Accurate Detection of Paroxysmal Atrial Fibrillation with Certified-GAN and Neural Architecture Search

Mehdi Asadi[1] Fatemeh Poursalim[2] Mohammad Loni[3]
Masoud Daneshtalab[3] Mikael Sjödin[3] Arash Gharehbaghi[4,*]

[1]Department of Electrical Engineering, Tarbiat Modares University, Tehran, Iran
mehdi.asadi@modares.ac.ir

[2]Shiraz University of Medical Science, Shiraz, Iran
fpoursalim72@gmail.com

[3]School of Innovation, Design and Engineering, Mälardalen University, Västerås, Sweden
{mohammad.loni, masoud.daneshtalab, mikael.sjodin}@mdu.se

[4]Department of Biomedical Engineering, Linköping University, Linköping, Sweden
arash.gharehbaghi@liu.se
* Corresponding Author

Abstract

This paper presents a novel machine learning framework for detecting PxAF, a pathological characteristic of Electrocardiogram (ECG) that can lead to fatal conditions such as heart attack. To enhance the learning process, the framework involves a Generative Adversarial Network (GAN) along with a Neural Architecture Search (NAS) in the data preparation and classifier optimization phases. The GAN is innovatively invoked to overcome the class imbalance of the training data by producing the synthetic ECG for PxAF class in a certified manner. The effect of the certified GAN is statistically validated. Instead of using a general-purpose classifier, the NAS automatically designs a highly accurate convolutional neural network architecture customized for the PxAF classification task. Experimental results show that the accuracy of the proposed framework exhibits a high value of 99.0% which not only enhances state-of-the-art by up to 5.1%, but also improves the classification performance of the two widely-accepted baseline methods, ResNet-18, and Auto-Sklearn, by 2.2% and 6.1%.

Keywords: Electrocardiogram (ECG), Paroxysmal Atrial Fibrillation (PAF), Data Augmentation, Neural Architecture Search

1 Introduction

Recent progresses in artificial intelligence and Deep Learning (DL) methods created a leap toward automatic decision-making in various domains including health and medicine. Sophisticated Deep Learning (DL) methods have been proposed for classifying biological signals [20], including heart sound [21, 22] and electrocardiogram [13]. Electrocardiograph (ECG) is a recording of the electrical activity of the heart. An ECG signal shows a rhythmic behavior identified by a sequence of the patterns in a cyclic manner, where the regularity of the rhythm along with the shape of the patterns convey important information about the electrical activity of the heart. Paroxysmal Atrial Fibrillation (PxAF) is a type of irregularity in heart rhythm, called cardiac arrhythmia, characterized by intermittent episodes of rapid and irregular heartbeat, originating in heart atria. PxAF can cause symptoms such as palpitations, shortness of breath, dizziness, and chest discomfort that can lead to fatal conditions like cardiac stroke [45]. PxAF episodes often occur spontaneously and can last from a few seconds to several days before spontaneously converting back to normal sinus rhythm. Screening patients with PxAF is currently performed by physicians in their clinical practice, and the development of a reliable system for automated detection of PxAF is a need for any healthcare system for patient monitoring purposes.

Several methods have been proposed for detecting PxAF on ECG signal [13, 23, 48, 53], from which the DL-based ones are considered as the state-of-the-art of this topic [48, 55]. Nevertheless, accurate detection of PxAF is still an open research question [48, 55]).

We hypothesize two bottlenecks in reaching accurate PxAF diagnosis. Firstly, the class imbalance is commonly seen in most of the public ECG databases, where the size of the class with PxAF arrhythmia is by far smaller than the one with normal cases. Secondly, the backbone architectures used in the state-of-the-art studies may not be optimal as they were manually designed for image classification tasks.

One solution to tackle the first issue is to increase the group size of the minority class, i.e., the PxAF class [31], by producing synthetic data from the real ones. Patients' real data are being recorded electronically by healthcare providers and private industries. However, the recorded data is hardly accessible to scientists due to patient privacy concerns. Even when researchers are able to access high-quality data, they must ensure that the data is properly used and protected in a legal and ethical manner which is a time-consuming process [24].

Generating synthetic medical data has been broadly explored for various sorts of medical data including physiological signals [58]. Synthetic ECG data has been reported as the case study in several reports (Section 3.2). Recently, Generative Adversarial Networks (GANs) have demonstrated impressive performance in medical data augmentation. However, the synthetic ECGs, generated by GAN, are mostly immature to be used as the training data due to morphological irrelevance, and thus, leveraging them in the training process can mislead the classifier. As we will see in the sequels, this important point is elaborately considered by the proposed method.

Neural Architecture Search (NAS), as an automated technique for designing artificial neural networks, has recently received attention from researchers and engineers. It provides a solid tool to achieve an optimized architecture for the problem of designing an optimal machine learning solution. Applicability of this technique has been explored in different domains such as biomedical engineering, in which classification of physiological signals is an important challenge [16, 36, 40, 44].

In this paper, we propose an original framework for detecting PxAF arrhythmia based on an enhanced combination of GAN and NAS. The framework is composed of three compartments: 1) data enrichment, 2) signal processing, and 3) machine learning compartments. The proposed framework introduces innovative ideas in the methodologies employed for this important research question. It proposes the use of a GAN architecture for data enrichment in a new manner, named certified-GAN, in conjunction with the original signal processing and machine learning methods. The performance of the framework is statistically evaluated both holistically and independently for each compartment. The accuracy of the framework in detecting PxAF was estimated to be 99%, exhibiting a considerable improvement in the state-of-the-art.

To the best of our knowledge, this paper is the first study proposing an automatic methodology for certified synthetic data generation and designing an accurate CNN architecture for PxAF detection. We name this combination of certified-GAN and NAS for PxAF detection as Deep-PxAF. The contributions of this paper are:

- A novel data enrichment method is proposed that enables the generation of the certified synthetic PxAF samples based on the recommendations of an expert physician (Section 4.2).
- A novel data pre-processing approach is proposed to improve the detection performance (Section 4.3).
- A cell-based neural architecture search method is employed to design a specialized CNN architecture for the PxAF detection task (Section 4.4).
- We provide extensive experiments to demonstrate the effectiveness of Deep-PxAF (Section 6). Plus, we discuss the reproducibility results of the proposed method (Section 7).

Results show that Deep-PxAF achieves higher accuracy compared to handcrafted DL architectures and automated machine learning (AutoML) tools on the PhysioNet PxAF database [8]. Moreover, Deep-PxAF shows stable results with marginal differences with multiple repetitions, confirming the reproducibility of the results. The database of certified labels is open-access and can be used by any researcher for scientific purposes.

2 Preliminaries

2.1 Paroxysmal Atrial Fibrillation (PxAF)

ECG is a registration of the electrical activity of heart cells. A normal ECG is a cyclic signal composed of several waves and peaks within each cycle from which the QRS complex, T-wave, and P-wave are mostly regarded as indicative patterns of the signal. Fig. 1.a depicts a normal ECG signal along with the indicative patterns occurring in a certain order in time. The cyclic behavior of the ECG signal comes from the fact that heart muscles have two phases of activity: contraction and relaxation. A contraction is normally followed by a relaxation, where the contraction is initiated from the right atrium down to the ventricles and returned to its initiating point to create a self-stimulating activity through the heart muscles with a rhythmic behavior. This rhythmic action is projected to the ECG signal. The P-wave and the QRS complex coincide with the atrial and ventricular contraction, respectively, while the T-wave results from the ventricular relaxation. In the cardiac investigation, a complete relaxation followed by a left ventricle contraction is known as the cardiac cycle. However, for simplicity in ECG signal processing, a cardiac cycle can be defined as the interval between two successive R-peaks for computerized processing.

The morphology of an ECG signal conveys important information about the heart's electrical activity and, to a lesser extent, about its mechanical activity. This includes not only the duration of the QRS complex and the time intervals between the waves and the complex but also the amplitude of the patterns. Deviation from the typical characteristics of ECG can be resulted either from a physiological condition such as sinus arrhythmia or from pathological conditions, e.g., arrhythmia. Sinus arrhythmia can be dominantly caused by respiration. PxAF is a pathological condition of the electrical heart action that can happen when the atrial contraction is performed inappropriately. PxAF can initiate an arrhythmia and requires medical considerations and sometimes appropriate management. Fig. 1.b shows a PxAF condition versus a normal sinus rhythm.

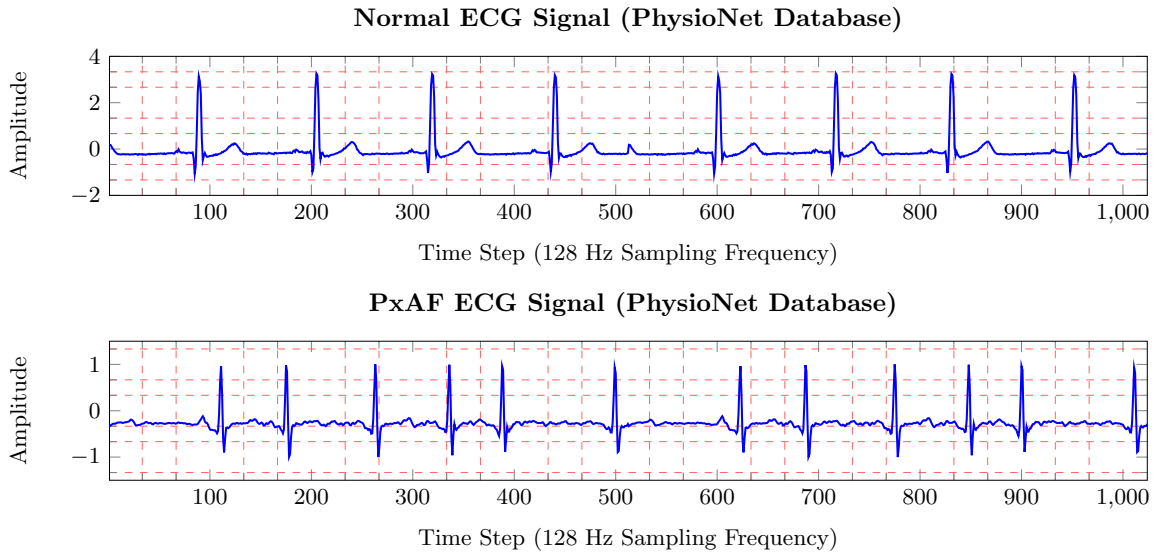


Figure 1: (a) Illustration of a sinus rhythm condition. Heart rate variation within 60-100 beats per minute. (b) PxAF condition. Heart rate variability in the form of arrhythmia and P-wave alterations.

As can be seen in Fig. 1, cardiac cycles show a physiological variation of sinus rhythm with clearly visible P-waves in all the cycles. In contrast, in the PxAF case, the P-waves show noticeable alterations over the cycles along with the arrhythmia. An association between PxAF and mortality has been previously demonstrated [19]. It is also studied that timely detection of PxAF can improve survival in this patient group by appropriate medical management [19].

2.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a class of deep learning architectures that have been successfully used to generate synthetic images, time-series data, and other data modalities [7, 29]. In general, GANs are comprised of two sub-networks: the generator (G) and the discriminator (D). G generates synthetic data that is as close as possible to the real data, while D determines whether the

generated data is real or not. These two sub-networks compete with each other in a two-player minimax game with a loss function of $V(G, D)$ (Eq. 1). The goal of solving Eq. 1 optimization problem is to reach Nash equilibrium [27].

$$\min_G \max_D V(G, D) = E_{x \sim p_{data}(x)}\big[logD(x)\big] + E_{z \sim p_z(z)}\big[log(1 - D(G(z)))\big] \quad (1)$$

Probability $D(x)$ determines whether x is generated data or real data.

3 Related Works

3.1 PxAF Diagnosis Using DL Methods

Previous studies on DL-based methods showed less attention paid to PxAF detection than other forms of arrhythmia [13]. Pourbabaee et al. [48] proposed a method for identifying patients with PxAF. Their proposed method employs raw ECG data as input; then, uses a CNN with one fully-connected layer to learn a discriminative pattern of data in the time domain. Plus, they manually tweaked various classification methods to achieve maximum performance. [53] proposed an attention-based DL method for detecting PxAF episodes from a synthetic database composed of 24-hour Holter ECG recordings. Time-frequency representations of 30-second windows are fed sequentially into the CNN. Then, the extracted features are presented to a bidirectional recurrent neural network with an attention layer. [23] constructed a new long-term ECG database (24 to 96 hours) for the purpose of detecting PxAF. After careful analysis by a cardiologist, 250 AF onsets of PxAF have been detected. They proposed a CNN followed by a bidirectional Gated Recurrent Units (GRU) network for PxAF detection. The network was trained to distinguish between RR intervals that precede an AF onset and RR intervals distant from any AF. They concluded that RR intervals contain information about the incoming AF episode. [59] proposed to predict the occurrence of PxAF by combining wavelet decomposition and a CNN classifier. [55] aimed to detect PxAF episodes before occurrence. [55] leveraged a CNN to process normalized heart rate variability features resulting in 87.76% accuracy and 87.50% f1-score in heart rate variability.

In this perspective, Atrial Fibrillation (AF), which is regarded as a case study with physiological similarities, has been reported in a large number of related studies. [65] developed a dual-domain attention cascade model called D2AFNet, which addresses the challenge of accurate AF detection. [41] introduced the utilization of a fully-connected network that incorporates diverse input ECG features and tested on ECG recordings obtained through portable devices. [30] introduced a method for detecting AF from Holter-ECG recordings using a CNN. After eliminating artifacts and noises, the proposed approach first extracts abnormal waveforms using a one-dimensional CNN, then identifies AF using a two-dimensional CNN trained with segmented ECG spectrograms. [28, 62] introduced the utilization of transformer models [60] for the purpose of detecting AF. Their aim was to enhance the capturing of inter-heartbeat dependencies by leveraging transformers in the detection process. Compared to state-of-the-art AF detection methods, Deep-PxAF stands out as the pioneering study that utilized neural architecture search on a synthetic-verified database.

3.2 Synthetic Data Generation for ECGs

Medical data tend to be highly sensitive by nature and are often subject to severe usage restrictions. As a result, it is difficult for researchers to collect and share this data. A possible alternative to address the problem of data scarcity is to generate realistic synthetic data [7]. [42, 50] proposed mathematical dynamical models to generate continuous ECG signals. These models, however, were limited to one lead signal and did not provide any insight into the mechanism of disease.

Recent studies have demonstrated that GANs are extremely effective at synthesizing ECG waveforms based on a prior distribution of data. Prior works are mainly focused on efficient GAN architecture [1, 3, 10, 33, 57, 61, 66]. [10] studied various GAN architectures by leveraging LSTM or BiLSTM as the generator and a CNN discriminator with single or multiple Convolution-ReLU-Pooling layer(s). Results show that a BiLSTM GAN with a single `Convolution-ReLU-Pooling` layer provides the best performance. [66] used a BiLSTM-CNN GAN model to generate synthetic ECG signals. A GAN architecture based on a four-layer generator and a five-layer fully-connected discriminator is proposed in [52]. [3] proposed a multi-GAN method to generate ECG waveforms for atrial fibrillation arrhythmia by combining the output of GAN models. [57] proposed two GAN architectures, WaveGAN$*$ and Pulse2Pulse, with

the ability to generate synthetic 10-s ECG waveforms. Pulse2Pulse, which is based on a U-net generative model, is superior to producing realistic ECGs. [33] was the first to propose a transformer-based conditional GAN architecture, named TTS-CGAN, to generate synthetic time-series with sequences of arbitrary length. Compared to popular RNN or LSTM-based GANs for generating time-series [10, 15, 64], TTS-CGAN has no difficulties in producing long synthetic sequences. In continuation, [61] proposed TCGAN, an architecture combined with a transformer generator and CNN discriminator.

Despite the success of these methods, they do not guarantee that the generated data is trustworthy, resulting in the failure of classifiers to make accurate predictions. This paper sheds light on the fact that synthesizing high-quality artificial data play a crucial role in accurate predictions. Thus, we propose a novel physician-certified synthetic data generation method that provides ECG samples indistinguishable from real ones.

3.3 Neural Architecture Search for ECG

Several DL models have been developed for detecting a variety of cardiac arrhythmias. However, increasing the complexity of manual-designed networks does not always lead to better performance. Moreover, the introduced deep neural networks mostly require a cumbersome phase of trial-and-error, which results in enormous computational costs [37]. Recent advances in Neural Architecture Search (NAS) have enabled the designing of scalable and resource-efficient neural architectures. Being inspired by the remarkable success of NAS in the computer vision domain [14], several techniques very recently proposed to leverage NAS for designing accurate architectures for arrhythmia detection [16, 17, 36, 40, 44].

Fayyazifar et al. [17] studied the impact of manually tweaking deep neural networks for cardiac abnormality classification. Additionally, they used wavelet decomposition to enhance the classification performance of the PhysioNet Challenge 2020 [2]. [16] employed a NAS method for AF classification where they achieved an accuracy of 84.15% on the PhysioNet challenge 2017 [8]. Heart-Darts [40] proposed a heartbeat classification method by automatically designing an efficient CNN architecture with a differentiable NAS method. Heart-Darts provides state-of-the-art performance, applied to the MIT-BIH arrhythmia database [43]. [36] developed a NAS-based learning method to detect cardiovascular diseases in 12-lead ECG data. In particular, they proposed a novel search strategy that optimizes different attention modules of the same network synchronously. EExNAS [44] designed energy-efficient CNN architectures for detecting Myocardial Infarction (MI) and Human Activity Recognition (HAR) on wearable devices.

These methods utilize NAS to design an efficient arrhythmia classifier; however, they are limited to optimizing the feature extraction part. Further, it is not conclusive that the findings of the prior studies are reproducible, especially since there is no comprehensive evaluation found in their report [34].

4 Methodology

4.1 Method Overview

We propose a novel method with three phases, comprising: 1) synthetic data generation, 2) ECG Signal Processing, and 3) CNN Architecture Search. Fig. 2 depicts the bird's eye view of the proposed method. In the first phase, we generate synthetic ECGs for the PxAF class using a GAN model. After the GAN creates synthetic ECGs, an expert physician evaluates them to identify high-quality training data. The second phase of the method employs the wavelet transform of an ECG signal along with the recurrence graph. Rhythmic information of an ECG within short length windows of 4 second is preserved in a recurrence graph. The outcome of the first stage is a sequence of the two-dimensional images, each incorporating rhythmic contents of a 4 second interval of an input ECG. In the last phase, a CNN is trained to classify the images where the architecture of the CNN is found using NAS. As we will see, the combination of these innovations noticeably improves the performance of the classification.

Phase 1: Certified Synthetic Data Generation. The public databases of ECG mostly contain a heavy class imbalance for the arrhythmia classes. The machine learning methods trained by such databases will be consequently biased for the normal classes. In order to cope with the shortage of signals from the minority class, i.e. the PxAF class, a structure GAN is invoked to create synthetic ECGs from the PxAF class. Obviously, inappropriate synthetic ECGs can mislead the classifier. Therefore, the synthetic ECGs created by the GAN are evaluated by an expert physician in terms of quality using a clearly-defined protocol. The disqualified ECGs will be discarded from the training and the synthetic ECGs certified by the expert physicians will be invoked for the learning process (Section 4.2).

Phase 2: ECG Signal Processing. ECG signal in its raw form is contaminated by different sources of noises and disturbances, such that the PxAF information can be fully concealed. In order to extract discriminant contents of PxAF from the pathological signals, a level of signal processing is required to purify indicative signal contents (Section 4.3). This processing yields a sequence of 2D images, each containing the dynamics of a few seconds of the signal, to a CNN architecture, in which the ultimate classification is performed.

Phase 3: CNN Architecture Search. Manual design of task-specific neural architectures requires tremendous human effort and domain expertise. In addition, the knowledge learned from designing a network cannot be directly transferred to another person. Neural Architecture Search (NAS) is the process of automatically optimizing a neural network architecture. NAS research has shown significant progress in enabling accurate neural architectures for computer vision applications [6, 14, 37, 38]. Because of this insight, we came up with the idea of leveraging NAS with the hope of improving the accuracy of PxAF detection (Section 4.4).

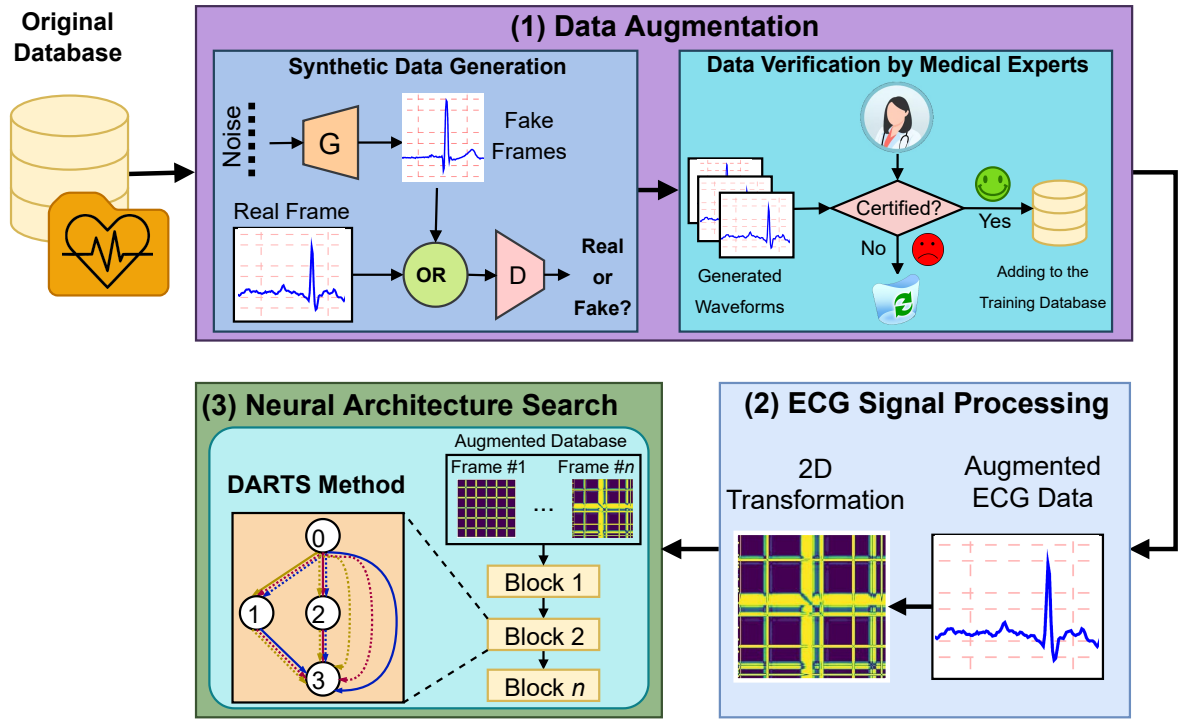


Figure 2: The bird's-eye view of the proposed method.

4.2 Certified Synthetic Data Generation

4.2.1 GAN Architecture

In this paper, we used the Pulse2Pulse GAN model proposed by [57]. Here, we briefly present generator and discriminator architectures. Then, we present the procedure of certifying the quality of generated data with the help of an expert physician.

Generator. The architecture of the generator is inspired by the U-Net architecture. The U-Net implementation uses 1D convolutional layers for ECG signal generation. The network takes a 2×5000 noise vector to generate a 2-lead signal, which is equal to the dimension of the output layer. The noise is passed through six down-sampling blocks followed by six up-sampling blocks. Each down-sampling block consists of a 1D-convolution layer followed by a Leaky ReLU activation. The deconvolution blocks were built from a series of four layers: an up-sampling layer, a constant padding layer, a 1D-convolution layer, and a ReLU activation function consecutively.

Discriminator. The discriminator takes an ECG as input and outputs a score indicating how close it is to a fake ECG. The architecture is composed of seven convolutional layers that follow the `Convolution+Leaky ReLU+Phase Shuffle` order. Using phase shuffle operation, each feature map's phase is uniformly perturbed [11]. Training specification is reported in Table2.

4.2.2 Synthetic Data Certification.

We observe that not all GAN-generated synthetic ECGs cannot be used as training segments due to their improper morphology, and thus, leveraging all GAN-generated segments in the training process will negatively affect the classification accuracy. Based on the morphological characteristics of ECG signal for PxAF cases, an expert physician manually verified all the synthetic ECGs and certified the valid ones (e.g., Fig. 3.a) based on the directives listed in Table 1. In this table, the bizarre shape implies the condition in which the sequence of the ECG peaks and waves, and/or their shapes fundamentally differ from the ones, seen in clinical practice. This condition might be seen in a segment (directive 2), or the entire of synthetic ECG. It was also observed that the QRS complexes of the synthetic ECG are inconsistent, or accompanied by extra weird morphology (directives 3, 4). The PxAF characteristics were inconsistently seen in some of the data, affecting the learning process, and thus were eliminated (directive 5).

Table 1: Directives for rejecting improper synthetic ECG segments.

Directive	Explanation	Plot
1. Bizarre Shape	Improper morphology with undetectable peaks or waves	Fig. 3.b
2. Distorted PxAF	There are distorted segments of the signal with bizarre shape	Fig. 3.c
3. Inconsistent QRS-complex	Heart beat exist, but the QRS-complexes are inconsistent in different beats	Fig. 3.d
4. Redundant/Noisy R peaks	Extra and noisy R peaks in the segment	Fig. 3.e
5. Partial PxAF	Segment partially include PxAF pattern	Fig. 3.f

4.3 ECG Signal Processing

Fig. 4 shows the major steps of the proposed signal processing pipeline. As shown, the input ECG signal is firstly decomposed to its constitutive components using wavelet transformation until the 10^{th} level using the Daubechies 3 wavelet family. The detail of the wavelet transforms at the levels $2, 3$ and 4 along with the approximation contents of the $10 * th$ level are reconstructed and added together, to eliminate the noises and the disturbances contaminating the signal. The resulting signal is then normalized by the absolute value of the points with the largest value. Next, the Shannon energy of the normalized signal is calculated using the following formula:

$$\mathbf{Y}_i(t) = x^2(t) log(x^2(t)) \tag{2}$$

where $x(t)$ is the normalized ECG signal which is positively biased to secure non-zero values. An envelope of the resulting Shannon energy signal is found by using a non-overlapping temporal window of length $100ms$, which slides over the signal. Lastly, a recurrence 2D function of the envelope is obtained. Calculational details of finding the recurrence plot are found in [25]. The output of the signal processing algorithm is a 2D representation of an input signal, which is discriminant for the PxAF and the normal classes. A CNN employs 2D images for classification.

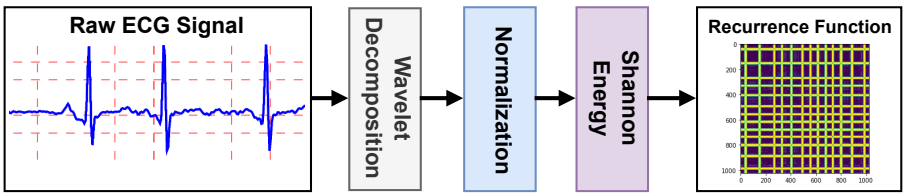


Figure 4: Illustration of the proposed signal processing pipeline.

4.4 CNN Architecture Search

In general, the learning proficiency of CNNs will be improved by increasing the number of network layers. However, simply stacking the network layers may cause accuracy degradation since the deeper networks will encounter a vanishing/explosion gradient problem. Neural Architecture Search (NAS) methods aim to help engineers to design highly efficient neural networks from scratch [35, 37, 38].

The NAS pipeline typically begins with a pre-defined space of network operators. Since the search space is often enormous (e.g., containing 10^{24} or even more possible architectures [38]), it is unlikely

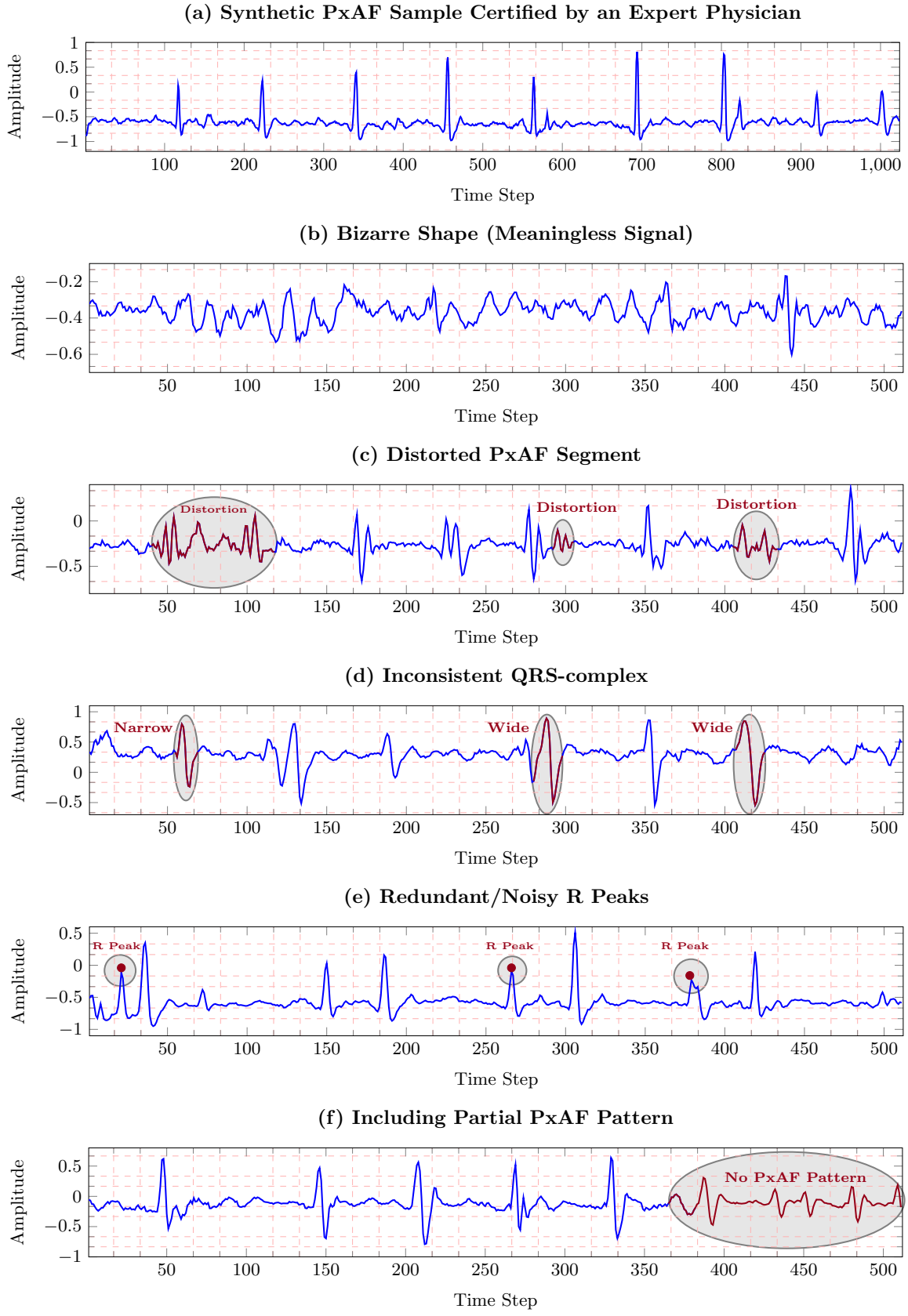


Figure 3: (a) Plotting a certified synthetic PxAF sample. Plotting PxAF synthetic samples rejected by an expert physician due to (b) bizarre shape, (c) distorted PxAF, (d) inconsistent QRS-complex, (e) redundant/noisy R peaks (showing with red points), and (f) partially existing PxAF pattern in the segment. The sampling frequency is 128 Hz.

that an exhaustive search is tractable. Thus, heuristic search methods are widely applied to speed up the search process. At an early age, each sampled architecture undergoes an individual training process from scratch, and thus the overall computational overhead is large, e.g., hundreds of GPU-days (e.g., [56] requires 3800 GPU days).

To alleviate the computing cost of NAS methods, researchers proposed to share computation among the sampled architectures, with the key idea of reusing network weights trained previously [5, 38] or starting from a well-trained super-network [47]. These efforts shed light on the one-shot NAS methods, which require training the super-network only once, and therefore run more efficiently (e.g., 2-3 orders of magnitude faster than conventional approaches).

One-shot NAS methods jointly formulate architecture search and network training [6, 12, 35]. Differentiable NAS methods solve this problem using gradient-based algorithms such as Stochastic Gradient Descent (SGD). DARTS [35] is a well-known differentiable NAS method that constructs a super-network with all possible operators. DARTS utilizes a cell-based design space to search for a well-behaved cell architecture [12, 35]. Then, the cell may be stacked any number of times to meet various hardware devices' resource requirements. In this paper, we utilize DARTS [35] to design CNN architectures due to significantly reducing the notorious design time of neural networks.

Mathematically, the final DARTS architecture is a function, $f(x;\omega,\alpha)$, where x is input, ω is network parameters (e.g., convolutional kernels), and α in architectural parameters (e.g., indicating the importance of each operator between each pair of layers). $f(x;\omega,\alpha)$ is differentiable to both ω and α could be optimized using the SGD algorithm. $f(x;\omega,\alpha)$ is composed of a few cells, where each cell of DARTS is defined by a directed acyclic graph with a pre-defined number of layers and a limited set of neural operators. Each cell contains N nodes, and there is a predefined set, E, which indicates connected pairs of nodes. For each connected node pair (i,j) and $i<j$, node j takes x_i as input and propagates it through a pre-defined operator set, O, and sums up all outputs (Eq. 3). O supports separable convolution (3×3, 5×5), dilated convolution (3×3, 5×5), max/average-pooling (3×3), and Identify operators.

$$y^{(i,j)}(x_i)=\sum_{o\in O}\frac{exp(\alpha_{o^{(i,j)}})}{\sum_{o'\in O}exp(\alpha_{o^{(i,j)}}}.o(X_i) \tag{3}$$

The normalization is performed by computing the Softmax function over the architectural weights. α and ω get optimized alternately in each search iteration. Afterward, the operator o with the maximum value is preserved for each edge (i,j), and all other network parameters ω are discarded. In DARTS, the type of each cell is either a normal cell for feature extraction or a reduction cell for both feature extraction and dimension reduction. After designing the optimal cell, we assemble the final network by stacking 18 normal cells with two reduction cells, where every six normal cells are followed by one reduction cell [39]. Last, the final architecture is re-trained from scratch to fine-tune the network parameters.

5 Experimental Setup

5.1 Database Preparation

Deep-PxAF identifies individuals who are at risk of PxAF. To this end, we utilized the PhysioNet PxAF prediction challenge database [8]. This database includes two-channel ECG recordings. The ECG signals were digitized with a 128 Hz sampling frequency, 16 bits per sample, and nominally 200 A/D units per millivolt. The database is divided into training and testing sets. The original train set consists of 100 records with a duration of 30 minutes that are collected for normal individuals and PAF patients, each with an equal number of recordings. The test set contains 50 records of 30 minutes duration in which 28 subjects are at risk of PxAF, and 22 subjects are healthy individuals. We completely isolate the training and testing sets. We also did not create a separate validation set to evaluate training performance since the size of the database is relatively small.

In this paper, we partitioned each 30-minute ECG signal into segments of four seconds duration resulting in 512 samples/segment. To build the original database ($D_{Original}$), we randomly select 4231, 906, and 906 segments for train, validation, and testing, respectively. We consider two classes for training and testing sets: normal (healthy) and PxAF patients. $D_{Original}$ contains 3545 and 2498 samples for normal and PxAF classes, respectively. The ECG data labeling tool is released alongside the codes.

We generate 10000 synthetic segments for the PxAF class using GAN. As we have data imbalance for the PxAF class, we only synthesize PxAF segments. The original training database has been augmented with 10000 synthetic segments (D_{GAN}). Due to the fact that most of the generated segments are not

of high quality, an expert physician evaluated all synthetic data and certified 539 segments containing PxAF. Then, we add the certified synthetic PxAF segments to the original training set to make the final synthetic database (D_{CGAN}). The synthetic data generation time takes ≈ 42 GPU hours on a single NVIDIA GTX 1080 Ti that produces ≈4.3 Kg CO_2 [32].

5.2 Configuration Setup

Table 2 summarizes the configuration setup of experiments. In this paper, each DARTS cell consists of seven nodes equipped with a depth-wise concatenation operation as the output node. The convolutional operations follow the `Convolution+Batch Normalization+ReLU` order. The network design time (search+re-training) takes ≈ 9 GPU hours on a single NVIDIA GTX 1080 Ti that produces ≈0.97 Kg CO_2 [32]. The rest of the setup follows [35].

Table 2: The configuration setup of the signal processing and neural architecture search hyper-parameters.

Signal Processing Pipeline	**Value**
Maximum wavelet scales	level 10
Shannon Window Length	0.1 second
Recurrence Length	4 second
Synthetic Data Generation	**Value**
# Epochs	8000
Optimizer	Adam
Learning Rate (lr)	1.0×10^{-4}
NAS Hyper-parameters: Design	**Value**
Train/Test Segments	5000/1000
# Epochs	50
Batch Size	6
Optimizer	SGD
Learning Rate (lr)	0.025
weight decay	3×10^{-4}
momentum	0.9
NAS Hyper-parameters: Fine-tuning	**Value**
# Epochs	200
Batch Size	10
Optimizer	SGD
Learning Rate (lr)	0.025
weight decay	3.0×10^{-4}
momentum	0.9
Color Noise Parameters	**Value**
Filter Type	Butterworth Lowpass Filter
Cutoff Frequency o Filter	50 Hz
Order of Filter	4
Hardware Specification	
GPU	NVIDIA GTX 1080 Ti (2.5 GHz)
GPU Compiler	cuDNN Version 7.1
Operating System	Ubuntu 18.04
Training System Memory	32 GB

5.3 Baseline for Comparison

Auto-Sklearn [18]. Auto-Sklearn is a state-of-the-art library for automated machine learning (AutoML) that is compatible with the scikit-learn library [46]. Auto-Sklearn automatically selects appropriate hyperparameters for a given database by leveraging Bayesian optimization [51] as the search method. Auto-Sklearn uses four data preprocessing techniques, 14 feature preprocessing techniques, 15 classifiers, and a structured hypothesis space with 110 hyperparameters. Auto-Sklearn considers the past performance of similar databases and constructs ensembles from the machine learning models evaluated during the optimization to improve the optimization quality. Due to the high efficiency of Auto-Sklearn in

customizing the machine learning pipeline [9, 49], we consider Auto-Sklearn as the second comparison baseline.

Deep Residual Network (ResNet) [26]. ResNet is a family of handcrafted architectures that won the ILSVRC competition challenge in 2015. ResNet is constructed by several back-to-back residual blocks connected to a final linear fully-connected layer. In this study, we used ResNet as the third comparison baseline since ResNet has been widely used in automated clinical diagnosis of various diseases [4, 13, 54, 63].

5.4 Performance Measurement

This section introduces common quantitative metrics used for presenting how well synthetic data generation and classification methods work.

GAN Performance. For evaluating the performance of GAN, we use a database containing GAN output data and original data to train a model, which is then tested on a held-out set of true examples. This requires the generated data to have labels - an expert physician provides labels to GAN output data. We statistically analyze the distribution of read ECGs and fake ECGs using Kolmogorov-Smirnov test (K-S test). Plus, we will show the Q-Q plot to look at the skewness of fake data from real data.

Classifier Performance. The formulas for quantifying measurements are listed below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

where TP, TN, FP, and FN denote True Positives, True Negatives, False Positive, and False Negative, respectively.

6 Experimental Results

6.1 The Synthetic ECGs

The previously-described GAN is trained with 8000 epochs and a learning rate of 0.0001. Fig. 5 depicts the loss function of the generator and the discriminator of the GAN. Both of the losses converge to a similar low margin implying the learning relevance. The outcomes of the GAN generator constitute our synthetic ECGs.

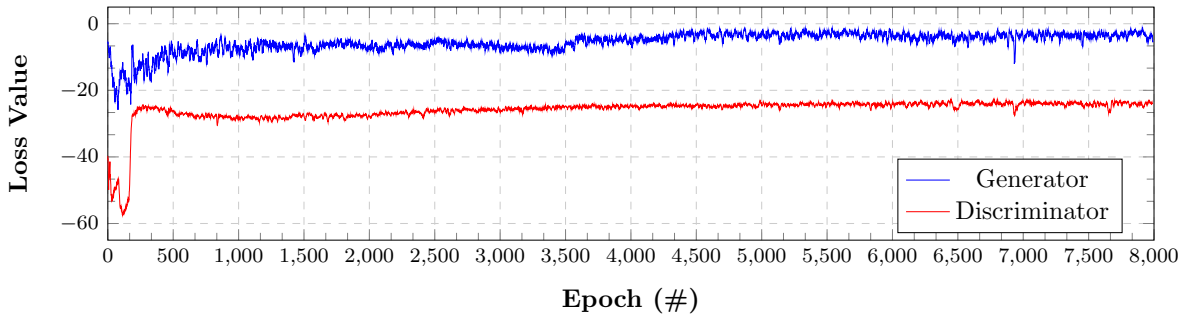


Figure 5: Loss of the discriminator and generator during GAN training.

The quality of the synthetic ECGs is evaluated based on the statistical measures, separately applied to the entire original and synthetic populations, once using the outcomes of the certified-GAN and once using the GAN without accreditation of the expert physician. In both cases, the fidelity of the synthetic ECGs is evaluated by using the three PxAF-related parameters of ECG: heart rate, R-peak to R-peak interval (RR Interval), and QRS interval. These three parameters are independently calculated for the populations using the signal processing algorithm described in Section 4.3. It is worth noting that these three parameters reflect the variation of the cardiac cycle and heart rate that is linked to arrhythmia.

In total, 10000 synthetic ECGs were generated using the previously-described GAN, from which 539 were accredited by the expert physician. Fig. 6 illustrates the histogram of the three PxAF-related parameters for the real and the synthetic ECGs resulting from the certified-GAN. The modal similarities are obviously seen for the synthetic and real populations.

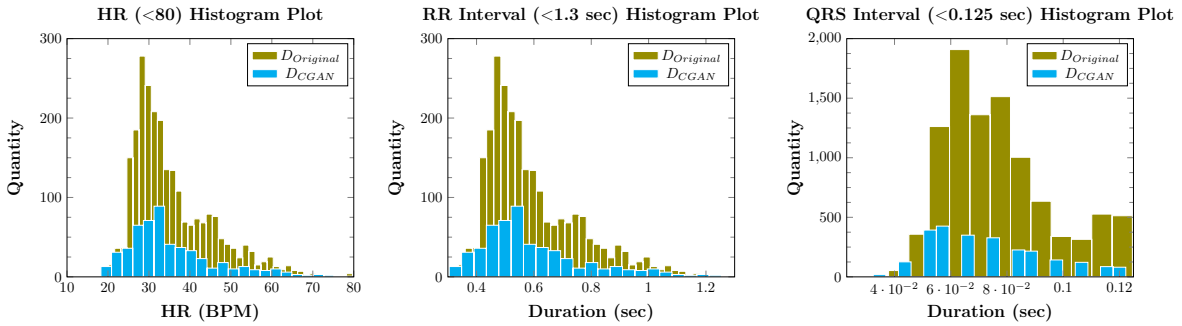


Figure 6: Distribution of (left) heart rates, (middle) RR interval, and QRS interval in all 539 certified segments (D_{CGAN}) compared to the original database ($D_{Original}$).

In order to explore the fidelity of the synthetic ECGs, descriptive statistics are calculated over the three populations: The real ECGs, the GAN, and the certified-GAN subjected to having PxAF condition. Table 3 represents the mean, standard deviation, and percentile values corresponding to the three populations. From the population perspective, the three PxAF-related parameters of the certified synthetic ECGs demonstrate very good fitness to the population of the real ECGs, with a marginal deviation of less than 2% for the mean value. This value is almost 4% for the data from the GAN. The deviation of the percentile values is less than 10%. The certified-GAN provides clear improvements in all the statistics, but the 2.5% percentile which corresponds to the outlier data.

Table 3: Mean, standard deviation (STD), 2.5%, and 97.5% percentile for HR, RR interval, and QRS interval parameters in real and synthetic ECGs. BPM stands for beats per minute.

Feature	**Database**											
	$D_{Original}$				D_{GAN}				D_{CGAN}			
	Mean	**STD**	**2.5%**	**97.5%**	**Mean**	**STD**	**2.5%**	**97.5%**	**Mean**	**STD**	**2.5%**	**97.5%**
RR Interval (sec)	0.5976	0.1640	0.3906	1.0078	0.621	0.218	0.376	1.164	0.604	0.203	0.351	1.101
HR (BPM)	35.86	10.0	23.43	60.46	37.25	13.0	21.56	69.84	36.26	12.0	21.09	66.09
QRS (sec)	0.07	0.0234	0.039	0.1172	0.069	0.0234	0.039	0.1172	0.069	0.0234	0.039	0.1172

In order to obtain a better understanding of the outperformance of the certified-GAN, the quantile distribution of the real and synthetic data, the so-called Q-Q plot, is investigated. Fig. 7 illustrates the resulting Q-Q plot.

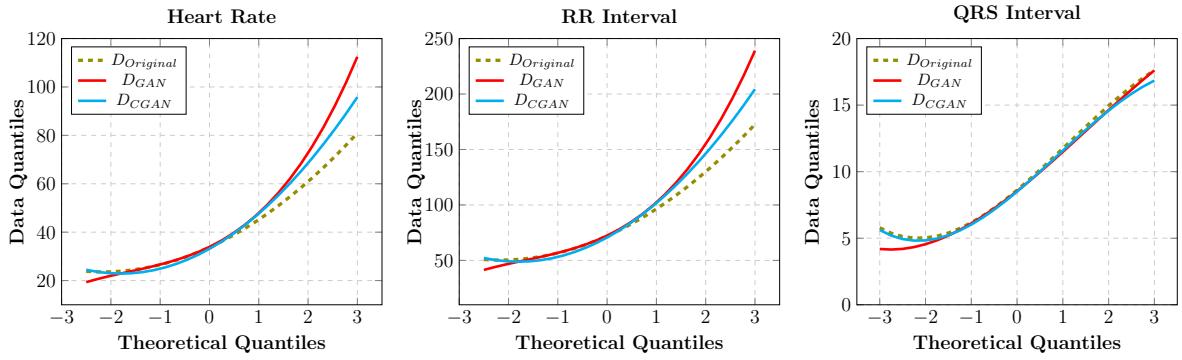


Figure 7: Illustration of the Q-Q-plot for (left) heart rate, (middle) RR interval, and QRS interval (right).

It is obviously seen that the certified-GAN provides closer statistical distribution to the real one, as compared to the plain GAN. This is also explored by using the Kolmogorov-Smirnov Test.

Table 4 presents the results of the Kolmogorov-Smirnov (K-S) test for heart rate and QRS interval. As seen in the table, the certified-GAN improves the K-S statistics as well as the p-value, showing a

closer distribution to the real population. This distribution is closer to the real population than the one for the GAN, confirming the effectiveness of the certified-GAN.

Table 4: Kolmogorov-Smirnov Test Results.

Feature	Parameter	Database	
		($D_{Original}$ & D_{GAN})	($D_{Original}$ & D_{CGAN})
Heart Rate	Statistic	0.0593	0.0758
	p-value	1.4897e-6	0.0116
QRS Interval	Statistic	0.0454	0.0431
	p-value	6.1256e-15	0.0016

6.2 PxAF Classification Performance

Table 5 compares the results of Deep-PxAF with the state-of-the-art and state-of-practice classification methods. Results show that Deep-PxAF provides the most accurate classification result with 99.0% accuracy compared to all counterparts. The analysis of the best DARTS cells searched by Deep-PxAF is provided in Section. A.2.

Table 5: Comparing the results of Deep-PxAF with state-of-the-art and state-of-practice methods.

Method	PhysioNet Classification Accuracy (%)	
Pourbabaee et al. [48]‡	91.0	
Surucu et al. [55]	93.88	
	$D_{Original}$ (%)	D_{CGAN} (%)
ResNet-18 [26]	95.2	97.0
Auto_Sklearn [18]	92.53	92.83
Deep-PxAF (Ours)	97.3	99.0

† Using the same search space as DARTS.
‡ Reporting the best results by CNN architecture with a K-nearest neighbor (KNN) classifier.

This study proposed an accurate method for screening PxAF. In this application, the trade-off between sensitivity and specificity is made by assigning the threshold of the output layer, where sensitivity and specificity are defined as:

- Sensitivity is the probability of PxAF condition when the classification result is positive
- Specificity is the probability of normal condition when the classification result is negative

Receiver Operating Characteristics (ROC) is a plot of the $Sensitivity$ against (1-$Specificity$), in which the optimal point is the point with maximal Sensitivity and specificity. Fig. 8 illustrates the ROC curve for the proposed method in comparison with the ResNet-18 classification method. As can be seen in Fig. 8, Deep-PxAF provides a more favorable characteristic in terms of the compromise between Sensitivity and Specificity with a closer curve to the ideal case of the straight angle. The Area Under the Curve (AUC) of ROC for Deep-PxAF trained on D_{CGAN} is improved by 0.32% and 0.47% compared to Deep-PxAF trained on $D_{Original}$ and ResNet-18 trained on D_{CGAN}, respectively.

6.3 Robustness against the Background Noise

The robustness of the classifiers is investigated by adding background noise to the input signals and calculating the accuracy of the classifiers. Two different background noises are simulated for the investigation: a white noise with normal distribution and a color noise which is indeed a filtered white noise. We employed a Butterworth lowpass filter with a cutoff frequency of about 50 Hz, as often used in ECG acquisition systems. The accuracy of the classifiers is explored for various Signal-to-Noise Ratios (SNR) of the white noise and the color noise, separately. Figure 9 demonstrates the profile of the accuracy and the SNR. As can be seen, the superiority of the Deep-PxAF is well preserved under the noisy conditions of the background noises.

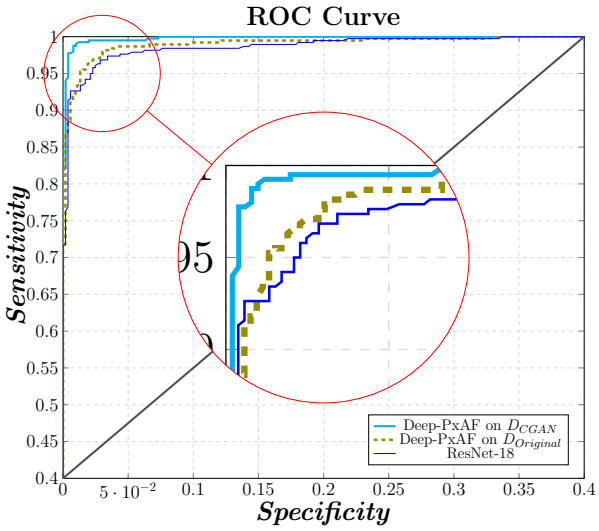


Figure 8: Comparing the ROC curve of Deep-PxAF trained on $D_{Original}$ and D_{CGAN} to the ResNet-18 trained on D_{CGAN} baseline method.

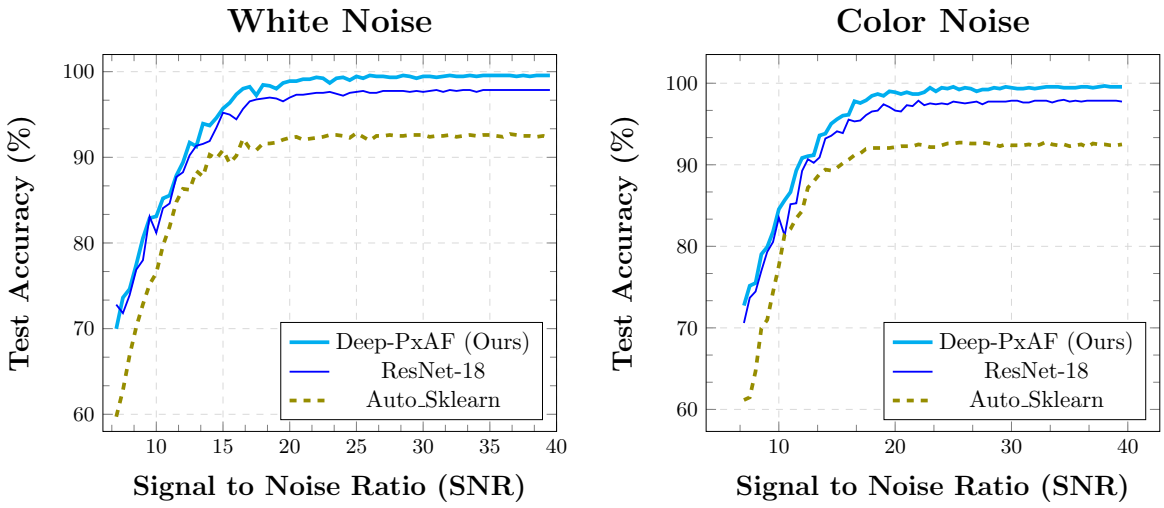


Figure 9: (left) Gaussian white noise and (right) noise with Butterworth filter with 50 Hz cut-off frequency.

7 Discussion

This study suggested an original framework for PxAF classification using a novel combination of a GAN and NAS in conjunction with an advanced signal processing method. The proposed framework introduces a phase of generating synthetic ECG using GAN, to enhance the accuracy of the classification method by enriching the training data and overcoming class imbalance. Generating valid synthetic ECGs through a certified procedure was the main objective of this phase. Using a rich training dataset with consistent class size can evidently enhance the learning process. The experimental results showed that the enhancement in the accuracy is considerable, which was confirmed by the ROC graph (see Figure 8). Besides, the classifier employs NAS, as a reliable architecture designer to boost the classification performance. The resulting classification method was optimized and implemented to detect patients with PxAF arrhythmia, which is regarded as an important case study with vital importance. The proposed method improved the screening accuracy by 6.1% compared to the state-of-the-art automated machine learning method [18]. The baseline for comparison was ResNet-18 and Auto-Sklearn which are well-known benchmarks for the machine learning method. These benchmarks were noticeably outperformed by the proposed method.

7.1 Synthetic Data Generation

This study employed a GAN architecture to generate synthetic ECGs and meanwhile invoked an expert physician to accredit the synthetic data. The application of GAN in generating synthetic ECG has been already explored [57, 66], however, the effectiveness of the generated ECGs in the training process is questionable since inappropriate synthetic data can evidently mislead the classifier. The certified-GAN which was proposed by this study effectively pruned the inappropriate signals. Results showed a noticeable improvement in the learning process using the certified-GAN. We made these synthetic signals publicly available to any researcher to explore for any scientific purposes.

Another exciting aspect of this study is the statistical techniques employed to study the fidelity of synthetic ECGs. Heart rate and R-R interval were employed as the measures for the PxAF. The statistical techniques mainly perform population-based evaluations which fit well into the scope of the learning process. The certified-GAN showed incapability to generate appropriate outliers, as reflected by the 2.5 percentile in Table3. Such outlier data cannot play an important role in the learning process performed by the proposed deep learning architecture.

7.2 ECG Signal Processing

In this study, the rhythmic contents of the heartbeats are innovatively preserved at the feature extraction level through signal processing and the recurrence images. Like other methods sufficing to the temporal features, there are a number of design parameters associated with the method at this level, such as the window's length for obtaining the recurrence graph as well as the wavelet transformation. These parameters were empirically obtained based on prior knowledge of the signal. Integration of finding the optimal values for these design parameters with the optimization process might provide further improvements.

7.3 CNN Architecture Search

Although several NAS methods have been proposed to detect various arrhythmias [16, 17, 36, 40, 44], the area is still unexplored for designing an efficient method for PxAF detection based on an optimized architecture of CNN. Moreover, the optimization process was not performed at the feature extraction level.

7.4 Design Parameters

Deep-PxAF learns the dynamic variation of the heartbeats at the feature learning level by designing customized architectures for recurrence images. Several design parameters are associated with the method at this level, such as the number of training epochs. We empirically obtained these parameters based on prior knowledge about the neural architecture search. PxAF yields higher performance compared to the results of conventional machine learning techniques that are automatically tuned by Auto-Sklearn. This primarily results from our custom-designed CNN architecture's higher feature extraction performance. On the other hand, manually tuning a generic CNN architecture [48] may result in lower accuracy in comparison with Auto-Sklearn.

7.5 Clinical Relevance

The classifier proposed by Deep-PxAF showed very high accuracy in detecting the pathological condition PxAF from the heart rate variability seen in normal conditions such as sinus rhythm. It is obvious that the classifier can be trained for detecting other pathological conditions. However, in order to be able to undertake the study in the patient level, we need a rich dataset of ECG signals in conjunction with comprehensive meta data of patient information. The resulting methods can be ultimately implemented in wearable ECG devices for detecting pathological conditions, i.e. PxAF, in a real-time manner. Nevertheless, a phase of clinical validation with a large number of individuals is necessitated to meet the standardization requirements. It is evident that pathological conditions like PxAF can lead to cardiac stroke, and hence, monitoring such a life-threatening condition can effectively reduce the aftermath consequences. Although the resulting classifier demands computational power in the training phase, the testing phase is light enough to be implemented in any kind of mobile technology, e.g. patch ECG, to be used for screening and patient monitoring in the clinical setting.

7.6 Statement of Reproducibility

To foster reproducibility:

- **Reproducibility analysis.** Many works on NAS have issues regarding reproducibility due to intrinsic stochasticity. To guarantee the reproducibility of results, we follow the Reproducibility checklist proposed by Lindauer et al. [34] (See Appendix A.1).
- **Code release.** Deep-PxAF is an open-source project. Code is made available on the GitHub repository through `www.github.com/0mehdi0/Deep-PxAF`.
- **Availability of database.** In this study, we evaluated our networks using the PhysioNet PAF database [8] that is available through `https://physionet.org/content/afpdb/1.0.0/`. Thus, this work does not involve any new data collection or human subject evaluation. The synthetic ECGs with the corresponding ground truth labels can be downloaded from the GitHub project repository: `www.github.com/0mehdi0/Deep-PxAF/tree/main/datasets`. The Deep-PxAF may be freely used for scientific use or commercial algorithm development if this paper is properly cited.

8 Conclusion & Future Work

This paper suggested an original combination of certified synthetic data generation in conjunction with the NAS method for classifying a vital pathological sign of ECG signal: Paroxysmal Atrial Fibrillation (PxAF). To overcome privacy and ethical concerns for data sharing, a GAN model was used to generate synthetic data. The synthetic ECGs were purified by an expert physician to discard the irrelevant ones. We employed a CNN for the classification, for which the optimal was found by the NAS. The input images to the CNN were extracted from the ECGs using recurrence graphs of the wavelet transform. It is found that the proposed framework offers a noticeable improvement in classification performance compared to the state-of-the-art as well as the existing benchmarks. In future work, the performance of the classifier resulting from this study will be practically explored on the general population after being implemented in an appropriate platform of wearable ECG.

References

[1] Edmond Adib, Fatemeh Afghah, and John J Prevost. Synthetic ecg signal generation using generative neural networks. *arXiv preprint arXiv:2112.03268*, 2021.

[2] Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, et al. Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. *Physiological measurement*, 41(12):124003, 2020.

[3] Rohan Banerjee and Avik Ghose. Synthesis of realistic ecg waveforms using a composite generative adversarial network for classification of atrial fibrillation. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 1145–1149. IEEE, 2021.

[4] Upasana Bhattacharjya and Kandarpa Kumar Sarma. Existing methods and emerging trends for novel coronavirus (covid-19) detection using residual network (resnet): A review on deep learning analysis. *Smart Healthcare Monitoring Using IoT with 5G*, pages 131–147, 2021.

[5] Han Cai, Tianyao Chen, Weinan Zhang, Yong Yu, and Jun Wang. Efficient architecture search by network transformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[6] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*, 2018.

[7] Zhipeng Cai, Zuobin Xiong, Honghui Xu, Peng Wang, Wei Li, and Yi Pan. Generative adversarial networks: A survey toward private and secure applications. *ACM Computing Surveys (CSUR)*, 54 (6):1–38, 2021.

[8] Gari D Clifford, Chengyu Liu, Benjamin Moody, H Lehman Li-wei, Ikaro Silva, Qiao Li, AE Johnson, and Roger G Mark. Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017.

[9] Felix Conrad, Mauritz Mälzer, Michael Schwarzenberger, Hajo Wiemer, and Steffen Ihlenfeldt. Benchmarking automl for regression tasks on small tabular data in materials design. *Scientific Reports*, 12(1):1–14, 2022.

[10] Anne Marie Delaney, Eoin Brophy, and Tomas E Ward. Synthesis of realistic ecg using generative adversarial networks. *arXiv preprint arXiv:1909.09150*, 2019.

[11] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018.

[12] Xuanyi Dong and Yi Yang. Searching for a robust neural architecture in four gpu hours. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1761–1770, 2019.

[13] Zahra Ebrahimi, Mohammad Loni, Masoud Daneshtalab, and Arash Gharehbaghi. A review on deep learning methods for ecg arrhythmia classification. *Expert Systems with Applications: X*, 7: 100033, 2020.

[14] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *The Journal of Machine Learning Research*, 20(1):1997–2017, 2019.

[15] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

[16] Najmeh Fayyazifar. An accurate cnn architecture for atrial fibrillation detection using neural architecture search. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 1135–1139. IEEE, 2021.

[17] Najmeh Fayyazifar, Selam Ahderom, David Suter, Andrew Maiorana, and Girish Dwivedi. Impact of neural architecture design on cardiac abnormality classification using 12-lead ecg signals. In *2020 Computing in Cardiology*, pages 1–4. IEEE, 2020.

[18] Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, and Frank Hutter. Auto-sklearn 2.0: Hands-free automl via meta-learning. *arXiv:2007.04074 [cs.LG]*, 2020.

[19] Leif Friberg, Niklas Hammar, Hans Pettersson, and Maå rten Rosenqvist. Increased mortality in paroxysmal atrial fibrillation: report from the stockholm cohort-study of atrial fibrillation (scaf). *Eur. Heart J.*, 28(19):2346–53, 2007.

[20] Arash Gharehbaghi and Maria Lindén. A deep machine learning method for classifying cyclic time series of biological signals using time-growing neural network. *IEEE Transactions on Neural Networks and Learning Systems*, 29(9):4102–4115, 2018. doi:10.1109/TNNLS.2017.2754294.

[21] Arash Gharehbaghi, Amir A Sepehri, Maria Lindén, and Ankica Babic. Intelligent phonocardiography for screening ventricular septal defect using time growing neural network. In *Studies in health technology and informatics*, volume 238, pages 108–111. IOC Press, 2017.

[22] Arash Gharehbaghi, Maria Lindén, and Ankica Babic. An artificial intelligent-based model for detecting systolic pathological patterns of phonocardiogram based on time-growing neural network. *Applied Soft Computing*, 83:105615, 2019. ISSN 1568-4946. doi:https://doi.org/10.1016/j.asoc.2019.105615. URL https://www.sciencedirect.com/science/article/pii/S1568494619303953.

[23] Cédric Gilon, Jean-Marie Grégoire, and Hugues Bersini. Forecast of paroxysmal atrial fibrillation using a deep neural network. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[24] Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. Generation and evaluation of synthetic patient data. *BMC medical research methodology*, 20(1):1–40, 2020.

[25] Bedartha Goswami. A brief introduction to nonlinear time series analysis and recurrence plots. *Vibration*, 2(4):332–368, 2019. ISSN 2571-631X. doi:10.3390/vibration2040021. URL https://www.mdpi.com/2571-631X/2/4/21.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[27] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[28] Rui Hu, Jie Chen, and Li Zhou. A transformer-based deep neural network for arrhythmia detection using continuous ecg signals. *Computers in Biology and Medicine*, 144:105325, 2022.

[29] Abdul Jabbar, Xi Li, and Bourahla Omar. A survey on generative adversarial networks: Variants, applications, and training. *ACM Computing Surveys (CSUR)*, 54(8):1–49, 2021.

[30] Hidefumi Kamozawa, Sho Muroga, and Motoshi Tanaka. A detection method of atrial fibrillation from 24-hour holter-ecg using cnn. *IEEJ Transactions on Electrical and Electronic Engineering*, 18 (4):577–582, 2023.

[31] Matloob Khushi, Kamran Shaukat, Talha Mahboob Alam, Ibrahim A Hameed, Shahadat Uddin, Suhuai Luo, Xiaoyan Yang, and Maranatha Consuelo Reyes. A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access*, 9:109960–109975, 2021.

[32] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

[33] Xiaomin Li, Anne Hee Hiong Ngu, and Vangelis Metsis. Tts-cgan: A transformer time-series conditional gan for biosignal data augmentation. *arXiv preprint arXiv:2206.13676*, 2022.

[34] Marius Lindauer and Frank Hutter. Best practices for scientific research on neural architecture search. *Journal of Machine Learning Research*, 21(243):1–18, 2020.

[35] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

[36] Zuhao Liu, Huan Wang, Yibo Gao, and Shunchen Shi. Automatic attention learning using neural architecture search for detection of cardiac abnormality in 12-lead ecg. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021.

[37] Mohammad Loni, Sima Sinaei, Ali Zoljodi, Masoud Daneshtalab, and Mikael Sjödin. Deepmaker: A multi-objective optimization framework for deep neural networks in embedded systems. *Microprocessors and Microsystems*, 73:102989, 2020.

[38] Mohammad Loni, Ali Zoljodi, Amin Majd, Byung Hoon Ahn, Masoud Daneshtalab, Mikael Sjödin, and Hadi Esmaeilzadeh. Faststereonet: A fast neural architecture search for improving the inference of disparity estimation on resource-limited platforms. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021.

[39] Mohammad Loni, Hamid Mousavi, Mohammad Riazati, Masoud Daneshtalab, and Mikael Sjödin. Tas:ternarized neural architecture search for resource-constrained edge devices. In *Design, Automation & Test in Europe Conference & Exhibition DATE'22, 14 March 2022, Antwerp, Belgium*. IEEE, March 2022. URL http://www.es.mdh.se/publications/6351-.

[40] Jindi Lv, Qing Ye, Yanan Sun, Juan Zhao, and Jiancheng Lv. Heart-darts: Classification of heartbeats using differentiable architecture search. *arXiv preprint arXiv:2105.00693*, 2021.

[41] Daniele Marinucci, Agnese Sbrollini, Ilaria Marcantoni, Micaela Morettini, Cees A Swenne, and Laura Burattini. Artificial neural network for atrial fibrillation identification in portable devices. *Sensors*, 20(12):3570, 2020.

[42] Patrick E McSharry, Gari D Clifford, Lionel Tarassenko, and Leonard A Smith. A dynamical model for generating synthetic electrocardiogram signals. *IEEE transactions on biomedical engineering*, 50 (3):289–294, 2003.

[43] George B Moody and Roger G Mark. The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001.

[44] Mohanad Odema, Nafiul Rashid, and Mohammad Abdullah Al Faruque. Eexnas: Early-exit neural architecture search solutions for low-power wearable devices. In *2021 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, pages 1–6. IEEE, 2021.

[45] Hisashi Ogawa, Yoshimori An, Syuhei Ikeda, Yuya Aono, Kosuke Doi, Mitsuru Ishii, Moritake Iguchi, Nobutoyo Masunaga, Masahiro Esato, Hikari Tsuji, Hiromichi Wada, Koji Hasegawa, Mitsuru Abe, Gregory Y.H. Lip, Masaharu Akao, and null null. Progression from paroxysmal to sustained atrial fibrillation is associated with increased adverse events. *Stroke*, 49(10):2301–2308, 2018. doi:10.1161/STROKEAHA.118.021396. https://www.ahajournals.org/doi/pdf/10.1161/STROKEAHA.118.021396. URL https://www.ahajournals.org/doi/abs/10.1161/STROKEAHA.118.021396.

[46] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[47] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In *International Conference on Machine Learning*, pages 4095–4104. PMLR, 2018.

[48] Bahareh Pourbabaee, Mehrsan Javan Roshtkhari, and Khashayar Khorasani. Deep convolutional neural networks and learning ecg features for screening paroxysmal atrial fibrillation patients. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(12):2095–2104, 2018.

[49] Pedro Henrique Ribeiro, Patryk Orzechowski, Joost Wagenaar, and Jason H Moore. Benchmarking automl algorithms on a collection of binary problems. *arXiv preprint arXiv:2212.02704*, 2022.

[50] Omid Sayadi, Mohammad B Shamsollahi, and Gari D Clifford. Synthetic ecg generation and bayesian filtering using a gaussian wave-based dynamical model. *Physiological measurement*, 31 (10):1309, 2010.

[51] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

[52] Abdelrahman M Shaker, Manal Tantawi, Howida A Shedeed, and Mohamed F Tolba. Generalization of convolutional neural networks for ecg classification using generative adversarial networks. *IEEE Access*, 8:35592–35605, 2020.

[53] Supreeth P Shashikumar, Amit J Shah, Gari D Clifford, and Shamim Nemati. Detection of paroxysmal atrial fibrillation using attention-based bidirectional recurrent neural networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 715–723, 2018.

[54] Feng Shi, Jun Wang, Jun Shi, Ziyan Wu, Qian Wang, Zhenyu Tang, Kelei He, Yinghuan Shi, and Dinggang Shen. Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for covid-19. *IEEE reviews in biomedical engineering*, 14:4–15, 2020.

[55] M Surucu, Y Isler, M Perc, and R Kara. Convolutional neural networks predict the onset of paroxysmal atrial fibrillation: Theory and applications. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(11):113119, 2021.

[56] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2820–2828, 2019.

[57] Vajira Thambawita, Jonas L Isaksen, Steven A Hicks, Jonas Ghouse, Gustav Ahlberg, Allan Linneberg, Niels Grarup, Christina Ellervik, Morten Salling Olesen, Torben Hansen, et al. Deepfake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. *Scientific reports*, 11(1):1–8, 2021.

[58] Allan Tucker, Zhenchen Wang, Ylenia Rotalinti, and Puja Myles. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *npj Digital Medicine*, 3(1):147, 2020.

[59] Heng-An Tzou, Shien-Fong Lin, and Peng-Sheng Chen. Paroxysmal atrial fibrillation prediction based on morphological variant p-wave analysis with wideband ecg and deep learning. *Computer Methods and Programs in Biomedicine*, 211:106396, 2021.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[61] Yi Xia, Yangyang Xu, Peng Chen, Jun Zhang, and Yongliang Zhang. Generative adversarial network with transformer generator for boosting ecg classification. *Biomedical Signal Processing and Control*, 80:104276, 2023.

[62] Min-Uk Yang, Dae-In Lee, and Seung Park. Automated diagnosis of atrial fibrillation using ecg component-aware transformer. *Computers in Biology and Medicine*, 150:106115, 2022.

[63] Lee-Ren Yeh, Yang Zhang, Jeon-Hor Chen, Yan-Lin Liu, An-Chi Wang, Jie-Yu Yang, Wei-Cheng Yeh, Chiu-Shih Cheng, Li-Kuang Chen, and Min-Ying Su. A deep learning-based method for the diagnosis of vertebral fractures on spine mri: retrospective training and validation of resnet. *European Spine Journal*, pages 1–9, 2022.

[64] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32, 2019.

[65] Peng Zhang, Chenbin Ma, Fan Song, Yangyang Sun, Youdan Feng, Yufang He, Tianyi Zhang, and Guanglei Zhang. D2afnet: A dual-domain attention cascade network for accurate and interpretable atrial fibrillation detection. *Biomedical Signal Processing and Control*, 82:104615, 2023.

[66] Fei Zhu, Fei Ye, Yuchen Fu, Quan Liu, and Bairong Shen. Electrocardiogram generation with a bidirectional lstm-cnn generative adversarial network. *Scientific reports*, 9(1):1–11, 2019.

A Supplementary materials

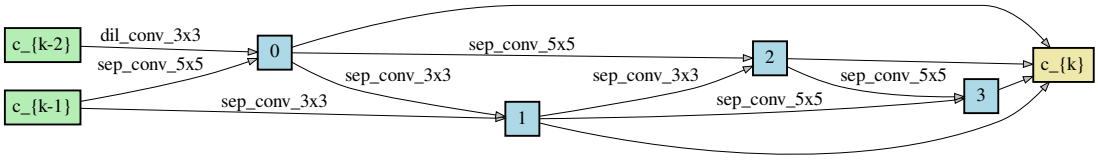
A.1 Reproducibility Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We discuss limitations of this work in Section 7
   (c) Did you discuss any potential negative social impacts of your work? [Yes] We discuss any societal impacts of this work in Section 7
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] All assumptions are described in the paper as well as the detail in the Appendix section.
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results, including all requirements (e.g., `requirements.txt` with explicit version), an instructive `README` with installation, and execution commands (either in the supplemental material or as a URL)? [Yes] Added all required hyper-parameters (Section 5), seeds, download links to databases, and GitHub repository.
   (b) Did you include the license of the datasets? [N/A] Our experiments were conducted on publicly available datasets and we have not introduced new datasets.
   (c) Did you include the raw results of running the given instructions on the given code and data? [Yes] All results are using the provided code.
   (d) Did you include scripts and commands that can be used to generate the figures and tables in your paper based on the raw results of the code, data, and instructions given? [Yes] See Code ReadMe file.
   (e) Did you ensure sufficient code quality such that your code can be safely executed and the code is properly documented? [Yes]
   (f) Did you specify all the training details (e.g., data splits, pre-processing, search spaces, fixed hyperparameter settings, and how they were chosen)? [Yes] For our experiments, we used the PhysioNet PxAF prediction challenge database (download link). Plus, all details are explained in Section 5 and Supplementary.
   (g) Did you ensure that you compared different methods (including your own) exactly on the same benchmarks, including the same datasets, search space, code for training and hyperparameters for that code? [Yes] Please see Section 6.2
   (h) Did you run ablation studies to assess the impact of different components of your approach? [Yes] Section 6.2 compares the classification results of the proposed method on three different datasets, including $D_{Original}$, D_{GAN}, and D_{CGAN}.
   (i) Did you use the same evaluation protocol for the methods being compared? [Yes]
   (j) Did you compare performance over time? [Yes] Anytime performance was assessed with the number of GPU hours as explained in Section 5.2.
   (k) Did you perform multiple runs of your experiments and report random seeds? [Yes] We re-ran the Deep-PxAF search procedure three more times with different random seeds to verify the reproducibility of the results. Results show that the average of multiple runs converges to neural architectures with similar results with the standard deviation (STD) of 0.2% for Deep-PxAF trained on D_{CGAN}. [Yes] Please check Table 5.
   (l) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

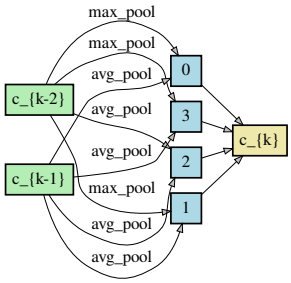
(m) Did you report how you tuned hyperparameters, and what time and resources this required (if they were not automatically tuned by your AutoML method, e.g. in a NAS approach; and also hyperparameters of your own method)? [N/A] In this paper, we did not use any method for optimizing learning hyperparameters. for the NAS, we use the DARTS method with default hyperparameters (please see Section 4.4).

A.2 Qualitative Analysis of the Searched Cells

Fig. 10 depicts the best cells searched by Deep-PxAF for the D_{CGAN} database. For the normal cell, DARTS tends to increase the portion of dilated convolution separable convolution (`sep_conv`) operations with the 5×5 kernel size. This is because larger kernel sizes (5×5) improve the representational power of the network. In contrast, the reduction cell has many average pooling operations for compressing the information across the spatial dimension. This is because pooling operations can increase the nonlinear representation ability of the network. Referring to the recurrence graphs in which rhythmic contents of ECG are preserved within the squares of 4 second (see Fig2, one can intuitively understand that an optimal kernel size is one that can include rhythms. A small kernel size can negatively impact the learning quality due to its failure to incorporate rhythmic content.



(a)



(b)

Figure 10: (a) Normal cell. (b) Reduction cell.

A.3 Details of Comparison Baselines

1. **Pourbabaee et al. [48]:** To obtain the optimal classification accuracy, a three-layer CNN architecture including convolutional, sub-sampling, and K-nearest neighbor (KNN) layers are utilized in which the optimal network parameters have been represented in Table 6.

2. **Surucu et al. [55]:** To achieve the best classification performance, a six-layer CNN architecture including three one-dimensional convolutional, dropout, pooling, and two fully-connected layers are utilized.

Table 6: The configuration setup of the PxAF detection method proposed by Pourbabaee et al. [48].

Parameter	**Value**
# Epochs	88
Optimizer	SGD
Learning Rate (lr)	0.09
Momentum	0.9
Sub-sampling Layer Kernel Size	128
# KNN Clusters	2