Ashraf Hafez, Christopher W. Good, Brandon K. Fornwalt, and Christopher M. Haggerty. Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ecg and help identify those at risk of atrial fibrillation–related stroke. *Circulation*, 143:1287 – 1298, 2021.

Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021–1028, 2018.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Data augmentation can improve robustness. *ArXiv*, abs/2111.05328, 2021a.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A. Mann. Fixing data augmentation to improve adversarial robustness. *ArXiv*, abs/2103.01946, 2021b.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34, 2021c.

Divya Shanmugam, Davis Blalock, and John Guttag. Multiple instance learning for ecg risk stratification. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pp. 124–139. PMLR, 2019.

Sandra Śmigiel, Krzysztof Pałczyński, and Damian Ledziński. Ecg signal classification using deep learning techniques based on the ptb-xl dataset. *Entropy*, 23(9):1121, 2021.

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for eeg decoding. *ArXiv*, abs/2106.11170, 2021.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE Journal of Biomedical and Health Informatics*, 25:1519–1528, 2021.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

Jenny Venton. Investigating the robustness of deep learning to electrocardiogram noise. *2021 Computing in Cardiology (CinC)*, 48:1–4, 2021.

Jenny Venton, Peter M. Harris, Ashish Sundar, Nadia A. S. Smith, and Philip J. Aston. Robustness of convolutional neural networks to physiological ecg noise. *ArXiv*, abs/2108.01995, 2021.

Cédric Villani. Topics in optimal transportation. 2003.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

Patrick Wagner, Nils Strodthoff, R. Bousseljot, D. Kreiseler, F. Lunze, W. Samek, and T. Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data*, 7, 2020.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models, 2021.

Maurice Weber, Linyi Li, Boxin Wang, Zhikuan Zhao, Bo Li, and Ce Zhang. Certifying out-of-domain generalization for blackbox functions. *arXiv preprint arXiv:2202.01679*, 2022.

Kuba Weimann and Tim O. F. Conrad. Transfer learning for ecg classification. *Scientific Reports*, 11, 2021.

Fan Wu, Linyi Li, Zijian Huang, Yevgeniy Vorobeychik, Ding Zhao, and Bo Li. Crop: Certifying robust policies for reinforcement learning through functional smoothing. *arXiv preprint arXiv:2106.09292*, 2021a.

Fan Wu, Linyi Li, Huan Zhang, Bhavya Kailkhura, Krishnaram Kenthapadi, Ding Zhao, and Bo Li. Copa: Certifying robust policies for offline reinforcement learning against poisoning attacks. In *International Conference on Learning Representations*, 2021b.

Chulin Xie, Yunhui Long, Pin-Yu Chen, Krishnaram Kenthapadi, and Bo Li. Certified robustness for free in differentially private federated learning.

Chulin Xie, Minghao Chen, Pin-Yu Chen, and Bo Li. Crfl: Certifiably robust federated learning against backdoor attacks. In *International Conference on Machine Learning*, pp. 11372–11382. PMLR, 2021.

Genshen Yan, Shen Liang, Yanchun Zhang, and Fan Liu. Fusing transformer model with temporal features for ecg heartbeat classification. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 898–905, 2019.

Zhuolin Yang, Bo Li, Pin-Yu Chen, and Dawn Song. Characterizing audio adversarial examples using temporal dependency. *arXiv preprint arXiv:1809.10875*, 2018.

Zhuolin Yang, Linyi Li, Xiaojun Xu, Bhavya Kailkhura, Tao Xie, and Bo Li. On the certified robustness for ensemble models and beyond. *arXiv preprint arXiv:2107.10873*, 2021a.

Zhuolin Yang, Linyi Li, Xiaojun Xu, Shiliang Zuo, Qian Chen, Pan Zhou, Benjamin Rubinstein, Ce Zhang, and Bo Li. Trs: Transferability reduced ensemble via promoting gradient diversity and model smoothness. *Advances in Neural Information Processing Systems*, 34, 2021b.

Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2018.

Rongzhi Zhang, Yue Yu, and Chao Zhang. Seqmix: Augmenting active sequence labeling via sequence mixup. *ArXiv*, abs/2010.02322, 2020.

Jiacheng Zhu, Aritra Guha, Mengdi Xu, Yingchen Ma, Rayleigh Lei, Vincenzo Loffredo, Xuan-Long Nguyen, and Ding Zhao. Functional optimal transport: Mapping estimation and domain adaptation for functional data. *ArXiv*, abs/2102.03895, 2021.

A Data Extraction

We carried out the experiments on the PTB-XL dataset Wagner et al. (2020), which contains clinical 12-lead ECG signals of 10-second length. There are five conditions in total, which include Normal ECG (NORM), Myocardial Infarction (MI), ST/T Change (STTC), Conduction Disturbance (CD), and Hypertrophy (HYP). The waveform files are stored in WaveForm DataBase (WFDB) format with 16-bit precision at a resolution of 1μV/LSB and a sampling frequency of 100Hz.

First, we read the raw data by wfdb library[1] and perform Fast Fourier transform (fft) to process the time series data into the spectrum, which is shown in Fig. 4. Then we perform n-points window filtering to filter the noise and adopt notch processing to filter power frequency interference (noise frequency: 50Hz, quality factor: 30), where the filtered result after n-points window filtering and notch processing is shown in Fig. 5.

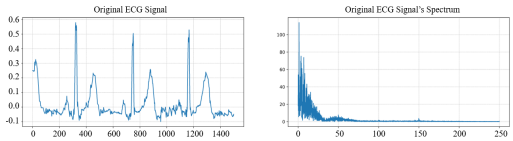

Figure 4: ECG data in time and spectrum.

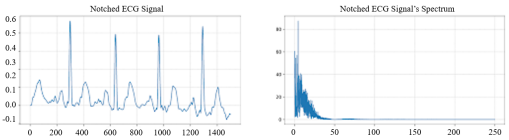

Figure 5: ECG filtered data.

We then detect the R peaks of each signal by ECG detectors[2], so the data can be sliced at the fixed-sized interval on both sides to obtain individual beats. The examples of detecting R peaks in ECG signals and divided pieces are shown in Fig. 6 and Fig. 7, respectively.

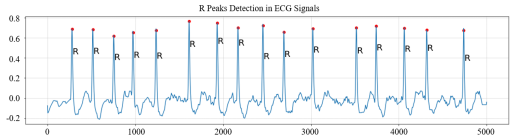

Figure 6: Detecting R peaks in the ECG signals.

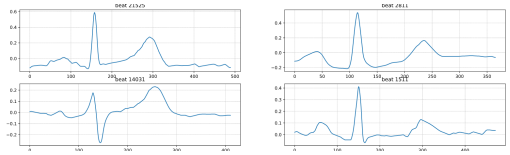

Figure 7: Extracted ECG pieces divided by R peaks.

To reduce the dimension of ECG features, we downsample the processed ECG signals to 50Hz. Then we extract more time domain features and frequency domain features to better represent the ECG signals. The time-domain features include: maximum, minimum, range, mean, median, mode, standard deviation, root mean square, mean square, k-order moment and skewness, kurtosis, kurtosis factor, waveform factor, pulse factor, margin factor. The frequency-domain features include: fft mean, fft variance, fft entropy, fft energy, fft skew, fft kurt, fft shape mean, fft shape std, fft shape skew, fft kurt, which are shown in Table 3.

There are five categories in total, including NORM, MI, STTC, CD, and HYP. In a balanced dataset, each category should occupy the same proportion. In the original dataset, the number of patients in the NORM category is much larger than the others. After dividing the ECG signals into individual beats, the portion of each category changed due to heartbeat variance among people. However, if we count the segmented ECG beats and compare different categories' data, the imbalance issue still exists, which is shown in Table 4. From Table 4, we can find out that NORM category and CD category is much larger than the other three categories, making the dataset unbalanced.

B Multi-Feature Transformer

For the classification model, we take advantage of the transformer encoder Vaswani et al. (2017), and proposed a Multi-Feature Transformer (MF-Transformer) model. The transformer is based on the attention mechanism Vaswani et al. (2017) and outperforms previous models in accuracy and performance. The original transformer model is composed of an encoder and a decoder. The encoder maps an input sequence into a latent representation, and the decoder uses the representation along

[1] https://pypi.org/project/wfdb/

[2] https://pypi.org/project/py-ecg-detectors/

Table 3: ECG signal statistical features in frequency domain.

Feature Symbol	Formula
Z_1	$\frac{1}{N}\sum_{k=1}^{N} F(k)$
Z_2	$\frac{1}{N-1}\sum_{k=1}^{N}\left(F(k)-Z_1\right)^2$
Z_3	$-1\times\sum_{k=1}^{N}\left(\frac{F(k)}{Z_1 N}\log_2\frac{F(k)}{Z_1 N}\right)$
Z_4	$\frac{1}{N}\sum_{k=1}^{N}(F(k))^2$
Z_5	$\frac{1}{N}\sum_{k=1}^{N}\left(\frac{F(k)-Z_1}{\sqrt{Z_2}}\right)^3$
Z_6	$\frac{1}{N}\sum_{k=1}^{N}\left(\frac{F(k)-Z_1}{\sqrt{Z_2}}\right)^4$
Z_7	$\frac{\sum_{k=1}^{N}(f(k)-F(k))}{\sum_{k=1}^{N}F(k)}$
Z_8	$\sqrt{\frac{\sum_{k=1}^{N}\left[(f(k)-Z_6)^2F(k)\right]}{\sum_{k=1}^{N}F(k)}}$
Z_9	$\frac{\sum_{k=1}^{N}\left[(f(k)-F(k))^3F(k)\right]}{\sum_{k=1}^{N}F(k)}$
Z_{10}	$\frac{\sum_{k=1}^{N}\left[(f(k)-F(k))^4F(k)\right]}{\sum_{k=1}^{N}F(k)}$

Table 4: Statistics of the data.

Category	Patients	Percentage	ECG beats	Percentage
NORM	9528	34.2%	28419	36.6%
MI	5486	19.7%	10959	14.1%
STTC	5250	18.9%	8906	11.5%
CD	4907	17.6%	20955	27.0%
HYP	2655	9.5%	8342	10.8%

with other inputs to generate a target sequence. Our model is mostly based on the encoder, since we aim at learning the representations of ECG features, instead of decoding it to another sequence.

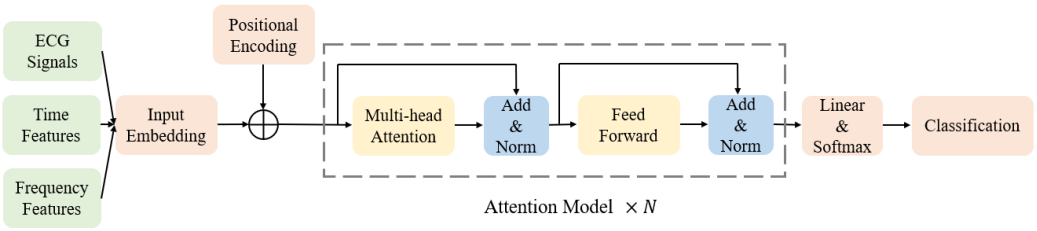

Figure 8: The architecture of the Multi-Feature Transformer model.

The input for the Multi-Feature Transformer is composed of three parts, including ECG raw features, time-domain features, and frequency domain features. The detailed feature pre-processing steps are introduced in Section **??**. First, we feed out the input into an embedding layer, which is a learned vector representation of each ECG feature by mapping each ECG feature to a vector with continuous values. Then we inject positional information into the embeddings by:

$$\begin{aligned} PE_{(pos,2i)} &= \sin\left(pos/10000^{2i/d_{\text{model}}}\right) \\ PE_{(pos,2i+1)} &= \cos\left(pos/10000^{2i/d_{\text{model}}}\right) \end{aligned} \tag{7}$$

The attention model contains two sub-modules, a multi-headed attention model and a fully connected network. The multi-headed attention computes the attention weights for the input and produces an output vector with encoded information on how each feature should attend to all other features in the sequence. There are residual connections around each of the two sub-layers followed by a layer normalization, where the residual connection means adding the multi-headed attention

output vector to the original positional input embedding, which helps the network train by allowing gradients to flow through the networks directly. Multi-headed attention applies a self-attention mechanism, where the input goes into three distinct fully connected layers to create the query, key, and value vectors. The output of the residual connection goes through a layer normalization.

In our model, our attention model contains $N = 5$ same layers, and each layer contains two sub-layers, which are a multi-head self-attention model and a fully connected feed-forward network. Residual connection and normalization are added in each sub-layer. So the output of the sub-layer can be expressed as:

$$\text{Output} = \text{LayerNorm}(x + (\text{SubLayer}(x))) \tag{8}$$

For the Multi-head self-attention module, the attention can be expressed as:

$$\text{attention} = \text{Attention}(Q, K, V) \tag{9}$$

where multi-head attention uses h different linear transformations to project query, key, and value, which are Q, K, and V, respectively, and finally concatenate different attention results:

$$\text{MultiHead(Q,K,V)} = \text{Concat}(head_1, ..., head_h)W^O \tag{10}$$

$$head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{11}$$

where the projections are parameter matrices:

$$\begin{aligned} W_i^Q &\in \mathbb{R}^{d_{\text{model}}\, d_k}, \qquad W_i^K \in \mathbb{R}^{d_{\text{model}}\, d_k} \\ W_i^V &\in \mathbb{R}^{d_{\text{model}}\, d_v}, \quad W_i^O \in \mathbb{R}^{hd_v \times d_{\text{model}}} \end{aligned} \tag{12}$$

where the computation of attention adopted scaled dot-product:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{13}$$

For the output, we use a 1D convolutional layer and softmax layer to calculate the final output.

B.1 Optimal Transport Based Data Augmentation

We use optimal transport to push forward samples from the distribution of a majority class to a minority class. We expect optimal transport to exploit global geometric information so that the synthetic samples match the real samples. In specific, we denote the data from a majority class to be $\mathbf{X}_s = \{x_{s,1}, ..., x_{s,n_s}\} \in \Omega_s$ and the minority class data to be $\mathbf{X}_t = \{x_{t,1}, ..., x_{t,n_t}\} \in \Omega_t$. We assume that they are subject to distributions $\mathbf{X}_s \sim \mu_s$ and $\mathbf{X}_t \sim \nu_t$, respectively, and we associate empirical measures to data samples:

$$\hat{\mu_s} = \sum_{i=1}^{n_s} p_{s,i}\delta_{x_{s,i}}\ , \hat{\nu_t} = \sum_{i=1}^{n_t} p_{t,i}\delta_{x_{t,i}}, \tag{14}$$

where δ_x is the Dirac function at location x and p_i are the probabilities masses associated to the samples. Solving the optimal transport objective give us the coupling:

$$\pi^* = \arg\min_{\pi \in \Pi} \sum_{i=1}^{n_s}\sum_{j=1}^{n_t} \pi_{i,j}C_{i,j}\ + \gamma H(\pi), \tag{15}$$

where $C_{i,j} = \|x_i - x_j\|_2^2$ is a cost matrix, γ is a coefficient, and $H(\pi) = \sum \pi_{i,j} \log \pi_{i,j}$ is the negative entropy regularization that enable us to employ the celebrated Sinkhorn algorithm Cuturi (2013). The solution to problem (15) actually express the barycentric mapping

$$\hat{x}_{s,i} = \arg\min_{x \in \Omega_t} \sum_{j=1}^{n_t} \pi^*(i,j)c(x, x_{t,j}), \tag{16}$$

where $x_{s,i}$ is source sample and $\hat{x}_{s,i}$ is the resulting mapped sample. When using l_2 norm as cost function, the barycenter has a convenient format that maps the source samples into the convex hull of target samples as $\hat{\mathbf{X}}_\mathbf{s} = n_s\pi^*\hat{\mathbf{X}}_\mathbf{t}$.

B.2 HEART DISEASE DETECTION

To evaluate our method, we used a MF-Transformer model as classifier. We trained the MF-Transformer model with (1) original PTB-XL data, (2) oversampling augmented data, and (3) our augmented data.

First, we trained the MF-Transformer model with the original PTB-XL data to obtain the baseline performance for different categories. Second, we used the oversampling strategy to augment the ECG signals for the minority categories, then we trained the MF-Transformer model from scratch to obtain the performance by oversampling data augmentation method. Third, we augmented the data with our data augmentation method, and trained the MF-Transformer model from scratch again to evaluate the performance of our method. Note that the augmented data is only used for training, and the testing set remains the same as for all the experiments, which only contain the real-world ECG signals to have a fair evaluation of the proposed method. The training and testing splitting strategy is the same as in Wagner et al. (2020); Strodthoff et al. (2021). The experiments are carried out on four Nvidia Tesla V100 GPUs. Tabele 2 and Fig. 3 show that compared with baseline results and oversampling results, our augmented method not only improved the classification accuracy of each category but also improved the average classification result from 71.80% (original) and 72.05% (oversampling) to 75.82% (ours). Each category's performance comes to be more balanced, showing the robustness improvement compared with the baseline results and oversampling results in Fig. 3(a) and Fig. 3 (b).

C MORE RELATED WORK

Traditional methods of data augmentation include sampling, cost-sensitive methods, kernel-based methods, active learning methods, and one-class learning or novelty detection methods (He & Garcia, 2009). Among them, sampling methods are mostly used, including random oversampling and undersampling, informed undersampling, synthetic sampling with data generation, adaptive synthetic sampling, sampling with data cleaning techniques, cluster-based sampling method, and integration of sampling and boosting. But traditional methods may introduce their own set of problematic consequences that can potentially hinder learning (Holte et al., 1989; Mease et al., 2007; Drummond & Holte, 2003), which can cause the classifier to miss important concepts pertaining to the majority class, or lead to overfitting (Mease et al., 2007; He & Garcia, 2009), making the classification performance on the unseen testing data generally far worse.

One motivation for data augmentation is to solve the data imbalance in ECG data. Martin et al. tried to use oversampling method to augment the imbalanced data (Martin et al., 2021). ClementVirgeniya & Ramaraj (2021) also addressed the ECG data imbalance problem, where instead of using synthetic models such as synthetic minority oversampling technique (SMOTE), SMOTEBoost, or DataBoostIM, they tried to feed the data into the adaptive synthetic (ADASYN) He et al. (2008) based sampling model, which utilized a weighted distribution for different minority class samples depending upon the learning stages of difficulty. Liu et al. (2021) augmented the ECG data by using band-pass filter, noise addition, time-frequency transform and data selection. The methods above showed that balanced dataset performance is superior than unbalanced one.

Optimal Transport (OT) is a field of mathematics that studies the geometry of probability spaces (Villani, 2003). The theoretical importance of OT is that it defines the Wasserstein metric between probability distributions. It reveals a canonical geometric structure with rich properties to be exploited. The earliest contribution to OT originated from Monge in the eighteenth century. Kantorovich rediscovered it under a different formalism, namely the Linear Programming formulation of OT. With the development of scalable solvers, OT is widely applied to many real-world problems (Zhu et al., 2021; Flamary et al., 2021).

With the development in machine learning, many models have been applied to ECG disease detection Kiranyaz et al. (2015); Nonaka & Seita (2021); Khurshid et al. (2021); Raghunath et al. (2021); Giudicessi et al. (2021); Strodthoff et al. (2021); Qiu et al. (2022). Al-Zaiti et al. (2020) predicted acute myocardial ischemia in patients with chest pain with a fusion voting method. Acharya et al. proposed a nine-layer deep convolutional neural network (CNN) to classify heartbeats in the MIT-BIH Arrhythmia database (Acharya et al., 2017; Moody & Mark, 2001). Shanmugam et al. (2019) estimate a patient's risk of cardiovascular death after an acute coronary syndrome by a multiple instance learning framework. Recently,Ravanelli & Bengio (2018) proposed models based on SincNet and used entropy-based features for cardiovascular diseases classification Śmigiel et al. (2021).

ECG signal can be considered as one type of sequential data, and Seq2seq models (Sutskever et al., 2014) are widely used in time series tasks. Since the attention mechanism was proposed (Bahdanau et al., 2015), the Seq2seq model with attention has been improved in various tasks, which outperformed previous methods. Then Transformer model Vaswani et al. (2017) was proposed to solve the problem in the Seq2Seq model, replacing Long Short-Term Memory (LSTM) models with an attention structure, which achieved better results in translation tasks. The transformer model has also recently been adopted in several ECG applications, i.e., arrhythmia classification, abnormalities detection, stress detection, etc (Yan et al., 2019; Che et al., 2021; Natarajan et al., 2020; Behinaein et al., 2021; Song et al., 2021; Weimann & Conrad, 2021). But those models take only ECG temporal features as input and haven't considered the frequency domain features. To take advantage of multiple features across time and frequency domains, we proposed a Multi-Feature Transformer as our classification model to predict the heart diseases with 12-lead ECG signals.