# Multiscaled Fusion of Deep Convolutional Neural Networks for Screening Atrial Fibrillation From Single Lead Short ECG Recordings

Xiaomao Fan, Qihang Yao, Yunpeng Cai, Fen Miao, Fangmin Sun, and Ye Li, *Member, IEEE*

***Abstract*—Atrial fibrillation (AF) is one of the most common sustained chronic cardiac arrhythmia in elderly population, associated with a high mortality and morbidity in stroke, heart failure, coronary artery disease, systemic thromboembolism, etc. The early detection of AF is necessary for averting the possibility of disability or mortality. However, AF detection remains problematic due to its episodic pattern. In this paper, a multiscaled fusion of deep convolutional neural network (MS-CNN) is proposed to screen out AF recordings from single lead short electrocardiogram (ECG) recordings. The MS-CNN employs the architecture of two-stream convolutional networks with different filter sizes to capture features of different scales. The experimental results show that the proposed MS-CNN achieves 96.99% of classification accuracy on ECG recordings cropped/padded to 5 s. Especially, the best classification accuracy, 98.13%, is obtained on ECG recordings of 20 s. Compared with artificial neural network, shallow single-stream CNN, and VisualGeometry group network, the MS-CNN can achieve the better classification performance. Meanwhile, visualization of the learned features from the MS-CNN demonstrates its superiority in extracting linear separable ECG features without hand-craft feature engineering. The excellent AF screening performance of the MS-CNN can satisfy the most elders for daily monitoring with wearable devices.**

***Index Terms*—Deep convolutional neural network, ECG, atrial fibrillation, deep learning, classification, biomedical monitoring.**

## I. Introduction

THE boom of the aging population has raised more concerns about the health problem of elders. Cardiovascular disease, which is the age-related disease, has been identified as the leading cause of morbidity and mortality in developed countries [1]. More importantly, atrial fibrillation (AF) which stems from disordered activation and irregular atrial contraction is associated with cardiovascular disease, with a substantial increase in the risk of stroke, heart failure, coronary artery disease, systemic thromboembolism etc. [2], [3]. The attack of AF is often paroxysmal and happens outside the hospital. Dynamic electrocardiogram (ECG), a long-term and continuous physiological signal collected by Holter, is one important tool to monitor AF. However, wearing Holter is cumbersome and has a great impact on people's daily activities. An alternative solution is to utilize wearable devices, with portability, usability, and comfortability, to monitor AF. Wearable devices conventionally acquire single lead dynamic ECG recordings, but there is much high noises and artifacts. Therefore, the early diagnosis of AF using such dynamic ECG recordings is of great significance.

Generally, an ECG recording, which is conventionally used for AF detection, consists of P, Q, R, S, T, and U wave. Fig. 1 shows an ECG recording acquired from an AF patient with characteristics of P-wave absence or irregular variability of RR intervals. In the past few decades, researchers have proposed lots of effective automatic detectors of AF using ECG recordings acquired by Holter. These automatic AF detectors can be summarized as being based on P-wave absence [4], [5] or variability of RR interval. The automatic detectors of AF based on P-wave absence has not gained wide application in Holter monitoring due to its significant limitations. It is a pretty difficult task to locate the fiducial position of a small P wave, especially in the case of noise. Subsequently, the variability of RR intervals has been more frequently used for automatic detection of AF instead. Dash *et al.* [6] proposed an automatic detection algorithm of AF based on the randomness, variability and complexity of the heartbeat interval time series. Lian *et al.* [7] proposed an AF detection algorithm based on a map that showed the scatter plot of RR intervals versus change of RR intervals. Roonizi

Manuscript received December 31, 2017; revised June 7, 2018 and July 10, 2018; accepted July 14, 2018. Date of publication August 6, 2018; date of current version October 15, 2018. This work was supported in part by the Major Special Project of Guangdong Province under Grant 2017B030308007, in part by the Shenzhen Basic Research Projects (JCYJ20170818163445670, JCYJ20170413161515911, and JCYJ20170412110753954), in part by the Guangdong Applied Special Project for Research and Development under Grant 2015B010129012, and in part by the Special Fund Project for Overseas High-level Talents Innovation and Entrepreneurship Grant KQJSCX20170731165939298. *(Corresponding author: Ye Li.)*

X. Fan is with the Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: xm.fan@siat.ac.cn).

Q. Yao, Y. Cai, F. Miao, and F. Sun are with the Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: qh.yao@siat.ac.cn; yp.cai@siat.ac.cn; fen.miao@siat.ac.cn; fm.sun@siat.ac.cn).

Y. Li is with Key Laboratory for Health Informatics of the Chinese Academy of Sciences, Shenzhen Institutes of advanced technology, Shenzhen 518055, China (e-mail: ye.li@siat.ac.cn).

Digital Object Identifier 10.1109/JBHI.2018.2858789

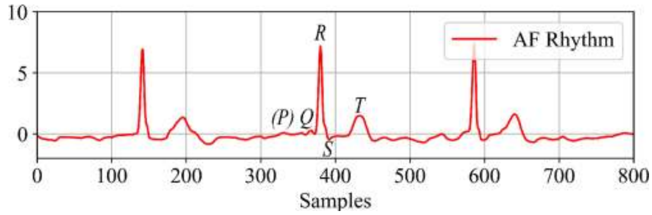


Fig. 1. AF patient ECG recording with characteristics of P-wave absence or irregular variability of RR interval.

*et al.* [8] extended nonlinear Bayesian filtering framework for analysis of AF in single channel ECG recordings. Huang *et al.* [9] proposed a novel method for detection of the transition between AF and sinus rhythm based on RR intervals. Even though existing automatic detection methods of AF based on the variability of RR intervals has achieved high accuracy on ECG recordings acquired by Holter. However, these kind of methods still has deficiencies to obtain a good performance on screening out AF from single lead ECG recordings collected by wearable devices due to its high noise and artifacts.

Recent years, deep convolutional neural networks, with a strong capability in feature extraction, have achieved a great success in computer vision [10]–[15]. Many researchers have tried to use deep convolutional neural networks to solve the problem of AF detection. Pourbabaee *et al.* [16] developed a novel computationally intelligence-based automatic AF detection method using deep convolutional networks. The features of AF were learned automatically and then utilized in classification module. This method simplified the feature extraction process without the help of feature engineering by expert to define appropriate and critical features. Limam *et al.* [17] proposed a convolutional recurrent neural network (CRNN). The CRNN consisted of two independent CNNs to extract related features, one from ECGs and the other from heart rates. The final performance was evaluated by support vector machine(SVM). Yao *et al.* [18] proposed a multi-scale convolutional neural networks (MCNN), which applied time scaling on input signals and detected AF based on scaled inputs. In their experiment a strong correlation between the depth of the MCNN and the detection performance was shown. Although the above mentioned methods are effective on solving the problem of AF detection, AF detection remains problematic as this kind of methods are generally limited in applicability. Their good performances are achieved based on carefully-selected often clean data and only a small number of patients were used. Reliable AF detection from single lead short ECG recordings is still a big challenge and particularly difficult due to the limited rhythms information.

In this study, a multi-scaled fusion of deep convolutional neural network (MS-CNN), inspired by [10], [14], [15], is proposed to detect AF signals based on single lead short ECG recordings. To summarize, the main contributions are described to be: (1) The proposed MS-CNN employs two streams of convolutional networks with different filter size, which could capture ECG features with different scales and achieve a excellent performance for screening out AF from single lead short ECG recordings. (2) The effect of the proposed feature learning mechanism is investigated extensively by visualizing learned features in different depths of layers.

This paper is organized as follows: Section II introduces the architecture of the MS-CNN. And the detailed experimental process is described in Section III. Section IV presents the experimental results and discussions. Section V concludes the paper.

## II. Methods

### A. Problem Formulation

The task of real-time AF detection is to identify the AF signals from single lead short ECG recordings. The training set $X = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}$ consists of input ECG recording $x^{(i)}$ and label $y^{(i)}$, where $x^{(i)} \in R^n$ and $y^{(i)} \in \{0, 1\}$. The label $y^{(i)} = 1$ means that the corresponding input ECG recording $x^{(i)}$ is the AF recording. The label $y^{(i)} = 0$ corresponds to the normal sinus ECG recording. The proposed MS-CNN accepts $x^{(i)}$ as input and $z^{(i)}$ as output. The details refer to equation (1) which can be defined as:

$$z^{(i)} = F(x^{(i)}; \theta) \tag{1}$$

where $F(\cdot)$ is the function which describes the process of an ECG recording from the input layer to the last fully connected layer. The $\theta$ is the related parameters of the MS-CNN. Equation (2) is the softmax function which transforms output value $z^{(i)}$ into possibility to categorize the input ECG recording as a normal or AF signal.

$$p(z^{(i)}) = \frac{\exp\left(z^{(i)}\right)}{\sum_j \exp\left(z^{(i)}\right)} \tag{2}$$

where $p\{\cdot\}$ is the probability which the MS-CNN assigns to the $i$-th output label taking the input value $z^{(i)}$. The cross entropy of these outputs with respects to their true labels is then calculated to generate a loss, which effectively evaluates the performance of the model on training data. For $m$ observable instances in the training set, the objective loss function can be defined as:

$$L(X) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=0}^{1} 1\{y^{(i)} = j\} \log p(z^{(i)}) \tag{3}$$

where $1\{\cdot\}$ is the indicator function, so that $1\{\cdot\} = 1$ if the statement is true, and $1\{\cdot\} = 0$ if the statement is false. $p\{\cdot\}$ is the function from equation (2).

### B. Model Architecture

The architecture of the MS-CNN is designed and built drawing on the experiences of VGGNet(VisualGeometry group network) [10], a mature neural network which has been proved to be effective in solving a variety of problems in computer vision field. The MS-CNN is composed of 2 streams of 13-layer convolutional neural networks (Conv) and 3 fully connected layers (FC) after them, as shown in Fig. 2. The nonparametric layers, including five pooling layers and one softmax layer, also play important roles in this architecture. Each fully connected layer provides a transformation function $f(\cdot)$ mapping input $\boldsymbol{x}$ to its

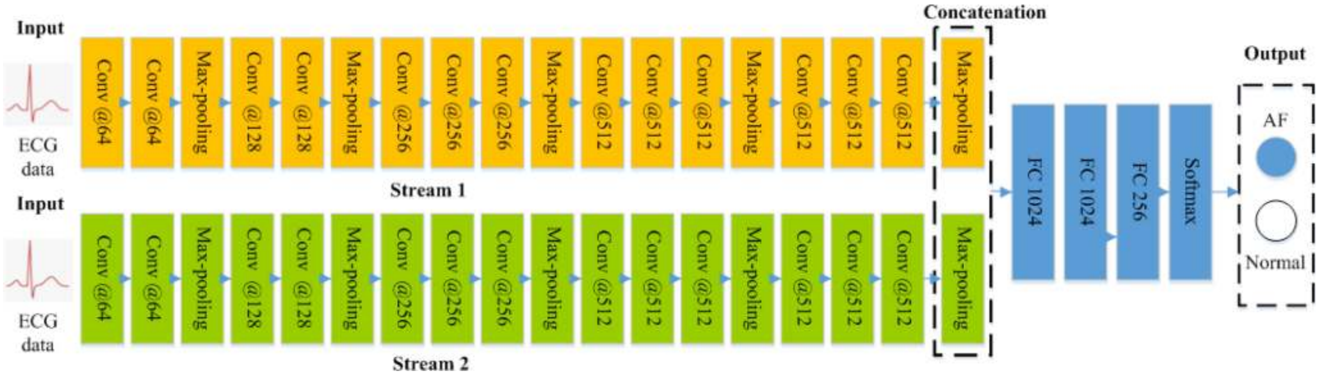


Fig. 2. The network architecture of the MS-CNN. The MS-CNN has two stream networks, which consist of 13 layer convolutional neural network (Conv), 5 maxpooling layers, and 3 fully connected layers (FC).

TABLE I
THE CONFIGURATION OF THE MS-CNN

| | MS-CNN(3, 3) | | MS-CNN(3, 5) | | MS-CNN(3, 7) | | MS-CNN(3, 9) | |
|---|---|---|---|---|---|---|---|---|
| | Input ECG recordings | | | | | | | |
| Layer 1 | Stream 1 | Stream 2 | Stream 1 | Stream 2 | Stream 1 | Stream 2 | Stream 1 | Stream 2 |
| Layer 2 | Conv(3)-64 | Conv(3)-64 | Conv(3)-64 | Conv(5)-64 | Conv(3)-64 | Conv(7)-64 | Conv(3)-64 | Conv(9)-64 |
| Layer 3 | Conv(3)-64 | Conv(3)-64 | Conv(3)-64 | Conv(5)-64 | Conv(3)-64 | Conv(7)-64 | Conv(3)-64 | Conv(9)-64 |
| | Max-pooling, Stride(3) | Max-pooling, Stride(3) | Max-pooling, Stride(3) | Max-pooling, Stride(3) | Max-pooling, Stride(3) | Max-pooling, Stride(3) | Max-pooling, Stride(3) | Max-pooling, Stride(3) |
| Layer 4 | Conv(3)-128 | Conv(3)-128 | Conv(3)-128 | Conv(5)-128 | Conv(3)-128 | Conv(7)-128 | Conv(3)-128 | Conv(9)-128 |
| Layer 5 | Conv(3)-128 | Conv(3)-128 | Conv(3)-128 | Conv(5)-128 | Conv(3)-128 | Conv(7)-128 | Conv(3)-128 | Conv(9)-128 |
| | Max-pooling, Stride(3) | Max-pooling, Stride(3) | Max-pooling, Stride(3) | Max-pooling, Stride(3) | Max-pooling, Stride(3) | Max-pooling, Stride(3) | Max-pooling, Stride(3) | Max-pooling, Stride(3) |
| Layer 5 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 |
| Layer 6 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 |
| Layer 7 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 | Conv(3)-256 |
| | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) |
| Layer 8 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 |
| Layer 9 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 |
| Layer 10 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 |
| | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) |
| Layer 11 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 |
| Layer 12 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 |
| Layer 13 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 | Conv(3)-512 |
| | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) | Max-pooling, Stride(2) |
| | Concatenation | | Concatenation | | Concatenation | | Concatenation | |
| Layer 14 | FC-1024 | | FC-1024 | | FC-1024 | | FC-1024 | |
| Layer 15 | FC-1024 | | FC-1024 | | FC-1024 | | FC-1024 | |
| Layer 16 | FC-256 | | FC-256 | | FC-256 | | FC-256 | |
| | Softmax | | Softmax | | Softmax | | Softmax | |

output $\boldsymbol{y}$ [19]. The detailed form of this function is given as:

$$\boldsymbol{y} = f(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \tag{4}$$

where $\boldsymbol{W}$ is a weight matrix and $\boldsymbol{b}$ is a bias vector. They are all trainable variables which help to form various linear transformations, and $f(\cdot)$ is the predefined activation function adding nonlinearity to this function.

Convolutional layers in the fully convolutional neural networks use convolution to replace the multiplication in the fully connected layers. As the outputs of convolutional layers maintain spatial locality, they are commonly referred to as feature maps. The corresponding function is defined to be:

$$\boldsymbol{y}_j = f\left(\sum_{i \in M} \boldsymbol{k}_{ij} * \boldsymbol{x}_i + \boldsymbol{b}_j\right) \tag{5}$$

where $M$ is the filter size, $j$ denotes the index of convolution kernels, $i$ denotes the index of input feature maps, and $\boldsymbol{k}_{ij}$ denotes the convolution kernel for the $i$-th input map and $j$-th output map. This transition greatly reduces the number of parameters needed to be trained, and leads to better utilities as convolution has already been widely used in signal filtering. Meanwhile, multiple feature maps are generated using different trainable kernels, in order to compensate for the reduced transformation complexity. Pooling layers reduce the size of input by assembling neighboring data points. The max-pooling layers used in the proposed model choose to output only the largest number in each size-predefined blocks. These layers help to reduce the complexity of following calculation, and also introduce translation invariance to the model. Each stream of 13-layer convolutional neural networks, which refers to the VGGNet, has 13 convolutional layers and 5 max-pooling layers. 2 or 3 convolutional layers are grouped together, sharing the same number of filters, and groups are separated by max-pooling layers. Therefore, 2 groups of 2 convolutional layers, and 3 groups of 3 convolutional layers form the entire convolutional network, together with max-pooling layer behind right after the group. The number of convolutional filters is 64 in the first group and then increases by a factor of two after each max-pooling layer, until it reaches 512.

In one stream of convolutional neural network, the kernel size in all layers is three. In another stream, the convolutional kernels sizes in first four layers are set larger, in order to capture features of different scales from the input ECG recordings. In this paper, the alternative filter sizes including 3, 5, 7, 9 are tested. In remaining layers, the kernel size remains to be three. The stride of the initial two max-pooling layers is three, which enables more aggressive computing complexity reduction while the feature extraction performance wouldn't be compromised much as these features are very shallow. In addition, a stride size of two is applied in the remained max-pooling layers due to the concentration of feature information in deeper layers. At the end of the MS-CNN, softmax layer is applied to transform real values into probabilities. The detailed configurations of the MS-CNN evaluated in this paper are outlined in Table I. The MS-CNN (3, $n$) denotes that the receptive field sizes of two stream neural networks in first four convolutional layers are 3 and $n$, respectively. The activation function that all hidden layers of the MS-CNN are equipped with is rectified linear unit (ReLU) [20], which has many advantages such as sparse activation, efficient gradient propagation, and efficient computation. Mini-batch gradient descent [21] is utilized as the optimizer, which computes the gradients and updates the network parameters based on a small batch of samples in each step. $L2$ regularization is a good way to improve generalization error and used in the last three fully connected layers of the MS-CNN [22].

## III. Experiment

The experiment on atrial fibrillation detection consists of data preprocessing, cross validation, and model parameter optimization. Details of this experiment process are illustrated in Fig. 3.

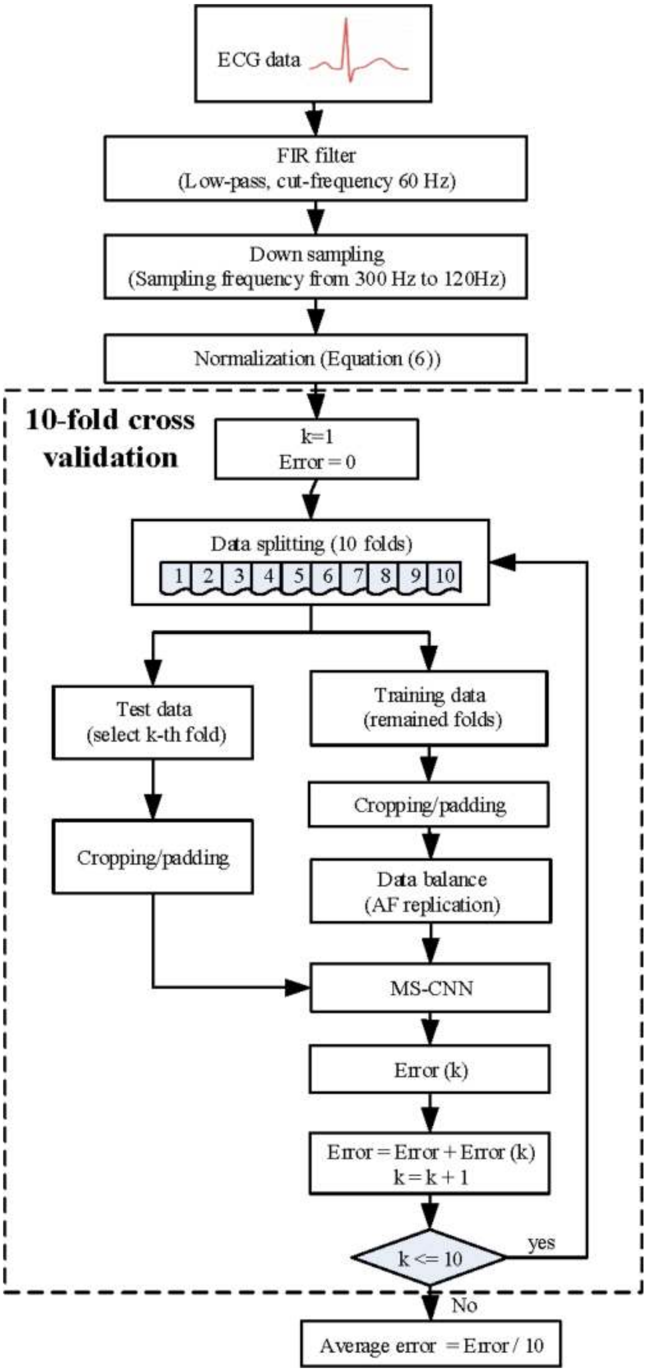


Fig. 3. Block diagram of AF screening experimental pipeline.

### A. Data Preprocessing

*1) Downsampling:* It is known that the complexity of neural networks is partially dependent upon the input size. In order to reduce the proposed model size and shorten the training time, downsampling technology is employed in this paper. According to the periodograms which represent input records' power spectral density showed in Fig. 4, ECG components with cut-off frequency higher than 60 Hz can be negligible on an AF detection problem. In accordance with the main frequency range presented, input ECG recordings are filtered with a 512-order lowpass finite impulse response (FIR) filter which has a cut-off frequency of 60 Hz, and then downsampled from original 300 Hz to 120 Hz. The FIR filter is applied to avoid aliasing, and the cut-off frequency is determined by the Nyquist Theorem [23].

*2) Normalization:* There is a large variation in terms of amplitudes of ECG recordings among different people, or even the same person with different lead positions. In practice, it is found that neural network models tend to converge better when all inputs have similar distributions. So we subtracted the mean value from each recording and then divided them with their standard deviation, by which the effects of different amplitudes among ECG recordings can be eliminated. The normalization function

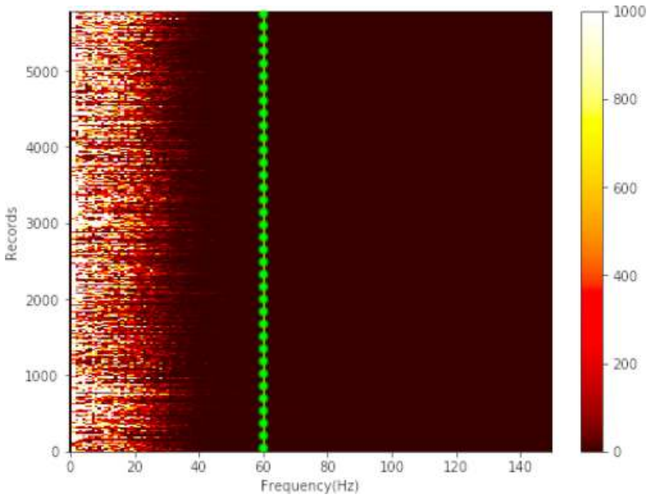


Fig. 4. Power spectral density estimate using periodogram. Cut-off frequency of FIR filter is 60 Hz.

TABLE II
CLASS DISTRIBUTION OF THE DATASET

| Type | Recording | Average Time Length(s) |
|---|---|---|
| Normal | 5154 | 31.9 |
| AF | 771 | 31.6 |
| Other rhythm | 2557 | 34.1 |
| Noisy | 46 | 27.1 |
| Total | 8528 | 32.5 |

TABLE III
CLASSIFICATION PERFORMANCE OF THE MS-CNN

| Model | Input length | $Sen$ | $Spe$ | $Pre$ | $Acc$ |
|---|---|---|---|---|---|
| MS-CNN(3, 3) | 5s | 91.6% | 97.52% | 84.39% | 96.77% |
| | 10s | 93.09% | 98.28% | 88.76% | 97.62% |
| | 20s | 93.09% | 98.63% | 90.94% | 98.03% |
| | 30s | 85.91% | 98.91% | 92.02% | 97.25% |
| MS-CNN(3, 5) | 5s | 92.41% | 97.70% | 85.43% | **96.99%** |
| | 10s | 93.36% | 98.24% | 88.56% | 97.62% |
| | 20s | 93.77% | 98.71% | 91.41% | 98.08% |
| | 30s | 89.16% | 97.54% | 84.14% | 96.48% |
| MS-CNN(3, 7) | 5s | 92.28% | 97.60% | 84.91% | 96.92% |
| | 10s | 94.31% | 98.22% | 88.55% | **97.72%** |
| | 20s | 93.77% | 98.77% | 91.78% | **98.13%** |
| | 30s | 89.92% | 98.83% | 91.38% | **97.75%** |
| MS-CNN(3, 9) | 5s | 92.41% | 97.62% | 85.04% | 96.96% |
| | 10s | 93.63% | 98.34% | 89.16% | 97.14% |
| | 20s | 92.68% | 98.61% | 90.72% | 97.86% |
| | 30s | 88.62% | 98.73% | 91.09% | 97.44% |

is defined as:

$$Normalized(X) = \frac{X - \overline{X}}{S} \tag{6}$$

where $X$ refers to the ECG recording values, and $\overline{X}$, $S$ refers to the average and standard error of these values, correspondingly.

*3) Data Balance:* For ECG datasets, it is common that the number of ECG recordings labeled with AF are less than that of ECG recordings labeled with normal. This biased distribution results in much higher difficulty to screen an AF patient than a normal case. Meanwhile, imbalanced training data has negative impacts on deep learning model as the training model would favor the dominating classes. Even worse, there would be only several AF recordings in a batch when mini-batch training scheme is applied.

To solve this problem, AF recordings can be replicated in the training dataset to enhance effects on the gradient direction. In this paper, the AF recordings are replicated three times in the training dataset. This replication operation makes AF recordings more than 30% in the input ECG recordings, which could balance the data-biased distribution and over-fitting problem.

*4) Cropping and Padding:* The input ECG recordings have varied lengths from 9 to 61 seconds. Varied-length input fails to fit with convolutional neural network models. Thus, short ECG recordings in fixed length of 5 seconds, 10 seconds, 20 seconds, and 30 seconds are selected as the input of the MS-CNN to verify the influence of different input sizes on the model performance. ECG recordings that are too long should be cropped and those that are too short should be padded with zeros in order to fit the model. However, fixed cropping fails to utilize the entire training dataset. Therefore, the MS-CNN is programed to be able to automatically fetch data batches from the dataset and crop or pad them randomly. In contrast, the test data are centrally cropped or padded to have fair performance measurement between models.

## B. Cross Validation

As the target dataset is relatively small, 10-fold cross validation is introduced. The original dataset is randomly partitioned into ten equal sized subsets. 10 models are trained with each of the 10 subsets used exactly once as the test data and remaining parts as training data. Then, classification results of these models are gathered and combined, to estimate the model's average performance on the entire dataset.

## C. Model Parameters Selection

*1) Activation Function:* Activation function plays a very important role since it introduces nonlinear factors into deep learning model. However, the gradient saturation problem exists in common used activation function, such as sigmoid [24] and tanh [25]. ReLU (Rectified Linear Unit) function is a popular activation function in convolutional network in recent years. Compared with sigmoid and tanh function, the gradient saturation problem can be avoided well in ReLU function when the input is positive. Therefore, the ReLU function is chosen as the activation function of the MS-CNN. And the ReLU function is defined as:

$$f(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases} \tag{7}$$

*2) Model Optimization:* Starting from the fundamental gradient descent algorithm, optimization algorithms of deep learning have been constantly developed. Mini-batch gradient descent, also known as stochastic gradient descent in deep learning, is introduced to reduce the time cost of each step of training. Momentum [26] is introduced to avoid the local optima problem. Adaptive gradient descent (Adagrad) [27] adapts the learning rate to different parameters, performing larger updates for infrequent parameters and smaller updates for frequent parameters. Adaptive moment estimation (Adam) [28], [29], which stores exponentially decaying averages of both past gradients and past squared gradients and updates parameters based on them, is usually considered having the best overall performance. Nevertheless, in many practices it has been found that stochastic gradient descent (SGD) with a simple learning rate decay schedule usually manages to find a minimum, though with other optimizers model converges faster. Therefore, in this experiment, simple stochastic gradient descent with batch size of 128, initial learning rate of 0.01, and exponential learning rate decay of 10 times in every 300 epoch, is applied. The learning rate $\eta$ is decayed as following:

$$\eta = \eta_0 \cdot 0.1^{\frac{N}{300}} \tag{8}$$

where $\eta_0$ is the initial learning rate, and $N$ is the number of epochs.

*3) $L2$ Regularization:* In the field of machine learning, the problem of over-fitting becomes more serious when the learned model is with high complexity, which tends to fit not only between data and label, but also between noise and random error. To avoid over-fitting, regularization is introduced to the MS-CNN. In this experiment, $L2$ norm of model parameters for the last three fully connected layers are added to the loss function, with a penalty factor λ.

$$l(x) = L(X) + \lambda \sum_{i=1}^{3} \|W_i\|^2 \tag{9}$$

where $l(\cdot)$ is the total loss function with $L2$ regularization. $L(\cdot)$ is the cross entropy loss from equation (3). $W_i$ is the weight parameters of the $i$-th fully connected layers.

## IV. Results and Discussion

### A. Experimental Environment

The MS-CNN runs on the prevailing deep learning framework Tensorflow1.2.1, using the CentOS7.3 operating system. The platform of Tensorflow is deployed on the high-performance computing server equipped with two 16-core Intel Xeon E5-2640 V3 processors with 128 GB memory. The computing server is also equipped with eight NVIDIA Tesla K80 with 13 multi-processors (total 19,968 cores) and 12 GB memory (total 96 GB memory).

### B. Data Source

Provided by the PhysioNet,[1] a total of 8,528 ECG records (about 23,878,400 heartbeats) lasting from 9 s to 61 s were gathered to support its Computing in Cardiology Challenge 2017. One of three generations of AliveCor's single-channel ECG device was utilized to collect ECG recordings from health-monitored users. These noise-contaminated recordings sampled at 300 Hz were labelled by an ECG expert in four classes, normal rhythm, AF rhythm, other rhythm and noisy recordings. While those recordings are suitable for practical analysis for their varied length and noise-contaminated feature, they are also seriously imbalanced, for only 771 in 8528 of the total records are of AF rhythm, which are of major concern. The class distribution of the dataset is shown in Table II.

In this paper, normal and AF recordings are selected in this work, as strong waveform variation lies in the group of ECG recordings annotated as other rhythms and noisy recordings. 5154 normal sinus recordings and 771 AF recordings are used in experiments, with aforementioned strategy to tackle with the problem of unbalanced distribution.

### C. Classification Performance

In this paper, sensitivity ($Sen$), specificity ($Spe$), precision ($Pre$) and accuracy ($Acc$) are four metrics for evaluating classification performance of the MS-CNN. Based on true positive ($TP$), true negative ($TN$), false positive ($FP$) and false negative ($FN$), the definition of this four metrics can be defined as:

$$Sen = \frac{\#(TP)}{\#(TP) + \#(FN)} \tag{10}$$

$$Spe = \frac{\#(TN)}{\#(TN) + \#(FP)} \tag{11}$$

$$Pre = \frac{\#(TP)}{\#(TP) + \#(FP)} \tag{12}$$

$$Acc = \frac{\#(TP) + \#(TN)}{\#(TP + TN + FN + FP)} \tag{13}$$

As shown in Table III, the MS-CNN with different configuration of convolutional filter sizes are evaluated. At the same time, input ECG recordings of different lengths (5 seconds, 10 seconds, 20 seconds, and 30 seconds) are tested. The experimental results show that the MS-CNN(3, 5) achieves the best classification accuracy, 96.99%, when the input ECG recording is 5 seconds long. A typical 5-second ECG signal contains about 6–7 heartbeats. For an ECG recording of this length, the classification accuracies of the MS-CNN(3, 5), MS-CNN(3, 7), and MS-CNN (3, 9) are almost the same but better than that of the MS-CNN(3, 3). It is likely that the MS-CNN(3, 3) can not capture multi-scaled features from ECG recordings due to its same filter size. Through the case of the MS-CNN(3, 7)'s performance in classification of longer input recordings, it is demonstrated that multi-scaled filters could increase the AF screening performance. For the ECG length of 10 seconds, 20 seconds, and 30 seconds, the MS-CNN(3, 7) has outperformed other configurations. Using this configuration, the highest classification accuracy of 98.13% is achieved when the length of input recordings is 20 seconds. In theory, a longer ECG signal, which covers more heartbeat rhythm information, would lead to better classification performance. However, in all configurations the best classification performance is achieved when input ECG recordings are 20 seconds long. This result is strongly related to the length distribution of the ECG dataset we used. 10.95% of ECG recordings are shorter than 30 seconds while about 6.77% of ECG recordings are shorter than 20 seconds. An ECG signal needs

[1] https://www.physionet.org/challenge/2017/

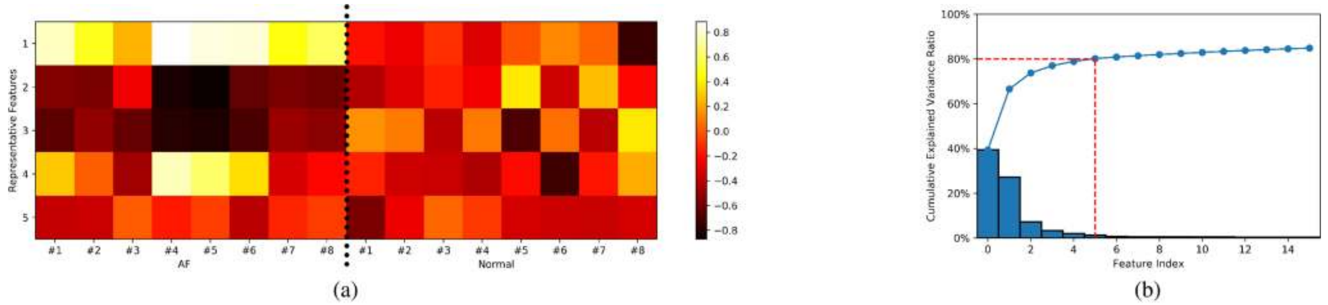


Fig. 5. Visualization of the learned features (a) Heat map of learned features from healthy and AF cases. The first eight columns correspond to the AF patient cases. The remaining columns are healthy cases. The rows are the representative features from the final layers features through PCA. The features between AF patients and healthy individuals are easily discriminated. (b) Pareto of learned features. Top 5 features take up over 80% information of all features.

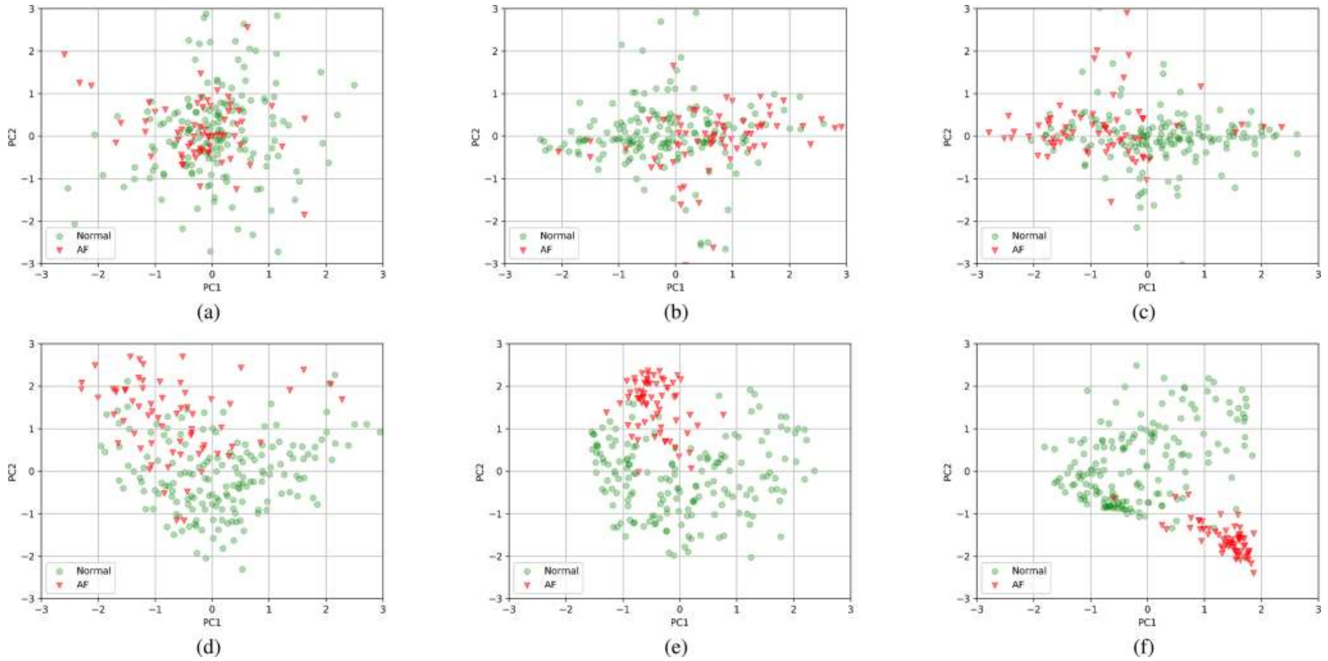


Fig. 6. Classification of learned features from representative layers in the MS-CNN(3, 7): (a) 2nd layer, (b) 4th layer, (c) 7th layer, (d) 10th layer, (e) 13th layer, (f) 16th layer. AF means AF cases. Normal means health cases. Both PC1 and PC2 are top two principle components.

to be padded with zeros when its length is shorter than the fixed length of the input ECG recordings for the MS-CNN. The number of input ECG recordings which need to be padded to 30 seconds long are larger than that of those which need to be padded to 20 seconds long. More noises are introduced to the input ECG recordings when more zeros are padded into an ECG recording, which also lead to worse AF screening performance. Meanwhile, The MS-CNN(3, 7), the filter sizes of which are much different in two-stream convolutional networks, has the ability to obtain the features with different scales. Therefore, the classification performance of the MS-CNN on 20-second-long ECG recordings is better than that on 30-second-long ECG recordings.

### D. Visualization of Learned Features

The MS-CNN(3, 7) which performs better in most of cases is selected to visualize the learned features. Eight healthy and eight AF cases are randomly selected from test set as the input. The input ECG recordings are cropped/padded to 20 seconds long. Instead of applying softmax function and outputting the classification results, the softmax layer is discarded, the remaining layers of the MS-CNN(3, 7) are utilized as a feature extractor. The generated feature vectors consist of 256 values, outputted by 256 neurons in the final fully connected layer. In order to get better representative features, our experiments take advantage of unsupervised principle component analysis (PCA) method to reduce the dimension of feature vectors from 256 to 5. The first 5 feature attributes cover over 80% information showed in Fig. 5(b). And the top two components take up 66% of feature information. Meanwhile, visualization of learned features on a heat map is drawn in Fig. 5(a). The 8 cases on the left side of the horizontal axis are healthy cases, and the 8 cases on the right side of the horizontal axis are AF cases. The top 5 components from the PCA method is on the vertical axis. As shown in Fig. 5(a), healthy cases and AF cases could be clearly discriminated using those 5 representative features. Associated with the top one feature (1st row in Fig. 5(a)), the feature colors

TABLE IV
CLASSIFICATION PERFORMANCE COMPARISON ON AF VERSUS NORMAL SINUS RHYTHM)

| Model | Input length | $Sen$ | $Spe$ | $Pre$ | $Acc$ |
|---|---|---|---|---|---|
| ANN[30] | 5s | 19.92% | 95.50% | 39.30% | 85.87% |
| | 10s | 21.27% | 93.82% | 33.48% | 84.57% |
| | 20s | 18.02% | 92.14% | 25.09% | 82.69% |
| | 30s | 18.16% | 90.57% | 21.97% | 81.34% |
| CNN[16] | 5s | 59.21% | 93.31% | 56.39% | 88.96% |
| | 10s | 67.07% | 93.19% | 59.00% | 89.86% |
| | 20s | 65.31% | 93.13% | 58.14% | 89.58% |
| | 30s | 59.89% | 93.37% | 56.89% | 89.10% |
| VGGNet[10] | 5s | 91.87% | 97.72% | 85.50% | 96.68% |
| | 10s | 90.92% | 98.57% | 90.31% | 97.60% |
| | 20s | 92.68% | 98.81% | 91.94% | 98.03% |
| | 30s | 87.88% | 98.05% | 85.29% | 96.89% |
| AliveCor[31] | 30s | 85.00% | 90.00% | - | - |
| The proposed | 5s | 92.41% | 97.70% | 85.43% | **96.99%** |
| | 10s | 94.31% | 98.22% | 88.55% | **97.72%** |
| | 20s | 93.77% | 98.77% | 91.78% | **98.13%** |
| | 30s | 89.92% | 98.83% | 91.38% | **97.75%** |

of healthy cases are more light than that of AF cases. Compared with AF cases, the second and third features (2nd and 3rd row) are in darker colors. There is no big difference on the 4th and 5th features which cover feature information lower than 4.5%. Even so, the clear difference of the top 3 learned features between healthy and AF cases help to explain why the proposed MS-CNN(3, 7)'s classification performs well.

To further capture and understand the learning procedure of the MS-CNN(3, 7) , some representative layers (2nd layer, 4th layer, 7th layer, 10th layer, 13th layer, and 16th layer) are selected to extract low dimensional features from feature mapping vectors using PCA method respectively. Fig. 6 shows the first two principal components of those aforementioned 16th recordings. It is demonstrated obviously that healthy and AF cases become more distinguishable as the depth of the MS-CNN (3, 7) increases. For example, AF and healthy cases are mingled together tightly in the 2nd layer (Fig. 6(a)). In the 4th and 7th layer (Fig. 6(a) and (b)), both AF and healthy cases are slowly clustering but not clearly separated. When the depth of the MS-CNN(3, 7) increases to 10 layers (Fig. 6(d)), over half of healthy cases can be separated from the AF cases. In the 13th layer (Fig. 6(e)), most of healthy and AF cases are separated well and just a few cases are mingled together. As shown in Fig. 6(f), AF and healthy cases can be well separated linearly in the 16th layer (final layer in the MS-CNN). It is also observed that AF cases are more concentrated than health cases, which can be ascribed to the facts that healthy cases have more waveform patterns such as P wave ,and the number of health cases are much more than that of AF cases. Nevertheless, the experiment results show that the MS-CNN(3, 7)'s classification performance is not affected by this biased distribution.

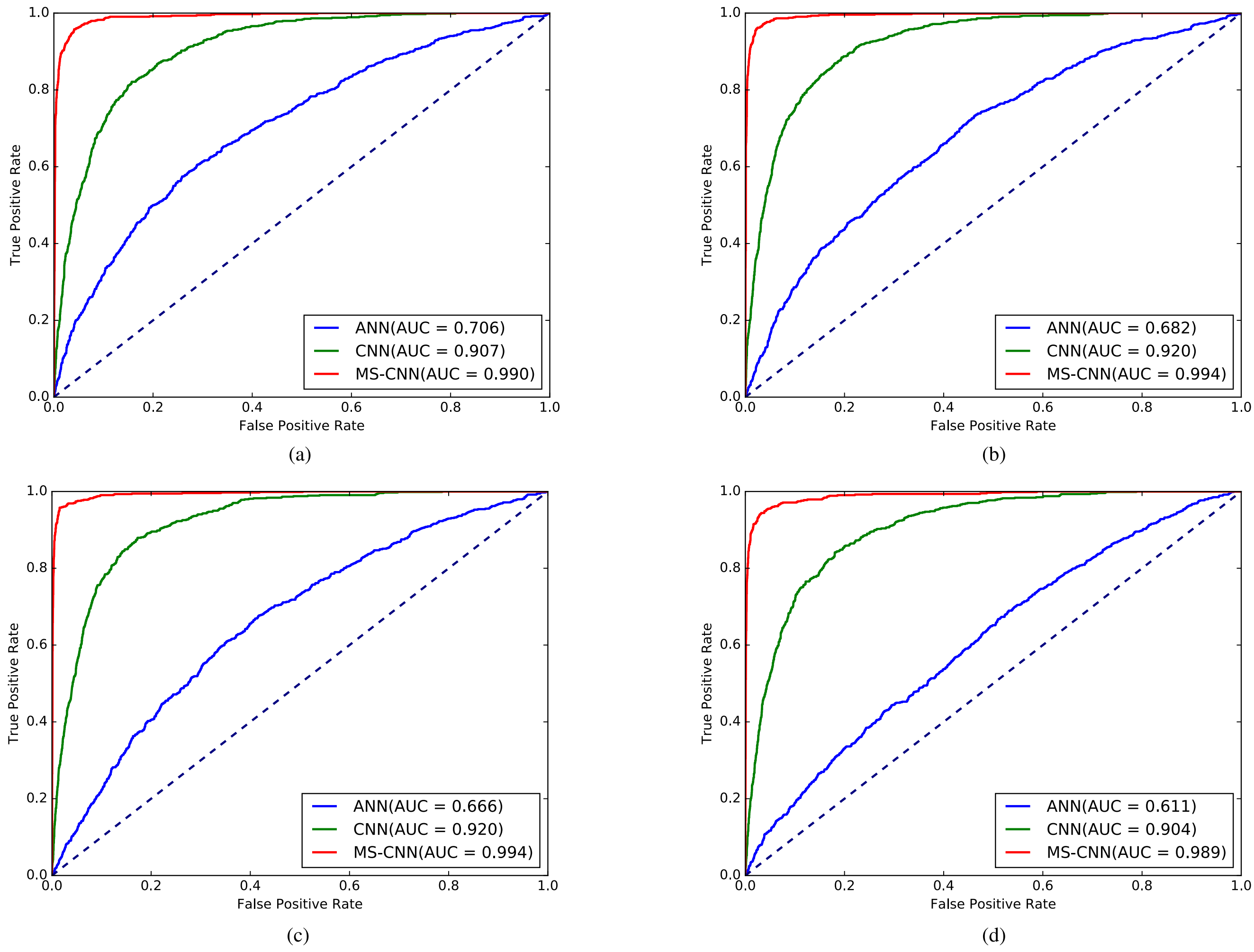
## E. Classification Performance Comparison

In this paper, three popular machine learning methods named artificial neural network (ANN) [30], convolutional neural network (CNN) [16], and VGGNet [10] are implemented to address AF screening problem for comparison. The implemented ANN is a very classical model, consisting of one input layer, one hidden layer, and one output layer. The sizes of input and output layers depends upon the problem, and they are set to be the length of input ECG recordings and the number of output classes correspondingly in this experiment. The implemented CNN consists of five layers, including one convolutional layer, one pooling layer, and one fully connected layer, except input and output layers. Compared with the proposed MS-CNN, this CNN model uses a very aggressive dimension reduction strategy in its pooling layer, and is proved to be successful in screening paroxysmal AF using long and clean ECG recordings [16]. It also has a comparatively larger filter kernel size, but not as deep as the proposed method. The implemented VGGNet, whose convolutional layers compose one stream of MS-CNN, consists of 13 convolutional layers with kernel size 3 and 3 fully connected layers. Meanwhile, the same data processing operations of the MS-CNN are done on the same training dataset and test dataset for models of the ANN, CNN, and VGGNet.

TABLE V
CLASSIFICATION PERFORMANCE COMPARISON ON AF VERSUS OTHER RHYTHM

| Model | $Sen$ | $Spe$ | $Pre$ | $Acc$ |
|---|---|---|---|---|
| VGGNet[10] | 67.11% | 96.01% | 63.75% | 93.67% |
| Behar[32] | 81.97% | 98.42% | 83.10% | 96.99% |
| Limam[17] | 72.70% | 98.60% | - | - |
| The proposed | 80.26% | 98.84% | 87.14% | 97.19% |

The experimental results show that there is an obvious difference of classification performance among the ANN, CNN, VGGNet, and MS-CNN. As shown in Table IV, the proposed MS-CNN outperforms the other three methods in all kinds of the fixed length of input ECG recordings. The ANN fails to recognize the underlying features inside ECG recordings, and tends to classify most input ECG recordings as healthy cases. It leads to low sensitivity and precision defined in Equation (10) and (13). The high specificity of the ANN (over 90%) results in 80% of its classification accuracy because of a high identification of health cases. Compared with the ANN, the classification metrics of sensitivity and specificity of the shallow CNN in detecting AF recordings are at least improved by 40% and 35%, respectively. The CNN has better classification performance than the ANN. However, the classification performance of the CNN in short ECG recordings is still not satisfying, demonstrating that a model with higher representing ability is needed to extract detailed features in ECG waveforms. When input ECG signal is short, it is vital to extract as many details as possible from the waveforms, while AF detection in longer recordings have abundant information concerning one's ECG rhythms. The proposed MS-CNN therefore outperforms CNN in all metrics, reaching nearly 30% increment in sensitivity and precision, and more than 8% reduction in overall error rate, compared to the CNN. Although classification performance on AF versus normal sinus rhythm of the VGGNet is not lower too much than the MS-CNN, the MS-CNN outperforms the VGGNet obviously in



(a) (b) (c) (d)

Fig. 7. Receiver operating characteristic curve (ROC): (a) 5 s, (b) 10 s (c) 20 s, (d) 30 s.

screening out AF from other rhythms showed in Table V. It is likely that the MS-CNN could capture more kinds of rhythm information than the VGGNet. At the same time, it is known that an AF detection method developed by AliveCor has FDA (Food and Drug Administration) clearance and CE (Conformity Europe) mark [31]. In order to detect AF episodes in real time, the length of the ECG signal analyzed by AliveCor is 30 seconds. This method, which requires a longer input ECG recording, identifies AF versus normal sinus rhythm with 85% sensitivity and 90% specificity. Compared with our proposed method with 94% sensitivity and 98% specificity in 5-second-long ECG recordings, the method developed by AliveCor performs worse in screening out AF recordings.

Receiver operating characteristic (ROC) analysis is related in a direct and natural way to compare the effectiveness of the aforementioned three methods (the ANN, CNN, and MS-CNN). As shown in Fig. 7, four ROC curves are drawn from the ANN, CNN, and MS-CNN in input ECG recordings of 5 seconds, 10 seconds, 20 seconds and 30 seconds. Compared with the ANN and CNN, our proposed MS-CNN obtains the best value of area under the curve (AUC) in all kinds of input ECG recordings. The AUC values of the MS-CNN for input ECG recordings are 0.990, 0.994, 0.994, and 0.989 respectively, which demonstrates that the MS-CNN has a higher classification performance in detecting short AF recordings.

Apart from validating the MS-CNN on screening AF versus normal sinus rhythm, we also evaluate the identification capability of the MS-CNN on AF versus other rhythm, which is another common clinical situation. As shown in Table V, the proposed MS-CNN obtains 80.26% sensitivity, 98.84% specificity, and 97.19% accuracy. Compared with the state of the art methods, the proposed method is more competitive in identifying AF from other rhythms.

## V. Conclusions

In this paper, we propose a multi-scaled fusion of deep convolutional neural network named MS-CNN for screening out AF recordings from single lead short ECG recordings. The proposed MS-CNN can capture different scaled feature information without hand-craft engineering by its specified network architecture, which consists of three fully connected layers and two streams of 13-layer convolutional neural networks with different filter sizes in first four hidden layers. By visualizing the representative layers, the learned features become linear separable as the depth of layers increases. The experimental results show that the proposed MS-CNN achieves 96.99% of classification accuracy on ECG recordings cropped/padded to 5 seconds. Especially, the best classification accuracy, 98.13%, is obtained on input ECG recordings of 20 seconds. Meanwhile, three popular machine learning methods (ANN, CNN, and VGGNet) are implemented to detect AF recordings on the same training and test dataset for comparison. Our proposed MS-CNN outperforms the ANN, CNN, and VGGNet in all classification metrics. What's more, the lowest AUC value of the MS-CNN is up to 0.989 which further demonstrates its high classification performance. However, the MS-CNN can still has the potential to improve the capability of screening AF from other rhythms, which we will focus on in our future research work.

## Acknowledgment

The authors would like to thank Dr. R. Wang and L. Yin who give their precious suggestions to improve the quality of this paper.

## References

[1] W. F. Andrawes, C. Bussy, and J. Belmin, "Prevention of cardiovascular events in elderly people," *Drugs Aging*, vol. 22, no. 10, pp. 859–876, 2005.

[2] A. J. Camm *et al.*, "Guidelines for the management of atrial fibrillation: The task force for the management of atrial fibrillation of the european society of cardiology (esc)," *Europace*, vol. 12, no. 10, pp. 1360–1420, 2010.

[3] J. Oster and G. D. Clifford, "Impact of the presence of noise on rr interval-based atrial fibrillation detection," *J. Electrocardiology*, vol. 48, no. 6, pp. 947–951, 2015.

[4] S. A. Guidera and J. S. Steinberg, "The signal-averaged p wave duration: A rapid and noninvasive marker of risk of atrial fibrillation," *J. Amer. College Cardiology*, vol. 21, no. 7, pp. 1645–1651, 1993.

[5] S. Mehta, N. Lingayat, and S. Sanghvi, "Detection and delineation of p and t waves in 12-lead electrocardiograms," *Expert Syst.*, vol. 26, no. 1, pp. 125–143, 2009.

[6] S. Dash, K. Chon, S. Lu, and E. Raeder, "Automatic real time detection of atrial fibrillation," *Ann. Biomed. Eng.*, vol. 37, no. 9, pp. 1701–1709, 2009.

[7] J. Lian, L. Wang, and D. Muessig, "A simple method to detect atrial fibrillation using rr intervals," *Amer. J. Cardiology*, vol. 107, no. 10, pp. 1494–1497, 2011.

[8] R. E. Kheirati and R. Sassi, "An extended bayesian framework for atrial and ventricular activity separation in atrial fibrillation," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 6, pp. 1573–1580, Nov. 2017.

[9] C. Huang, S. Ye, H. Chen, D. Li, F. He, and Y. Tu, "A novel method for detection of the transition between atrial fibrillation and sinus rhythm," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 4, pp. 1113–1119, Apr. 2011.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Comput. Sci.*, pp. 1409–1556, 2014.

[11] Z. Zhong, L. Jin, and Z. Xie, "High performance offline handwritten chinese character recognition using googlenet and directional feature maps," in *Proc. Int. Conf. Document Anal. Recognit.*, 2015, pp. 846–850.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[13] P. Ballester and R. M. Araujo, "On the performance of googlenet and alexnet applied to sketches," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 1124–1128.

[14] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Adv. Neural Inf. Process. Syst.*, vol. 1, no. 4, pp. 568–576, 2014.

[15] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 1933–1941.

[16] B. Pourbabaee, M. J. Roshtkhari, and K. Khorasani, "Deep convolution neural networks and learning ECG features for screening paroxysmal atrial fibrillatio patients[J]," *IEEE Trans. Syst. Man Cybern. Syst.*, no. 99, pp. 1–10, 2017.

[17] M. Limam and F. Precioso, "AF detection and ECG classification based on convolutional recurrent neural network," in *Proc. Comput. Cardiology Conf.*, 2017.

[18] Z. Yao, Z. Zhu, and Y. Chen, "Atrial fibrillation detection by multi-scale convolutional neural networks," in *Proc. 2017 20th Int. Conf. Inf. Fusion*, 2017, pp. 1–6.

[19] S. J., "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, 2014.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[22] J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, Q. V. Le, and A. Y. Ng, "On optimization methods for deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 265–272.

[23] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[24] T. M. Mitchell, J. G. Carbonell, and R. S. Michalski, *Machine Learning*. New York, NY, USA: McGraw-Hill, 2003.

[25] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2067–2075.

[26] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Netw.*, vol. 12, no. 1, pp. 145–151, 1999.

[27] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[28] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[29] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*.

[30] R. Couceiro, P. Carvalho, J. Henriques, M. Antunes, M. Harris, and J. Habetha, "Detection of atrial fibrillation using model-based ECG analysis," in *Proc. 19th Int. Conf. Pattern Recognit.*, 2008, pp. 1–5.

[31] AliveCor, "Alivecor heart monitor and aliveecg app (kardia mobile) for detecting atrial fibrillation," 2015. [Online]. Available: https://www.nice.org.uk/advice/mib35/chapter/technology-overview

[32] J. A. Behar, A. Rosenberg, Y. Yaniv, and J. Oster, "Rhythm and quality classification from short ECGs recorded using a mobile device," in *Proc. Comput. Cardiology Conf.*, 2017.