Computers in Biology and Medicine 116 (2020) 103378

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Accurate detection of atrial fibrillation from 12-lead ECG using deep neural network

Wenjuan Cai[a], Yundai Chen[b], Jun Guo[b], Baoshi Han[b], Yajun Shi[b], Lei Ji[b], Jinliang Wang[c], Guanglei Zhang[d,e,**], Jianwen Luo[a,*]

[a] Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing, 100084, China
[b] Chinese PLA General Hospital (301 Hospital), Beijing, 100084, China
[c] CardioCloud Medical Technology (Beijing) Co. Ltd, Beijing, 100084, China
[d] Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing, 100191, China
[e] School of Biological Science and Medical Engineering, Beihang University, Beijing, 100191, China

ARTICLE INFO

Keywords:
12-Lead ECG
Atrial fibrillation detection
Deep learning
Densely connected neural network

ABSTRACT

Atrial fibrillation (AF) is the most common heart arrhythmia, and 12-lead electrocardiogram (ECG) is regarded as the gold standard for AF diagnosis. Highly accurate diagnosis of AF based on 12-lead ECG is valuable and remains challenging. In this paper, we proposed a novel method with high accuracy for AF detection based on deep learning. The proposed method constructed a novel one-dimensional deep densely connected neural network (DDNN) to detect AF in ECG waveforms with a length of 10s. A large set of 16,557 12-lead ECG recordings collected from multiple hospitals and wearable ECG devices were used to evaluate the performance of the DDNN. In the test dataset (3312 12-lead ECG recordings), the DDNN obtained high performance with an accuracy of 99.35 ± 0.26%, a sensitivity of 99.19 ± 0.31%, and a specificity of 99.44 ± 0.17%. Its high performance and automatic nature both demonstrate that the proposed network has a great potential to be applied to clinical computer-aided diagnosis of AF or future screening of AF in wearable devices.

1. Introduction

Atrial fibrillation (AF) is the most common heart arrhythmia condition, affecting around 2.3 million patients in the US and 4.5 million in the European Union [1]. Approximately a third of all strokes is associated with AF [2]. However, the AF in over one-third patients is asymptomatic and undiagnosed, which greatly increases the risk of stroke even death [3]. Early and accurate detection of AF and its treatment (e.g. anticoagulation) could significantly help in prevention of the complications associated with the stroke event [4].

The common approaches for AF diagnosis include pulse palpation, photoplethysmography, oscillometric blood pressure, and electrocardiogram (ECG). Only standard 12-lead ECG recording is commonly used as gold standard to confirm the diagnosis among these methods [5]. Other methods are neither adequately sensitive nor specific for accurate AF diagnosis [5]. In AF, the electrical impulses of the atria are disorganized and unsynchronized with the ventricles. Fig. 1 shows examples of electrical impulses from a healthy heart and a heart with AF. It can be found that the electrical impulses from the heart with AF are chaotic compared with that from the healthy heart. A normal ECG waveform and an ECG waveform with AF extracted from the dataset used in our experiments are also shown in Fig. 1. A normal waveform is composed of the P wave, QRS complex, and T wave. However, in an AF waveform, the P wave is replaced by many inconsistent fibrillatory waves (F waves) and R–R irregularities (irregularities of intervals between successive R waves on the ECG) appear [6]. Many automatic algorithms based on the detection of absence of P-waves [7] or R–R irregularities [8,9] have been developed in place of the time-consuming and subjective manual inspection of long-term of ECG recording. However, P-waves are prone to contamination with noise artifacts, which makes accurate AF detection based on these morphological characteristics difficult. Besides, many methods based on feature-extraction have also been proposed. These methods start with feature extraction step using feature extractors such as independent component analysis [10], discrete wavelet transform [11], higher-order spectra [10], and entropy [11]. Usually, a large number of features are generated during the feature extraction process. Therefore, feature selection is required to remove irrelevant and redundant features. A few

* Corresponding author.
** Corresponding author. Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing, 100191, China.
E-mail addresses: guangleizhang@buaa.edu.cn (G. Zhang), luo_jianwen@tsinghua.edu.cn (J. Luo).

https://doi.org/10.1016/j.compbiomed.2019.103378
Received 21 May 2019; Received in revised form 1 August 2019; Accepted 1 August 2019

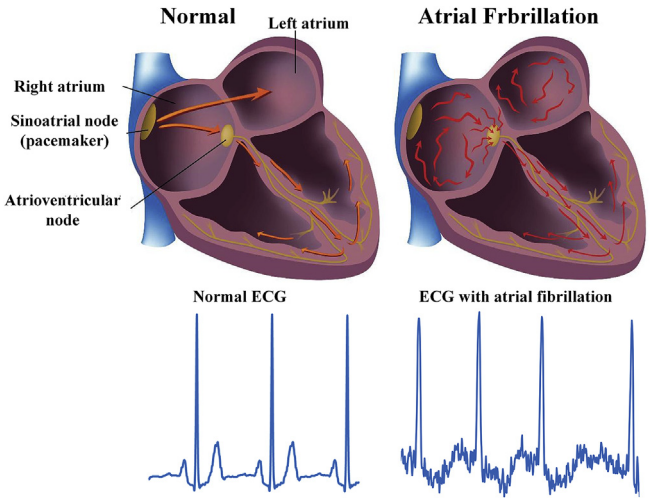

Fig. 1. Representations of a healthy heart and a heart with atrial fibrillation, a normal ECG waveform and an ECG waveform with atrial fibrillation (image adapted from Centers for Disease Control and Prevention) [22].

techniques are used including analysis of variance [12] random forest [13], and evolutionary algorithms [14]. After the selection, classifiers including support vector machine [6], k-nearest neighbor [10], and random forest [13] are employed to make classification. These methods based on feature extraction require multiple steps including feature extraction, feature selection and classification. And compared to these traditional methods with multiple steps, the end-to-end diagnostic methods make the AF diagnosis more direct and simpler. Moreover, multiple steps mean that many errors have the chance to be superimposed on the diagnosis results.

With the rapid development of deep learning, deep neural networks have become popular to solve the classification, segmentation, and detection problems. The convolutional neural network (CNN) [15–18], recurrent neural network (RNN) [19], and autoencoder [20] are a few deep learning algorithms that have been implemented for AF diagnosis. With deep learning, R-peak detection and removal of noises and artifacts from the ECG signals could be minimal, and the extraction of features and the selection of hand-crafted features are not necessary. Although previous works on AF detection have achieved good performance, there also exist a few problems: 1) Some methods making classification based on ECG beats require firstly to split ECG recordings to beats [12,16,21], which is troublesome to implement in real world. 2) Many methods are only trained and tested on dataset from single source like MIT/BIH arrhythmia database or a small dataset [19,21]. A dataset from single source may not be enough for comprehensive assessment of the performance of an algorithm. 3) A high sensitivity and a high specificity are both required to achieve accurate diagnosis or screening of AF. However, some methods can only satisfy one of these two requirements [15,18,20].

In this work, we designed and trained a novel one-dimensional deep densely connected neural network (DDNN) to distinguish among AF, normal and other abnormalities (labeled as other) with a large set of 12-lead ECG signals (length = 10 s) from different sources. In contrast to handcrafted features, the DDNN automatically learn the most predictive features directly from the raw 12-lead ECG waveform based on the training examples. Additionally, we employed several visualization methods to gain insight on what the network had learnt.

2. Materials and methods

2.1. Dataset and preprocessing

In this study, we merged 12-lead ECG datasets from 3 different sources: data source 1, the Chinese PLA General Hospital (301 Hospital); data source 2, wearable ECG devices (Mason-Likar ECG 12-lead system, this device includes 4 limb leads and chest leads integrated into a chest band) (CardioCloud Medical Technology (Beijing) Co. Ltd); and data source 3, other 11 hospitals (The China Physiological Signal Challenge 2018) (see Fig. 2(a)). All ECG data from 3 different sources are standard 12-lead ECG recordings with high quality and same sampling frequency. A total of 16,557 samples of 12-lead ECG recordings from 11,994 unique subjects were adopted and assigned to three classes: AF (n = 3353), normal (n = 5650) and other abnormalities (n = 7554) (see Table 1). Other abnormalities include premature atrial contraction, premature ventricular contraction, ST-segment depression, flat T waves, T wave inversions and so on. Mean age of subjects in the dataset was 57.0 ± 18.7 years and 47.4% were men. All ECG recordings were sampled at 500 Hz and divided into segments of 10 s. A committee of three experienced cardiologists served as gold-standard annotators for the ECG recordings including data source 1 and 2. Each ECG recording was labeled by both two cardiologists, and then the label results were checked by the third one. When the two labels from the cardiologists did not agree with each other, the third one will relabel this ECG recording. According to these three label results, three cardiologists will discuss on the disagreement and then give final result. The Kappa score for the labeling results between first two cardiologists is 0.9985. All ECG recordings (n = 16,557) were randomly split into training subset (n = 10,663, 64%), validation subset (n = 2,582, 16%) and test subset (n = 3,312, 20%) for 5-fold cross-validation (shown in Fig. 2(b)). The five results of the 5-fold cross-validation on the same test dataset were averaged. The ECG recordings from the same subject can only be grouped into the same subset (training/validation/test subset) and there was no overlap among training, validation and test subjects, which can effectively reduce model overfitting.

All 12-lead ECG recordings were filtered by using a band-pass filter

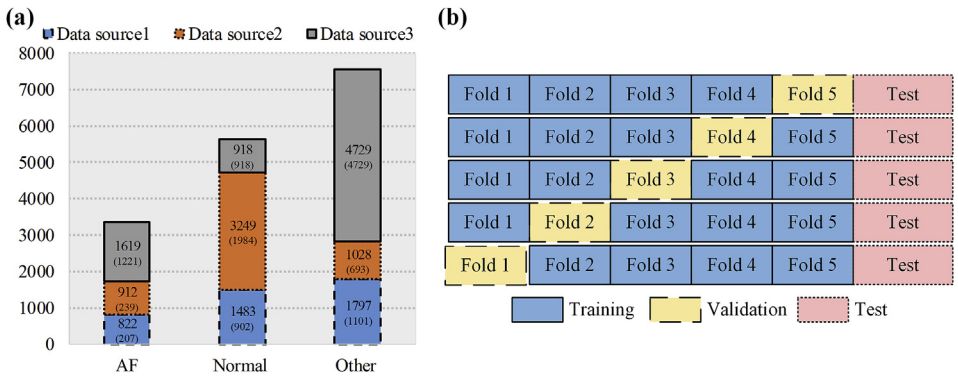

Fig. 2. Data sources and workflow diagram. (a) Data sources for AF, normal, and other classes, the numbers in each block denote the number of samples and subjects in each label. (b) 5-fold cross-validation, blue part as training subset; yellow part as validation subset; red part as test subset. Five results on test dataset are averaged.

Table 1
Summary of the subjects and the ECG recordings.

Characteristics	n
Number of ECG recordings	16,557
Patient demographics	
Number of unique subjects	11,994
Age, mean ± SD	57.0 ± 18.7
Male, n_1/N_1 (%)	5620/11,856 (47.4%)
ECG rhythm diagnosis, n_2/N_2 (%)	
Normal	5650/16,557 (34.1%)
Atrial fibrillation	3353/16,557 (20.3%)
Other	7554/16,557 (45.6%)

n_1, number of males in total subject; N_1, total subject (loss of gender information in some subjects); n_2, number of subsamples; N_2, total sample.

(0.5–35 Hz, an elliptical band-pass filter with filter order of 10) to remove baseline wander and high frequency components. The ECG signal in each lead is normalized with Z-score normalization to address the problem of amplitude scaling and eliminate the offset effect.

$$z = (x - \mu) / \sigma$$

where μ denotes the mean value of the data x and σ denotes its standard deviation.

2.2. Neural network architecture and training

Our neural network takes as input a 12-lead ECG waveform of 10 s long (12 × 10 s, the 12 channels are fed in parallel) and outputs a label prediction of two (AF and normal, AF and non-AF) or three (AF, normal, and other) classes. 12-lead ECG predictions of normal and other were considered as a non-AF label. We designed a densely connected DDNN architecture with four dense blocks (a total of 36 layers) and a growth rate of 6 (see Fig. 3(a)). Residual neural networks [23,24] always have good performance on the tasks based on image data. Compared to residual neural networks, densely connected neural networks [25] can alleviate the vanishing-gradient problem, encourage feature reuse, and substantially reduce the number of parameters to be learnt, which can decrease the probability of overfitting. Inspired by these studies, we have firstly introduced the densely connected network into the field of signal processing and carefully designed a one-dimensional densely connected neural network used for AF detection in this work. The diagram of the dense blocks is shown in Fig. 3(b). The block sizes (the numbers of convolutional layers) of the four dense blocks are 2, 4, 6, and 4, respectively. Inspired by inception module in the GoogLeNet, Szegedy et al. [26] used a series of filters of different sizes (1 × 1, 3 × 3, 5 × 5, filter concatenation module) to handle multiple scales. We use a similar strategy here, the network starts with the filter concatenation module (shown in Fig. 3(a)), which uses convolutions of different sizes in parallel to capture details at varied scales (1 × 8, 1 × 16, 1 × 24), and the other convolutional layers all have a filter length of 16. To improve the computational efficiency and model compactness, we used bottleneck (1 × 1 convolution) and transition layers (convolution and pooling, placed between two adjacent dense blocks) in the DDNN. Before the fully connected layer we applied global average pooling (GAP) [27]. The squeeze-and-excitation (SE) module [28] was employed to improve the representational power of a network by enabling it to perform dynamic channel-wise feature recalibration. The SE module includes squeeze operation (F_{sq}) implemented by GAP, excitation operation (F_{ex}) realized by two fully connected layers, and rescale operation (F_{scale}) (see Fig. 3(c)). The specific implementation is shown in Fig. 3(c) and Eqs. (1)–(3). In the DDNN, the SE module was placed before each dense block and each transition layer.

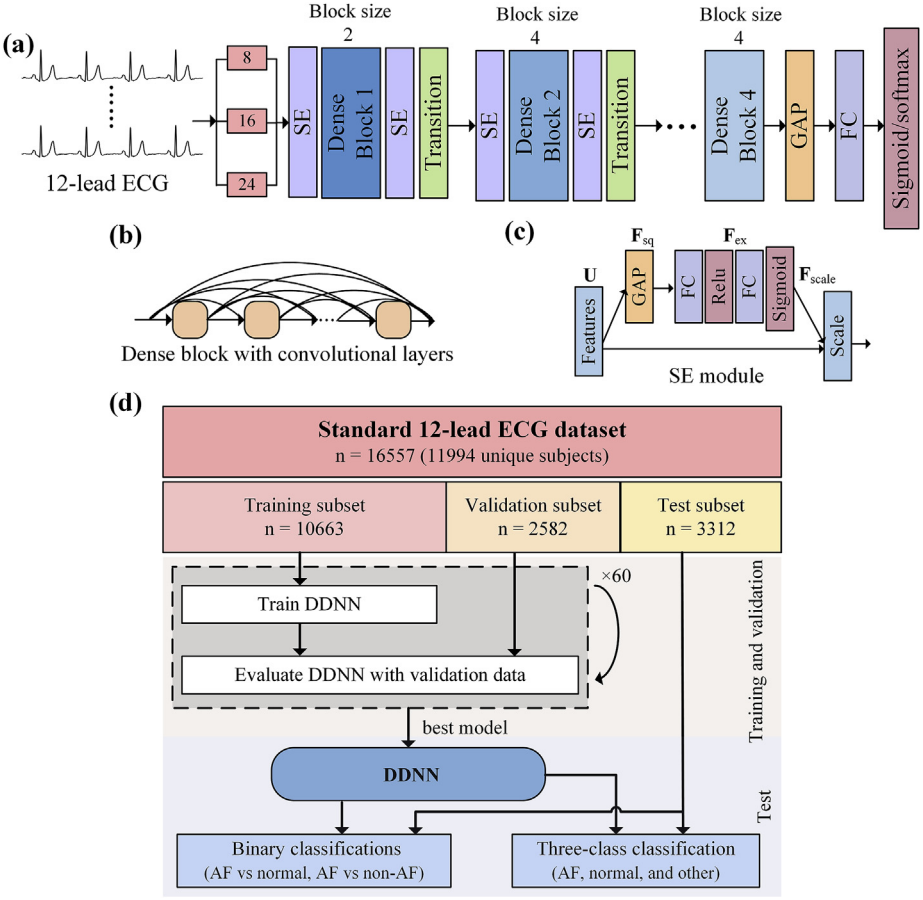

Fig. 3. Architecture of the DDNN. (a) The DDNN has 4 blocks, each of which consists of multiple densely connected convolutional layers. (b) Diagram of the SE module. (c) Diagram of the dense block with multiple convolutional layers in the DDNN. SE, squeeze-and-excitation; FC, fully connected layer; GAP, global average pooling; U, feature map; F_{sq}, squeeze operation; F_{ex}, excitation operation; F_{scale}, rescale operation. (d) Workflow diagram showing the subsets used to train, validate and test the DDNN.

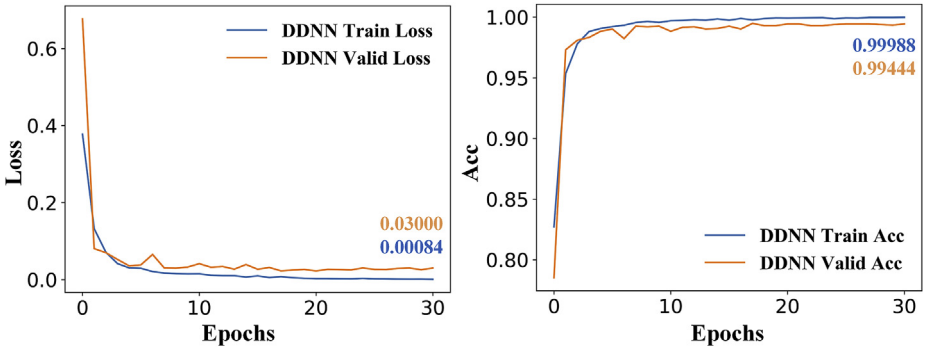

Fig. 4. Loss and accuracy of the DDNN (AF and normal) on training subset and validation subset. Loss and accuracy on training and validation subset at 30th epoch are marked, Acc, accuracy.

$$z_c = \mathrm{F}_{sq}(\mathrm{u}_c) = \frac{1}{W}\sum_{i=1}^{W} u_c(i) \tag{1}$$

where z_c is the c-th element of $\mathbf{z}$, u_c is the c-th element of $\mathbf{U}$, $z \in \mathbb{R}^K$ is generated by shrinking feature map $\mathbf{U}$ ($U \in \mathbb{R}^{K\times W}$) through length W, K is the number of filters, F_{sq} is implemented by GAP.

$$s = \mathrm{F}_{ex}(z, W) = \sigma_2(g(z, W)) = \sigma_2(W_2\sigma_1(W_1 z)) \tag{2}$$

where $\mathbf{s}$ is scaling weight, g refers to two fully connected layer operations, σ_1 refers to the ReLU function [29], σ_2 refers to the sigmoid function, fully connected layer parameters $W_1 \in \mathbb{R}^{\frac{K}{r}\times K}$ and $W_2 \in \mathbb{R}^{K\times\frac{K}{r}}$, r is the reduction ratio, F_{ex} is realized by two fully connected layers.

$$\tilde{x}_c = \mathrm{F}_{scale}(\mathrm{u}_c, s_c) = s_c \cdot \mathrm{u}_c \tag{3}$$

where $\tilde{x}_c$ is the c-th element of $\tilde{X}$, $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_K]$, s_c is the c-th element of $\mathbf{s}$, $\mathrm{F}_{scale}(\mathrm{u}_c, s_c)$ refers to channel-wise multiplication between the feature map $\mathrm{u}_c \in \mathbb{R}^W$ and the scalar s_c.

We used the Keras (Tensorflow as backend) to implement our model and adopted Tensorboard to monitor the training process. The training, validation and test were implemented on a NVIDIA TITAN XP GPU with 12G memory, Intel Xeon E5-2620v4 and python 3.6. The workflow for training, validating and testing the DDNN is shown in Fig. 3(d). We trained our model from scratch on the 12-lead ECG training subset by adopting weight initialization [30] and using Adam optimizer [31] with the default parameters for a total of 60 epochs. We used an exponential learning rate schedule and it decayed every 4 epochs using an exponential rate of 0.8. The selection of hyper-parameters including the size of the bottleneck, the filter length and the growth rate was conducted by Bayesian optimization.

The model with best performance on the 12-lead ECG validation subset was saved and used for the subsequent test. The 12-lead ECG test subset was used to characterize the accuracy of the DDNN for AF classification. Three experiments including two binary classifications (AF and normal, AF and non-AF) and a three-class classification (AF, normal, and other classes) were conducted in our study. These three classifiers were trained and tested independently. The accuracy, sensitivity, specificity and F_1 score (computed by Eqs. (4)–(7)) of DDNN for AF detection were evaluated in our experiments. Receiver operating characteristic (ROC) curves were generated by varying the operating threshold for DDNN. The performance of DDNN on binary classifications and three-class classification was compared using accuracy, sensitivity, specificity, F_1 score and area under the ROC curve (AUC). The means ± standard deviations (means ± SDs) of these metrics in 5-fold cross-validation were calculated. The confusion matrixes of the DDNN were plotted to further show classification results. To visualize what the model has learnt, we computed saliency map [32] which is used to highlight input regions that cause the most change in the output, and visualized the last hidden layer representations using principal component analysis (PCA) [33] and t-distributed stochastic neighbor embedding (t-SNE) [34].where TP denotes the number of cases correctly identified as AF, FP denotes the number of cases incorrectly identified as AF, TN denotes the number of cases correctly identified as normal/non-AF and FN denotes the number of cases incorrectly identified as normal/non-AF. True positive rate = sensitivity, false positive rate = 1-specificity.

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

$$sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$specificity = \frac{TN}{FP + TN} \tag{6}$$

$$F_1\ score = \frac{2{*}TP}{2{*}TP + FP + FN} \tag{7}$$

3. Results

3.1. Experimental results

In order to showcase the details of training process, we took the example of the binary classification of AF and normal to show the properties of the proposed model. The DDNN model consists of a total of 69,087 trainable parameters and only requires 0.6 MB of storage space. The training time of the DDNN on the binary classification of AF and normal is 19.7 min, and the test time is 5.9 s (2160 recordings of AF and normal, 2.7 ms/recording). Therefore, the well-trained DDNN can make fast detection for a new recording. The accuracy and loss of the DDNN on training and validation set are shown in Fig. 4. It can be observed that with the increase of epoch number, both training and validation losses of the DDNN converge to a very low value and their corresponding accuracies converge close to 1. Specifically, the differences of accuracy and loss of the DDNN between training and validation are both very small. This proves that the DDNN has strong generalization capability.

We conducted the experiments to compare the performance of the DDNN at different settings with the performance of the original 121-layer densely connected neural network (DenseNet-121) [25] (see Table 2). The performance of our DDNN greatly outperforms DenseNet-121. Compared to the original DenseNet-121, we changed a two-dimensional network into a one-dimensional network, decreased the number of network layers, and selected the appropriate size for the dense blocks to construct the DDNN. Besides, subject's age information, the filter concatenation module, the SE modules were added into the DDNN to try to further improve the performance of AF detection. From Table 2, it is observed that adding subject's age in the last layer (fully connected layer) of the DDNN did not help improve the performance of AF detection. In contrast, the filter concatenation module and the SE module contributed to better performance of the DDNN (by almost

Table 2
DDNN performance for detection of AF (AF and normal) on test subset at different settings.

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)
DenseNet-121	88.25	90.58	86.10
DDNN	96.73	96.67	96.93
DDNN-Age	96.49	96.71	96.91
DDNN-FC	97.44	97.54	97.60
DDNN-FC-SE	**99.35**	**99.19**	**99.44**

FC, filter concatenation module; SE, squeeze-and-excitation module. The highest score is indicated in bold.

0.7% and 2% on accuracy, respectively). These results indicate that appropriate network structures including the filter concatenation module and the SE modules can help obtain a better performance of AF detection.

3.2. Binary and three-class classifications

The performance of the DDNN for classifying 12-lead ECG recordings on binary and three-class classifications is presented in Table 3. On the binary classification of AF and normal, the accuracy, sensitivity, specificity of the DDNN were 99.35 ± 0.06%, 99.19 ± 0.24%, 99.44 ± 0.06%, respectively. On binary classification of AF and non-AF, the accuracy, sensitivity and specificity of the DDNN were 98.21 ± 0.09%, 97.04 ± 0.88%, 98.63 ± 0.28%, respectively. On the three-class classification, AF achieved relatively high metrics (accuracy: 97.74 ± 0.30%, sensitivity: 98.85 ± 0.89%, specificity: 98.38 ± 0.14%) and other abnormalities had relatively low metrics (accuracy: 91.12 ± 0.18%, sensitivity: 85.47 ± 2.13%, specificity%: 94.12 ± 0.70). The DDNN had the best performance on the binary classification of AF and normal among these three classifications. The performances of different methods (convolutional neural network, long short-term memory, autoencoder) based on deep learning for AF detection are presented in Table 4. The DDNN outperformed all other methods across all metrics. This indicates that the DDNN can achieve high performance for AF detection without data processing (like stationary wavelet transform, ECG beat detection) and feature extraction.

The AUCs of the DDNN on binary and three-class classifications were 0.999 (for binary classification of AF and normal), 0.994 (for binary classification of AF and non-AF) (see Fig. 5(a)), and 0.979, 0.994 and 0.954 (for three-class classification of normal, AF and other) (see Fig. 5(b)), respectively. The effect of recording length on the performance of the DDNN on binary classification of AF and normal was evaluated by testing recording lengths of 0.25–12 s (see Fig. 5(c)). The metrics of accuracy, sensitivity, and specificity of the DDNN increased slightly as the recording length increased from 2 s to 8 s, and kept steady when the recording length increased from 8 s to 12 s. While the performance of the DDNN deteriorated rapidly for recording lengths shorter than 2 s. In order to ensure the stability of the test results, we chose input length of 10 s that contain richer information for AF detection compared to 8 s.

The confusion matrixes of the DDNN on binary and three-class classifications are shown in Fig. 6(a). On the binary classifications, the numbers of false positive and false negative were small. Especially on binary classification of AF and normal, only eight false positives and six false negatives were made by the DDNN. On the three-class classification, most of the mistakes were concentrated in misclassifications between the normal and other classes. Examples of the ECG waveforms incorrectly classified (AF misclassified into non-AF and non-AF misclassified into AF) by the DDNN on binary classification of AF and normal are shown in Fig. 6(b) and (c). It can be observed that these misclassifications may be caused by the excessive noise or non-obvious F waves in the ECG waveforms.

3.3. Visualizing the network

To gain insight on what the DDNN had learnt, we visualized the first-layer weights that represent the learnt convolutional filters (see Fig. 7(a)). The learnt filters appear to be suitable for detecting low-level features such as peaks, troughs, and upward and downward slopes. We also computed the saliency map which is used to highlight input regions that cause the most change in the output (see Fig. 7(b)). The distribution of the saliency map is mainly concentrated in the regions between T wave and next QRS complex, which seem to be exactly where the F waves occur.

The internal features automatically learnt by the DDNN were visualized using PCA and t-SNE (see Fig. 8). We plotted three main components of the multi-dimensional features in the DDNN's last hidden layer. In the binary classification, points belonging to AF and non-AF classes were clearly distinguished in the three-dimensional space of PCA. The distribution of points belonging to AF class was more concentrated than that of points belonging to non-AF class. In the three-class classification, points belonging to the normal, AF and other classes were distributed in different regions in three-dimensional space of PCA. Each point in figure of t-SNE represents a 12-lead ECG recording projected from the output of the DDNN's last hidden layer into two dimensions. Points belonging to the same class were clustered together. Some points from normal and other classes were overlapped, which is consistent with the results of confusion matrix.

4. Discussion

In this paper, we designed and trained a novel neural network DDNN to explore the utilization of deep neural networks for automatic AF detection from 12-lead ECG recordings. The results demonstrate that through the innovative design of network structures and the unique utilization of the filter concatenation module and the SE module, the proposed DDNN achieved accurate detection of AF from short ECG waveforms without need to extract and select features. The DDNN achieved promising performance with an accuracy of 99.35 ± 0.26%, a sensitivity of 99.19 ± 0.31%, and a specificity of 99.44 ± 0.17% for AF/normal classification. The ability of the DDNN to outperform the methods based on explicit features highlights the value of information captured from raw data. The deep learning-based approaches can reduce efforts of feature engineering and may allow learning novel and hidden features that might not be discovered using traditional machine learning methods [41]. Additionally, the selection of appropriate network structures including dense blocks and SE modules may contribute

Table 3
Performance of the DDNN for binary and three-class classifications on test subset (means ± SDs).

Classification		Accuracy (%)	Sensitivity (%)	Specificity (%)	F_1 score (%)
Binary	AF and Normal	99.35 ± 0.06	99.44 ± 0.06	99.19 ± 0.24	99.06 ± 00.09
	AF and Non-AF	98.21 ± 0.09	98.63 ± 0.28	97.04 ± 0.88	96.45 ± 00.12
Three-class	Normal	92.96 ± 0.28	93.18 ± 0.97	92.56 ± 1.05	90.86 ± 00.34
	AF	97.74 ± 0.30	98.38 ± 0.14	95.85 ± 0.89	95.36 ± 00.41
	Other	91.12 ± 0.18	94.12 ± 0.70	85.47 ± 2.13	85.90 ± 00.69

Non-AF, normal and other, the highest score is indicated in bold.

Table 4
Performance of the DDNN for detection of AF (AF and normal) versus several state-of-the-art methods.

Authors	Method	Data type	Accuracy (%)	Sensitivity (%)	Specificity (%)	Ave.F_1 score (%)
Oh et al. [35]	CNN + LSTM	ECG beats	98.10	97.50	98.70	–
Erdenebayar et al. [36]	CNN	ECG beats	98.70	98.60	98.70	–
Acharya et al. [18]	CNN	5s	94.90	99.13	81.44	–
Yao et al. [16]	Multi-scale CNN	ECG beats	98.11	98.22	98.18	–
Xia et al. [15]	SWT + CNN	5s	98.63	98.79	97.87	–
Yuan et al. [20]	Autoencoder	10s	98.31	95.56	98.04	–
Xiong et al. [37]	CNN	ECG beats	–	–	–	81.8
Zihlmann et al. [38]	FE + LSTM	ECG beats	–	–	–	82.1
Warrick et al. [39]	Ensembled CNN + LSTM	ECG beats	–	–	–	84.5
Parvaneh et al. [40]	DenseNet	ECG beats	–	–	–	82.0
Our method	DDNN	10s	**99.35**	**99.19**	**99.44**	**90.7**

CNN, convolutional neural network; LSTM, long short-term memory; SWT, stationary wavelet transform; FE, Feature Extractor; SVM, support vector machine; DenseNet, Densely connected convolutional network; Ave.F1 score, average F1 score across normal, AF and other classes. The highest score is indicated in bold.

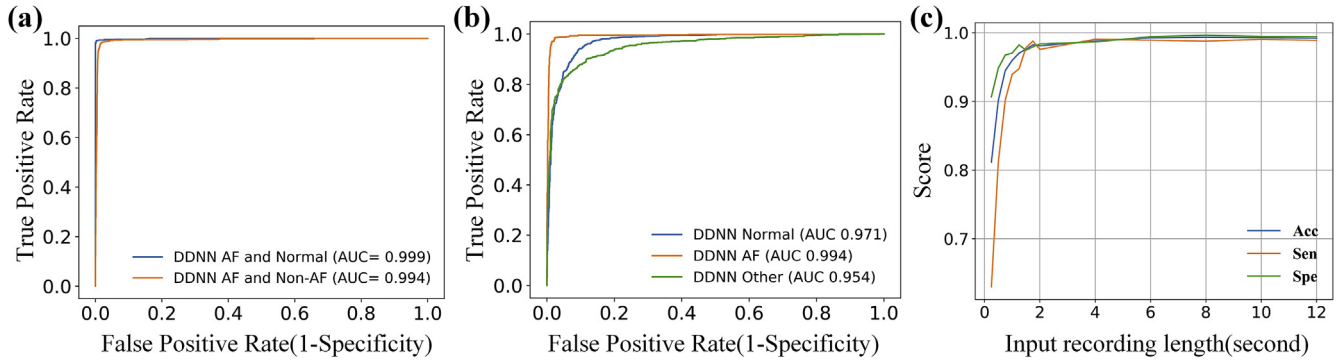

Fig. 5. Results of the DDNN on binary and three-class classifications. ROC and corresponding AUC of the DDNN for (a) the two binary classifications (AF and normal, AF and non-AF), (b) the three-class classification. (c) Effect of input recording length on the performance of the DDNN for the binary classification of AF and normal. AUC, area under the curves; Acc, accuracy; Sen, sensitivity; Spe, specificity.

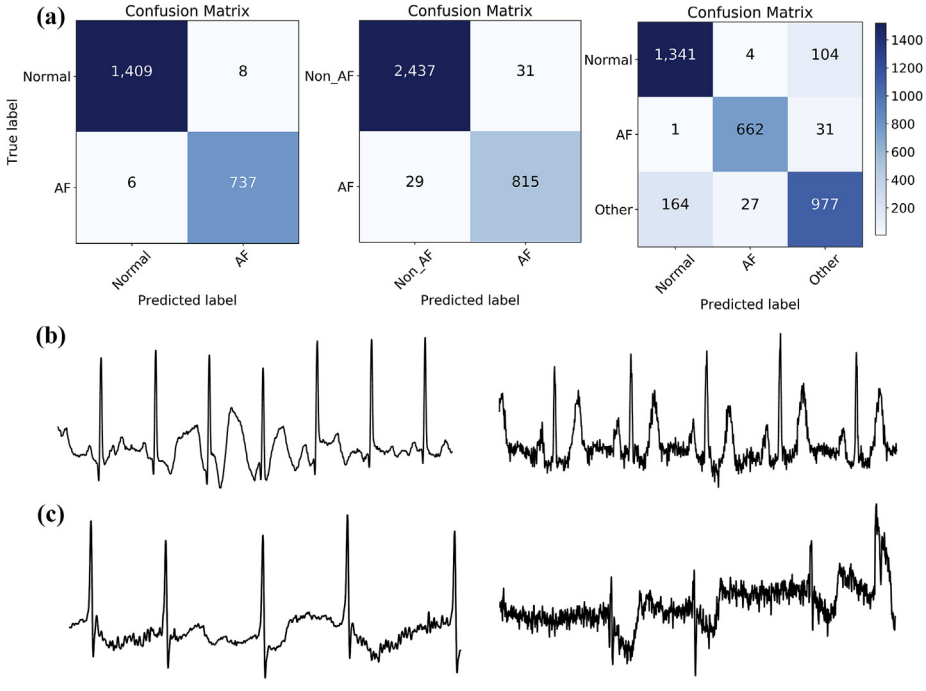

Fig. 6. Confusion matrixes and misclassified examples for the DDNN on test subset. (a) The confusion matrixes for the DDNN on the binary classification of AF and normal, the binary classification of AF and non-AF, and the three-class classification of normal, AF and other classes. Examples of ECG waveforms (lead II) incorrectly classified by the DDNN: examples of (b) two examples of non-AF misclassified as AF, and (c) two examples of AF misclassified as non-AF.

to better performance compared with other deep neural networks.

To have a better understanding about what the DDNN had learnt, we visualized the saliency map and the visualization maps using PCA and t-SNE. From the results of saliency map, it can be observed that the DDNN can learn to use F waves-related features to make classification, which indicates that the DDNN is able to capture features with underlying physiological meaning related to AF. As for other features such as R–R irregularities used to detect AF, we did not find appropriate visualization method to display whether the network has learnt them. The visualization maps using PCA and t-SNE demonstrated that the

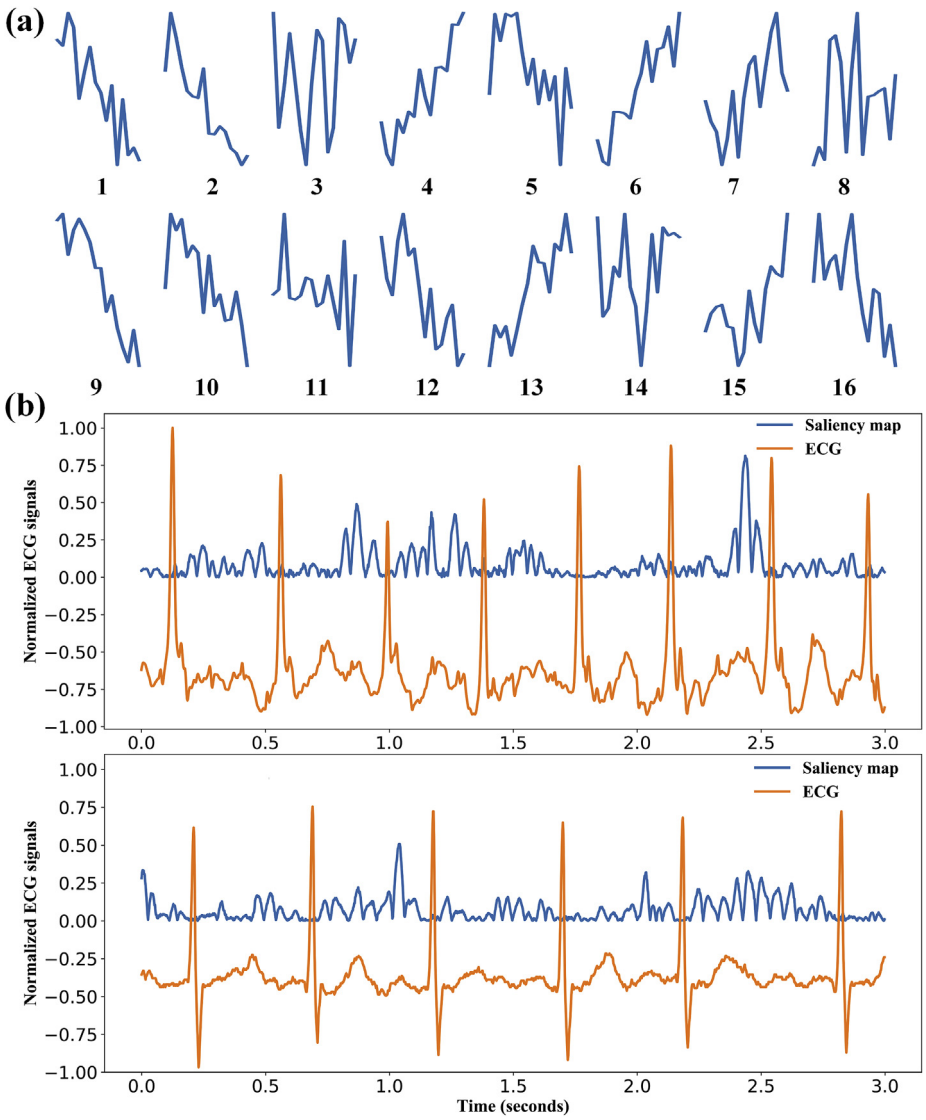

Fig. 7. Visualization of what the DDNN learnt on the binary classification of AF and normal. (a) First-layer 16 learnt filters of the DDNN. (b) Two examples of saliency map of the DDNN (lead II ECG was chosen to display).

DDNN did learn some valuable features used to make classification of AF.

The 12-lead ECG dataset used in our study was constructed with data from different sources. The classification results of the proposed network demonstrated that it had good performance on multi-center data. The generalization capability of the DDNN makes it promising for AF detection with different devices in different centers. Moreover, the network input only requires an ECG recording as short as 10 s, which makes the acquisition of a low-noise and high-quality ECG waveform easier and makes ECG measurements more convenient.

The American College of Cardiology/American Heart Association stated that early detection and treatment of asymptomatic AF before the first complications occur is a recognized priority for the prevention of stroke [42]. A high sensitivity is necessary for an AF screening tool. Meanwhile, a high specificity is preferred for mass screening to avoid generating many false negatives that can trigger unnecessary anxiety in patients and extra costs of follow-up examinations. Some guidelines recommend AF screening using pulse palpation for patients ≥65 years of age [43]. Although pulse palpation is a readily accessible method for screening in primary care, it is not common practice and has a lower sensitivity of 87.2% and specificity of 81.3% [44]. The capability of the DDNN in achieving both high sensitivity and specificity makes it promising for precise screening of AF in basic medical institutions.

The dataset used in our study is composed of the standard 12-lead ECG recordings. As single-lead ECG may not always show F waves, the AF diagnosis based on only single-lead ECG may result in false negatives. The standard 12-lead ECG is commonly taken as gold standard for diagnosis of AF. Recently, some wearable or hand-held 12-lead ECG devices are particularly designed to make daily self-measurements for patients at home. After patients' ECG recordings are sent to the related management application in the smartphone, using the proposed DDNN to screen AF from these recordings will be an attractive replacement for the ECG measurements in hospitals given its ease of use, timeliness and superior accuracy. We anticipate that the DDNN can be built into various application terminals to achieve automatic and accurate diagnosis of AF.

There are also several aspects where our proposed network can be further improved. First, to our knowledge, some characteristics of subjects like age should be useful for AF detection. In our study, however, adding this characteristic in our network was not helpful for improving the performance of AF detection. This could be due to the age information loss in some subjects (about 14.6%). Second, the proposed network could make accurate AF classification on most of the data with noise. However, there exist a few ECG recordings with excessive noise

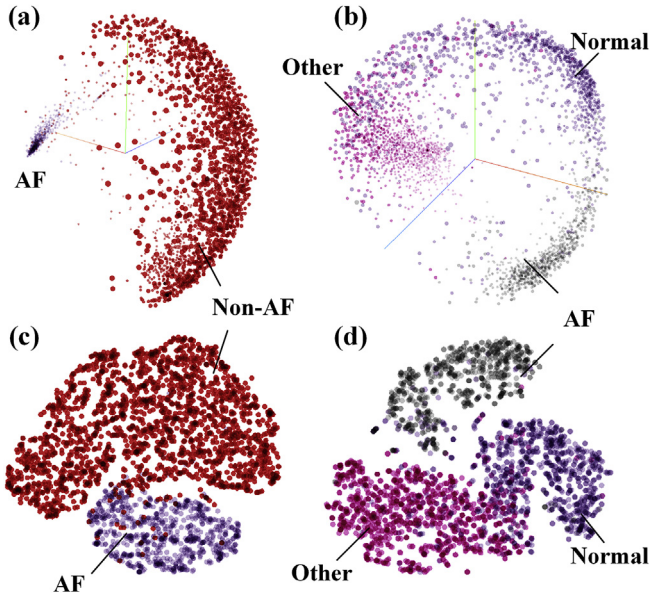

Fig. 8. PCA and t-SNE visualization. PCA visualization in three-dimension space (three principal components) for (a) the binary classification of AF and non-AF, (b) the three-class classification of AF, normal and other classes. T-SNE visualization in two-dimension map for (c) the binary classification of AF and non-AF, (d) the three-class classification of AF, normal and other classes. The colored clusters represent the different classes: AF (purple), non-AF (red) for binary classification; normal (purple), AF (grey) and other (red) for three-class classification. PCA, principal component analysis; t-SNE, t-distributed stochastic neighbor embedding.

or disturbance caused by casual improper operation during measurement, which can lead to misclassifications of network. Multiple measurements of the ECG recording may improve the performance of AF diagnosis. Third, the metrics for other class of three-class classification are lower, since other class includes many other abnormalities which require more data to train the model. Our proposed network is currently only used to detect AF due to lack of sufficient data of other categories. However, 12-lead ECG can be used for diagnosis of hundreds of cardiac diseases that are difficult or even impossible to detect on a single lead ECG. With the collection of more abnormal ECG recordings in the future, we hope to achieve multi-class classification for diagnosing more cardiac diseases with our proposed network.

5. Conclusions

We have proposed a novel DDNN to detect AF from raw 12-lead ECG recordings. It has been demonstrated that the DDNN can achieve promising performance. By applying the DDNN to a large dataset from different sources, we have shown its strong generalization capability on multi-center data. This work suggests that this new network is very suitable for automatic and accurate diagnosis of AF using 12-lead ECG, which may have important applications for future precise screening of AF in a real-world primary care.

Conflicts of interest

There is no conflict of interest in this work.

Acknowledgments

This work was supported in part by National Key R&D Program of China (No. 2018YFC2001200), in part by National Natural Science Foundation of China (Nos. 61871251, 81561168023, 61601019, and 61871022), in part by Chinese “111 Project” (No. B13003).

References

[1] M. Lainscak, N. Dagres, G.S. Filippatos, et al., Atrial fibrillation in chronic non-cardiac disease: where do we stand? Int. J. Cardiol. 128 (2008) 311–315.
[2] L. Friberg, M. Rosenqvist, A. Lindgren, et al., High prevalence of atrial fibrillation among patients with ischemic stroke, Stroke 45 (2014) 2599–2605.
[3] J.S. Healey, S.J. Connolly, M.R. Gold, et al., Subclinical atrial fibrillation and the risk of stroke, N. Engl. J. Med. 366 (2012) 120–129.
[4] L.A. Sposato, L.E. Cipriano, G. Saposnik, et al., Diagnosis of atrial fibrillation after stroke and transient ischaemic attack: a systematic review and meta-analysis, Lancet Neurol. 14 (2015) 377–387.
[5] B. Freedman, J. Camm, H. Calkins, et al., Screening for atrial fibrillation: a report of the AF-SCREEN international collaboration, Circulation 135 (2017) 1851–1867.
[6] S. Asgari, A. Mehrnia, M. Moussavi, Automatic detection of atrial fibrillation using stationary wavelet transform and support vector machine, Comput. Biol. Med. 60 (2015) 132–142.
[7] L. Clavier, J.-M. Boucher, R. Lepage, et al., Automatic P-wave analysis of patients prone to atrial fibrillation, Med. Biol. Eng. Comput. 40 (2002) 63–71.
[8] S. Dash, K. Chon, S. Lu, et al., Automatic real time detection of atrial fibrillation, Ann. Biomed. Eng. 37 (2009) 1701–1709.
[9] J. Lee, Y. Nam, D.D. McManus, et al., Time-varying coherence function for atrial fibrillation detection, IEEE Trans. Biomed. Eng. 60 (2013) 2783–2793.
[10] H. Prasad, R.J. Martis, U.R. Acharya, et al., Application of higher order spectra for accurate delineation of atrial arrhythmia, Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, IEEE, 2013, pp. 57–60.
[11] J. Lee, B.A. Reyes, D.D. McManus, et al., Atrial fibrillation detection using an iPhone 4S, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 60 (2013) 203–206.
[12] U.R. Acharya, H. Fujita, M. Adam, et al., Automated characterization of arrhythmias using nonlinear features from tachycardia ECG beats, 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2016, pp. 000533–000538.
[13] A. Kennedy, D.D. Finlay, D. Guldenring, et al., Automated detection of atrial fibrillation using RR intervals and multivariate-based classification, J. Electrocardiol. 49 (2016) 871–876.
[14] S. Parvaneh, M.R. Golpayegani, M. Firoozabadi, et al., Predicting the spontaneous termination of atrial fibrillation based on poincare section in the electrocardiogram phase space, Proc. Inst. Mech. Eng. H 226 (2012) 3–20.
[15] Y. Xia, N. Wulan, K. Wang, et al., Detecting atrial fibrillation by deep convolutional neural networks, Comput. Biol. Med. 93 (2018) 84–92.
[16] Z. Yao, Z. Zhu, Y. Chen, Atrial fibrillation detection by multi-scale convolutional neural networks, Information Fusion (Fusion), 2017 20th International Conference on, IEEE, 2017, pp. 1–6.
[17] P. Rajpurkar, A.Y. Hannun, M. Haghpanahi, et al., Cardiologist-level Arrhythmia Detection with Convolutional Neural Networks, (2017) arXiv preprint arXiv:1707.01836.
[18] U.R. Acharya, H. Fujita, O.S. Lih, et al., Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network, Inf. Sci. 405 (2017) 81–90.
[19] J. Ji, X. Chen, C. Luo, et al., A deep multi-task learning approach for ECG data analysis, Biomedical & Health Informatics (BHI), 2018 IEEE EMBS International Conference on, IEEE, 2018, pp. 124–127.
[20] C. Yuan, Y. Yan, L. Zhou, et al., Automated atrial fibrillation detection based on deep learning network, Information and Automation (ICIA), 2016 IEEE International Conference on, IEEE, 2016, pp. 1159–1164.
[21] S. Kiranyaz, T. Ince, M. Gabbouj, Real-time patient-specific ECG classification by 1-D convolutional neural networks, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 63 (2016) 664–675.
[22] Atrial Fibrillation: National Center for Chronic Disease Prevention and Health Promotion , Division for Heart Disease and Stroke Prevention, (August 22, 2017) https://www.cdc.gov/dhdsp/data_statistics/fact_sheets/fs_atrial_fibrillation.htm.
[23] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[24] K. He, X. Zhang, S. Ren, et al., Identity Mappings in Deep Residual Networks, European Conference on Computer Vision, Springer, 2016, pp. 630–645.
[25] G. Huang, Z. Liu, L. Van Der Maaten, et al., Densely Connected Convolutional Networks, CVPR, 2017, p. 3.
[26] C. Szegedy, W. Liu, Y. Jia, et al., Going deeper with convolutions, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[27] M. Lin, S. Chen QYan, Network in Network, (2013) arXiv preprint arXiv:1312.4400.
[28] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation Networks, vol. 7, (2017) arXiv preprint arXiv:1709.01507.
[29] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, Proceedings of the 27th International Conference on Machine Learning, ICML-10, 2010, pp. 807–814.
[30] K. He, X. Zhang, S. Ren, et al., Delving deep into rectifiers: surpassing human-level performance on imagenet classification, Proc. IEEE Int. Conf. Comput. Vis. (2015) 1026–1034.
[31] D.P. Kingma, J. Ba, Adam. A Method for Stochastic Optimization, (2014) arXiv preprint arXiv:1412.6980.
[32] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, (2013) arXiv preprint arXiv:1312.6034.

[33] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1987) 37–52.
[34] L.V.D. Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[35] S.L. Oh, E.Y.K. Ng, R.S. Tan, et al., Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats, Comput. Biol. Med. 102 (2018) 278–287.
[36] U. Erdenebayar, H. Kim, J.U. Park, et al., Automatic prediction of atrial fibrillation based on convolutional neural network using a short-term normal electrocardiogram signal, J. Korean Med. Sci. 34 (2019) e64.
[37] Z. Xiong, M. Stiles, J. Zhao, Robust ECG signal classification for the detection of atrial fibrillation using novel neural networks, 2017 Computing in Cardiology Conference, CinC, 2017.
[38] M. Zihlmann, D. Perekrestenko, M. Tschannen, Convolutional recurrent neural networks for electrocardiogram classification, 2017 Computing in Cardiology Conference, CinC, 2017.
[39] P.A. Warrick, M.N. Homsi, Ensembling convolutional and long short-term memory networks for electrocardiogram arrhythmia detection, Physiol. Meas. 39 (11) (2018) 114002.
[40] S. Parvaneh, J. Rubin, A. Rahman, et al., Analyzing single-lead short ECG recordings using dense convolutional neural networks and feature-based post-processing to detect atrial fibrillation, Physiol. Meas. 39 (8) (2018) 084003.
[41] S. Parvaneh, J. Rubin, Electrocardiogram monitoring and interpretation: from traditional machine learning to deep learning, and their combination, 2018 Computing in Cardiology Conference, CinC, 2018.
[42] C.T. January, L.S. Wann, J.S. Alpert, et al., AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American heart association task force on practice guidelines and the heart rhythm society, J. Am. Coll. Cardiol. 64 (2014) e1–e76 2014.
[43] F.R. Hobbs, D. Fitzmaurice, J. Mant, et al., A randomised controlled trial and cost-effectiveness study of systematic screening (targeted and total population screening) versus routine practice for the detection of atrial fibrillation in people aged 65 and over. The SAFE study, Health Technol. Assess. 9 (2005) 93.
[44] I. Willits, K. Keltie, J. Craig, et al., WatchBP Home A for opportunistically detecting atrial fibrillation during diagnosis and monitoring of hypertension: a NICE medical technology guidance, Appl. Health Econ. Health Policy 12 (2014) 255–265.