Title

Log-Spectral Matching GAN: PPG-based Atrial Fibrillation Detection can be Enhanced by GAN-based Data Augmentation with Integration of Spectral Loss.

Authors

Ding, Cheng
Xiao, Ran
Do, Duc
et al.

Publication Date

2023-01-06

DOI

10.1109/JBHI.2023.3234557

Peer reviewed

HHS Public Access

Author manuscript

IEEE J Biomed Health Inform. Author manuscript; available in PMC 2024 July 26.

Published in final edited form as:

IEEE J Biomed Health Inform. ; PP: . doi:10.1109/JBHI.2023.3234557.

Log-Spectral Matching GAN: PPG-based Atrial Fibrillation Detection can be Enhanced by GAN-based Data Augmentation with Integration of Spectral Loss

Cheng Ding[1], **Ran Xiao**[2], **Duc Do**[3], **David Scott Lee**[4], **Randall J Lee**[5], **Shadi Kalantarian**[5], **Xiao Hu**[2,6,7]

[1]Department of Biomedical Engineering, Georgia Institute of Technology & Emory University, Atlanta, Georgia, United States

[2]Nell Hodgson Woodruff school of Nursing, Emory University, Atlanta, GA, USA

[3]Duc Do is with David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, USA

[4]Department of Otolaryngology, Washington University in St. Louis St Louis, MO, USA

[5]School of Medicine, University of California San Francisco, San Francisco, CA, United States

[6]Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA, USA

[7]Department of Computer Science, College of Arts and Sciences, Emory University, Atlanta, GA, USA

Abstract

Photoplethysmography (PPG) is a ubiquitous physiological measurement that detects beat-to-beat pulsatile blood volume changes and hence has a potential for monitoring cardiovascular conditions, particularly in ambulatory settings. A PPG dataset that is created for a particular use case is often imbalanced, due to a low prevalence of the pathological condition it targets to predict and the paroxysmal nature of the condition as well. To tackle this problem, we propose log-spectral matching GAN (LSM-GAN), a generative model that can be used as a data augmentation technique to alleviate the class imbalance in a PPG dataset to train a classifier. LSM-GAN utilizes a novel generator that generates a synthetic signal without a up-sampling process of input white noises, as well as adds the mismatch between real and synthetic signals in frequency domain to the conventional adversarial loss. In this study, experiments are designed focusing on examining how the influence of LSM-GAN as a data augmentation technique on one specific classification task - atrial fibrillation (AF) detection using PPG. We show that by taking spectral information into consideration, LSM-GAN as a data augmentation solution can generate more realistic PPG signals.

I. Introduction

The currently estimated prevalence of AF in adults is between 2% and 4% , and a significant rise is expected owing to extended longevity in the general population and intensifying search for undiagnosed AF [1]. If left untreated, AF confers various significant

health risks. AF is linked to a 5-fold increase in the risk of ischemic stroke, a 3-fold increase in the risk of heart failure, and a 2-fold increase in the risk of mortality from heart disease [2]. Therefore, it marks great clinical and economic significance to have an affordable, portable, and continuous AF screening tool that patients with AF can access at scale. In the past, AF detection has been mainly relying on analysis and interpretation of electrocardiogram (ECG). Recent advancement in wearable technologies, such as fitness bands and smartwatches, offers convenient and continuous recordings of photoplethysmography (PPG), which demonstrates to be a potential alternative to ECG for AF detection [3], [4]. Current wearables with continuous monitoring of PPG offer many benefits, such as friendly user interface, low cost, and portability, making it a promising platform to achieve an AF screening tool accessible to the general population. Therefore, PPG garnered tremendous research interest in recent years for a reliable and accurate AF detection solution. For PPG-based AF detection, it has been shown that deep learning (DL) algorithms achieve better results than traditional machine learning algorithms [5]–[8]. Recent studies benefit from a balanced sample setup between AF and non-AF classes that offer promising AF detection performance [6]–[11]. Abundant data can be obtained through either in-hospital settings or from free-living subjects outside of the hospital. However, it remains a challenge to obtain a balanced dataset with a large number of samples in both classes, owing to the low prevalence of AF in the general population [5].

Data augmentation has become a standard approach to handle the sample-imbalance issue in machine learning, particularly in image classification tasks. Not only can it help mitigate overfitting when training supervised learning models [12]–[14], but also can it increase the sample size by generating synthetic samples from real ones so that machine learning models can be developed based on a dataset of limited sample size. However, not many studies have investigated the implementation of data augmentation on PPG data. Gotlibovych et al. [15] applied data augmentation by random selection of raw PPG segments, performing signal processing with scaling and additive shifts, and finally including the augmented segments into training samples to train an AF detection model. The majority of the data in the study were recorded during sleep, so it warrants further validation on how the model performs in ambulatory settings, where PPG signals are more susceptible to artifact. The study only investigated the effect of data augmentation on the model training process and showed that augmented data could help smooth the change of training loss over epochs. However, the impact of data augmentation on the testing performance still needs to be further explored. Another study, PlethAugment [16], implemented a more advanced technique, generative adversarial network (GAN) [17], for PPG data augmentation. Three different conditional GANs were tested on various public datasets for different classification tasks, showing that GAN can help generate realistic PPG signals and improve the performance of PPG-based models. This study also sheds light on the effect of synthetic data on class imbalance and the influence of different ratios of real-world to artificial training data on classification performance. However, the performance comparison of GANs with traditional augmentation techniques, such as shifting and cropping, was not conducted in the study, which is a missed opportunity to inform whether the optimal choice of data augmentation solutions is task-specific. Furthermore, the tasks investigated in the study did not include AF detection.

To investigate suitable augmentation techniques for the PPG-based AF detection task, the present study started with various popular GAN constructs, including deep convolutional GAN (DCGAN) [18], Wasserstein DCGAN (W-DCGAN) [19]. However, we quickly realized that these off-the-shelf techniques do not work well for PPG-based AF detection as they show a very limited amount of improvement over simple data augmentation techniques. Therefore, we propose a new GAN to generate synthetic PPG signals. In this new GAN, we adopted a different generator architecture as well as a new loss function that measures how close a synthetic signal is to a PPG in the frequency domain. The objective of this new GAN is to generate synthetic PPGs with a power spectrum close to the real ones. Therefore, we call this GAN Log-Spectral matching (LSM)-GAN and the new loss function LSM-loss, as shown in Fig. 1. Our results show that the LSM-GAN generates synthetic AF signals with a distribution closest to the real ones and achieves the greatest performance gain in AF detection, compared to the other two GANs and two conventional data augmentation techniques. The main contributions of this paper include:

- To our knowledge, this is the first work that incorporates spectral information into the loss function for data augmentation of PPG signals. And we are the first to consider using GAN to generate synthetic PPG data for the task of AF detection.
- A weighting mechanism is implemented to balance the LSM-loss and adversarial loss, and an algorithm to automatically search for the optimal weighting parameter is proposed.
- Besides testing on the internal dataset, we evaluate trained models from each augmentation method on two public PPG datasets, which validates the generalizability of the proposed LSM-GAN.

The rest of this paper is organized as follows: section 2 describes prior related works on AF detection and physiological signal augmentation. Details of our experiments, including datasets and training procedures, as well as our proposed method, are provided in Section 3. Experiments are presented in Section 4, followed by a Result section 5, then discussion and summary of our work in Section 6.

II. Related Work

A. Traditional data augmentation techniques for PPG

To realize its full potential for PPG-based AF detection, deep learning algorithms as reviewed above require a large amount of training data. Because the available datasets are not always enough in quantity and diversity, various data augmentation methods are explored to tackle the data shortage problem. Soonil et al. [9] implemented a 20-second overlap between consecutive PPG segments when splitting the continuous PPG recording, which can be considered as a basic data augmentation method. Although it enlarges the training sample size, no new information is introduced into the training data. Similar to the work [15], Cheng et al. [22] adopted three traditional data augmentation methods: scaling, adding Gaussian noise, randomly changing the amplitude, and random combinations of these three methods are applied to the PPG signal to enlarge the training sample size. These

methods might either disrupt the signal continuity, introduce non-physiologic patterns into the waveform, or even alter the waveform into the opposite class, which will introduce additional label noise into the training set and compromise the model performance. Andrius et al. [23] developed a model to simulate AF-PPG signals from ECG. The model takes the RR intervals calculated from ECG as input and characterizes pulse width, amplitude, and scale of PPG signals through a linear combination of one log-normal and two Gaussian functions. Therefore, the model is capable of simulating PPG signals during AF or with premature beats. Although results demonstrate that the synthetic PPG is close to the real PPG morphologically, it remains to be studied whether those synthetic PPGs are helpful for downstream tasks such as AF detection. Another study [25] managed to model the PPG with modeling the finite element method and monte carlo. However it can only generate single pulse, the change of interval between consecutive pulses can not be modeled. Mazumder et al. [27] propose a physical model of the cardiac system to generate synthetic PPG, which considers pathophysiological features. They also conduct the comparison experiment with DCGAN [18] and claims the GANs is less effective to help improve the downstream classification task. Qunfeng et al. [28] develop a matlab toolbot to generate synthetic PPG template with regular, irregular, fast rhythm and motion artifacts.

B. GAN based PPG synthesis

Besides the traditional augmentation techniques mentioned above, several studies have developed GANs to generate synthetic PPG signals. Heean et al. [24] introduced a GAN to generate high-quality PPG signal from the simultaneously recorded ECG. The architecture contains a Bi-LSTM based generator and 1D-CNN based discriminator. However, the proposed GAN can only take 1-second ECG and generate 1-second PPG at a time. One has to stitch consecutive one-second of PPGs in order to get a longer duration signal. In SynSigGAN [29], a GAN model was proposed to generate four kinds of physiological signals (electrocardiogram (ECG), electroencephalogram (EEG), electromyography (EMG), PPG). In the preprocessing stage, each signal goes through a discrete wavelet transformation and an inverse discrete wavelet transformation, and the signal denoising process takes place in between. As the last part of preprocessing, automatic segmentation is applied to set the length for each type of physiological signal from GAN. Again, no downstream use case is investigated to evaluate the efficacy of the synthetic signals. Seyed et al. [30] proposed a cycleGAN [31] based approach to generate PPG signal for respiratory rate (RR) estimation. A novel loss function was introduced, which takes the RR of synthetic signal into account. Results showed that, by adding the synthetic PPG signals, the accuracy of RR estimation outperformed other state-of-the-art methods using an identical experiment setting and dataset. The study suggests that introducing task-related information into the loss function is a promising solution to improve many GAN-based tasks.

III. Methods

A. Data Sets

Training data: Continuous fingertip PPG (fPPG) recordings were collected with pulse oximeters from 126 in-hospital patients aged between 18 and 95 years (median 63) who were admitted to UCLA Medical Center between April 2010 and March 2013. AF episodes

were annotated by one board-certified cardiac electrophysiologist by marking the start and end of each episode based on co-registered ECG recordings. In the training data, 104 of 126 patients (83%) had recordings with Non-AF rhythms, 14 of 126 patients (11%) had recordings with persistent AF rhythms, and 8 of 126 (6%) patients had recordings with mixed (AF and non-AF) rhythms. Continuous PPG recordings were divided into consecutive non-overlapping 30-second records. Each record was labeled “AF” or “non-AF” depending on if it was extracted within or outside an AF marked episode, respectively. The IRB authorized the retrospective use of this dataset for this investigation with a waiver of patient permission (IRB approval number: 16-18764).

Testing data—A set of continuous PPG data were collected from wearable devices (Empatica E4) worn by 13 acute stroke patients admitted into the neurological intensive care unit (NICU) of UCSF Medical Center between October 2016 and January 2018. Patients’ age was between 19 to 91 (median = 73.5). Within these patients, 8 of the 13 (61%) patients had recordings with AF episodes. The collected PPG recordings lasted between 3h and 22h (median = 10.5h). With the same method used in the training data, the continuous signals in the test set were segmented into consecutive non-overlapping 30-second segments (5831 in total). A subset of the test set (2683 out of 5831 segments) with signal quality labels was built where a congruent agreement was obtained among three annotators who labeled the records with respect to signal quality (good vs. bad). The details on signal quality annotation can be found in the work [34]. Table 1 shows the distribution of AF and Non-AF segments in the training and testing sets. The protocol to collect the test set was approved by UCSF IRB, and written informed consent was obtained from all patients.

Artifact segmentation: Each 30-second record from the test set was annotated to identify onset and offset of each artifactual region following the protocol described in the work [34]. The proportion of artifacts was calculated by summing up the duration of all segments in a PPG record that were considered artifacts and divided by the total length (30 seconds) of the signal. Artifacts within 30-sec PPG signals were highlighted in examples shown in Figure 2.

AF Annotation: The testing set was annotated with respect to AF presence by seven clinicians as described in a previous study [34]. Guided by 7-lead ECG recordings, simultaneously recorded PPG signals were labeled ”AF”, ”Not AF”, or ”Not Sure” if they were respectively identified to contain AF rhythm, other rhythms other than AF, or ambiguous/unidentifiable rhythms. Figure 3 shows the distributions of the number of records with respect to artifact proportions for AF and Non-AF conditions.

B. Baseline data augmentation methods

Data-copying: As the most straightforward data augmentation method, data-copying simply duplicates randomly selected signals and then adds them into the training set.

Permutation: A 30-second PPG signal is divided into five equal-length sub-segments, then all five sub-segments are rearranged by randomly permuting their orders and concatenated to form a new 30-second signal.

Deep convolutional generative adversarial network (DCGAN) and DCGAN with Wasserstein distance (W-DCGAN): GAN has two components [16]: a generator network G and a discriminator network D, as shown in Fig. 1. G receives a random signal z and generates a 'fake' PPG signal. D is a binary classifier to determine whether the input signal is a real or fake PPG. The training process of GAN is a zero-sum game, where generator and discriminator aim to minimize the objective function:

$$\min_{G}\max_{D} V(D,G) = E_{x\sim p_{data(x)}}[D(x)] + E_{z\sim p_{z(z)}}[\log(1 - D(G(z)))] \tag{1}$$

where x is the real PPG signal and G(z) is the synthetic data generated from random signal z. D(x) and D(G(z)) are the discriminator's estimates of the probability of x and G(z) being real, respectively.

Convolutional neural network (CNN) based models are capable of learning good feature representations of the input data, leading to state-of-the-art performance in many classification tasks. Compared to vanilla GAN, DCGAN adopts the structure of deep CNN to improve the quality of generated signal and accelerate the converging process.

The Wasserstein deep convolutional generative adversarial network (W-DCGAN) is a further extension of DCGAN. In DCGAN, we only change the model structure and keep the same cost function as vanilla GAN. By introducing the Wasserstein distance, W-DCGAN not only improves the training stability but also has a cost function that is related to the quality of the generated signal. The new cost function is

$$\min_{G}\max_{D} L_{WGAN}(D,G) = -E_{x\sim p_{data(x)}}[D(x)] + E_{z\sim p_{z(z)}}[(D(G(z)))] \tag{2}$$

C. Proposed Architecture

Integrating additional loss into the cost function of the generator has been proven to be helpful for generating synthetic biomedical signals [29]. Also, in the speech signal area, introducing information from spectral-domain has brought significant improvement in speech recognition and speech waveform generation tasks [32], [33]. In the present study, we propose to include spectral information from PPG waveforms into the cost function for GAN. The hallmark of AF compared to normal sinus rhythm lies in the irregular irregularity in the rhythm. Therefore, synthetic signals that retain spectral characteristics of real ones will likely improve the model performance. Additionally, because our AF detector and most reported deep neural network approaches process PPG signals in time domain, matching in spectral domain still allows randomness in the phases of synthetic signals and hence enriches training data in a profound way. We hypothesize that such randomness will be beneficial for training AF detectors.

GAN is composed of two neural networks: **generator(G)** and **discriminator(D)**, the generator takes random noise as input to generate synthetic signals. However, the length

of the input noise is an arbitrary choice and few studies have investigated its influence. In the present study, instead of following a conventional choice – 100 – as the dimension of random noise, we choose an identical length to the output dimension as the input (1200 in this study). To achieve this, we build a new architecture for the generator which is able to generate the synthetic signal with an identical length as input. The comparison of two different generators is shown in Fig.4 . To better distinguish GANs with different generator architectures, we use DCGAN-100 and DCGAN-1200 to name two baseline methods with different input dimensions in later sections

D. LSM-Loss

We divide a PPG strip into successive blocks and calculate the periodogram of each block separately. To achieve this, we define two distances: matching distance M and self-consistency distance C. Matching distance is designed to measure the difference between blocks in real and synthetic signals with the same index. Self-consistency distance measures the difference between blocks within one synthetic signal. Two distances are defined as:

$$M_i = \|log(psd(b_i)) - log(psd(\hat{b}_i))\|^2 \tag{3}$$

$$C_{i,j} = \|log(psd(\hat{b}_i)) - log(psd(\hat{b}_j))\|^2, i < j \tag{4}$$

where $\hat{b}_i$ and b_j represent i-th block from synthetic and real PPG, respectively. At the same time, $\hat{b}_i$ and $\hat{b}_j$ are the two different blocks within one synthetic PPG, $i, j \in \{1, 2, 3, \ldots, N\}$ and N is the number of blocks for one PPG. psd (•) is the function that returns the magnitude of spectrum for the input time sequence. As illustrated in Fig 5 , the matching distance is calculated by the L2 norm between the spectra of a matched real and a synthetic PPG signal block. Self-consistency distance is calculated by measuring the L2 norm among different blocks of a synthetic PPG signal itself. LSM-loss also considers aggregating M and C for blocks in one PPG through a weighted paradigm. However, directly averaging may not be the best approach to aggregate each block. For example, irregular pulses can exist anywhere in one AF-PPG segment, and the spectral value of blocks having irregular pulses would be significantly different from other blocks. But averaging all blocks will dilute the contribution from those irregular pulses. To avoid this situation, we utilize the aggregation function $F \in \{Mean, Max\}$ as a hyperparameter to accommodate signals from AF or Non-AF. The adversarial loss is defined as,

$$L_{adv}(G, D) = E_{z \sim N(0,1), x \sim p(data)}\left[(1 - D(G(Z)))^2\right], \tag{5}$$

where z represents the input noise, and G(z) is the synthetic signal. The final LSM-loss is defined as the linear combination of three losses described above,

$$L_{lsm} = L_{adv}(G, D) + \lambda_1 \times F(M_1, \ldots.., M_N) + \lambda_2 \times F(C_{1,2}, C_{1,3}, \ldots.) \quad (6)$$

Where N represents number of blocks for one PPG, λ_1 and λ_2 are the hyperparameters balancing the three losses, and F is either Mean or Max function, in order to better aggregate losses from each block

E. Hyperparameters selection

There are three hyperparameters in equation 4 that need to be optimized including λ_1 and λ_2 and F. We design the loss function in this flexible way because we anticipate the different choices of hyperparameters in this loss function would be needed to accommodate training different GAN to generate AF vs Non-AF signals. For an AF PPG, self-consistency loss would be less important but the matching of spectra at a block level would be critical. On the other hand, for a non-AF PPG (most of which correspond to sinus rhythm), self-consistency would be needed to ensure a more realistic artificial signal Different from many other studies which set the weights manually,

Algorithm 1

Hyperparameter selection process for LSM-GAN. We select k = 300 samples from each signal set. ***Least distance*** = infinity, ***Best set*** = {}

for λ_1 in [0,3] with step size 0.1 **do**

  for λ_2 [0,3] with step size 0.1 **do**

    for F in {**Mean, Max**} **do**

- Sample k signal $x^1, \ldots, x^k$ from data generating distribution $p_{data}(x)$
- Sample k noise samples $z^1, \ldots, z^k$ from noise prior $p_g(z)$
- Generate $\hat{x}^1, \ldots, \hat{x}^k$ synthetic signals with trained LSM-GAN G (z) with $\{\lambda_1, \lambda_2, F\}$
- Calculate the MMD Distance

$$\textbf{current} = \frac{1}{k}\sum_{i=1}^{k} MMD\,[autocorrelation\,(x^i), autocorrelation\,(\hat{x}^i)]$$

- **Best set** = $\{\lambda_1, \lambda_2, F\}$ if ***current*** < ***Least distance***

    end for

  end for

end for

Return the ***Best Set***.

we select the three hyper-parameters λ_1, λ_2 and F by optimizing a guided grid search process, as shown in algorithm 1. First, we choose a large range of [0, 3] for λ_1 and λ_2 with a step size of 0.1, and Mean, Max for F. Second, for each combination of [λ_1, λ_2 ,F], a set of 300 synthetic PPG signals from LSM-GAN will be generated. Another set of 300 real signals will be randomly selected from the training set. Third, autocorrelations are calculated for both real and synthetic signals, then the maximum mean discrepancy (MMD) distances

of autocorrelations. between the real signal set and each of the synthetic signal sets are calculated. Lastly, the best combination of [λ_1, λ_2,F] will be chosen based on the shortest MMD distance between synthetic signals and real signals.

F. Preprocessing and Training

Raw PPG signals collected in this study are at 240 sampling frequencies, we first down-sampling the PPG signals to a sampling rate of 40 Hz. Then, we apply a band-pass FIR filter with a pass-band frequency of 0.9 Hz and stop-band frequency of 5 Hz on the PPG signals. Finally, the min-max normalization is performed on PPG segments to ensure all signals are in the same scale.

The proposed LSM-GAN and other GANs are trained from scratch on AF signals and Non-AF signal separately. We use 200 as the batch size and then train the GANs for epochs at a learning rate of 0.001. The learning rate decays 0.0001 for each epoch. Early stopping is performed when the loss on validation does not improve in 6 epochs. In this study, we used 10% of the training data as validation set, and the augmented training data are also included. We trained each classifier 5 times and made sure that the PPG signals in validation sets are not overlapped for each time.

The Adam optimization algorithm and cross-entropy loss function are used to train the ResNet-50 with 512 mini-batch size, 50 epochs, and a learning rate of 0.001. To avoid overfitting, we also employed an early stopping procedure which stops the training procedure if the validation loss does not improve in 6 epochs. The GANs are implemented using Pytorch and Resnet-50 is implemented in Keras using TensorFlow backend, and the experiments were performed on a workstation with one NVIDIA RTX 1080Ti GPU and 64 GB memory.

IV. Experiment design

Three experiments are designed to evaluate the proposed LSM-GAN against baseline methods. Each experiment tests a specific aspect of practical relevance when considering a GAN-based data augmentation strategy, including inter-class balancing, resilience to artifacts and the training sample size. Resnet-50 is used here as the classifier for all the experiments.

Experiment 1:

A well-justified use of GAN is to focus on augmenting data from minority classes to balance the numbers of samples across different classes. In our application, the original training data is highly unbalanced, in which AF is the minority class and Non-AF is the majority class. The ratio between AF sample size and that of Non-AF is approximately 1:7. To investigate whether a more balanced training data would help improve the final classification accuracy, we augmented only the AF cases by 6 folds to achieve the inter-class balance in the first experiment. The proposed LSM-GAN and baseline models are all trained based on the same set of AF data in the training set and then the resultant models are used to generate 212,423 synthetic AF-PPG strips per each model to augment the training data. Various performance metrics, including accuracy, sensitivity, specificity, positive prediction

value (PPV) and negative prediction value (NPV), are used to compare the classifiers that are trained with the augmented dataset from different data augmentation approaches.

Experiment 2:

The issue of signal quality cannot be overlooked for PPG-based studies, which is especially true in ambulatory settings. Therefore, we dedicate this experiment to evaluating the performance of different augmentation techniques at various levels of artifacts by investigating the relationship between F1 scores and the proportion of artifacts within the PPG signals. We split the testing set into four groups based on the percentage of artifacts: clean (0%), (0% - 20%), [21%–60%) and [60%–100%]. Then we pick models from experiment 1 for each augmentation method and test them on those four groups separately.

Experiment 3:

From the results of experiment 1, we observe that an AF detector benefits from a balanced training set. However, another factor, the total number of training samples, has not been investigated. To test the effect of training sample size, we constructed a series of balanced training sets with an increasing sample size from 300,000 to 2,000,000. For each training set, if the number of required cases (e.g., 15,000 AF cases are needed for the total sample size of 30,000) is less than the original cases, then they are randomly selected within the existing data. When the required case exceeds the original size, additional samples will be generated by different augmentation methods for both AF and non-AF cases. All the trained models are tested on the same four signal quality groups as in experiment 2.

V. Results

A. Experiment 1: Performance comparison between data augmentation methods

Table 2 summarizes the performance of AF detection for different combinations of original and augmented data sets. A cutoff probability threshold of 0.5 was used to calculate different performance metrics.

Compared to training with the original dataset, a balanced training set by simply duplicating all the AF cases 6 times through data-copying would increase 11% of accuracy and 25% of sensitivity. Permutation, DCGAN and W-DCGAN achieve similar performance gain to data-copying, while the proposed LSM-GAN offers the most performance improvement, with a 24% gain in accuracy and 32% in sensitivity with around 1% reductions in specificity and PPV. We notice that traditional data augmentation methods and basic GAN models with conventional generator achieve similar performance, which all arrive at an accuracy of around 93%. With the new architecture design for the generator, DCGAN-1200 and WDCGAN-1200 both perform better than the ones with the conventional generator. Built on top of the new architecture of the generator, the proposed LSM-GAN integrates the LSM-loss component and offers an additional 3% improvement in accuracy and 5% in sensitivity over other GANs.

Table 3 summarizes the hyper-parameters selected by the guided grid search. For AF signal, match-loss and self-consistency loss share the same weight of 1.5. This result indicates that

two losses play a similar role when generating realistic AF signals. For Non-AF signals, self-consistency loss weights more than the other two losses, which is expected since the key characteristic of non-AF (Sinus rhythm most of the time) signals is periodicity. And the weight of 0.8 indicates the matching-loss is less important than the other two loss terms. In terms of F, results show that selecting the block with maximum loss value can help generate more realistic non-AF signals, while taking the average of all blocks is better for generating AF signals

Table 4 reports the performance for classifiers trained on augmented data generated from LSM-GAN under different hyperparameters. Two different hyperparameter set are investigated, {1,1,Mean} and {2.2,0.6,Max}. {1,1,Mean} represents no weight mechanism which three loss terms are treated equally, and {2.2,0.6,Max} is selected by the largest MMD distance calculated according to Algorithm 1. Results show that synthetic signals generated from LSM-GAN with optimized hyperparameters will help increase 1.5% accuracy compared to LSM-GAN without weight mechanism is introduced and help increase 3% in accuracy compared to LSM-GAN with the worst hyperparameter set.

B. Experiment 2: Resilience to artifacts

In addition to overall performance gain, the second experiment investigates the relationship between signal quality and the performance of different augmentation methods. Fig. 6 compares each method's performance across the presence of different proportions of artifacts in PPG signals. F1 score is used because AF and Non-AF classes become unbalanced within each signal quality group.

It can be seen that all methods perform poorly in the poor-quality group (more than 60% of artifacts), although substantial performance improvement can still be observed by different augmentation methods compared to the original dataset. On the other hand, all methods can achieve over 90% F1 score for the excellent quality group (0% of artifacts), especially for LSM-GAN which achieves 99% F1 score (Sensitivity : 0.98%; Specificity : 0.99%; PPV : 0.99%; NPV : 0.98%).

We can observe that F1 score decreases with the increasing artifacts portion in PPG, except Permutation, it achieves a F1 score of 96% in signal group (0-20%] which is better than 95% in the perfect signal quality group. Also, model trained on the original dataset has the least performance at each signal quality group, with F1 scores of 91%, 76%, 73% and 38%. At the same time, LSM-GAN achieves 99%, 96%, 90% and 72% separately in each signal quality group, which improves 8.%, 26%, 23% and 89%, respectively. Other baseline methods also show improvement but not as significantly as LSM-GAN. However, Permutation achieves 75% of F1 score in the signal quality group (60-100%], which is higher than 73% obtained by LSM-GAN. Also, Permutation has better F1 score compared to baseline GANs except in the signal quality group (20%-60%].

C. Experiment 3: Data augmentation at different sample sizes

To evaluate the effect of training sample size on the model performance, experiment 3 is conducted. A series of training sets with an increasing sample size from 300,000 to 2,000,000 is constituted. The same Resnet-50 is trained based on each training set separately

and tested on the same four signal quality groups as experiment 2. This process is repeated for all augmentation methods. Fig. 7 compares the F1 score from models trained with data of different sample sizes that are generated with different data augmentation methods. We can observe that in Fig. 7a when the signal quality is perfect, all the methods have a clear increasing trend with the added number of training samples except Data-copying. However, when the signal quality gets worse, in Fig. 7b, only WGAN-1200 and LSM-GAN still keep the increasing trend. Moreover, only LSM-GAN shows the steady uptrend in Fig.7c and 7d, where the signal quality is even worse and baseline methods have no positive relationship with added sample size. Another interesting observation is Permutation (orange curve). In Fig. 7a, Permutation has the least F1 score most of the time. While in Fig. 7(b)–(d), when there are artifacts in PPG signals, Permutation shows great resilience to artifacts and keeps the leading performance along with LSM-GAN.

D. External validation

A public dataset (DeepBeat dataset) released in 2020 [20] with both signal quality and AF annotations was adopted to test the generalizability of the proposed approach. A data harmonization process was designed in the study, given the following three differences between our data and the DeepBeat data. First, the DeepBeat data is collected from wrist-type watches, while our data is collected from fingertips in the ICU setting. Second, the signal length of one segment is 25 seconds with a sampling rate of 32 Hz for DeepBeat, while ours is 30 seconds with a sampling rate of 240 Hz. Third, the preprocessing details were not reported clearly in the original paper and may differ from ours. Based on the above discrepancies in data, we first extended the 25-second segment to 30-second by stitching the first five seconds to the end of each signal (which may cause phase discontinuity). We then upsampled the signal length in the DeepBeat test set to the same as ours and adopted the min-max normalization on the data. Models from experiment 3 for each data augmentation method were selected to test on the DeepBeat dataset. A summary of performance can be found in Fig.8 (a). Among our models, the proposed LSM-GAN consistently shows a leading performance in terms of F1 score, which is improved by 15% compared to data-copying.

The second external dataset was shared with us by authors in the work [35]. They selected around 60 hours of PPG and synchronized ECG from 60 patients (containing both AF and non-AF patients) in the MIMIC-III waveform database. All the PPG signals were annotated by cardiologists from Guilin Medical University. However, PPG signals were segmented into 10-second strips in the study. To accommodate the dataset to our model, we repeat each 10-second signal two times, producing 30-second signals. Then same models from experiment 3 are tested on this dataset, results are reported in Fig 8 (b). Although the proposed LSM-GAN does not lead at the early stage, it presents a consistently increasing trend with added samples and eventually arrives at the best F1 score across all data augmentation techniques.

E. Visualization

To further characterize signals generated from different GANs, an additional experiment was conducted to visualize the distribution of the synthetic data generated by different GANs for both AF and Non-AF cases. We first randomly select 300 real signals and 300 synthetic

signals from each GAN, then use Pairwise Controlled Manifold Approximation Projection (PaCMAP) [36] to visualize the real signals together with synthetic signals in a 2D fashion. Fig. 9 presents the distribution of real and synthetic signals through PaCMAP calculated from signal amplitudes for both AF and Non-AF. Four sub-plots in each row contain the same dots, and each sub-plot only highlights dots from the related category.

Compared to the distribution of real AF signals, DCGAN and W-DCGAN only capture part of the distribution of real AF signals, and they do not share much overlap. On the other hand, the proposed LSM-GAN generates signals with similar distribution as real AF signals. Similar pattern can be observed in the Non-AF situation, DCGAN and WGAN only learned partial distribution of real Non-AF signals, while signals generated from LSM-GAN are more dispersed and distribute in a way similar to real ones.

VI. DISCUSSION

We proposed a novel GAN-based data augmentation technique, LSM-GAN, that offers several innovations in the architecture design and achieves the best data augmentation effect. First, LSM-GAN integrates spectral information from the PPG waveform into the loss function to train generator in addition to the commonly used cross-entropy loss. Second, LSM-GAN uses a new architecture of generator which avoids up-sampling and aims to mimic a more well-understood filtering process to transform a white-noise into a narrow band signal like PPG. With such a design, LSM-GAN aims to generate PPG signals that possess main characteristics of real PPG at the population level and truly enrich training data to train PPG-based AF detection algorithms. The performance of LSM-GAN for AF detection was compared with six baseline data augmentation methods, including two traditional approaches: data-copying and permutation, and four other GAN-based approaches: DCGAN, W-DCGAN and their variants with the new generator architecture. Three experiments were conducted to probe various key factors in the PPG-based AF detection task, including inter-class sample balance, resilience to noise, and training sample size. The generalizability to the external test sets is also evaluated to further establish the efficacy of LSM-GAN. The analysis of the results is discussed as follows:

A. Introducing spectral information in generated samples improves the AF detection accuracy

Among all the data augmentation models reported in table 2, the proposed LSM-GAN achieves the best performance in accuracy and sensitivity with less than 1% reduction on specificity and PPV. One plausible reason is that PPG signal is only considered in time domain in previous reported networks, including our AF model. The randomness of phases of synthetic signals is allowed when matching in the spectral domain. Such randomness will enrich the training data and hence is beneficial for training AF models. This hypothesis can be partly supported by the PaCMAP visualization of generated signals from different data augmentation techniques in Fig. 9, which reveals the PPG signals generated with LSM-GAN present a more similar distribution to the real signals than the other two GANs.

B. Increasing the dimension of random noise for GAN helps improve the AF detection accuracy

Instead of directly adopting a conventional choice of the input length (100) for the generator of GAN, we proposed a new generator architecture, which outputs the same length of the signal as the input (1200 in this study). To evaluate the effect of the new architecture, we added two more models: DCGAN-1200 and W-DCGAN-1200, which only changed the generator compared to the original DCGAN and W-DCGAN. Results from Experiment 1 (see Table 2) show that the new architecture does provide performance gain, as evidenced by the performance improvement from DCGAN-100 to DCGAN-1200 and W-DCGAN-100 to W-DCGAN-1200. The key differentiator between the conventional architecture and the proposed architecture is whether an extra upsampling step is needed to ensure the length of a synthetic signal to be equal to a desired value. Learning GAN is essentially learning a series of transformations that convert random input noise into realistic synthetic data. In this study, when the generator does not alter the length of the input noise, the learned transformations can be better explained by a more well-understood filtering process. However, interpreting deep networks is still an ongoing effort and it remains interesting to uncover characteristics of the filters that are learned by LSM-GAN.

C. Performance of data augmentation versus signal quality

As reported in Fig. 6, the increasing artifacts in PPG signals have a negative influence on the AF detection performance. Because the AF model we used in this study – Resnet-50, is a generic model and we did not add any specific modification to handle poor quality signals. Because our training dataset also contains imperfect PPG signal strips that are from both AF and non-AF classes, it is likely that the trained classifier will be confused when the artifacts in the signal are learned as a pattern to be randomly associated with either AF or non-AF. However, the performance reduction can be alleviated through data augmentation methods. One plausible reason is because we only augmented good quality AF signals, in which the AF pattern is clear. Through data augmentation, real and clear AF patterns can be enhanced in the training data which reduce the negative influence of artifacts.

D. LSM-GAN maintains superiority on external datasets

In Fig. 8, after evaluating our classifiers on two external datasets, we observe that our proposed LSM-GAN helps improve performance compared to using original data to train the model, and we can also observe the superiority of LSM-GAN compared to other data augmentation methods. However, there is a performance reduction on these datasets compared to results reported in the original publications.Exact reasons for this difference cannot be established and they are not the focus of this study. However, because our study shows that performance of a generic deep neural network architecture as used in this study is particularly sensitive to the quality of a PPG signal, we speculate that DeepBeat database may contain large portion of poor quality PPG. Furthermore, DeepBeat algorithm explicitly incorporates PPG signal quality into AF detection and is expected to perform better on imperfect PPG signals. The original algorithm that was developed and tested on the MIMIC dataset did not consider PPG signal quality but it was developed to process 10-second PPG

strips while our network was designed for processing 30-second PPG strips, which may be the likely reason for differences in performances.

E. Limitations and future works

Both AF and Non-AF PPG samples generated from LSM-GAN have a more similar distribution to that of real signals than samples generated from other data augmentation methods, which in turn helps improve the performance of downstream AF classification tasks. However, what an adequate amount of data to train an effective GAN is not investigated. We only test the effectiveness of LSM-GAN on the AF detection task, it remains interesting to test LSM-GAN on other tasks.

The present study focuses on the comparison of different DA techniques, so the same deep learning architecture, i.e., Resnet-50, is adopted to achieve a fair comparison. This goal prevents us from exploring various deep learning models and customizations that may further improve the classification performance.

Moreover, we can observe that from experiment 3 and external validation, although LSM-GAN boosts the performance, it still reaches a plateau of performance after one million training samples. This phenomenon indicates the limitation of current LSM-GAN, which warrants future work that integrates other AF characteristics into the loss design for improvement. In parallel to the proposed GAN-based approach, another plausible future direction worth exploring is through a model-based approach that models various characteristics of AF signals, such as slopes, amplitude fluctuations and their changes corresponding to the change of pulse rate to synthesize AF signals. The limitation of LSM-GAN is also lead from the small size of patient cohort (126 patients in this study), which indicates that more real-world data is necessary to further improve the performance.

VII. Conclusion

In the field of AI health, it remains difficult to obtain both large and well-annotated datasets. Data shortage and class-imbalance issues are standing challenges to properly train high-performing machine learning algorithms. In this study, we showed that properly designed GAN can potentially be used to augment and re-balance training data and improve classifiers solely trained on the original dataset that is imbalanced and contains fewer samples.

Acknowledgement

This work has been partially supported by NHLBI award R01HL166233.

References

[1]. Hindricks Gerhard, et al. “2020 ESC Guidelines for the Diagnosis and Management of Atrial Fibrillation Developed in Collaboration with the European Association for Cardio-Thoracic Surgery (Eacts).” European Heart Journal, vol. 42, no. 5, 2020, pp. 373–498., 10.1093/eurheartj/ehaa612.

[2]. Camm AJ et al. , “2012 focused update of the esc guidelines for the management of atrial fibrillation,” European heart journal, vol. 33, no. 21, pp. 2719–2747, 2012 [PubMed: 22922413]

[3]. Carpenter A and Frontera A, “Smart-watches: a potential challenger to the implantable loop recorder?” Europace, p. euv427, 2016.

[4]. Nemati S. et al., “Monitoring and detecting atrial fibrillation using wearable technology,” in Engineering in Medicine and Biology Society (EMBC), 2016 Annual International Conference of the IEEE. IEEE, 2016

[5]. Pereira T. et al. , “Photoplethysmography based atrial fibrillation detection: a review,” npj Digit. Med, 2020

[6]. Voisin M, Shen Y, Aliamiri A, Avati A, Hannun A, Ng A. Ambulatory atrial fibrillation monitoring using wearable photoplethysmography with deep learning. arXiv preprint arXiv:1811.07774. 2018 Nov 12.

[7]. Aliamiri A and Shen Y, “Deep learning based atrial fibrillation detection using wearable photoplethysmography sensor,” 2018 IEEE EMBS Int. Conf. Biomed. Heal. Informatics, BHI 2018, vol. 2018-Janua, no. March, pp. 442–445, 2018.

[8]. Poh MZ et al. , “Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms,” Heart, pp. 1921–1928, 2018. [PubMed: 29853485]

[9]. Kwon S. et al. , “Deep Learning Approaches to Detect Atrial Fibrillation Using Photoplethysmographic Signals: Algorithms Development Study,” JMIR mHealth uHealth, 2019

[10]. Tison GH et al. , “Passive Detection of Atrial Fibrillation Using a Commercially Available Smartwatch,” JAMA Cardiol, pp. 1–8, 2018.

[11]. Shashikumar SP, Shah AJ, Clifford GD, and Nemati S, “Detection of Paroxysmal Atrial Fibrillation using Attention-based Bidirectional Recurrent Neural Networks,” arXiv, 2018

[12]. Luis P, Jason W. The efectiveness of data augmentation in image classifcation using deep learning. In: Stanford University research report, 2017.

[13]. Lemley J, Barzrafkan S, Corcoran P. Smart augmentation learning an optimal data augmentation strategy. In: IEEE Access. 2017.

[14]. Ekin DC, Barret Z, Dandelion M, Vijay V, Quoc VL. AutoAugment: learning augmentation policies from data. ArXiv preprint. 2018

[15]. Gotlibovych I, Crawford S, Goyal D, Liu J, Kerem Y, Benaron D, Yilmaz D, Marcus G, & Li YI (2018). End-to-end Deep Learning from Raw Sensor Data: Atrial Fibrillation Detection using Wearables. ArXiv, abs/1807.10707..

[16]. Kiyasseh D. et al. , ”PlethAugment: GAN-Based PPG Augmentation for Medical Diagnosis in Low-Resource Settings,” in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.2979608.

[17]. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y, 2014. Generative adversarial nets. Advances in neural information processing systems, 27.

[18]. Radford A, Metz L and Chintala S, 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

[19]. Arjovsky M, Chintala S and Bottou L, 2017, July. Wasserstein generative adversarial networks. In International conference on machine learning (pp. 214–223). PMLR.

[20]. Torres-Soto J and Ashley EA, 2020. Multi-task deep learning for cardiac rhythm detection in wearable devices. NPJ digital medicine, 3(1), pp.1–8. [PubMed: 31934645]

[21]. Das SSS, Shanto SK, Rahman M, Islam M, Rahman A, Masud MM and Ali ME, 2020. BayesBeat: A Bayesian deep learning approach for atrial fibrillation detection from noisy photoplethysmography data. arXiv preprint arXiv:2011.00753.

[22]. Cheng P, Chen Z, Li Q, Gong Q, Zhu J and Liang Y, 2020. Atrial fibrillation identification with PPG signals using a combination of time-frequency analysis and deep learning. IEEE Access, 8, pp.172692–172706.

[23]. Sološenko A, Petrėnas A, Marozas V, & Sörnmo L (2017). Modeling of the photoplethysmogram during atrial fibrillation. Computers in Biology and Medicine, 81(October 2016), 130–138. 10.1016/jxompbiomed.2016.12.016 [PubMed: 28061368]

[24]. Shin H, Sun S, Lee J, & Kim HC (2021). Complementary Photoplethysmogram Synthesis from Electrocardiogram Using Generative Adversarial Network. IEEE Access, 9, 70639–70649. 10.1109/ACCESS.2021.3078534

[25]. Boonya-Ananta T, Rodriguez AJ, Ajmal A, Du Le VN, Hansen AK, Hutcheson JD and Ramella-Roman JC, 2021. Synthetic photoplethysmography (PPG) of the radial artery through parallelized Monte Carlo and its correlation to body mass index (BMI). Scientific reports, 11(1), pp.1–11. [PubMed: 33414495]

[26]. Wu BF, Chiu LW, Wu YC, Lai CC and Chu PH, 2022. Contactless Blood Pressure Measurement via Remote Photoplethysmography With Synthetic Data Generation Using Generative Adversarial Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2130–2138).

[27]. Mazumder O, Banerjee R, Roy D, Bhattacharya S, Ghose A and Sinha A, 2022. Synthetic PPG Signal Generation to Improve Coronary Artery Disease Classification: Study With Physical Model of Cardiovascular System. IEEE Journal of Biomedical and Health Informatics, 26(5), pp.2136–2146. [PubMed: 35104231]

[28]. Tang Q, Chen Z, Menon C, Ward R and Elgendi M, 2021. PPGTempStitch: A MATLAB Toolbox for Augmenting Annotated Photoplethsmogram Signals. Sensors, 21(12), p.4007. [PubMed: 34200635]

[29]. Hazra D, & Byun YC (2020). Synsiggan: Generative adversarial networks for synthetic biomedical signal generation. Biology, 9(12), 1–20. 10.3390/biology9120441

[30]. Aqajari SAH, Cao R, Zargari AHA,]& Rahmani AM (2021). An End-to-End and Accurate PPG-based Respiratory Rate Estimation Approach Using Cycle Generative Adversarial Networks. http://arxiv.org/abs/2105.00594

[31]. Zhu JY, Park T, Isola P and Efros AA, 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2223–2232).

[32]. Kingsbury BE, Morgan N and Greenberg S, 1998. Robust speech recognition using the modulation spectrogram. Speech communication, 25(1-3), pp.117–132.

[33]. Yamamoto R, Song E and Kim JM, 2020, May. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6199–6203). IEEE.

[34]. Pereira T, Ding C, Gadhoumi K, Tran N, Colorado RA, Meisel K and Hu X, 2019. Deep learning approaches for plethysmography signal quality assessment in the presence of atrial fibrillation. Physiological measurement, 40(12), p.125002. [PubMed: 31766037]

[35]. Cheng P, Chen Z, Li Q, Gong Q, Zhu J and Liang Y, 2020. Atrial fibrillation identification with PPG signals using a combination of time-frequency analysis and deep learning. IEEE Access, 8, pp.172692–172706.

[36]. Wang Y, Huang H, Rudin C and Shaposhnik Y, 2021. Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. J Mach. Learn. Res, 22, pp.1–73.

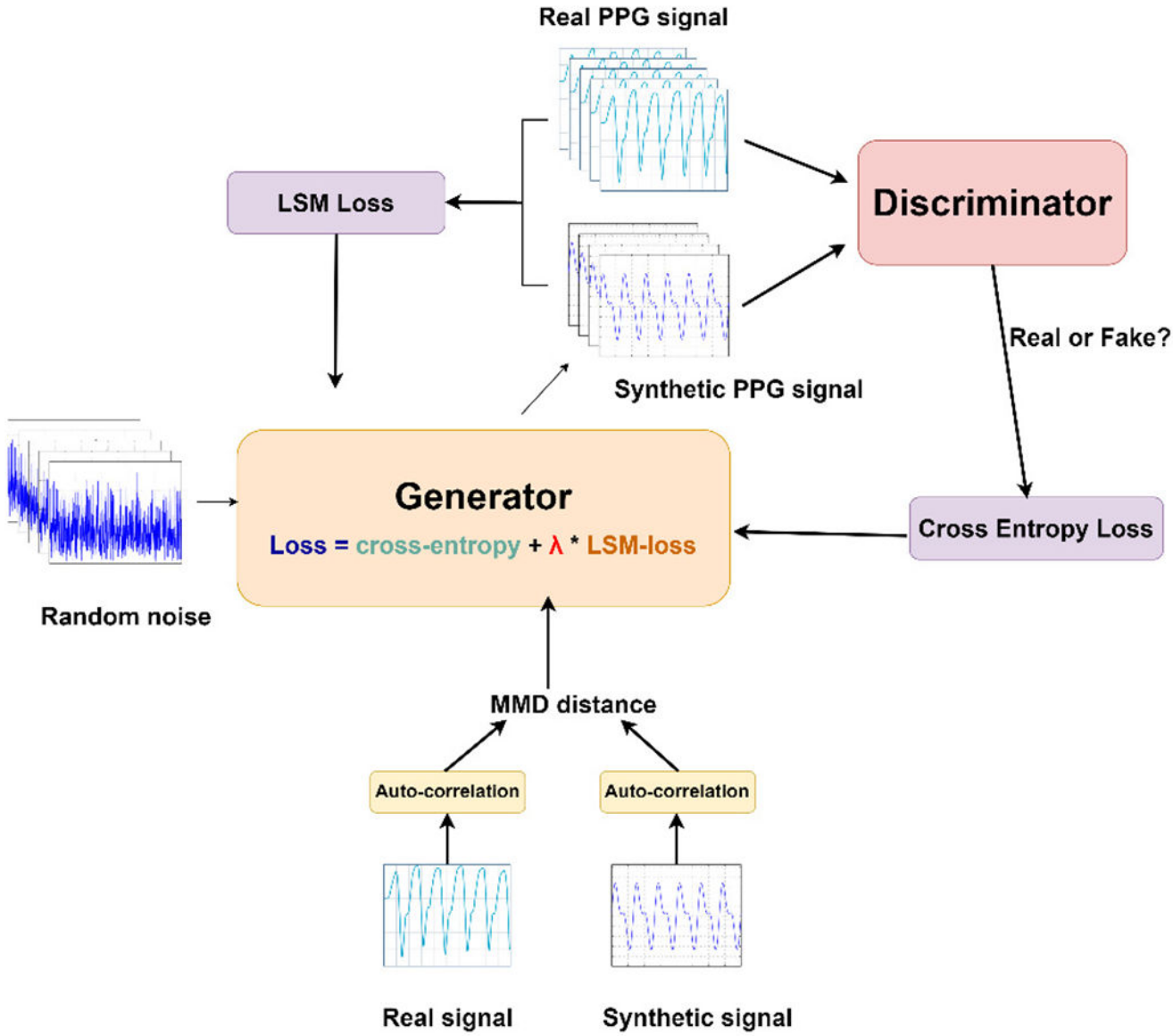


Fig. 1:

The overall workflow of LSM-GAN. LSM-loss is integrated with adversarial loss in a weighting paradigm. The weight parameter is selected by the shortest MMD distance between the autocorrelation of real and synthetic signals.

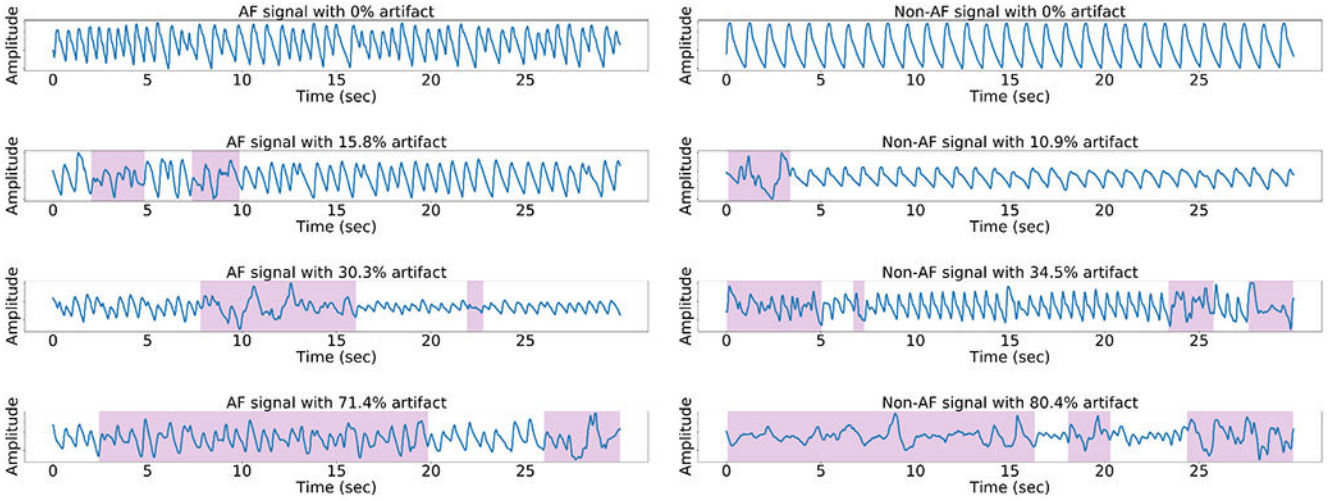


Fig. 2:
PPG segments with artifacts marked in purple.

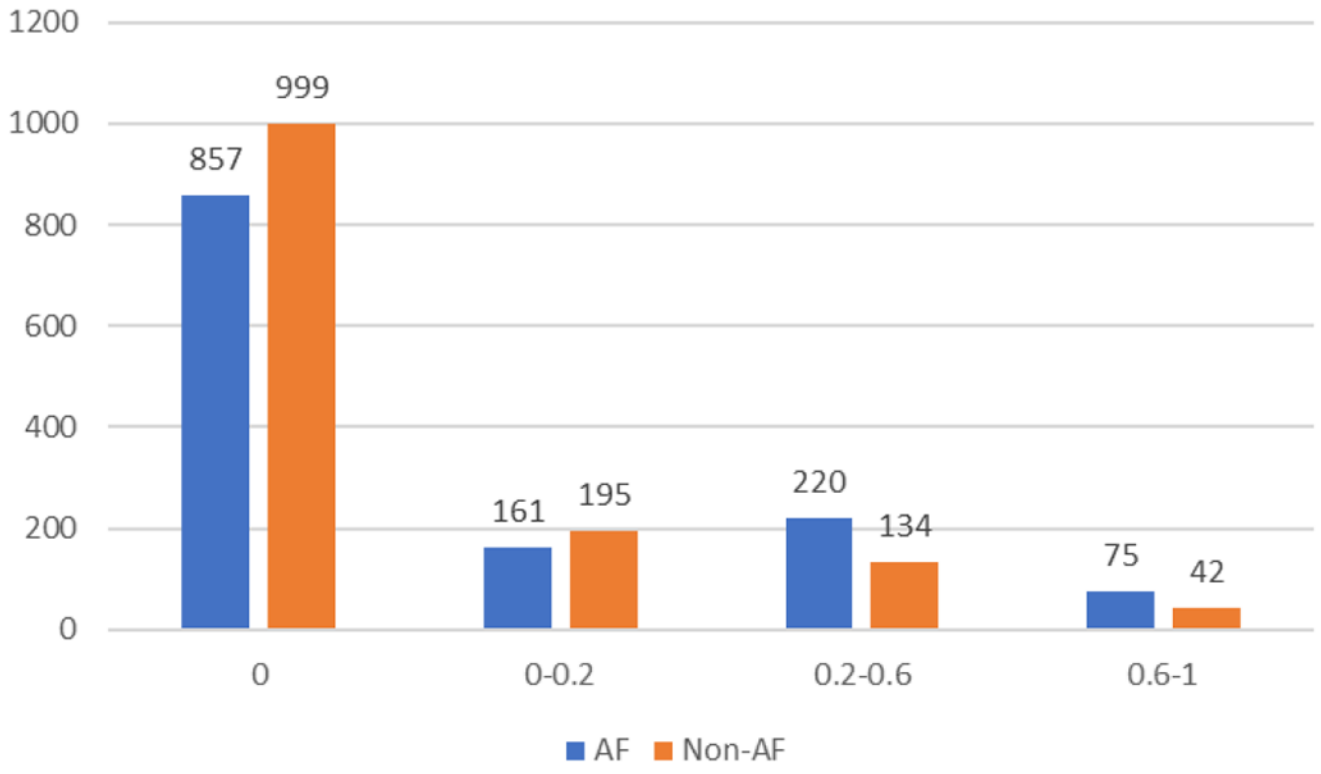


Fig. 3:

Distributions of the number of records with respect to artifact proportions for AF and Non-AF conditions in the testing set. The test set are divided into four signal quality groups: 0%, (0 - 20%], (20% - 60%] and (60% - 100%].

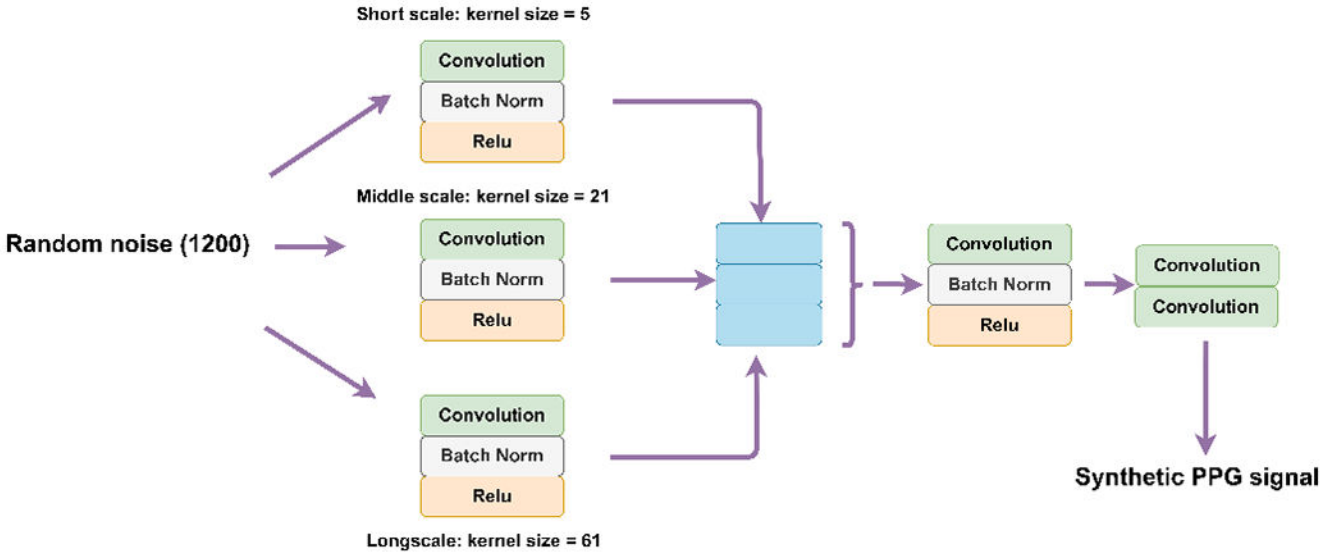


Fig. 4:

The network architecture: discriminator and generators with conventional and new architectures.

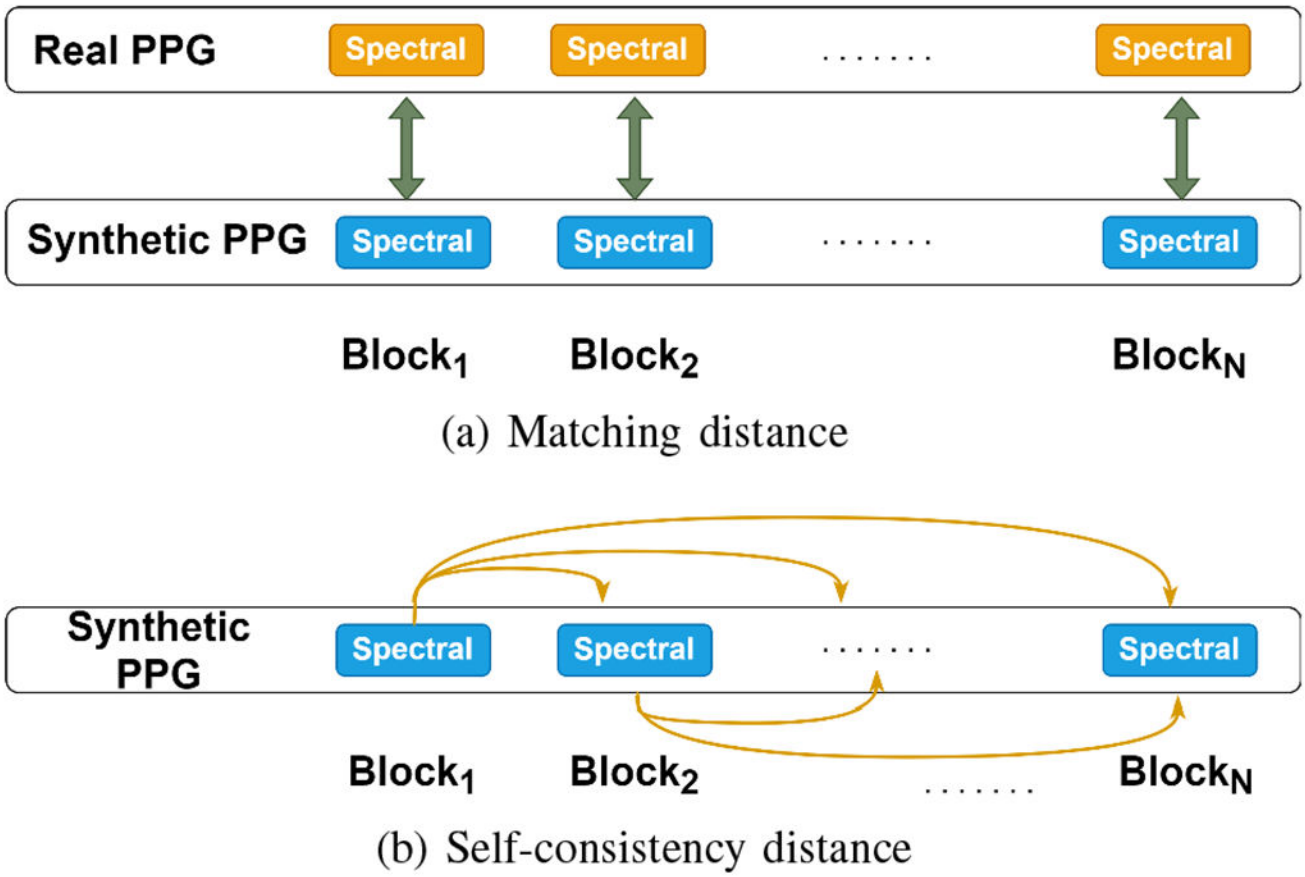


Fig. 5:

The proposed two distances: matching distance (*M*) and self-consistency distance (*C*). Matching distance is designed to measure the difference between blocks in real and synthetic signals with same index. Self-consistency distance measures the difference between blocks within one synthetic signal.

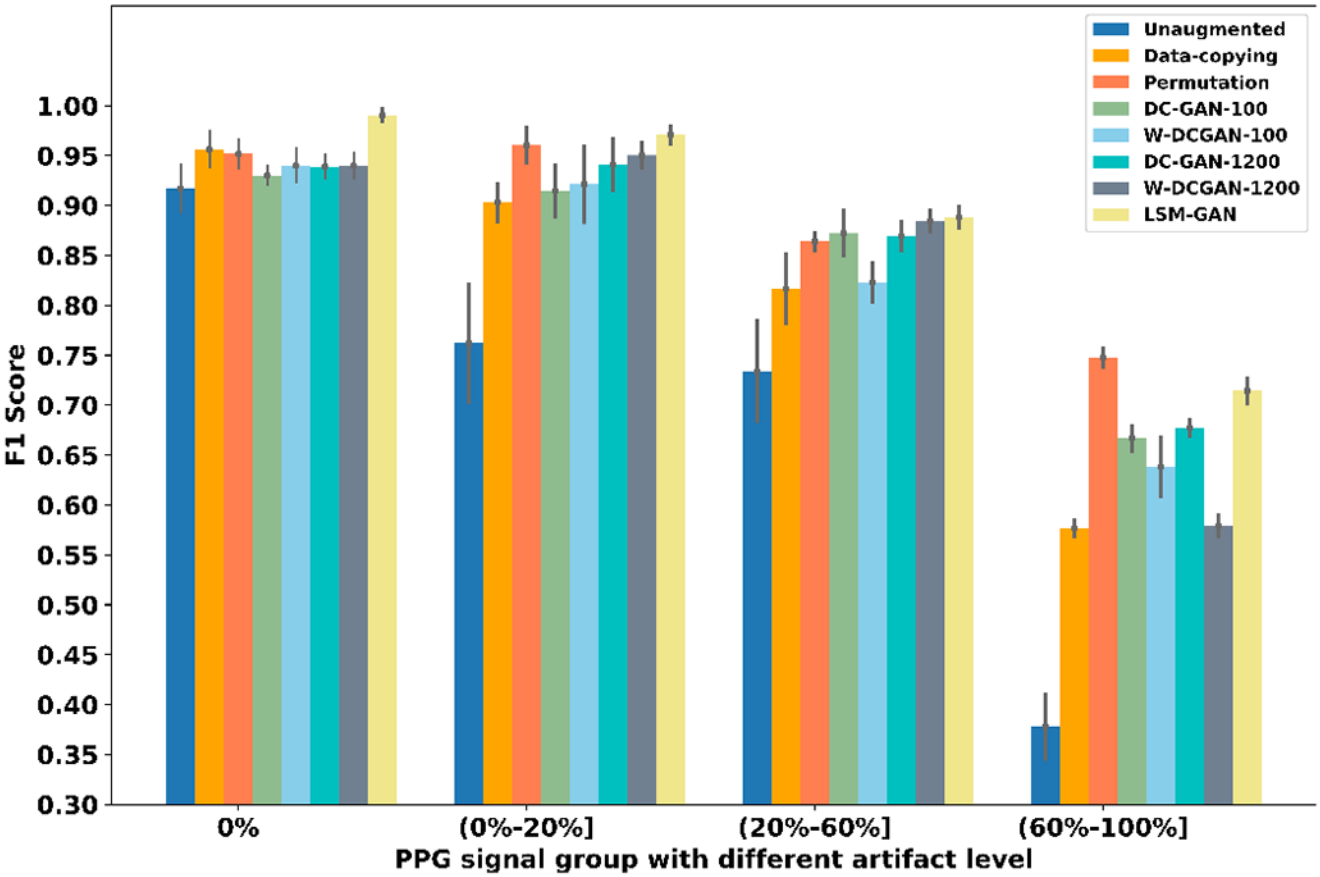
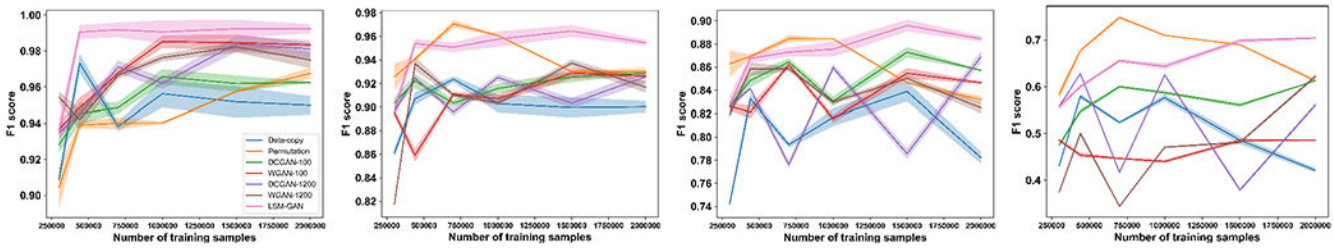


Fig. 6:
Performance tested on PPG records with different percentage artifact level



(a) Signal quality group - 0% (b) Signal quality group - (0% -20%] (c) Signal quality group - (20% -60%] (d) Signal quality group - (60% -100%]

Fig. 7:

Comparison of F1 score for different training sample sizes on different signal quality groups. Four subplots have different scales for *Y*-axis in order to demonstrate the results in a better resolution.

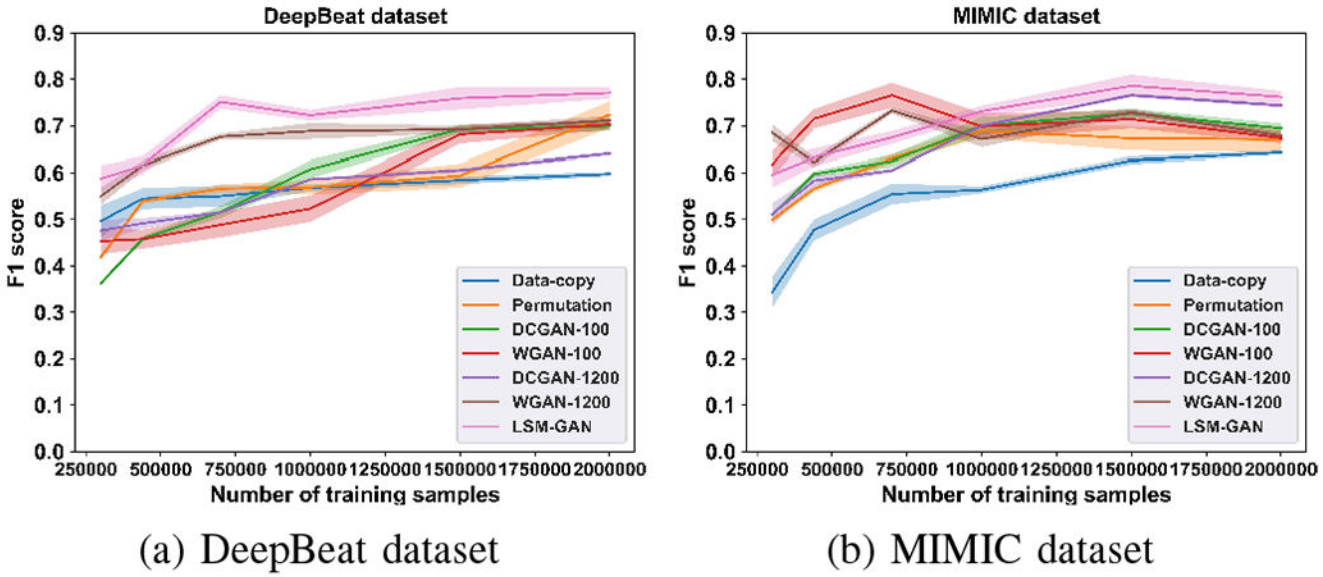


Fig. 8:
External validation of models from Experiment 3

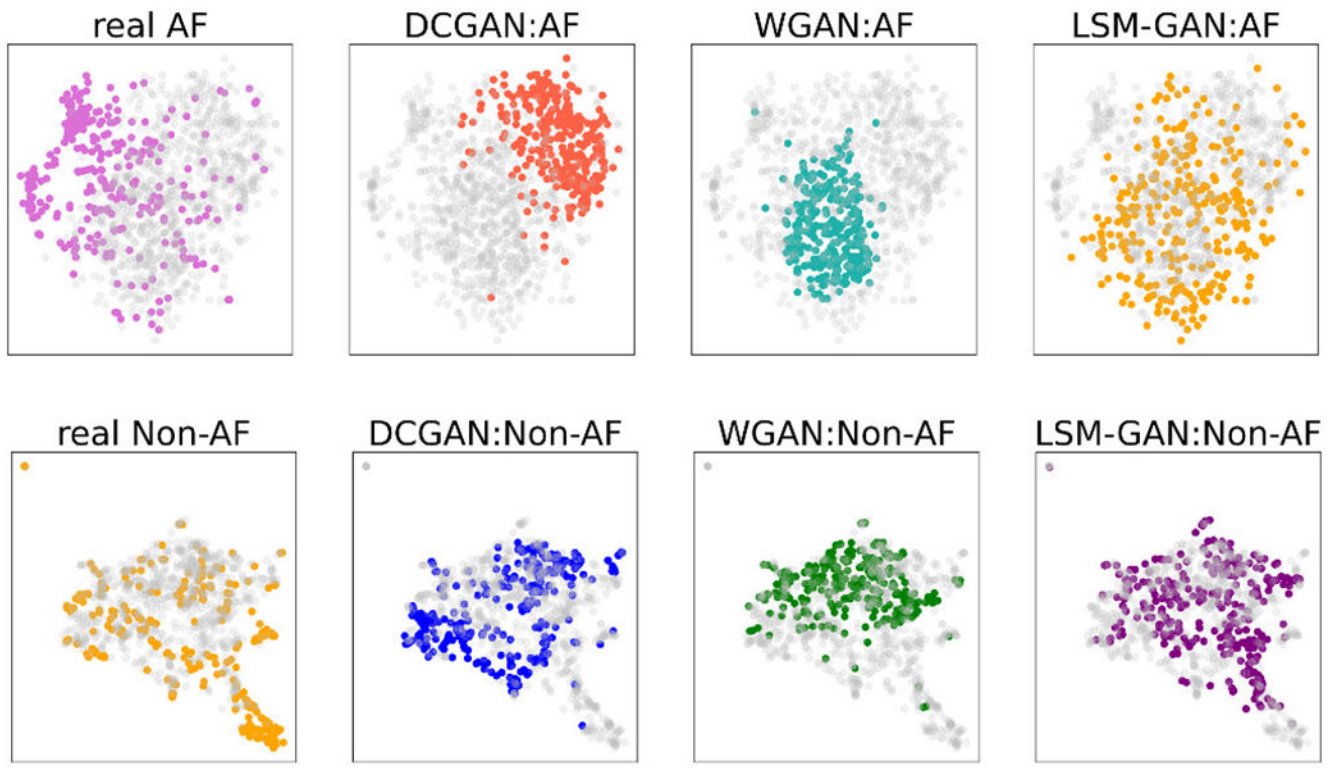


Fig. 9:
Visualization of distribution for real and synthetic AF signal.

TABLE I:

The number of records in the training and test sets

	Training Set		**Test Set**	
Center	UCLA Medical Center		UCSF Neuro	
Number of Patients	126		13	
Age	18 to 95 years (median 63)		19 to 91 (median = 73.5)	
Number of records	***AF***	***NonAF***	***AF***	***NonAF***
	36855	248278	1216	1467
Total	276133		2683	
percentage	15.24%	84.75%	45.32%	54.68%

TABLE II:

The performance of AF detection from each augmentation method

	Accuracy	Sensitivity	Specificity	PPV	NPV
Original	0.8210	0.6060	**0.9993**	**0.9986**	0.7537
Data-Copying	0.9288	0.8593	0.9863	0.9812	0.8943
Permutation	0.9369	0.8848	0.9850	0.9799	0.9117
DCGAN-100	0.9213	0.8470	0.9829	0.9763	0.8857
DCGAN-1200	0.9284	0.8684	0.9781	0.9705	0.8996
W-DCGAN-100	0.9250	0.8511	0.9863	0.9810	0.8888
W-DCGAN-1200	0.9370	0.8799	0.9843	0.9789	0.9081
LSM-GAN	**0.9612**	**0.9284**	0.9884	0.9851	**0.943**

TABLE III:

The selected optimal hyper-parameters for LSM-GAN data augmentation model

	λ_1	λ_2	**F**
Af	1.5	1.5	Mean
Non AF	0.8	3.0	Max

TABLE IV:

AF detection performance under different hyperparameters for AF signals

	Accuracy	Sensitivity	Specificity	PPV	NPV
$\lambda_1 = \lambda_2 = 1$,F=Mean(no weight mechanism)	0.9467	0.9095	0.9775	0.9710	0.9287
$\lambda_1 = 2.2, \lambda_2 = 2$, F=Max(Largest MMD)	0.9306	0.8651	0.9850	0.9795	0.8980
$\lambda_1 =$, $\lambda_2 = 1.5$,F=Mean (optimized hyperparameters)	0.9612	0.9284	0.9884	0.9851	0.9433

IEEE J Biomed Health Inform. Author manuscript; available in PMC 2024 July 26.