GeoECG: Data Augmentation via Wasserstein Geodesic Perturbation for Robust Electrocardiogram Prediction

Jiacheng Zhu* JZHU4@ANDREW.CMU.EDU
Department of Mechanical Engineering, Carnegie Mellon University
Pittsburgh, PA 15213, USA

Jielin Qiu* JIELINQ@ANDREW.CMU.EDU
Computer Science Department, Carnegie Mellon University
Pittsburgh, PA 15213, USA

Zhuolin Yang ZHUOLIN5@ILLINOIS.EDU
Computer Science Department, University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

Douglas Weber DOUGWEBER@CMU.EDU
Department of Mechanical Engineering, Carnegie Mellon University
Pittsburgh, PA 15213, USA

Michael A. Rosenberg MICHAEL.A.ROSENBERG@CUANSCHUTZ.EDU
University of Colorado Denver - Anschutz Medical Campus
Aurora, CO 80045, USA

Emerson Liu EMERSONLIU@MSN.COM
Allegheny Health Network
Pittsburgh, PA 15212, USA

Bo Li LBO@ILLINOIS.EDU
Computer Science Department, University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

Ding Zhao DINGZHAO@CMU.EDU
Department of Mechanical Engineering, Carnegie Mellon University
Pittsburgh, PA 15213, USA

* authors of equal contribution

Abstract

There has been an increased interest in applying deep neural networks to automatically interpret and analyze the 12-lead electrocardiogram (ECG). The current paradigms with machine learning methods are often limited by the amount of labeled data. This phenomenon is particularly problematic for clinically-relevant data, where labeling at scale can be time-consuming and costly in terms of the specialized expertise and human effort required. Moreover, deep learning classifiers may be vulnerable to adversarial examples and perturbations, which could have catastrophic consequences, for example, when applied in the context of medical treatment, clinical trials, or insurance claims. In this paper, we propose a physiologically-inspired data augmentation method to improve performance and increase the robustness of heart disease detection based on ECG signals. We obtain augmented samples by perturbing the data distribution towards other classes along the geodesic in Wasserstein space. To better utilize domain-specific knowledge, we design a ground metric that recognizes the difference between ECG signals based on physiologically determined 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

Heart and cardiovascular diseases are the leading global cause of death, with 80% of cardiovascular disease-related deaths due to heart attacks and strokes. The 12-lead ECG can be considered as the foundation of cardiology and electrophysiology. It provides unique information about the structure and electrical activity of the heart as well as systemic conditions, through changes in timing and morphology of the recorded waveforms. Consequently, the clinical 12-lead ECG when correctly interpreted, remains a primary tool for detecting cardiac abnormalities and screening at-risk populations for heart-related issues.

Accurate ECG interpretations of acute cardiac conditions are critical for timely, efficient, and cost-effective interventions. Consequently, achievement of reliable machine assisted ECG interpretations could significantly impact patient outcomes (Zhu et al., 2022). With the development of machine learning and deep learning methods, it may also 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 diagnostic performance of developed algorithms is not always robust. Also, it has been shown that deep learning models for ECG data could be susceptible to adversarial attacks. (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). 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., 2021). However, different types of data usually contain critical domain-specific properties.

While people have effectively applied different neural network architectures to ECG classification problems, it has become an increasing concern that these neural networks are susceptible to adversarial examples (Han et al., 2020). Several studies (Raghu et al., 2022; Nonaka and Seita, 2020) have explored data augmentation techniques for ECG datasets. Nevertheless, unlike other fields such as computer vision (Zhao et al., 2020) or NLP (Morris et al., 2020), the effect of data augmentation on ECG deep learning robustness is less explored.

In this paper, we propose a data augmentation method from a probability and geometric perspective. Following the notion of optimal mass transportation theories (Villani, 2009), we perturb the data distributions along the geodesic in a Wasserstein space. Specifically, the ground metric of this Wasserstein space is computed via comparing the geometry of ECG signals, which exploits the cardiovascular properties. In summary, our contribution is threefold:

1. Our proposed data augmentation method augments samples by perturbing an empirical distribution towards samples of other classes. This data augmentation scheme can preserve the local structure on the data manifold.

2. To perform the computation of Wasserstein barycenters, we propose a similarity metric for ECG signals by comparing their shapes, where we consider each beat of an ECG as a *continuous probability* and compute the corresponding Wasserstein distance.

3. We validate our method on the PTB-XL dataset, which covers a variety of conditions, including Myocardial Infarction (MI), ST/T Change (STTC), Conduction Disturbance (CD), and Hypertrophy (HYP), collected from subjects of different ages and gender. We compare our methods with a list of baseline methods in terms of the standard prediction performances and robust prediction under adversarial attacks.

Generalizable Insights about Machine Learning in the Context of Healthcare

ECG signals can be treated as continuous sequential data and directly fed into deep learning models, since some neural networks architectures, such as ResNet, can effectively capture information from raw ECG signals. However, our study emphasizes the potential importance of utilizing physiologically informed features that are intrinsically encoded in the structure or shape of the ECG waveforms. This statement is motivated by the following: (1) Specific components of ECG signals are generated from different parts of the cardiac cycle and represent different physiology (2) The periodicity (or absence) of expected ECG waveforms contains valuable information beyond that contained within the waveforms themselves. (3) Human experts specify ECG categories predominantly based on coarser visual features (shape and morphology) that can reflect a broad variety of structural or conduction abnormalities. (4) The advantages of incorporating some prior knowledge into models is supported by the fact that the leading method of the 2020 PhysioNet challenge only used handcrafted features rather than raw signals.

Hence, we suggest exploring and developing algorithms based on electrocardiograms' properties and physiological features. While decompiling ECG signals into individual wave components (P, QRS, T waves) is challenging, our method of comparing ECG beats with respect to their underlying geometry offers a principle approach to discriminating ECG signals.

2. Related Work

ECG Robustness The 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 differing signal-to-noise ratios (SNRs) of noise on ECG classification performance.

Deep learning in ECG Deep learning approaches have been rapidly adopted across a wide range of fields due to their accuracy and flexibility but require large labeled training sets. With the development in machine learning, many models have been applied to ECG disease detection (Kiranyaz et al., 2015; Nonaka and Seita, 2021; Khurshid et al., 2021; Raghunath et al., 2021; Giudicessi et al., 2021; Strodthoff et al., 2021). Al-Zaiti et al. predicted acute myocardial ischemia in patients with chest pain with a fusion voting method (Al-Zaiti et al., 2020). 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 and Mark, 2001). Shanmugam et al. estimate a patient's risk of cardiovascular death after an acute coronary syndrome by a multiple instance learning framework (Shanmugam et al., 2019). Recently, Smigiel et al. proposed models based on SincNet (Ravanelli and Bengio, 2018) and used entropy-based features for cardiovascular diseases classification (Śmigiel et al., 2021). 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 and Conrad, 2021).

Data augmentation for ECG The data augmentation task has also been explored for ECG applications in the previous studies. Martin et al. (2021) tried to use oversampling method to augment the imbalanced data. ClementVirgeniya and Ramaraj (2021) 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, instead of using synthetic models such as synthetic minority oversampling technique (SMOTE). Liu et al. (2021) augmented the ECG data by using a band-pass filter, noise addition, time-frequency transformation, and data selection. Data augmentation is also a good method to deal with imbalanced ECG dataset (Qiu et al., 2022a). Recently, a task-dependent learnable data augmentation policy (Raghu et al., 2022) has been developed for 12-lead ECG detection. This study showed that data augmentation techniques are not always helpful.

Data augmentation & robustness in ML: A promising way to enable robust learning is to provide adversarially perturbed samples with data augmentation. It has already been shown that data augmentation (Rebuffi et al., 2021c; Volpi et al., 2018) or more training data (Carmon et al., 2019; Deng et al., 2021) can improve the performance and increase the robustness of deep learning models.

Zhang et al. (2018) proposed Mixup, an effective model regularizer for data augmentation that encourages linear interpolation in-between training examples, which has been applied in sequential data. 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 a 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. Perhaps there is one recent work that is conceptually close to our study where a k-mixup data augmentation (Greenewald et al., 2021), guided by Optimal Transport (OT), is proposed to improve the generalization and robustness of neural networks. This method uses optimal coupling to interpolate for vicinal data samples that respect local structures. Our method also enjoys this benefit

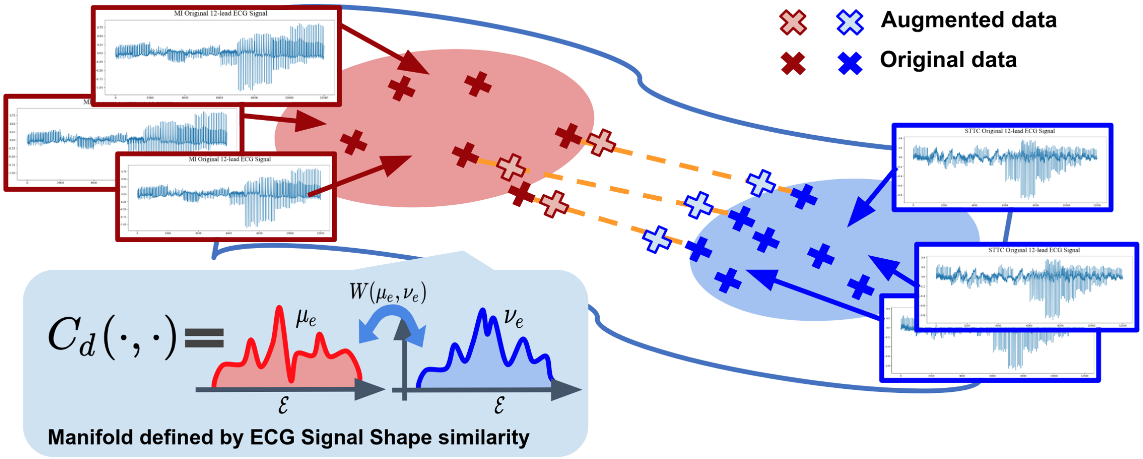
Augmented data
Original data
MI Original 12-lead ECG Signal
MI Original 12-lead ECG Signal
STTC Original 12-lead ECG Signal
STTC Original 12-lead ECG Signal
$C_d(\cdot,\cdot)=$
μ_e
$W(\mu_e,\nu_e)$
ν_e
$\mathcal{E}$
$\mathcal{E}$
Manifold defined by ECG Signal Shape similarity

Figure 1: Our data augmentation creates perturbed samples toward vicinal other-class samples. The perturbation lies on the geodesic connecting two distributions on a Wasserstein space, whose ground cost metric is computed via another level Wasserstein distance that compares the geometric shape of ECG signals.

since the Wasserstein barycenter also exploits local distribution structures. However, this study uses $l2$ cost as the ground metric, which could be ineffective when dealing with high-dimensional data. On the contrary, our study utilizes a ground metric that compares ECG signals according to their cardiovascular characteristics.

3. Methods

In this work, we focus on the model's performance on adversarial examples, which are generated by adding imperceptible noises on clean inputs to mislead ML models' predictions through well-designed attack algorithms (Szegedy et al., 2013; Goodfellow et al., 2014; Eykholt et al., 2018). Given such malicious scenarios and ML security considerations, robust deep learning has been studied extensively (Salman et al., 2019; Zhu and Li, 2022), to develop effective learning algorithms to build robust models.

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 shifts in data distribution. A common way to describe this procedure is through the framework of distributional robust optimization (Weber et al., 2022). Specifically, 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 turn 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 aims at finding the worst-case optimal predictor h_θ^* when the data distribution P is perturbed towards some distribution $\mathcal{U}_P$.

In this work, we follow the distribution perturbing adversary frameworks (Sinha et al., 2017; Mehrabi et al., 2021) wherein the adversarial distributions can be viewed as the neighbor distribution of clean data distribution, characterized by certain distribution distance metrics (e.g., Wasserstein distance (Villani, 2009)). It is challenging to access this adversarial distribution explicitly. Thus, inspired by recent studies (Carmon et al., 2019; Zhai et al., 2019; Dan et al., 2020), we make further assumptions for the distribution of the data by writing out the joint data distribution $P(X, Y)$ as the product of conditional distributions $P(X, Y) = P(X|Y)P(Y)$. Since we focus on the multi-classification task, we denote $P_k(X) = P(X|Y_k)$ as the data distribution of one class k. During the data augmentation, we aim to perturb the class i's data distribution $P_i(X)$ towards class j's distribution $P_j(X), i \neq j$ since we believe the data samples lying on the geodesic can be served as adversarial samples.

We can illustrate our intuitions as follows: **(1)** Instead of the datapoint-specific adversarial perturbations that are aimed to attack one specific sample, the directed augmented data distribution can be considered as universal perturbations (Moosavi-Dezfooli et al., 2017) that cause label change for a set of samples from the perturbed distribution $\mathcal{U}_P$. **(2)** Such perturbation matches the global manifold structure of the dataset (Greenewald et al., 2021), therefore promoting a smoother decision boundary. **(3)** It is shown in Wei et al. (2020) that this augmentation strategy improves the expansion of the neighborhood of class-conditional distributions. More significantly, this formulation allows us to employ the results from OT theories (Villani, 2009) and Wasserstein Barycenter (Agueh and Carlier, 2011) thus firmly estimating the perturbed distribution $\mathcal{U}_P$.

3.2. Data Augmentation by Perturbation on the Geodesic

Let $\mathcal{X}$ be an arbitrary space. Assume $d(\cdot, \cdot) : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}^+$ is the ground metric cost function. The well-known Wasserstein distance originated from the Optimal Transport (OT) problem which aims at finding an optimal coupling π that minimizes the transportation cost.

Definition 1 *(Wasserstein Distances). For $p \in [1, \infty]$ and probability measures μ and $\nu \in \mathcal{M}(\mathcal{X})$. The $p-$Wasserstein distance between them is defined as*

$$W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi} \int_{\mathcal{X} \times \mathcal{X}} d^p(x, y) d\pi(x, y) \right)^{1/p}, \ (x, y) \in \mathcal{X} \times \mathcal{X} \tag{2}$$

where Π is the set of all probability measures on $\mathcal{X} \times \mathcal{X}$.

Considering the path of distributions (a geodesic) p_t that interpolates between two distributions μ and ν, one of the most intriguing properties of this interpolation is that it will preserve the basic structure of μ and ν. In other words, such perturbation can be viewed as an optimal transport map that pushes forward μ along the geodesic that connects μ and ν.

Definition 2 *(Geodesics in Wasserstein space). Let μ and ν be two distributions. Consider a map $m : [0, 1] \mapsto \mathcal{M}(\mathcal{X})$ taking $[0, 1]$ to the set of distributions, such that $m(0) = \mu$ and*

$m(1) = \nu$, where $\mathcal{M}(\mathcal{X})$ is the set of Borel measures on $\mathcal{X}$. Thus $(p_\alpha : 0 \leq \alpha \leq 1)$ is a path connecting μ and ν, where $p_\alpha = m(\alpha)$. The length of m — denoted by $L(m)$ — is the supremum of $\sum_{i=1}^{K} W\left(m\left(\alpha_{i-1}\right), m\left(\alpha_i\right)\right)$ over all m and all $0 = \alpha_1 < \cdots < \alpha_K = 1$. Therefore, there exists such a path m such that $L(m) = W(\mu, \nu)$ and $(p_\alpha : 0 \leq \alpha \leq 1)$ is the geodesic connecting μ and ν.

The definition of the geodesic in Wasserstein space provides us a roadmap to obtain the perturbed distributions, as it boils down to the Wasserstein Barycenter problem.

Definition 3 *(Wasserstein Barycenter). The Wasserstein barycenter of a set of measures $\{\nu_1, ..., \nu_N\}$ in a probability space $\mathbb{P} \subset \mathcal{M}(\mathcal{X})$ is a minimizer of objective f_{wb} over $\mathbb{P}$, where*

$$f_{wb}(\mu) := \frac{1}{N} \sum_{i=1}^{N} \alpha_i W(\mu, \nu_i), \tag{3}$$

where α_i are the weights such that $\sum \alpha_i = 1$ and $\alpha_i > 0$.

If we are using uniform weights and consider all the distributions, the barycenter is the Fréchet mean, or the Wasserstein population mean (Bigot et al., 2017). Also, it is known that when we have only two samples, the barycenter corresponds to the interpolation between two distributions along the geodesic.

In this case, given class-conditional data distributions P_i and P_j, the perturbed augmentation which interpolates along the geodesic can be obtained via

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

Then the augmented samples can be obtained by sampling $(\tilde{x}_i, y_i) \sim \tilde{P}_{ij}$. Later on, we will provide an algorithmic derivation of this augmentation procedure for discrete data samples.

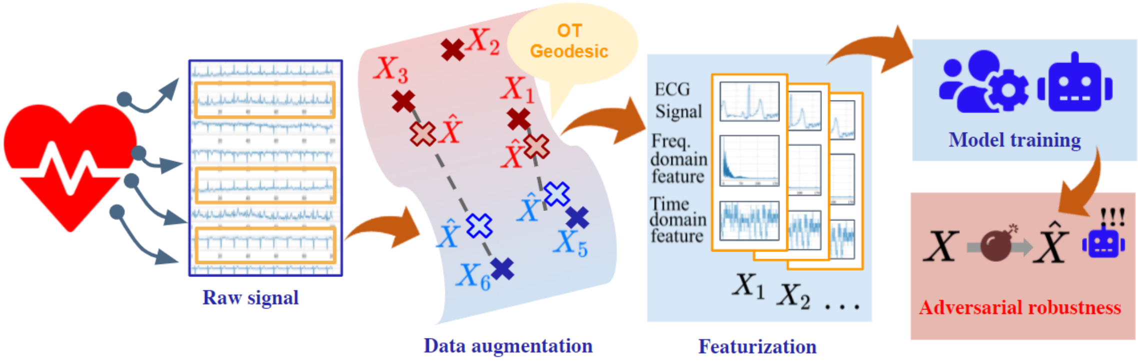

Figure 2: The semantic representation of our pipeline.

4. Algorithm & Computation

4.1. Computational Optimal Transport

In practice, we only observe discrete training samples that represents empirical distribution of P_i and P_j. Consider $\mathbf{X}_i = \{\mathbf{x}_l^i\}_{l=1}^{n_i}$ and $\mathbf{X}_j = \{\mathbf{x}_l^j\}_{l=1}^{n_j}$ are two set of features from class i and j respectively. The empirical distributions are written as $\hat{P}_i = \sum_{l=1}^{n_i} p_l^i \delta_{x_l^i}$ and $\hat{P}_j = \sum_{l=1}^{n_j} p_l^j \delta_{x_l^j}$ where δ_x is the Dirac function at location $x \in \Omega$, p_l^i and p_l^j are probability mass associated to the sample. Then the Wasserstein distance, or equation (2), between empirical measures $\hat{P}_i$ and $\hat{P}_j$ becomes

$$W(\hat{P}_i, \hat{P}_j) = \inf_{\pi \in \hat{\Pi_{ij}}} \sum_{l=1,k=1}^{n_i,n_j} c(\mathbf{x}_l^i, \mathbf{x}_k^j)\pi_{l,k}, \tag{5}$$

where $\hat{\Pi_{ij}} := \{\pi \in (\mathbb{R}^+)^{n_i \times n_j} | \pi \mathbf{1}_{n_j} = \mathbf{1}_{n_i}/n_i, \pi^\top \mathbf{1}_{n_i} = \mathbf{1}_{n_j}/n_j\}$ with $\mathbf{1}_n$ a length n vector of ones. $c(x, y)$ is the ground cost function that specifies the actually cost to transport the mass, or probability measure, from position x to y. Most studies use l_2 norm as the ground metric as there are a lot of desirable properties. However, here we emphasis that it is not appropriate to compare ECG signals with l_2 metrics.

4.2. A Physiology Inspired Metric for ECG

In practice, the accurate decomposition (Kanjilal et al., 1997) of ECG is a crucial step in providing medical diagnosis and services. For example, ventricular heart rate (Kundu et al., 2000) is the most common information that is extracted by measuring the time interval between two successive R peaks. While in most computer vision tasks, it is hard to describe features explicitly, informative characteristics are defined the wave features of the ECG signal, as illustrated in Fig. 3. A great deal of work (Zhong et al., 2020; Rasti-Meymandi and Ghaffari, 2021) focused on extracting or decomposing the wave components from ECG signals. However, it is still challenging.

In this work, we propose to directly compare the shape of two ECGs rather than parsing the ECG into the P wave, QRS, and T wave since it is challenging to process the noisy signals. Specifically, we first (1) treat them as probability densities, and then (2) compute a Wasserstein distance between these two densities. Formally, consider two individual ECG beat signals as two density function of time $\mu_e = \mu_e(t)$ and $\nu_e = \nu_e(t)$. The Wasserstein distance is obtained via:

$$W_e(\mu_e, \nu_e) := \inf_{\pi_e \in \Pi_e} \int_{[0,1]\times[0,1]} \|x - y\|_2^2 d\pi_e(x, y), \tag{6}$$

where Π_e is the joint distribution which has marginals as μ_e and ν_e. Therefore, now we have a reasonable metric that measures the pairwise similarity between ECG signals $C_d(\cdot, \cdot) = W_e(\cdot, \cdot)$, which can serve as the ground metric in the computation of the Wasserstein barycenter data augmentation procedure.

Computation concerns: batch OT and entropic OT Discrete optimal transport involves a linear program that has an $O(n^3)$ complexity. Our framework requires the computation of optimal transport in two levels: (1) Use Wasserstein distance to obtain the pairwise similarity of ECG signals. (2) Use Wasserstein Barycenter, which also computes Wasserstein distances, to interpolate between two sets of ECG signal samples from different conditions. Hence, the potential computation issues can not be ignored.

First of all, we adopted the celebrated entropic optimal transport (Cuturi, 2013) and used the Sinkhorn algorithm to solve for OT objectives and Barycenters (Janati et al., 2020).

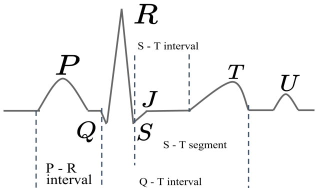

Figure 3: ECG components.

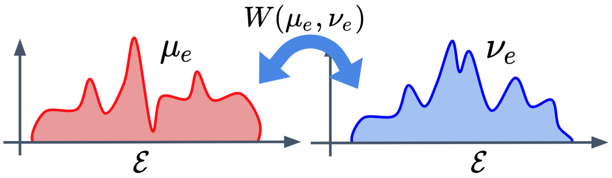

Figure 4: Treat ECG as continuous densities.

The Sinkhorn algorithm has a $O(n \log n)$ complexity, thus it can ease the computation burden. In addition, the pairwise Wasserstein distance of ECG signals can be precomputed and stored. Last but not least, we follow the concept of minibatch optimal transport (Fatras et al., 2021) where we sample a batch of ECG samples from each condition during the data augmentation procedure. Whereas minibatch OT could lead to non-optimal couplings, our experimental results have demonstrated that our data augmentation is still satisfactory.

4.3. Backbone Model

Nonaka and Seita (2021) used raw ECG signal as input, however, it is shown that the optimal predictive performance can be achieved by transformers trained with hand-crafted features (Natarajan et al., 2020). So 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 on many tasks (Xu et al., 2022; Qiu et al., 2022b). 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 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.

As shown in Fig. 5, the input for the Multi-Feature Transformer is composed of three parts, including ECG raw features, time-domain features, and frequency domain features. First, we feed out the input into an embedding layer and then inject positional information into the embeddings. In our model, the attention model contains $N = 5$ same layers, and each layer contains two sub-layers: a multi-head self-attention model and a fully connected feed-forward network. Residual connection and normalization are added in each sub-layer. We use a 1D convolutional and softmax layers for the output to calculate the final output. More details of the MF-Transformer model is introduced in Appendix A.

5. Cohort

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, including 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.

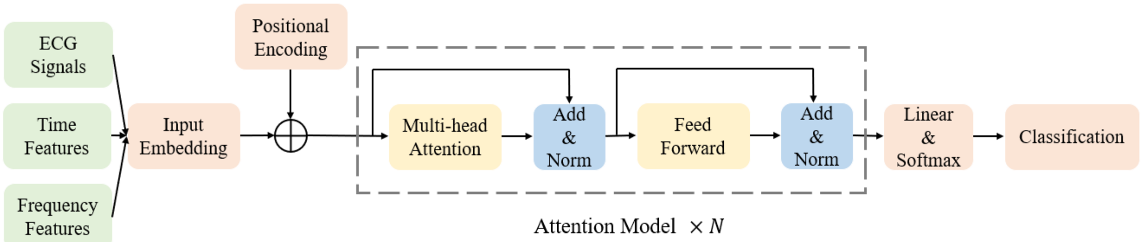

Figure 5: The architecture of the Multi-Feature Transformer model.

5.1. Signal Pre-processing

First, the raw ECG signals are processed by the wfdb library[1] and Fast Fourier transform (fft) to process the time series data into the spectrum, which is shown in Fig. 6. Then we perform n-points window filtering to filter the noise within the original ECG signals and adopt notch processing to filter power frequency interference (noise frequency: 50Hz, quality factor: 30). An example of the filtered ECG signal result after n-points window filtering and notch processing is shown in Fig. 7.

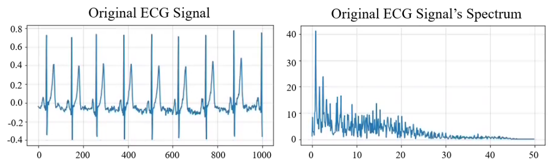

Figure 6: ECG data in the format of time series and spectrum.

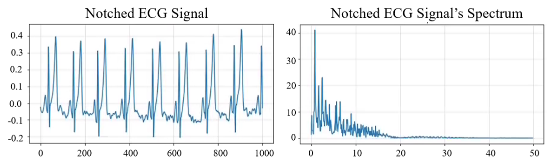

Figure 7: Filtered ECG data in the format of time series and spectrum.

Then we perform ECG segmentation by dividing the 10-second ECG signals into individual ECG beats. We first detect the R peaks of each signal by ECG detectors[2], and then slice the signal at a fixed-sized interval on both sides of the R peaks to obtain individual beats. Examples of R peak detection results and segmented ECG beats are shown in Fig. 8 and Fig. 9, respectively.

5.2. Feature Extraction

Instead of directly using the time-series signals, we extract time domain and frequency domain features to better represent 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, and 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, where the function of each component is shown in Table 1.

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

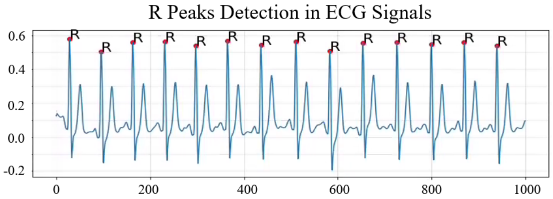

Figure 8: Detecting R peaks in the ECG signals.

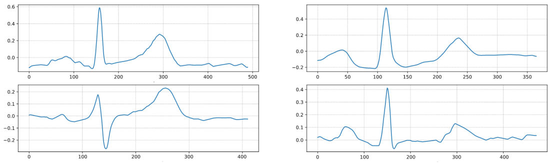

Figure 9: Extracted ECG beats divided by R peaks.

Table 1: ECG statistical features in frequency domain.

Feature Symbol	Formula	Feature Symbol	Formula
Z_1	$\frac{1}{N}\sum_{k=1}^{N} F(k)$	Z_6	$\frac{1}{N}\sum_{k=1}^{N}\left(\frac{F(k)-Z_1}{\sqrt{Z_2}}\right)^4$
Z_2	$\frac{1}{N-1}\sum_{k=1}^{N}\left(F(k)-Z_1\right)^2$	Z_7	$\frac{\sum_{k=1}^{N}(f(k)-F(k))}{\sum_{k=1}^{N} F(k)}$
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_8	$\sqrt{\frac{\sum_{k=1}^{N}\left[(f(k)-Z_6)^2 F(k)\right]}{\sum_{k=1}^{N} F(k)}}$
Z_4	$\frac{1}{N}\sum_{k=1}^{N}(F(k))^2$	Z_9	$\frac{\sum_{k=1}^{N}\left[(f(k)-F(k))^3 F(k)\right]}{\sum_{k=1}^{N} F(k)}$
Z_5	$\frac{1}{N}\sum_{k=1}^{N}\left(\frac{F(k)-Z_1}{\sqrt{Z_2}}\right)^3$	Z_{10}	$\frac{\sum_{k=1}^{N}\left[(f(k)-F(k))^4 F(k)\right]}{\sum_{k=1}^{N} F(k)}$

After processing the ECG signals, we analyzed the statistics of the processed ECG data, and the result is shown in Table 2, where there are five categories in total, including NORM, MI, STTC, CD, and HYP.

Table 2: Statistics of the processed ECG 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%

6. Experiments

6.1. Experimental Setup

We use the MF-Transformer model as our classifier, where the input contains three parts: the ECG signals, the time domain features, and the frequency domain feature. To reduce the dimension of ECG signals for the convenience of computation, we downsample the processed ECG signals to 50Hz. We computed the ECG features in Section 5.2 for each ECG beat for all 12 leads, and concatenated them with the downsampled and de-noised ECG signals. The dimension of the final features vector of each ECG beat is 864, where the dimensions for the ECG signals, time-domain features, and frequency domain features are 600, 156, and 108, respectively. Our experiments are carried out on two NVIDIA RTX A6000 GPUs.

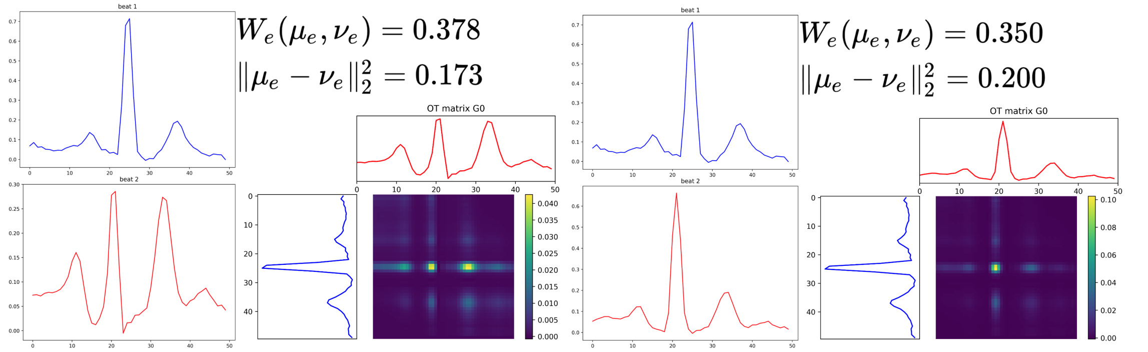

Figure 10: Difference between W-distance and l2 distance. We overlaid two ECG signals from different conditions on the left side, while on the right side, we put two ECG signals from the same condition. The Wasserstein distance correctly describes the similarity, but the l_2 norm indicates a large distance between two signals comes from the same condition but has a slight time shift.

6.2. Data Augmentation by Wasserstein Geodesic Perturbation

Our data augmentation strategy through Wasserstein Geodesic Perturbation aims at improving the robustness of heart disease diagnosis. In specific, (1) We use NORM individual beats as the source and transport the samples from the NORM into each other minor categories. (2) In the augmentation procedure, we randomly sample a batch of ECG signals from both the source and target categories and then use formulation in Section 3.2 to get the barycentric mapping samples. The label of augmented samples is set to be the target categories. (3) We mix the original data and augmented data together as input data for the MF-Transformer.

Examples of augmented data are shown in Fig. 11. The quality of the augmented data is also confirmed to preserve the semi-periodic nature. The augmentation results of each lead fit well with the ECG pattern compared with the original ECG signals.

6.3. Evaluation of Heart Disease Detection

We used the MF-Transformer model as the classifier to evaluate our methods to detect the heart conditions based on the ECG 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 different data augmentation strategies to augment the ECG signals for the minority categories, then trained the MF-Transformer model from scratch to obtain the performance by different data augmentation methods. Third, we augmented the ECG data with our data augmentation method and trained the MF-Transformer model from scratch again to evaluate the performance of our method. The augmented data is only used for training, the testing set remains the same as for all the experiments, which only contain the

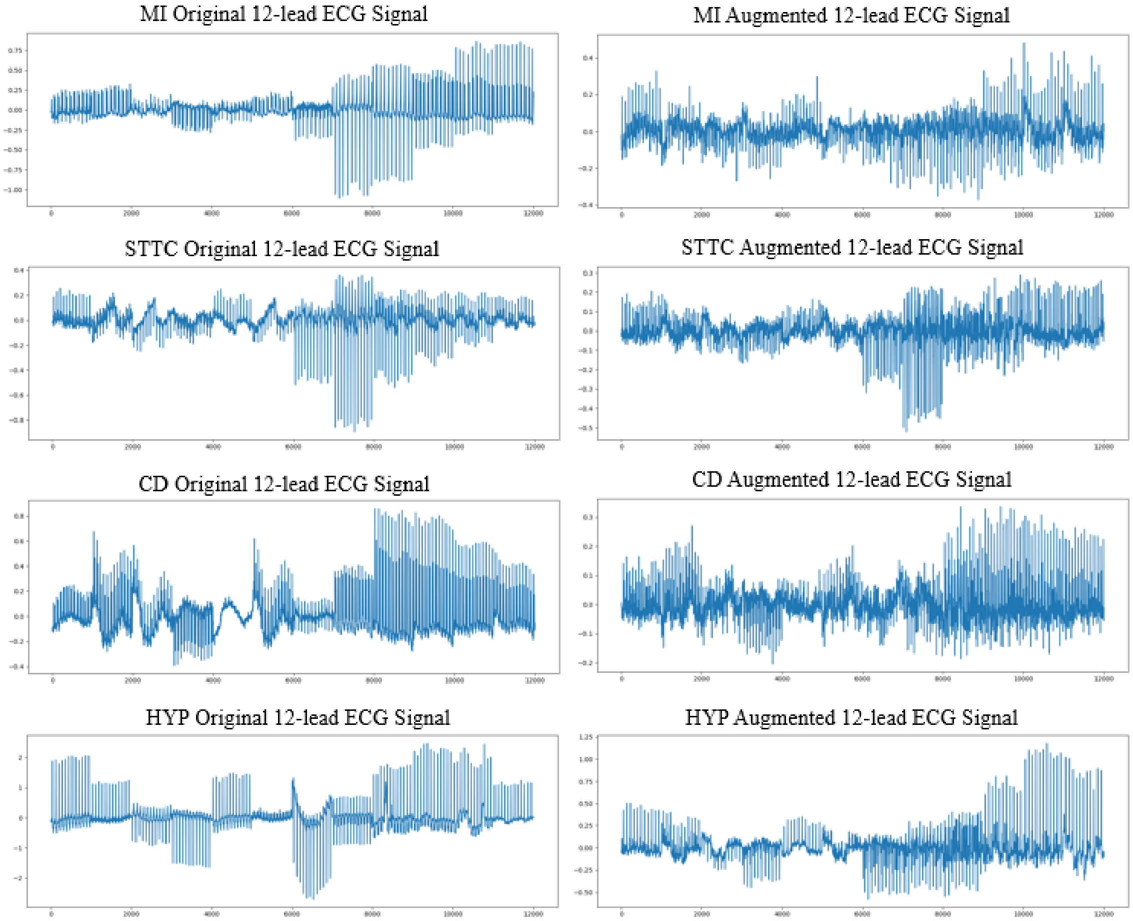

Figure 11: Example comparisons of 10-second 12-lead ECG signals within different conditions. Left column: original signals; Right column: augmented signals.

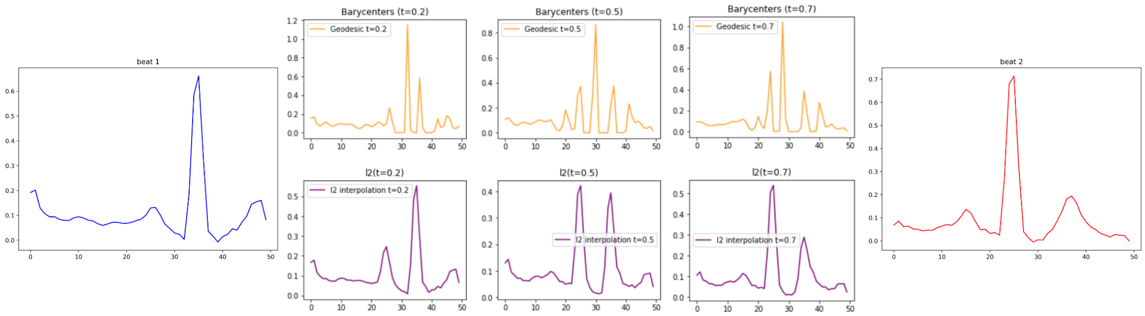

Figure 12: The difference between perturbation along the Wasserstein geodesic and on l_2 space. The geodesic perturbation keeps the structure of ECG beats while l_2 interpolation leads to situations that violates common senses (two QRS waves in one beat).

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).

Table 3: Comparison results of heart disease diagnosis (HYP, CD, STTC, MI) by different data augmentation methods, where the evaluation metrics are AUROC and F1-score.

Methods	AUROC	F1-score
No augmentation	0.843	0.575
Random Oversampling	0.820	0.536
SMOTE (Chawla et al., 2002)	0.799	0.534
ADASYN (He et al., 2008)	0.820	0.546
TaskAug (Raghu et al., 2022)	0.842	–
Ours	**0.931**	**0.707**

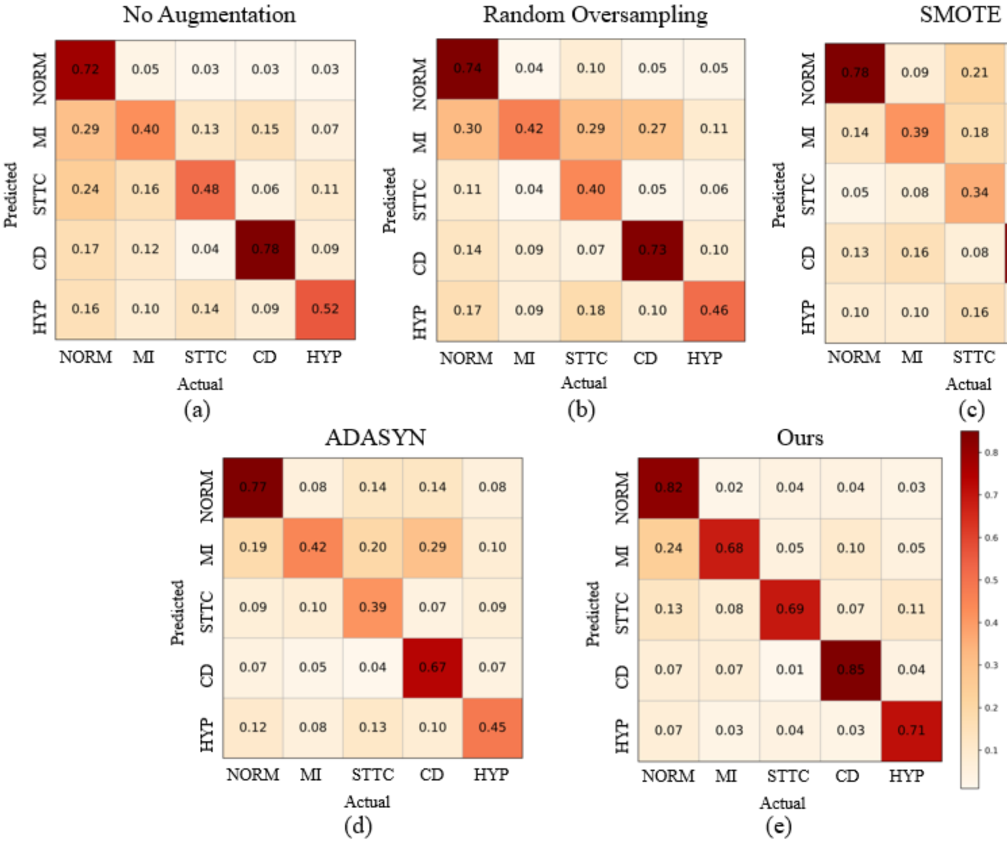

Figure 13: Confusion matrix of prediction results.

Table 3 shows the results of different data augmentation approaches, where standard evaluation metrics AUROC and F1-score are used to compare the performance of different strategies. We can find that even when some data augmentation methods are applied, i.e., random oversampling, SMOTE (Chawla et al., 2002), or ADASYN (He et al., 2008), the classification performance is not significantly improved (or even slightly worse) than using only ECG data without any augmentation. We also compared with TaskAug (Raghu et al., 2022), a new ECG data augmentation strategy which takes raw ECG signal as input. We can find that by using no augmentation strategy, our result is already better than TaskAug, showing: (1) high-level features are useful for the diagnosis task for learning the ECG patterns; (2) some data augmentation methods' performance improvement may be due to underfitting instead of learning additional patterns.

To have a more quantitative comparison of the classification results of each heart condition by different data augmentation methods, we compute the confusion matrix for each data augmentation method, as shown in Fig. 13. Our data augmentation method not only improves the classification accuracy of each category, but also improves the average classification result. Each category's performance becomes more balanced, showing that the robustness of the diagnosis result is improved.

Robust prediction: Following the pipeline of previous works (Han et al., 2020), we evaluate the robustness of our model as well as baseline method on the adversarial examples generated by Projected Gradient Descent (PGD) (Kurakin et al., 2016). PGD is a white-box attack methods that seeks adversarial samples with an ϵ-ball based on the gradient of a trained model. In our experiment, we gradually increase the capability of the adversarial by increasing the ϵ.

Table 4: AUROC result on clean and adversarial samples, for Myocardial Infarction (MI).

Myocardial Infarction (MI)	Clean AUROC	$\epsilon = 0.001$	$\epsilon = 0.002$	$\epsilon = 0.003$	$\epsilon = 0.004$
No augmentation	0.742	0.563	0.386	0.265	0.185
Random Oversampling	0.701	0.459	0.276	0.175	0.121
SMOTE (Chawla et al., 2002)	0.706	0.596	0.485	0.403	0.317
ADASYN(He et al., 2008)	0.726	0.621	0.497	0.405	0.323
Ours	**0.910**	**0.823**	**0.749**	**0.686**	**0.615**

Table 5: AUROC result on clean and adversarial samples for ST/T Change (STTC).

ST/T Change (STTC)	Clean AUROC	$\epsilon = 0.001$	$\epsilon = 0.002$	$\epsilon = 0.003$	$\epsilon = 0.004$
No augmentation	0.847	0.769	0.681	0.577	0.481
Random Oversampling	0.835	0.583	0.375	0.247	0.170
SMOTE (Chawla et al., 2002)	0.761	0.708	0.609	0.524	0.443
ADASYN (He et al., 2008)	0.824	0.727	0.619	0.525	0.433
Ours	**0.935**	**0.857**	**0.795**	**0.734**	**0.680**

Table 6: AUROC result on clean and adversarial samples for Conduction Disturbance (CD).

Conduction Disturbance (CD)	Clean AUROC	$\epsilon = 0.001$	$\epsilon = 0.002$	$\epsilon = 0.003$	$\epsilon = 0.004$
No augmentation	0.883	0.799	0.695	0.603	0.533
Random Oversampling	0.885	0.748	0.598	0.491	0.408
SMOTE (Chawla et al., 2002)	0.866	0.832	0.786	0.734	0.689
ADASYN (He et al., 2008)	0.869	0.780	0.690	0.615	0.552
Ours	**0.952**	**0.915**	**0.863**	**0.811**	**0.766**

Table 7: AUROC result on clean and adversarial samples for Hypertrophy (HYP).

Hypertrophy (HYP)	Clean AUROC	$\epsilon = 0.001$	$\epsilon = 0.002$	$\epsilon = 0.003$	$\epsilon = 0.004$
No augmentation	0.842	0.724	0.599	0.501	0.396
Random Oversampling	0.799	0.588	0.398	0.271	0.194
SMOTE (Chawla et al., 2002)	0.787	0.705	0.616	0.532	0.464
ADASYN (He et al., 2008)	0.806	0.647	0.514	0.419	0.364
Ours	**0.966**	**0.919**	**0.862**	**0.804**	**0.746**

7. Conclusion, Limitation, and Future Work

In this paper, we propose a new method for electrocardiograms data augmentation. We perturb the dataset along the geodesic in a Wasserstein space. We show that after data augmentation, there are both accuracy and robustness improvements in the classification results over five ECG categories, which demonstrate the effectiveness of our method. Although we focus on ECG prediction in this work, our proposed data augmentation method could be applied to sequential data in other healthcare applications.

The computational inefficiency might still be one of the significant obstacles with a large scale dataset. We would like to explore more advanced Wasserstein Barycenter algorithms that can improve efficiency.

Acknowledgments

The work is partly supported by the Allegheny Health Network and Mario Lemieux Center for Innovation and Research in EP.

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.

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

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.

Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Geodesic pca in the wasserstein space by convex pca. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 53, pages 1–26. Institut Henri Poincaré, 2017.

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.

N. Chawla, K. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357, 2002.

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, pages 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*, pages 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.

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

Chen Dan, Yuting Wei, and Pradeep Ravikumar. Sharp statistical guaratees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pages 2345–2355. PMLR, 2020.

Zhun Deng, Linjun Zhang, Amirata Ghorbani, and James Zou. Improving adversarial robustness via unlabeled out-of-domain data. In *International Conference on Artificial Intelligence and Statistics*, pages 2845–2853. PMLR, 2021.

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

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

Kilian Fatras, Younes Zine, Szymon Majewski, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Minibatch optimal transport distances; analysis and applications. *arXiv preprint arXiv:2101.01792*, 2021.

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)*, pages 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.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

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)*, pages 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)*, pages 50–56. IEEE, 2021b.

Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Debiased sinkhorn barycenters. In *International Conference on Machine Learning*, pages 4692–4701. PMLR, 2020.

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.

Partha Pratim Kanjilal, Sarbani Palit, and Goutam Saha. Fetal ecg extraction from single-channel maternal ecg using singular value decomposition. *IEEE Transactions on Biomedical Engineering*, 44(1):51–59, 1997.

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)*, pages 2608–2611, 2015.

Mahantapas Kundu, Mita Nasipuri, and Dipak Kumar Basu. Knowledge-based ecg interpretation: a critical review. *Pattern Recognition*, 33(3):351–373, 2000.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

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*, pages 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.

Mohammad Mehrabi, Adel Javanmard, Ryan A Rossi, Anup Rao, and Tung Mai. Fundamental tradeoffs in distributionally adversarial training. In *International Conference on Machine Learning*, pages 7544–7554. PMLR, 2021.

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.

John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909*, 2020.

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*, pages 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. Electrocardiogram classification by modified efficientnet with data augmentation. In *2020 Computing in Cardiology*, pages 1–4. IEEE, 2020.

Naoki Nonaka and Jun Seita. In-depth benchmarking of deep neural network architectures for ecg diagnosis. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, Proceedings of Machine Learning Research, pages 414–439. PMLR, 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, 2022a.

Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, Bo Li, Ding Zhao, and Hailin Jin. Mhms: Multimodal hierarchical multimedia summarization. *arXiv preprint arXiv:2204.03734*, 2022b.

Aniruddh Raghu, Divya Shanmugam, Eugene Pomerantsev, John Guttag, and Collin M Stultz. Data augmentation for electrocardiograms. In *Proceedings of the Conference on Health, Inference, and Learning*, Proceedings of Machine Learning Research, pages 282–310. PMLR, 2022.

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.

Arash Rasti-Meymandi and Aboozar Ghaffari. Aecg-decompnet: abdominal ecg signal decomposition through deep-learning model. *Physiological Measurement*, 42(4):045002, 2021.

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

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in Neural Information Processing Systems*, 32, 2019.

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*, pages 124–139. PMLR, 2019.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

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.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

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, PM Harris, A Sundar, NAS Smith, and PJ Aston. Robustness of convolutional neural networks to physiological electrocardiogram noise. *Philosophical Transactions of the Royal Society A*, 379(2212):20200262, 2021.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

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.

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.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.

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

Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, Joshua Tenenbaum, and Chuang Gan. Prompting decision transformer for few-shot policy generalization. In *International Conference on Machine Learning*, pages 24631–24645. PMLR, 2022.

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)*, pages 898–905, 2019.

Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

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.

Long Zhao, Ting Liu, Xi Peng, and Dimitris N. Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. *ArXiv*, abs/2010.08001, 2020.

Wei Zhong, Xuemei Guo, and Guoli Wang. Maternal ecg removal using short time fourier transform and convolutional auto-encoder. *International Journal of Data Mining and Bioinformatics*, 23(2):160–175, 2020.

Allen-Zeyuan Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 977–988. IEEE, 2022.

Jiacheng Zhu, Aritra Guha, Dat Do, Mengdi Xu, XuanLong Nguyen, and Ding Zhao. Functional optimal transport: map estimation and domain adaptation for functional data. *arXiv preprint arXiv:2102.03895*, 2021.

Jiacheng Zhu, Gregory Darnell, Agni Kumar, Ding Zhao, Bo Li, Xuanlong Nguyen, and Shirley You Ren. Physiomtl: Personalizing physiological patterns using optimal transport multi-task regression. In *Conference on Health, Inference, and Learning*, pages 354–374. PMLR, 2022.

Appendix A. Multi-Feature Transformer

The input for the Multi-Feature Transformer is composed of three parts, including ECG raw features, time-domain features, and frequency domain features. 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 network training 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:

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

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.

Appendix B. More Related Work

Traditional data augmentation methods include sampling, cost-sensitive methods, kernel-based methods, active learning methods, and one-class learning or novelty detection methods (He and 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 and 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 and Garcia, 2009), making the classification performance on the unseen testing data generally far worse.

Optimal Transport (OT) is a field of mathematics that studies the geometry of probability spaces (Villani, 2009). 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).

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 and 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.