# Empirical Study of Mix-based Data Augmentation Methods in Physiological Time Series Data

Peikun Guo
*Rice University*
Houston, USA
pg34@rice.edu

Huiyuan Yang
*Rice University*
Houston, USA
hy48@rice.edu

Akane Sano
*Rice University*
Houston, USA
akane.sano@rice.edu

***Abstract*—Data augmentation is a common practice to help generalization in the procedure of deep model training. In the context of physiological time series classification, previous research has primarily focused on label-invariant data augmentation methods. However, another class of augmentation techniques (*i.e., Mixup*) that emerged in the computer vision field has yet to be fully explored in the time series domain. In this study, we systematically review the mix-based augmentations, including mixup, cutmix, and manifold mixup, on six physiological datasets, evaluating their performance across different sensory data and classification tasks. Our results demonstrate that the three mix-based augmentations can consistently improve the performance on the six datasets. More importantly, the improvement does not rely on expert knowledge or extensive parameter tuning. Lastly, we provide an overview of the unique properties of the mix-based augmentation methods and highlight the potential benefits of using the mix-based augmentation in physiological time series data. Our code and results are available at https://github.com/comp-well-org/Mix-Augmentation-for-Physiological-Time-Series-Classification.**

***Index Terms*—Data augmentation, mixup, physiological time series**

## I. Introduction

Data augmentation is a crucial regularization technique for deep neural network models, as it serves to inform the network of potential variations in the input data during the training stage while preserving the integrity of the labels. This technique has been shown to improve network generalization, [1] by not only artificially increasing the size of the dataset but also imparting inductive bias through the encoding of information related to data invariances.

Traditional data augmentation techniques aim to increase the statistical support of the training data distribution by utilizing human knowledge and adding additional virtual samples from the vicinity distribution of training samples. This approach has been shown to improve generalization, as demonstrated in previous literature. Such data augmentations have been employed actively and effectively in computer vision [1]–[4] and speech recognition and synthesis [5]–[7]. The selection of specific augmentation methods remains a challenging task, as it is often based on heuristics and is highly dependent on the dataset, task, and even model architecture [8]. However, unlike in other domains, time series, particularly physiological data, do not follow a straightforward rule for label-invariant transformation. Methods such as jittering, rotation, scaling, permutation, magnitude warping, time warping, window slicing, and window warping have been shown to have unstable performance across datasets and tasks, or require human understandings of the data [9]–[11]. Two significant limitations of traditional transformation-based augmentations on physiological time series data are that: (1) certain transformations can be detrimental to the integrity of the physiological signal, and (2) the majority of traditional augmentations are data-dependent, lacking generalization and consistency across different datasets and tasks.

An alternative approach is represented by mixup regularization [4], which is based on the assumption that linear interpolations of feature vectors should lead to linear interpolations of the associated targets. Despite its simplicity, mixup has been shown to be effective across different domains (computer vision [4], [12], [13] and speech [14]–[16]) and different tasks. For time series classification tasks, previous studies also employed mix-based augmentations to enhance model representation and generalization [17]–[19]. However, none of the previous works provide a thorough empirical study of mix-based augmentations across various types of physiological times series, regarding both the quantitative gain in the metrics and the benefits of feature representation.

This study aims to evaluate the efficacy of mix-based data augmentation in the context of time series classification. Our evaluation compares mix-based augmentation against traditional data augmentation techniques, as classified in a previous survey, focusing on basic label-invariant time-domain transformations commonly used in time series classification. The baseline augmentations evaluated include jittering, rotation, scaling, permutation, magnitude warping, time warping, window slicing, and window warping. The unique benefits of mix-based augmentation are evaluated, and the contributions of this paper can be summarized as follows:

- We present an empirical study of three mixup-based data augmentation methods (i.e., mixup, cutmix, and manifold mixup) in the context of time series classification. We provide detailed formulations of these methods and evaluate their performances on six physiological and biobehavioral datasets.
- Our experiments reveal two significant distinctions between mixup-based augmentations and traditional data transformations. First, mixup-based methods do not re-

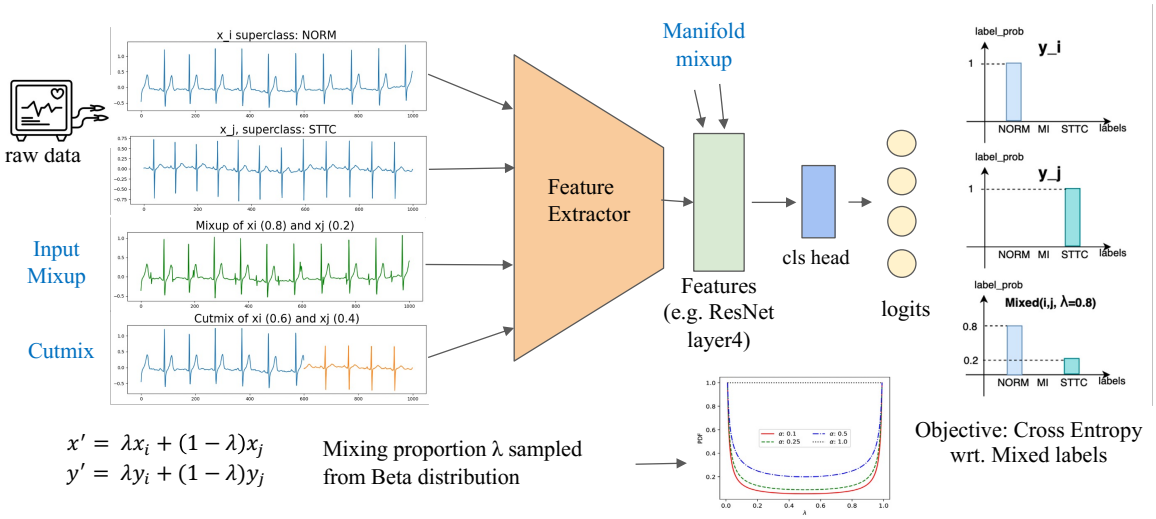



Fig. 1. The overview of mix-based augmentation procedure for physiological time series classification. From left to right: two sequences $x_i, x_j$ are shown as raw input signals. For mixup or cutmix, in each training epoch, virtual samples are created from the mini-batches; for manifold mixup, the linear combination is applied at the feature map level. For any of the three mix-based augmentations, the labels for the virtual samples are also mixed with the corresponding weights, shown on the right. NORM, MI, and STTC are representative classes of the PTB-XL dataset. See details in Figure 3.

quire human expert priors, making them more practical in various applications. Second, mixup-based methods consistently achieve higher or comparable performance compared to traditional methods.

## II. Related Works

### A. Traditional label-invariant time series data augmentation

For physiological time series data, the time domain transforms manipulate the original time series directly, as compared to more advanced augmentations using generative approaches. In this paper, we focus the evaluations on the following traditional augmentations: jittering, rotation, scaling, permutation, window warping, and window slicing.

Jittering refers to the injection of Gaussian or more sophisticated noise patterns, such as spikes and slope-like trends, into the raw signals. The schemes are introduced in [20]. Rotation for time series is achieved by multiplying the signal by a random rotation matrix. As it is not as suitable for time series data as in the image domain, one commonly used special case is flipping (changing the sign of the original time series). Permutation is a transformation that randomly shuffles segments of a time series. This operation does not preserve the sequential information of the original data. Window slicing or cropping, introduced in [21], randomly extracts continuous slices from the original samples. The window warping method is uniquely applicable to time series. It randomly selects a time interval, then upsamples or downsamples the segment, while keeping the rest of the time ranges unaltered. Window warping changes the total length of the original signal, therefore it is usually used along with window cropping. For time series classification tasks, all the above-mentioned augmentations do not change the labels of the altered training samples.

The effectiveness of these augmentations has been previously investigated in the literature. An empirical study about biobehavioral time series data augmentation [9] concludes that, while some augmentations are beneficial for biobehavioral classification tasks, their effectiveness varies across different datasets and model architectures. This finding agrees with [11], which also highlights the inconsistency of transformation-based augmentations across different non-physiological time series datasets.

Based on the analysis of the literature, we claim that two significant limitations exist when applying traditional transformation-based augmentations to physiological time series data:

- Certain transformations can be detrimental to the integrity of the physiological signals. For instance, rotation, permutation, and warping can be more harmful to ECG signals, as ECG beats possess relatively fixed patterns, such as the order, width, and intensity of the wave components.
- The majority of traditional augmentations are data-dependent, meaning that they require expert prior knowledge or multiple trials to select an appropriate transformation for a specific problem, due to the diverse properties of biobehavioral time series.

### B. Mixup: Vicinal Risk Minimization

Mixup is a data augmentation technique introduced by [4] to train neural networks by constructing virtual training examples using convex combinations of pairs of examples and their labels.

The methodology behind Mixup, as described in this paper, is rooted in Vicinal Risk Minimization (VRM), which diverges from the conventional Empirical Risk Minimization (ERM) by drawing examples from a vicinity distribution of the training examples. This aims to enlarge the support for the training distribution. As stated in Section II, prior knowledge has been

traditionally required for the identification of the vicinity or neighborhood. However, Mixup provides a more practical and data-agnostic alternative, as it does not necessitate domain expertise. Despite its simplicity, Mixup presents the following distinct advantages:

1. **Regularization**: The linear relationship established by mixup transformations between data augmentation and the supervision signal results in a strong regularization of the model's state, leading to improved performance.
2. **Generalization**: The authors of previous studies have reported improvements in generalization error for state-of-the-art models trained on ImageNet, CIFAR, speech, and tabular datasets when using mixup. Theoretical analysis [13] suggests that the soft targets of mixup virtual samples aid in model generalization in a manner similar to label smoothing and knowledge distillation. Interpolation/extrapolation of nearest neighbors in feature space can also improve generalization [22].

Since mixup was proposed, many incremental works have emerged and demonstrated improvements with different focuses, such as cutmix [23], manifold mixup [12], MixMatch [24], and AlignMix [25]. In this project, we choose to evaluate mixup, cutmix, and manifold mixup, as they have not been investigated in the context of time series classification in literature.

### C. Mixup for physiological time series data

In the domain of physiological time series classification, mixup has been employed to enhance generalization during training as demonstrated in prior studies. [17], [18] employ mixup for better generalization in ECG classification task, in the training batches of CNN models. [19] states that mixup improves the generalization performance of the ECG classification model regardless of leads and evaluation metrics. However, these studies lack a thorough examination of mixup's mechanism and reasons for performance improvement, and none of them have conducted ablation studies about mixup. Furthermore, although the 1D variant of vanilla Mixup [4] has been utilized in previous studies, the empirical results for cutmix [23] and manifold mixup [12] are currently lacking. In [26], the performance of mixup and cutmix were evaluated on the UEAMTSC dataset [27] using InceptionTime [28] as the baseline model. However, the subsets of UEAMTSC in that study were of small scale and not strictly comprised of time series data (*e.g.*image contours). This paper, on the other hand, aims to investigate the effectiveness of the mix-based methods on various physiological time series datasets, utilizing a higher capacity residual network structure.

## III. Augmentation methods

In this section, we provide a formal introduction and implementation of mix-based augmentations. Figure 1 shows the overall paradigm of how the mix-based augmentation is applied during the training process. The intrinsic difference between mix-based and traditional augmentations will be discussed in section V. The technical details for the implementation of the augmentations can be found in section IV-B.

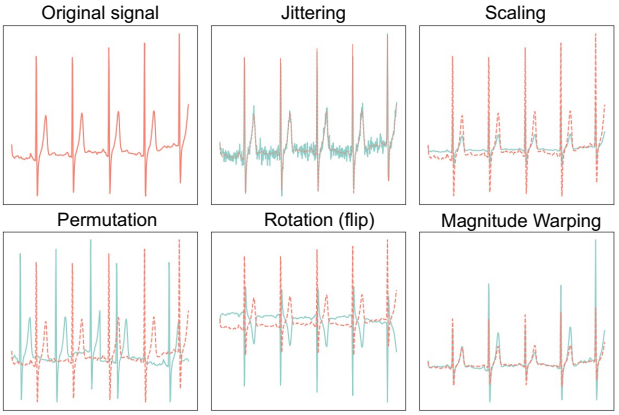



Fig. 2. Illustrations of traditional time series data augmentations. Orange: original signal. Green: augmented signals.

### A. Cutout

In this study, the Cutout augmentation serves as a benchmark for comparison against mix-based data augmentation methods. Cutout, originally introduced in the computer vision field to address occlusion issues, involves the removal of contiguous sections of data. To evaluate its effectiveness in the context of time series classification, a single-item transformation, similar to traditional label-invariant regularizations, is applied. Specifically, a random contiguous section of a time series is replaced with zeros, which can be considered a dropout (zero-masking) operation at the input layer. The size of the random time segment is fixed, while the starting index of the interval is randomly drawn from a uniform distribution, and applied to all channels of a single data point.

### B. Mixup

The time series mix-based augmentations in this study leverage multiple data samples from one training minibatch to generate virtual data points. We examine three commonly utilized configurations of mix-based augmentations from the computer vision domain for time series classification problems, namely mixup [4], cutmix [23], and manifold mixup (layer mixup, [12]).

The mixup augmentation blends random pairs of time series from the training data. Let $(x, y)$ denote a time series data instance, where $x \in \mathbb{R}^{L \times C}$, with $L$ representing the length of the sequence and $C$ denoting the number of channels, and $y \in \mathbb{R}^K$ being the class label with $K$ classes. The mixing ratio randomly drawn from a Beta distribution is denoted as $\lambda$. Given two samples $(x_i, y_i)$ and $(x_j, y_j)$, the mixup augmentation generates virtual training examples through the following formulation:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \tag{1}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j \tag{2}$$

Mixup extends the training distribution by incorporating the the assumption that linear interpolations of feature vectors should lead to linear interpolations of the associated targets [4]. Based on the formulation, the label for a mixup virtual sample is also a mixture of two original one-hot labels weighted by $\lambda$. The potential effect and benefit of the soft target will also be discussed in section V-B.

The mixing ratio, $\lambda$, in the mix-based data augmentation approach is sampled from a Beta distribution. The value of $\lambda$ close to 0 or 1 results in the created virtual time series being more similar to one of the raw data points, whereas a value of $\lambda$ close to 0.5 results in a more blended representation of the raw data points. For physiological time series data, it is desirable to have virtual time series that are similar to one of the raw data points, as these signals contain delicate features, such as the intensity of R peaks in ECG signals, which could be easily destroyed with random mixing.

Despite its simplicity, in the computer vision domain, mixup has allowed consistently superior performance in the CIFAR-10, CIFAR-100, and ImageNet image classification datasets [4]. As we will show in the results section, mixup can also improve classification metrics in time series classification problems, along with other desirable features.

### C. Cutmix

The image augmentation technique Cutmix, proposed in [23], shares similarities with the technique Mixup. The authors claim a key advantage of Cutmix is its ability to prevent the occurrence of ambiguous components in the generated samples caused by mixing, such as blurred image regions [23]. For time series cutmix, we select a random time segment from a pair of multivariate time series, then the values of the pair of time series within the segment are exchanged across all channels. The length of the segment is also determined randomly through the mixing ratio $\lambda$ drawn from a Beta distribution.

### D. Manifold Mixup

Manifold Mixup, presented in [12], demonstrates consistently superior performance across various computer vision tasks when compared to the original input-data-mixup approach. Unlike the original mixup, manifold mixup trains neural networks on linear combinations of hidden representations of training samples. The literature suggests that higher-level representations obtained from intermediate layers of the neural network feature extractor are low-dimensional, therefore, linear interpolations of hidden representations should cover meaningful regions of the feature space. In this paper, we implement Layer Mixup on layer 4 of the ResNet, prior to pooling and the classification head. The labels are also mixed in the same way as mixup and cutmix.

## IV. Experiments

### A. Datasets

We conduct experiments on six biomedical time series datasets, encompassing diverse data types and varying sizes. The datasets include two ECG datasets, PTB-XL for cardiac condition classification and Apnea-ECG for sleep apnea detection, two EEG datasets, Sleep-EDF for sleep stage recognition and MMIDB for sleep movement detection, and two IMU datasets, PAMAP2 and UCI-HAR for human activity recognition. Table I provides a summary of the datasets. Note that in column "Periodic", the IMU datasets are tagged as "motion", because IMU data may contain periodic patterns when the recorded activity contains periodic motion (*e.g.*walking).

*1) PTB-XL:* The PTB-XL dataset [29] is an ECG database of 12-lead recordings, containing 44 diagnostic statements grouped into 5 superclasses (normal, conduction disturbance, myocardial infarction, hypertrophy, and ST-T change). In this study, we formulate a five-class cardiac abnormality classification problem. The data is divided into 10 balanced folds, with the first 8 used for training, the 9th for validation, and the 10th for testing. The data we use is sampled at 100Hz. The same stratification process is applied to the other datasets if no validation/test set is provided.

The dataset is highly class-imbalanced as shown in Figure 3, with over half of the samples labeled normal and the least-represented class (HYP, hypertrophy in left ventricle) only constituting 3.3% of the data. This challenge is addressed in Section V through in-batch resampling and mix-based augmentations.

*2) Apnea-ECG:* The Apnea-ECG dataset examines the connection between sleep apnea symptoms and heart activity in humans [30], as monitored through ECG. This dataset has 70 records, sampled at 100 Hz, with 35 records designated for training and the remaining 35 for testing. Each record is 7-10 hours in length and includes a continuous ECG signal along with per-minute apnea annotations indicating the presence or absence of sleep apnea. We segmented the ECG recordings into 60-second frames at 100 Hz, resulting in 17233 samples for the training set and 17010 samples for the test set.

*3) Sleep-EDFE:* The Sleep-EDFE dataset is sourced from the publicly accessible Sleep European Data Format (EDF) database [31] on Physionet [32]. This database contains full-night PSG sleep records that include two-channel EEG (Fpz-Cz and Pz-Oz), a horizontal EOG, and EMG signal records, along with corresponding hypnograms (sleep stage annotations). The EEG signals have a sampling frequency of 100 Hz, and they are divided into 30-second epochs and normalized to have zero mean and unit standard deviation.

*4) MMIDB-EEG:* The EEGMMIDB (EEG Motor Movement/Imagery Database) from PhysioNet is collected using the BCI200 EEG system4 [33]. It records 64 channels of brain signals at a sampling rate of 160 Hz, totaling over 1500 recordings of 1-2 minutes each. Subjects are instructed to wear the EEG device and sit in front of a computer screen, performing specific typing tasks in response to on-screen prompts.

*5) PAMAP2:* The PAMAP2 dataset [34] comprises recordings from 9 participants who were asked to perform 12 daily activities, including household tasks and various exercises (e.g., Nordic walking, playing soccer). Data from accelerom-

TABLE I
DATASET SUMMARY

| DATASET | CATEGORY | # CHANNELS | # CLASSES | SAMPLE LENGTH | PERIODIC | # SAMPLES |
|---|---|---|---|---|---|---|
| PTB-XL | ECG | 12 | 5 | 1000 | YES | 17962 |
| APNEA-ECG | ECG | 1 | 2 | 6000 | YES | 34243 |
| SLEEP-EDFE | EEG | 1 | 5 | 3000 | NO | 42308 |
| MMIDB-EEG | EEG | 64 | 2 | 640 | NO | 4635 |
| PAMAP2 | IMU | 52 | 12 | 1000 | (MOTION) | 5452 |
| UCI-HAR | IMU | 9 | 6 | 128 | (MOTION) | 10299 |

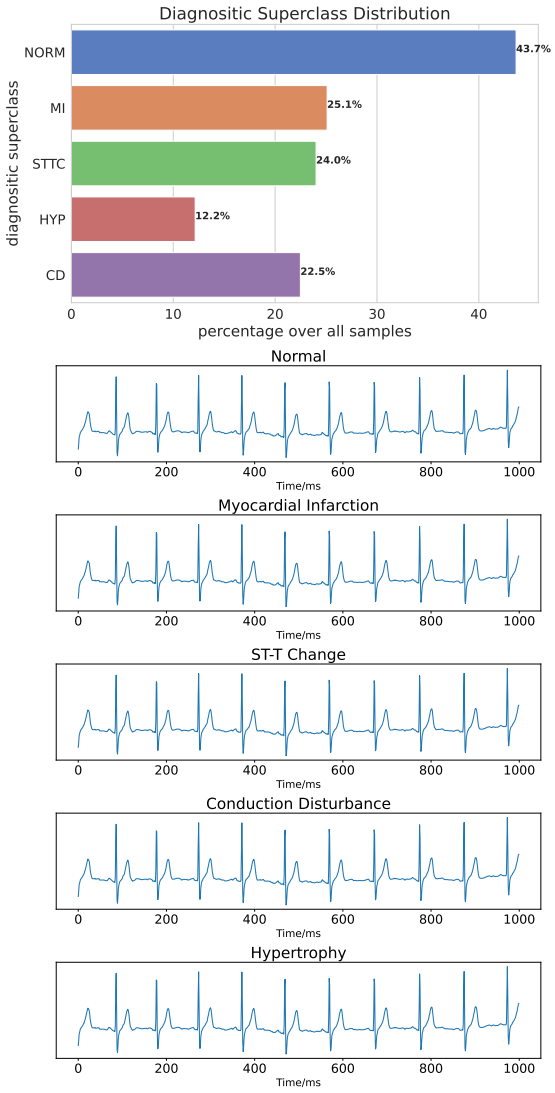



Fig. 3. PTB-XL dataset overview. Top: The five-superclass distribution of PTB-XL (NORM, MI, STTC, CD, HYP). In this paper, we only train with single-label samples. Bottom: Example waveforms of PTB-XL classes (lead IV), the x-axis denotes time in ms, and the y-axis is the normalized voltage.

eters, gyroscopes, magnetometers, temperature sensors, and heart rate monitors are recorded from inertial measurement units placed on the hand, chest, and ankle over 10 hours, resulting in a 52-dimensional dataset.

*6) UCI-HAR:* The UCI-HAR dataset [35] was collected from 30 volunteers aged 19 to 48 years. Participants were instructed to engage in six basic activities, which included three static postures (standing, sitting, lying) and three dynamic activities (walking, walking downstairs, walking upstairs). 3-axial accelerometer and gyroscope signals were recorded at a constant rate of 50 Hz. The data was collected using smartphones carried by the participants.

TABLE II
STRUCTURE OF THE RESNET-18 BACKBONE. $B$: BATCH SIZE; $L$: LENGTH OF SEQUENCE; $C$: NUMBER OF CHANNELS.

| Layer Name | Input Shape | Output Shape | Parameter |
|---|---|---|---|
| Reshape Conv1d | [B, L, C] | [B, C, L] | – |
| Layer1 | [C, L] | [B, 64, L] | 1x3, 64, max pool |
| Layer2 | [B, 64, L/2] | [B, 64, L/2] | [[1x3, 64], [1x3, 64]]x2 |
| Layer3 | [B, 128, L/4] | [B, 128, L/4] | [[1x3, 128], [1x3, 128]]x2 |
| Layer4 | [B, 256, L/8] | [B, 512, L/16] | [[1x3, 512], [1x3, 256]]x2 |
| Average pool | [B, 512, L/16] | [B, 512, 1] | – |
| FC | [B, 512] | [B, Classes ] | [512, Classes ] |

## B. Experimental Setup

*1) Network architecture:* In all experiments, we use a 1D-CNN-based ResNet-18 [36] as the backbone. This model has convolutions with a kernel size of 3, and stride 2. The blocks in the ResNet architecture have convolutional layers with 32, 64, 128, and 256 channels respectively. The output after the final block is average pooled in the temporal dimension, and then a linear layer is applied to predict the probability of the positive class. Details of the ResNet-18 structure are summarized in Table.II.

For the manifold mixup experiments, we take the output of layer 4 of ResNet18 as the input for mixing. We also perform t-SNE visualization on the features extracted from Layer 4 as validation for the quality of class representation.

*2) Augmentation Implementation:* Following the previous empirical studies for mix-based regularization: mixup [4] and MixMatch [24], we use $\alpha = 0.4$ and $\alpha = 0.75$ for mix-related hyperparameters in our experiments. For cutmix, the ratio of the random segment length to the signal length is set to 0.2, following [23]. For the baseline augmentations, the hyperparameters, such as the intensity of scaling and jittering, are manually chosen following [9].

Following [4], the implementation of the training step is based on the mini-batches sampled by a data loader. For each minibatch, random shuffling is applied, and the mixing operations can all be performed in a vectorized manner, incurring minimal computation overhead. In our PTB-XL profiling experiments, the mean increase in the processing time of each

mini-batch of size 128 is less than 0.001 second. We also report results for baselines (denoted as vanilla) that does not use any data augmentation.

*3) Optimization:* We use the AdamW optimizer with learning rate 0.001 for all datasets. For the profiling experiments on PTB-XL dataset, we also tested with Adam optimizer and 0.005 learning rate. We use a step decay learning rate scheduler with step size 5, and a decay rate of 0.9 across all experiments. Training takes 50 epochs, which was observed to be sufficient for convergence in all datasets.

*4) Computing resource:* : All model training was performed on a single NVIDIA GTX 2080Ti GPU.

## V. Results and discussion

### A. Quantitative performance of mix-base augmentations

Experiments were conducted on six datasets of diverse categories with a ResNet18-1D backbone. The performance of vanilla (no augmentation), cutout, and three mix-based augmentations are presented in Table III. We summarize our results as follows.

I).**The mix-based data augmentations can achieve superior accuracy in comparison to traditional augmentation methods.** All three mix-based augmentation techniques were found to outperform the baseline (no augmentation) in 16 out of 18 experimental trials, with only two exceptions (mixup for Apnea-ECG and cutmix for MMIDB-EEG). Furthermore, among the six datasets examined, the majority of the highest accuracy results were achieved through the use of cutmix and layer mixup.

Overall, the mix-based augmentations outperform the baselines, but the performance of augmentation methods varies across different datasets. This is likely due to the unique characteristics of each dataset and the strengths of each augmentation method. For instance, datasets containing complex temporal patterns or high levels of noise may benefit from the use of certain mix-based augmentation methods that are particularly effective at enhancing classification accuracy. The results in Table III also show that more comprehensive mixup schemes (cutmix, manifold mixup) help yield better accuracy.

II).**The mix-based augmentations deliver robust performance, and the accuracy gain is steady.** In addition to the quantitative performance advantages, these techniques are notable for their low dependency on expert knowledge and parameter tuning. Across all 18 mix-based experimental trials (excluding cutout), no significant reduction in accuracy was observed in comparison to the baseline. In contrast, traditional data transformation techniques can often result in a drastic decrease in accuracy if not implemented appropriately. For example, applying scaling in Apnea-ECG resulted in a 7.5% reduction in accuracy, permutation in MMIDB-EEG resulted in a 9.0% reduction, and jittering in UCI-HAR resulted in a 4.8% reduction. This finding implies that traditional transformations can undermine crucial features in physiological signals, such as wave intensity in ECG data and temporal correlations in EEG data. As a result, these augmentations can generate virtual samples that deviate from the actual data distribution, potentially compromising the generalization performance of the model.

Note that in the experiments, we compared the mix-based augmentations against some individual traditional augmentations. However, to fully harness the potential of data augmentation, it is compatible to apply mix-based augmentations in conjunction with one or more traditional augmentations.

### B. Profiling mix-based augmentations on PTB-XL

The PTB-XL dataset [29] is a well-studied ECG dataset for cardiac condition classification, with the current SOTA accuracy of recognizing five classes being less than 80% [9]. To evaluate the effectiveness of mix-based data augmentation methods, extensive profiling experiments were conducted on the PTB-XL dataset, using over 80 combinations of settings and hyperparameters (Section IV-B3). The scatter plot of the best validation accuracy vs the best F1 score is shown in Figure 4(a). Identical combinations of learning rates, optimizers, etc. were used for each of the four augmentation setups (vanilla, input mixup, cutmix, layer mixup). The scatter plot illustrates that mix-based augmentations produce the best results, with all top-performing results (1st to 28th) obtained from mix-based augmentations, supporting the conclusions drawn in Section V-A.

The PTB-XL dataset presents an imbalanced class distribution for the single-label data samples. To mitigate the high false negative rate for the minority class, a batch class-balanced data sampler was utilized in the data loader during the training process. The effect of the class-balanced sampler on the model's performance is shown in Figure 4 (c) and (d), by comparing the confusion matrices with and without the balanced sampling. As observed in the bottom row of Figure 4, which corresponds to the data samples of the minority class hypertrophy (HYP), the model without the balanced sampler was prone to producing the most false negative (NORM) predictions for the HYP cases. However, by incorporating both the class-balanced sampler and cutmix, virtual samples containing the information of the minority class were generated during the training, reducing the false negative rate. Additionally, the recall of the other three cardiac condition classes also received similar improvement as shown in the plot.

### C. Feature representations

The advantages of mix-based augmentations, or vicinal risk minimization, include the provision of more distinguishable representations for different classes. We present the results of t-SNE dimensional reduction of two models trained with cutmix and a baseline, on the training and test sets of the PTB-XL dataset. The feature vectors were calculated using layer 4 of ResNet18. The visualizations of the training set (Figure 5(a) and (c)) indicate that, compared to the baseline, cutmix gives more discriminative representations between classes. The projections of the test set (Figure 5(b) and (d)) demonstrate that the classes are more distinguishable with mix-based augmentation and balanced sampling.

TABLE III

CLASSIFICATION ACCURACIES FOR MIX-BASED AUGMENTATION METHODS AND THREE TRADITIONAL AUGMENTATIONS ARE REPORTED ON SIX DATASETS. THE WARP COLUMN REFERS TO WINDOW WARPING. BOLD NUMBERS INDICATE THE BEST PERFORMANCE.

| DATASET | BASELINE | MIXUP | CUTMIX | LAYER MIXUP | CUTOUT | JITTER | SCALE | PERMUTE | WARP |
|---|---|---|---|---|---|---|---|---|---|
| PTB-XL | 77.94 | 78.91 | **79.64** | 78.97 | 77.61 | 77.60 | 78.03 | 76.57 | 75.67 |
| APNEA-ECG | 78.10 | 74.73 | 78.30 | **79.69** | 78.27 | 76.44 | 70.64 | 77.26 | 76.34 |
| SLEEP-EDFE | 83.40 | 84.27 | 84.40 | 84.48 | 83.84 | 81.55 | 83.40 | **85.01** | 83.98 |
| MMIDB-EEG | 78.64 | 79.83 | 77.99 | **80.58** | 80.04 | 79.07 | 80.15 | 69.58 | 77.99 |
| PAMAP2 | 93.62 | 95.31 | **95.83** | 93.88 | 86.98 | 86.85 | 88.18 | 85.82 | 87.59 |
| UCI-HAR | 92.77 | 93.62 | **94.26** | 93.11 | 93.58 | 87.95 | 88.63 | 91.89 | 92.94 |

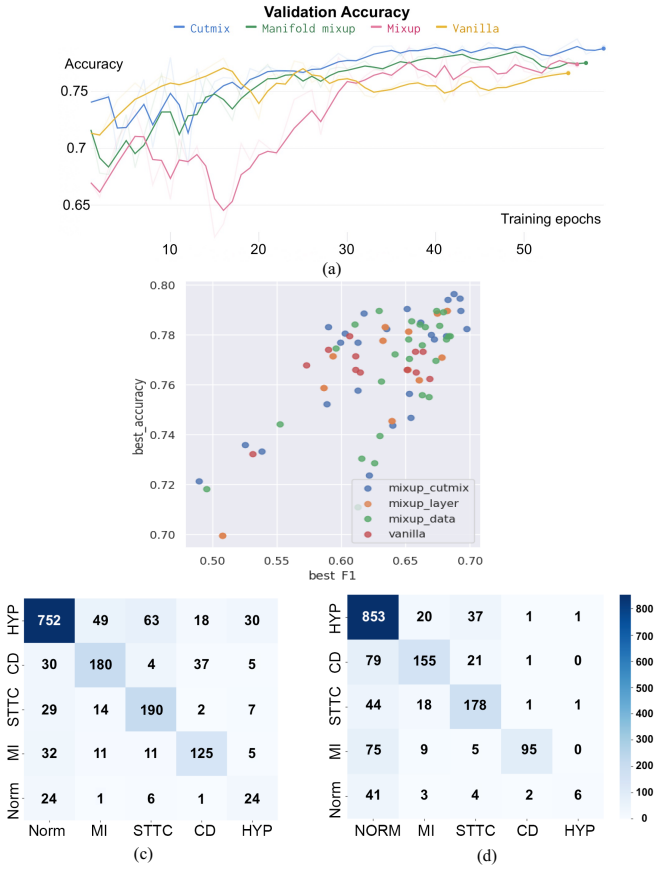
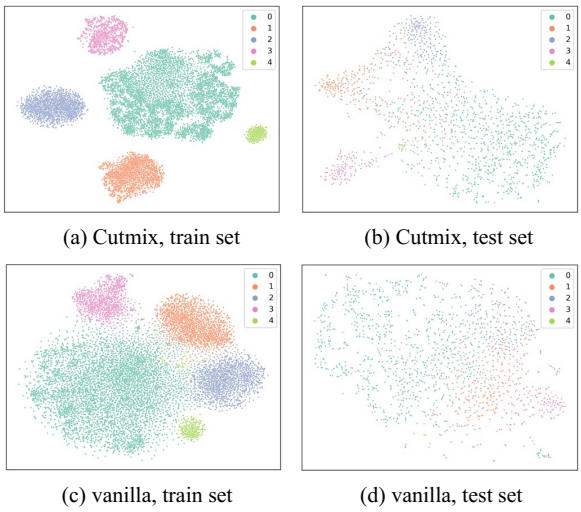



Fig. 4. (a): The PTB-XL validation accuracy computed after each training epoch, with baseline and variations of mixup. (b): The scatter plot of validation accuracy again F1 score for all 80 profiling experiments on PTB-XL dataset, different colors annotate different augmentations. (c) and (d): The confusion matrix of cutmix training, with or without a balanced sampler, respectively. The y-axis represents the true labels, and the x-axis is the predicted labels.



(a) Cutmix, train set (b) Cutmix, test set

(c) vanilla, train set (d) vanilla, test set

Fig. 5. The t-SNE visualizations of cutmix and vanilla settings after training on PTB-XL. In both experiments, a balanced data sampler is used during training to increase the performance of minority classes.

## VI. CONCLUSION AND FUTURE WORK

Inspired by the success of mixup in other domains, we investigate mixup and its variants, cutmix and manifold mixup for physiological for the time series classification task. This paper empirically shows that mix-based data augmentation techniques can achieve superior accuracy in comparison to traditional augmentation methods in the context of time series classification. In the experiments, the majority of the highest accuracy results were achieved through the use of cutmix and layer mixup, and these augmentations were found to deliver robust performance with a steady accuracy gain across various physiological and biobehavioral datasets. The low dependency on expert knowledge and parameter tuning, in addition to the quantitative performance advantages, makes mix-based augmentations more practical and effective in various applications. These findings highlight the effectiveness of mix-based, dataset-agnostic augmentations and the importance of appropriately choosing traditional data transformations, as they can compromise the generalization performance of the model.

We plan to explore the combination of mix-based augmentation and traditional time series augmentations, as the effectiveness of well-composited transformations has been shown in previous studies [11]. Furthermore, we aim to extend the applicability of mix-based augmentation to the frequency domain, following the success of such an approach in acoustic data classification tasks [37].

## REFERENCES


[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[2] P. Y. Simard, Y. A. LeCun, J. S. Denker, and B. Victorri, "Transformation invariance in pattern recognition—tangent distance and tangent

propagation," in *Neural networks: tricks of the trade*. Springer, 1998, pp. 239–274.

[3] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[4] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[5] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013, p. 21.

[6] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[7] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[8] D. Ho, E. Liang, X. Chen, I. Stoica, and P. Abbeel, "Population based augmentation: Efficient learning of augmentation policy schedules," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2731–2741.

[9] H. Yang, H. Yu, and A. Sano, "Empirical evaluation of data augmentations for biobehavioral time series data with deep learning," 2022. [Online]. Available: https://arxiv.org/abs/2210.06701

[10] Q. Wen, L. Sun, F. Yang, X. Song, J. Gao, X. Wang, and H. Xu, "Time series data augmentation for deep learning: A survey," *arXiv preprint arXiv:2002.12478*, 2020.

[11] B. K. Iwana and S. Uchida, "An empirical survey of data augmentation for time series classification with neural networks," *Plos one*, vol. 16, no. 7, p. e0254841, 2021.

[12] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6438–6447.

[13] L. Carratino, M. Cissé, R. Jenatton, and J.-P. Vert, "On mixup regularization," *arXiv preprint arXiv:2006.06049*, 2020.

[14] L. Meng, J. Xu, X. Tan, J. Wang, T. Qin, and B. Xu, "Mixspeech: Data augmentation for low-resource automatic speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7008–7012.

[15] N. Jia and C. Zheng, "Emotion speech synthesis method based on multichannel time–frequency domain generative adversarial networks (mc-tfd gans) and mixup," *Arabian Journal for Science and Engineering*, vol. 47, no. 2, pp. 1749–1762, 2022.

[16] K. Xu, D. Feng, H. Mi, B. Zhu, D. Wang, L. Zhang, H. Cai, and S. Liu, "Mixup-based acoustic scene classification using multi-channel convolutional neural network," in *Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part III 19*. Springer, 2018, pp. 14–23.

[17] L. Antoni, E. Bruoth, P. Bugata, P. Bugata Jr, D. Gajdoš, Š. Horvát, D. Hudák, V. Kmečová, R. Staňa, M. Staňková *et al.*, "Automatic ecg classification and label quality in training data," *Physiological Measurement*, vol. 43, no. 6, p. 064008, 2022.

[18] H. Han, S. Park, S. Min, H.-S. Choi, E. Kim, H. Kim, S. Park, J. Kim, J. Park, J. An *et al.*, "Towards high generalization performance on electrocardiogram classification," in *2021 Computing in Cardiology (CinC)*, vol. 48. IEEE, 2021, pp. 1–4.

[19] H. Han, S. Park, S. Min, E. Kim, H. Kim, S. Park, J. Kim, J. Park, J. An, K. Lee *et al.*, "Improving generalization performance of electrocardiogram classification models," *Physiological Measurement*, 2023.

[20] T. Wen and R. Keyes, "Time series anomaly detection using convolutional neural networks and transfer learning," *arXiv preprint arXiv:1905.13628*, 2019.

[21] Z. Cui, W. Chen, and Y. Chen, "Multi-scale convolutional neural networks for time series classification," *arXiv preprint arXiv:1603.06995*, 2016.

[22] T. DeVries and G. W. Taylor, "Dataset augmentation in feature space," *arXiv preprint arXiv:1702.05538*, 2017.

[23] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.

[24] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in neural information processing systems*, vol. 32, 2019.

[25] S. Venkataramanan, E. Kijak, L. Amsaleg, and Y. Avrithis, "Alignmixup: Improving representations by interpolating aligned features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 174–19 183.

[26] H. Yang and T. Desell, "Robust augmentation for multivariate time series classification," *arXiv preprint arXiv:2201.11739*, 2022.

[27] A. Bagnall, H. A. Dau, J. Lines, M. Flynn, J. Large, A. Bostrom, P. Southam, and E. Keogh, "The uea multivariate time series classification archive, 2018," *arXiv preprint arXiv:1811.00075*, 2018.

[28] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020.

[29] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "Ptb-xl, a large publicly available electrocardiography dataset," *Scientific data*, vol. 7, no. 1, pp. 1–15, 2020.

[30] T. Penzel, G. B. Moody, R. G. Mark, A. L. Goldberger, and J. H. Peter, "The apnea-ecg database," in *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)*. IEEE, 2000, pp. 255–258.

[31] B. Kemp, A. Zwinderman, B. Tuk, H. Kamphuisen, and J. Oberyé, "Sleep-edf database expanded," *physionet. org*, 2018.

[32] G. B. Moody, R. G. Mark, and A. L. Goldberger, "Physionet: a web-based resource for the study of physiologic signals," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 70–75, 2001.

[33] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system," *IEEE Transactions on biomedical engineering*, vol. 51, no. 6, pp. 1034–1043, 2004.

[34] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *2012 16th international symposium on wearable computers*. IEEE, 2012, pp. 108–109.

[35] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, and D. Anguita, "Transition-aware human activity recognition using smartphones," *Neurocomputing*, vol. 171, pp. 754–767, 2016.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] G. Kim, D. K. Han, and H. Ko, "Specmix: A mixed sample data augmentation method for training withtime-frequency domain features," *arXiv preprint arXiv:2108.03020*, 2021.