Deep Convolutional Neural Networks and Learning ECG Features for Screening Paroxysmal Atrial Fibrillation Patients

Bahareh Pourbabaee, Mehrsan Javan Roshtkhari, and Khashayar Khorasani, *Member, IEEE*

***Abstract*—In this paper, a novel computationally intelligent-based electrocardiogram (ECG) signal classification methodology using a deep learning (DL) machine is developed. The focus is on patient screening and identifying patients with paroxysmal atrial fibrillation (PAF), which represents a life threatening cardiac arrhythmia. The proposed approach operates with a large volume of raw ECG time-series data as inputs to a deep convolutional neural networks (CNN). It autonomously learns representative and key features of the PAF to be used by a classification module. The features are therefore learned directly from the large time domain ECG signals by using a CNN with one fully connected layer. The learned features can effectively replace the traditional *ad hoc* and time-consuming user's hand-crafted features. Our experimental results verify and validate the effectiveness and capabilities of the learned features for PAF patient screening. The main advantages of our proposed approach are to simplify the feature extraction process corresponding to different cardiac arrhythmias and to remove the need for using a human expert to define appropriate and critical features working with a large time-series data set. The extensive simulations and case studies conducted indicate that combining the learned features with other classifiers will significantly improve the performance of the patient screening system as compared to an end-to-end CNN classifier. The effectiveness and capabilities of our proposed ECG DL classification machine is demonstrated and quantitative comparisons with several conventional machine learning classifiers are also provided.**

***Index Terms*—Biomedical monitoring, deep convolution neural network, electrocardiogram (ECG), feature extraction, neural network architecture, paroxysmal atrial fibrillation (PAF).**

I. Introduction

BIOMEDICAL signal analysis, and in particular electrocardiogram (ECG) analysis, has many useful applications, including activity recognition, biometric identification and more importantly, patient screening and diagnosis [1]. Atrial fibrillation (AF) is one of the life threatening arrhythmias which may lead to a major risk of stroke or heart failure if it is not properly and promptly detected and diagnosed. This occurs when the atria loses the normal rhythm and beat chaotically. Over 2.3 million people in the U.S. and 6 million Europeans suffer from some type of AF [2]. In addition to inheriting heart problems, different external factors, such as consumption of nicotine, cocaine, caffeine, and alcohol may also lead to the AF. It is well-known that high stress, overactive thyroid, and low potassium may also increase the risk of AF.

AF may lead to an increased risk of heart failure and five-times-more-likely incidence of stroke. This necessitates an early diagnosis of AF. Although, the definitive diagnosis of AF is through the 12-lead ECG, it is more cost effective to screen the AF patients through the use of one lead of the ECG signal.

The paroxysmal AF (PAF) as one of the AF types is an episode of uncoordinated movement of atria that occurs occasionally and takes a few minutes to days to stop. Oftentimes, the PAF does not have any obvious symptoms for the patient and may not be detected during a clinical monitoring. In order to avoid any delay in the PAF diagnosis and to diminish the subsequent risks, physicians generally recommend that patients use an ECG wearable device to have their daily pervasive monitoring.

During an extended ECG pervasive monitoring, a significantly large data set is collected which renders it practically infeasible to employ a visual inspection by a physician. Therefore, a computer-based arrhythmia screening system needs to be designed to autonomously detect any abnormal beat. The provided system should work based on a set of subject-independent general fixed rules that are defined by physicians. One of the major drawbacks of such a system is that, in general, the rules do not consider variations in the arrhythmia patterns among the subjects or among different activities that are performed by the subjects during a particular test.

In general, the ECG signals consist of six components that are designated as P, Q, R, S, T, and U. Fig. 1 depicts the ECG signals for a healthy person as well as that of an AF patient. It can be observed from this figure that for an AF patient there are tiny irregular fluctuations in the *P*-wave and QRS complexes as compared to the healthy individual. Given the presence of these irregularities, for the purpose of arrhythmia screening various morphological features, including the peaks and widths corresponding to different ECG segments are typically used.

Manuscript received February 28, 2017; accepted April 25, 2017. This paper was recommended by Associate Editor J. Wu. *(Corresponding author: Khashayar Khorasani.)*

B. Pourbabaee and K. Khorasani are with the Department of Electrical and Computer Engineering, Concordia University, Montreal, QC H3G1M8, Canada (e-mail: b_pourba@ece.concordia.ca; kash@ece.concordia.ca).

M. J. Roshtkhari is with SPORTLOGiQ, Montreal, QC H2T3B3, Canada (e-mail: mehrsan@sportlogiq.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2017.2705582

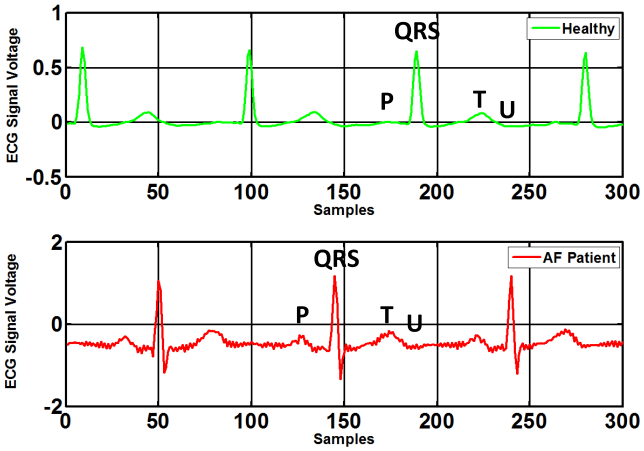

Fig. 1. ECG signal of (top) healthy person and (bottom) AF patient.

To date, most of the reported approaches in the literature for classifying the ECG signals solely rely on extracting hand-crafted features from the ECG signals. This is accomplished either by using conventional feature extraction algorithms or by taking advantage of human expert knowledge. The extracted features are then fed either to generative or discriminative models to predict or classify the ECG signals. Among these methods, support vector machines (SVMs) and hidden Markov models (HMMs) are commonly used with the hand-crafted features that have produced acceptable results in [3]–[7].

The quality of extracted features has the most significant impact on the reliability and performance of the classification/prediction strategy that is invoked and its outcome. Hence, it is always desired to extract the most representative, critical, dominant, and relevant features for a particular arrhythmia.

However, expert-knowledge-based feature extraction techniques are *ad hoc*, time-consuming and error-prone processes and the extracted features are not generally robust with respect to many variations, such as translations, noise, scaling, displacement, etc. Moreover, the ECG signal characteristics are highly subject-dependent and extracting effective features usually requires a deep domain knowledge and expertise [8], [9].

Alternatively, other approaches, such as the wavelet transform, discrete Fourier, and cosine transforms have been used in the literature for extracting features from the ECG signals in both the time and the frequency domains [10]–[14]. The feature extraction process is generally followed by a feature selection algorithm to choose the most relevant features having a higher discrimination power in order to reduce the dimensionality of the feature vector space [15], [16].

Recently, deep learning (DL) neural networks and in particular convolutional neural networks (CNNs) have gained significant interest in multidimensional signal processing problems due to their strong capabilities and functionalities for different applications, such as object detection and classification in computer vision [17]–[19], natural language processing [20], and time-series data analysis [21]–[25].

One of the recent applications of deep neural networks is in time-series classification problems. Time-series classification problems that specifically deal with a large amount of data are used in various applications in health care systems, bioinformatics, activity recognition, etc. Traditional time-series classification methods highly depend on the extracted features, however, it is difficult to extract all the fundamental, proper, and key features for capturing intrinsic properties of a time-series data [21].

Shallow networks that contain only a small number of nonlinear operations do not have sufficient capacity and representational capability to accurately model complex time-series and high dimensional data sets that are subject to noise effects [26]. One method to model such data is to extract relevant features from available data which is a time-consuming and complex process and commonly needs the domain expertise. Moreover, the extracted features need to be ensured to be invariant with respect to translations, scaling, and noise.

Therefore, recently researchers have been active in applying DL methods to solve the existing challenging time-series classification problems. Various methods, such as the deep belief networks, conditional and gated restricted Boltzmann machines, auto encoders, recurrent neural networks, HMMs, and also CNN have been utilized and developed in the literature to address various time-series data classification problems [27], [28].

The strong feature learning capabilities of CNNs make them an ideal and suitable choice for multidimensional signal processing applications. However, these techniques are not well exploited in the biomedical signal processing domain for patient screening due to inherent challenges for deep neural networks in learning from a small number of positive examples.

In general, there are two different approaches for employing deep neural networks for multidimensional signal analysis problems. The first framework is to use deep neural networks as a feature extractor that is combined with discriminative/generative models [18], [19], [21], [24], and the other framework is to use an entire neural network-based classification pipeline [22], [29].

In this paper, we propose a PAF patient screening system which replaces the hand-crafted features by learning them directly from the ECG time-series signals by using the CNNs.

More closely related to our approach is the work in [22], in which the electroencephalogram (EEG) signal is classified for detecting the P300 wave for the brain–computer interactions. Alternatively, in this paper we demonstrate how deep CNNs can be utilized as feature learning mechanisms that can operate with *only a limited* number of labeled data in order to classify biomedical signals, and in particular, the ECG-based patient screening.

Our proposed feature extraction and patient screening methodology takes the raw ECG time-series signals as inputs and learns a discriminative representation of them in the time domain. The raw ECG time-series signals are fed through a convolution neural network with only one fully connected layer. The network parameters are then learned by minimizing the classification error. The output of the fully connected

layer is then considered as a feature representation of the input signal to be used in conjunction with different classification schemes.

To summarize, the *first contribution* of this paper is the adoption of CNNs as a feature learning mechanism that is applied to ECG signals for patient screening. The feature learning mechanism is shown to be capable of generating robust features *without* requiring the domain knowledge and a feature selection algorithm, as opposed to conventional feature extraction schemes [30]–[32]. In our *second contribution*, the proposed CNN-based feature learning mechanism is integrated with other standard classifiers, namely the K-nearest neighbor (KNN), SVM, and the multilayered perceptron (MLP) networks to improve the accuracy of the patient screening procedure when compared to an end-to-end CNN structure that is used for both feature extraction and classification tasks. The effects of various learning and structural parameters on the training and testing data classification errors and the patient screening accuracy are also investigated through extensive simulation case studies.

It must be noted that our proposed framework can also be applied for screening other types of cardiac arrhythmias through fine-tuning the weights of our proposed CNN network. This is left as a topic for our future research.

II. Methodology

CNNs were initially developed in the 1980s by Fukushima [33]. It is the first DL approach, where hierarchical layers are trained robustly by means of the stochastic gradient decent algorithm. It is also a popular method for feature extraction and time-series data classification. It hardly needs any data preprocessing and pretraining algorithms [34].

The CNN has been widely used for feature extraction and signal classification problems in the literature. In [22], a CNN with two convolutional layers and one subsampling layer is used to solve two classification problems, namely the P300 wave detection via the EEG signal classification and the character recognition problems. Moreover, in [35] a Fourier transform is applied between the convolutional and subsampling layers to switch the EEG signals from the time-domain to the frequency-domain for detecting the steady-state visually evoked potential.

In [21], a multichannel CNN is used for both feature extraction and classification tasks that is applied to different time-series data. In this paper, the CNN is first proposed for the ECG signal feature extraction, and then it is utilized as a classifier with other types of classification algorithms. The detail on the general structure of our proposed CNN-based feature extraction method are provided below.

A. CNN Structure

In general, the CNN is composed of a number of convolutional layers, where each layer is usually followed by a subsampling (pooling) layer followed by one or more fully connected layers. There are also certain number of feature maps including a number of neurons that are located in both the convolutional and the subsampling layers. Also, a weight sharing mechanism exists among the neurons that are located in the same feature map, which implies that they use a similar set of weights. This can reduce the risk of overfitting problem due to use of a fewer number of network parameters.

In convolutional layers, the network is locally connected to extract and convolve the features with their associated weights. The weights that constitute the parameters of the convolutional kernels in each layer are trained by means of the backpropagation error algorithm. The higher the number of the convolutional layers, the more relevant and higher level features can be extracted.

An activation function which is either a *sigmoid* or a *tanh* function is applied to the convolved features as follows:

$$\begin{aligned} \sigma_i^k &= W^k x_i + b^k \\ h_i^k &= \frac{1}{1+e^{-\sigma_i^k}} \end{aligned} \tag{1}$$

where σ_i^k denotes the convolution result that is associated with the ith input and the kth feature map, W^k and b^k denote the corresponding weights and bias terms for the kth feature map, respectively, and h_i^k denotes the output of a nonlinear activation function that is applied to the kth feature map outputs. Moreover, x_i denotes the ith training data that has a dimension n ($x_i \in \mathbb{R}^n$). The label that is associated with x_i is represented by y_i, which has a dimension L.

In order to reduce the dimensionality of the convolved extracted features and to increase the speed of the learning process, either the max or the mean pooling operation is applied to the following hidden layer that is called the subsampling layer. The pooling mechanism is performed by computing the maximum or the average of convolved features among the adjacent neurons that are located in the preceding convolutional layer. Following a given set of convolutional and subsampling layers, one or more fully connected layers are used whose neurons are connected to all the neurons from the preceding layer. Most of the CNN parameters are generally occupied by the fully connected layer parameters.

In this paper, a five-layer CNN architecture including the input, convolutional, subsampling, fully connected, and output layers is utilized in which the optimal kernel size and the feature maps are obtained in Section III. The overall structure of our proposed CNN for the ECG signal feature extraction and classification tasks is depicted in Fig. 2.

B. CNN Learning Mechanism

The deep CNN network utilizes a supervised feature learning mechanism, which is known as the stochastic gradient descent (SGD) algorithm. The SGD scheme works more promptly and efficiently than the batch learning methods, specially when it is applied to a large data set. It is a robust learning method with respect to distortions, displacements, translations, and noise effects on the input signal data.

Similarly, as in ordinary gradient descent algorithms, the SGD method iteratively descends on an error surface that is specified by a loss function. However, unlike the standard gradient descent algorithm in which the gradients are computed

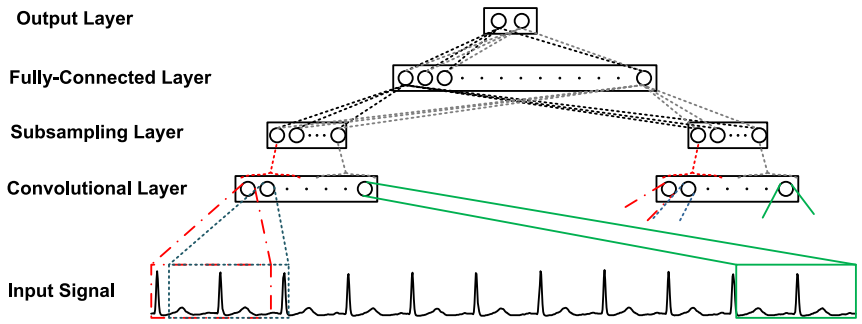

Fig. 2. Overall structure of our proposed CNN network that is developed for the ECG time-series signal feature extraction and classification problem. The network consists of convolutional, subsampling, fully connected, and output layers. In the training process the network is considered as a classifier. The learned features from the ECG signals are then encoded as the output of a fully connected layer.

for the entire training set, the SGD method computes the gradients by solely using a single/few examples of the training data at a time.

During the learning process and mechanism, an *a priori* selected loss or cost function is minimized. In this paper, the loss function that is selected and considered is the negative log-likelihood (NLL) function that is given by

$$\text{NLL}(\theta, \mathcal{D}) = -\sum_{i=0}^{|\mathcal{D}|} \log P(Y = y_i | x_i, \theta) \tag{2}$$

where θ and $\mathcal{D}$ denote the sets of CNN parameters and training data, respectively. To avoid overfitting and to improve the generalization capabilities of the CNN network, an L_2 regularization mechanism is also utilized here. For this purpose, an extra term is added to the loss function above which penalizes large values of the parameters. Therefore, the regularized loss function is given by

$$E(\theta, \mathcal{D}) = NLL(\theta, \mathcal{D}) + \lambda R(\theta) \tag{3}$$

where $R(\theta) = ||\theta||_2$ denotes the L_2 norm of the CNN parameter θ. Furthermore, λ denotes a design parameter which controls the regularization weight of the above loss function. Due to presence of the regularization process, a smooth mapping is generated by using the CNN network. Hence, minimizing $E(\theta, \mathcal{D})$ corresponds to obtaining a tradeoff between the network training data fitting accuracy and the network generalization capabilities to unseen data.

III. Simulation and Comparative Case Study Results

In this paper, a five-layer CNN network that includes the input, convolutional, subsampling, fully connected, and output layers is developed to extract features from the ECG signals and to solve a dichotomy classification problem. Our main classification objective is to diagnose and detect individuals who are at the risk of PAF. For this purpose, the PAF prediction *challenge database* that is available in [36], and which includes two-channel ECG recordings each having a 30 min duration is used in this paper.

The database is divided into two sets: 1) the training set which contains 100 signals of 30-min ECG duration that are collected for two classes of normal/healthy individuals and PAF patients each with equal number of recordings and 2) the testing set which contains 50 signals of 30-min ECG duration recordings in which 28 subjects are at the risk of PAF. In all the experiments in this paper, the test set was kept completely isolated from the training set, i.e., the same training data set is used to learn the ECG features and build the classifiers.

Due to the *small size* of the publicly available data set, we did not create a separate validation set. However, it is generally proposed to divide the data into three sets including the training, validation, and testing in case sufficient amount of data is available. Note that the ECG signals are recorded with a sampling frequency of 128 Hz resulting in 38 400 samples corresponding to a 5 min ECG recorded data (by partitioning the 30 min ECG signal into six segments).

The database in [36] that we are utilizing was collected for an announced challenge in 2001 in which the top score reported was a correct classification rate (CCR) of 82% [3]. In most of the reported results in the literature for this challenge, the researchers have extracted various types of features that are associated with the *P*-wave and the RR interval of the ECG time-series signals.

In this paper, as stated earlier our proposed CNN network consists of one convolutional layer, one subsampling (pooling) layer, and a fully connected layer. Different structural and learning parameters are investigated to achieve an *almost* optimal structure for the proposed CNN network to achieve and yield the minimum loss function. It is also possible to use a larger number of convolutional and subsampling layers to obtain a better classification performance in case a larger data set is available. However, increasing the number of fully connected layers will lead to a larger number of weights that cannot necessarily improve the network performance, since a vanishing gradient problem may occur during the backpropagation error process.

This paper investigates the PAF classification problem under two experiments: 1) an end-to-end CNN network is applied to extract the features from the normalized ECG signals and to classify them into the two classes and 2) only the first four layers of the CNN network are used to first obtain the feature vectors and then we utilize *other conventional* classifiers.

1) The KNN.
2) The SVM.
3) The MLP.

TABLE I
OPTIMAL CNN NETWORK SPECIFICATIONS DESIGNED FOR THE ECG CLASSIFICATION PROBLEM

Learning rate initial value	0.09
Momentum coefficient	0.9
Convolutional layer kernel size	32
No. of feature maps in the convolutional and sub-sampling layers	128
Sub-sampling layer kernel size	128
No. of neurons in the fully connected layer	64
Epoch number	88

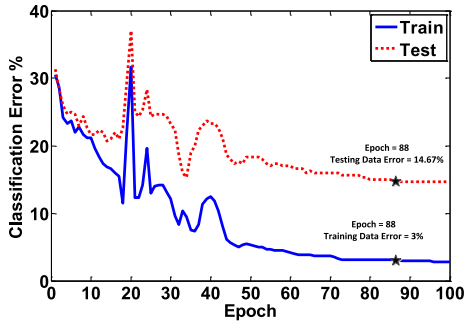

Fig. 3. Classification errors trends associated with the training and testing ECG data by using our proposed end-to-end CNN network with the optimal set of parameters as specified in Table I.

to classify the extracted feature vectors. The results corresponding to the above simulation case studies are quantitatively compared subsequently in terms of the classification error corresponding to the total number of the testing ECG data samples as well as the CCR.

A. Experiment I: End-to-End CNN Network

In order to obtain an optimal structure for our proposed CNN network, the effects of various structural and learning parameters are quantitatively investigated in this section. For this purpose, a trial and error parameter selection technique is used. The final error values that are associated with the training and testing ECG sample data are evaluated under multiple comparative case studies. In all the conducted simulation scenarios, the 30-min ECG signal is divided into six shorter length signals each of 5 min duration (38 400 samples) in order to decrease the number of necessary convolutions and to speed-up the learning process. The set of CNN network parameters that lead to the minimum training error is shown in Table I.

The number of feature maps in the subsampling layer is similar to the convolutional layer. The learning rate is initialized at 0.09 and is decreased by a factor of 0.0975 in the subsequent epochs. The proposed CNN network is capable of classifying the test subjects with a CCR of 85.33%.

The classification errors associated with the training and testing ECG data trends are also depicted in Fig. 3. According to results shown in Fig. 3, the classification error is stabilized and the CNN network has converged after 88 epochs. Due to the limited size of the available training data, the performance of the proposed CNN network *cannot* be further improved by increasing the number of convolutional and/or subsampling layers.

TABLE II
PERCENTAGE OF CLASSIFICATION ERRORS ASSOCIATED WITH THE TRAINING AND TESTING ECG DATA THAT ARE APPLIED TO OUR PROPOSED CNN WITH VARIOUS INITIAL LEARNING RATE VALUES

Learning Rate Initial Value	Training Data % Classification Error	Testing Data % Classification Error
0.01	30.5	31.5
0.05	26.83	29.33
0.07	8.17	18.33
0.09	3	14.67
0.1	12.75	20.5
0.2	7.25	21.5
0.5	48	47.33

TABLE III
PERCENTAGE OF CLASSIFICATION ERRORS ASSOCIATED WITH THE TRAINING AND TESTING ECG DATA THAT ARE APPLIED TO OUR PROPOSED CNN WITH VARIOUS MOMENTUM COEFFICIENT VALUES

Momentum Coefficient	Training Data % Classification Error	Testing Data % Classification Error
0.1	12.5	31
0.5	4.33	16.67
0.7	3.83	16.33
0.9	3	14.67
0.99	4.17	12.33

In the following sections, the effects of various learning and structural parameters on the performance of our proposed end-to-end CNN network classification methodology are studied.

1) Learning Rate Effects: In this experiment, the initial value of the learning rate is changed whereas a similar structure as shown in Table I is used for the CNN network. Table II displays the training and testing ECG data classification errors associated with various initial values of the learning rate during the 88 epochs.

According to the results shown in Table II, the performance of our proposed CNN network is improved neither by increasing nor by decreasing the learning rate initial values. Hence, the optimal value of the proposed CNN network structure is the one that is provided in Table I.

2) Momentum Coefficient Effects: In this scenario, the effects of varying the momentum coefficient on the performance of our proposed CNN network are investigated. The percentage of the classification errors associated with the training and testing ECG data corresponding to the 88 epochs are computed and shown in Table III for five different momentum coefficient values. The other CNN network structural parameters are the same as those that are shown in Table I.

According to the results that are shown in Table III, changing the momentum coefficient does not necessarily improve the performance of our proposed CNN network. The optimal momentum coefficient value is set to 0.9 for the selected ECG data set. The momentum coefficient prevents the neural network from converging to a local minimum or a saddle point. However, if it is set to a very high value, it may lead to an unstable learning system. For instance, a value set to one makes the network training to become infeasible since the gradient is then ignored.

3) CNN Network Structural Parameters Effects: In this paper, the initial values of the learning rate and the momentum

TABLE IV
PERCENTAGE OF CLASSIFICATION ERRORS ASSOCIATED WITH THE TRAINING AND TESTING ECG DATA FOR VARIOUS CNN STRUCTURAL PARAMETERS

Convolutional layer kernel size	32	32	32	32	32
Convolutional layer map size	32	32	64	64	64
Sub-sampling layer kernel size	32	128	32	32	64
Fully connected layer neurons	32	32	32	64	64
% Training Error	15.5	21	5.5	4.5	11
% Testing Error	25	33	28	33	30

coefficient are, respectively, set to 0.09 and 0.9 as indicated in Table I. However, the effects of changes in the other CNN network structural parameters, namely.

1) The kernel size.
2) The number of feature maps.
3) The number of neurons in the fully connected layer.

on the classification errors corresponding to both the training and testing ECG data are obtained and shown in Table IV.

It must be noted that the number of epochs is set to 88 for all the conducted simulation scenarios. According to the results that are shown in Table IV, changing the structural parameters of our proposed CNN *do not* necessarily decrease the classification error and improve the classification performance. Therefore, the parameters that are indicated in Table I are still considered to be as optimal selection for our application.

B. Experiment II: KNN Classifier

In this section, the outputs that are obtained from our proposed CNN network fully connected layer are now used as extracted features corresponding to the ECG signals that are subsequently applied to a KNN classifier subject to different values of k. The feature vectors are generated by the CNN network having the parameters that are specified in Table I. The dimension of the feature vectors is 64, which is the same as the number of neurons in the fully connected layer. This will lead to a more accurate classification result as confirmed by the conducted simulation experiments. The KNN classifier uses the Mahalanobis distance function and in case of undecided labels for even values of k, they are treated as positives. Table V shows the classification error percentages that are associated with the training and testing ECG data corresponding to different values of k. The highest CCR that is obtained for the testing ECG data is 91% corresponding to $k = 2$, which is significantly greater than the 85.33% CCR that is obtained by the end-to-end CNN architecture.

C. Experiment II: SVM Classifier

In this section, the outputs that are obtained from the CNN network fully connected layer are now used to represent the extracted features corresponding to the ECG time-series signals. These signals are now directly applied to an SVM classifier. The classification performance of this scheme is now compared with the previous classification methods. Both linear and Gaussian kernel SVMs are designed in this section subject to different parameters. The optimal learning result is obtained for a Gaussian kernel SVM network with the variance of 2.85 and the regularization parameter of 11.

The classification errors associated with the training and testing ECG time-series data for the Gaussian kernel SVM are obtained as 0.83% and 10%, respectively. Specifically, by using the Gaussian SVM, the CCR associated with the testing ECG data is obtained as 90%, which is far greater than the CCR that is obtained by the end-to-end CNN architecture.

It should be noted that linear SVM does not work as effectively as the Gaussian kernel SVM, however, the results that are obtained from the linear SVM are also compared in Table VI under different regularization parameter values.

According to the results that are shown in Table VI, the minimum training data classification error that is obtained corresponds to the regularization parameter of 25, for which the testing data CCR is 87.67%. However, the CCR is still less than that was generated by the Gaussian kernel SVM.

In another experiment, the regularization parameter is fixed at 11 and the variance of the Gaussian kernel is changed to investigate the classification performance of the Gaussian kernel SVM. The results are shown in Table VII. According to the results shown in Table VII, the classification performance and the learning mechanism are *not* necessarily improved by changing the variance of the Gaussian kernel as compared with those that are generated by the previous Gaussian kernel SVM.

In the last experiment of this section, the variance of the Gaussian kernel is fixed at 2.85, however, the regularization parameter is changed to investigate the classification performance of the Gaussian kernel SVM. The results are provided in Table VIII.

According to the results shown in Table VIII, the classification errors associated with the training and testing ECG data are reduced by increasing the regularization parameter. The training data classification error will not change if the regularization parameter exceeds 11, although the testing data classification error is increased, and therefore the SVM classification performance cannot be improved further by choosing a regularization parameter beyond 11.

D. Experiment II: MLP Classifier

In this experiment, the extracted features from the fully connected layer of the designed end-to-end CNN are used to train another MLP classifier to accomplish and tackle the classification problem. The MLP classifier consists of only one hidden layer. The best selection for the number of neurons in the hidden layer is obtained as 37 with the learning rate of 0.09. Fig. 4 depicts the training and the testing data error classification trends that are generated by the MLP classifier.

The learning process is terminated at the Epoch = 73, when the minimum training error is achieved. The resulting CCR for the testing ECG data reaches 86.33% by using the above MLP classifier.

To investigate the MLP classifier performance, Table IX shows the effects of changes in the number of hidden layer neurons on the classification errors associated with the training and testing ECG time-series data. The learning rate is fixed at 0.09 and the classification errors are finally measured at Epoch = 73. According to the results shown in Table IX,

TABLE V
PERCENTAGE OF THE CLASSIFICATION ERROR ASSOCIATED WITH THE TESTING ECG DATA FOR DIFFERENT VALUES OF k BY USING THE KNN CLASSIFIER

k	1	2	3	4	5	6	7	8	9	10
% Training Data Classification Error	7.17	5.67	6.33	6.67	7.5	6.17	8.33	8.83	9.5	8.17
% Testing Data Classification Error	10	9	11	10.66	12	11.66	13	12	12.33	12

TABLE VI
PERCENTAGE OF CLASSIFICATION ERROR ASSOCIATED WITH THE TRAINING AND TESTING ECG DATA BY USING THE LINEAR SVM WITH DIFFERENT VALUES OF THE REGULARIZATION PARAMETER

Regularization Parameter	Training Data % Classification Error	Testing Data % Classification Error
0.01	5.83	19
0.1	2.66	13.33
1	2	13.66
2	2	14.33
3	2	14.33
5	1.66	13.33
10	1.66	13
15	1.5	13.66
25	0.83	12.33

TABLE VII
PERCENTAGE OF CLASSIFICATION ERROR ASSOCIATED WITH THE TRAINING AND TESTING ECG DATA BY USING THE GAUSSIAN KERNEL SVM HAVING DIFFERENT VARIANCES AND FIXED REGULARIZATION PARAMETER

Gaussian Kernel Variance	Training Data % Classification Error	Testing Data % Classification Error
2.55	1	11.33
2.65	1	10.66
2.75	1	10.33
2.85	0.83	10
2.95	1.167	10.33
3.05	1.167	11.33
3.15	1.5	11.33
3.25	1.5	11.66

TABLE VIII
PERCENTAGE OF CLASSIFICATION ERROR ASSOCIATED WITH THE TRAINING AND TESTING ECG DATA BY USING THE GAUSSIAN KERNEL SVM HAVING DIFFERENT REGULARIZATION PARAMETERS AND FIXED VARIANCE

Regularization Parameter	Training Data % Classification Error	Testing Data % Classification Error
0.1	9.5	22.66
1	2	12.33
3	1.83	11.33
5	1.5	10.33
7	1	10
9	1	10.66
11	0.83	10
13	0.83	10.33
15	0.83	10.33

the minimum training data classification error is obtained by utilizing 37 neurons in the hidden layer.

IV. DISCUSSION

In this paper, we have proposed a deep CNN to be used for accomplishing two objectives, namely feature extraction and feature classification. The extracted features that are generated by the fully connected layer in the CNN are first classified through augmenting an output layer with two neurons into a CNN structure that results in an end-to-end CNN network. In the second framework that we have developed, the CNN network features are applied to other types of neural network classifiers, namely the KNN, linear SVM, Gaussian kernel SVM, and MLP.

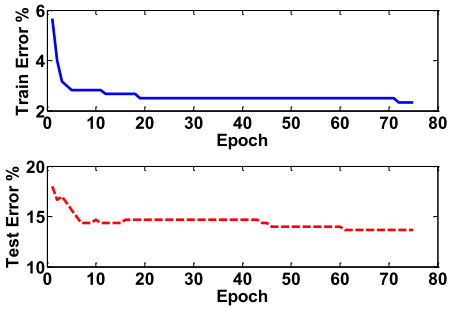

Fig. 4. Classification error changes associated with the training and testing ECG data by using an optimal MLP classifier.

TABLE IX
PERCENTAGE OF CLASSIFICATION ERRORS ASSOCIATED WITH THE TRAINING AND TESTING ECG DATA BY USING THE MLP CLASSIFIER WITH DIFFERENT NUMBERS OF NEURONS IN THE HIDDEN LAYER

Number of Hidden Layer Neurons	Training Data % Classification Error	Testing Data % Classification Error
5	2.83	13.33
9	2.66	15.33
13	2.66	13.36
17	3	13
21	2.66	14.66
25	2.66	14.66
29	2.83	14
33	2.83	14
37	2.5	13.66
41	2.83	14
64	2.66	15.33

TABLE X
CONFUSION MATRIX DEFINITION FOR THE PAF PATIENT SCREENING PROBLEM

Cases	Classified as Patient	Classified as Normal/Healthy
Actual Patient	TP	FN
Actual Normal/Healthy	FP	TN

The classification results that are obtained by applying the above two methodologies are described in detail in Section III. To perform a better comparison among various performance indices and metrics, the notion of the confusion matrix is utilized below and quantitative results are shown in Table XI. Furthermore, the true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN) metrics are defined according to Table X. It must be noted that positive (P) and negative (N) categories are associated with the actual AF case and the healthy one, respectively.

TABLE XI
PR, RC, TNR, FNR, FPR, CCR, NPV, AND DOR ASSOCIATED WITH END-TO-END CNN, KNN, LINEAR SVM, GAUSSIAN SVM, AND MLP CLASSIFICATION METHODS

Classification Method	PR	RC	TNR	FNR	FPR	CCR	NPV	DOR
End-to-End CNN	0.9360	0.7647	0.9456	0.2353	0.0544	0.8533	0.7943	56.4908
CNN with KNN	0.9079	0.9020	0.9048	0.0980	0.0952	0.9100	0.8987	87.4774
CNN with Linear SVM	0.8758	0.8758	0.8707	0.1248	0.1293	0.8767	0.8707	47.4847
CNN with Gaussian SVM	0.9296	0.8627	0.9320	0.1373	0.0680	0.9000	0.8671	86.1185
CNN with MLP	0.9065	0.8235	0.9116	0.1765	0.0884	0.8633	0.8323	48.1139

The performance indices that are identified in Table XI for various classifiers are defined as follows.

1) Precision (PR), $PR = [TP/(TP + FP)]$.
2) Recall (RC), $RC = [TP/(TP + FN)]$.
3) TN rate (TNR), $TNR = [TN/(TN + FP)]$.
4) FN rate (FNR), $FNR = [FN/(FN + TP)]$.
5) FP rate (FPR), $FPR = [FP/(FP + TN)]$.
6) Correct classification rate, $CCR = [(TP + TN)/(P + N)]$.
7) Negative predictive value (NPV), $NPV = [TN/(TN + FN)]$.
8) Diagnostic odds ratio (DOR), a measure of the effectiveness of the diagnostic test for a medical test, namely $DOR = [(RC/FPR)/(FNR/TNR)]$.

The integration of our proposed CNN network as a feature extractor and the KNN as a classifier provides and yields the *highest* RC, CCR, DOR, and the *lowest* FNR among *all* the other examined deep neural network classifier methods. Therefore, the KNN is capable of screening the PAF patients more accurately than the other classifier methods, and it can be proposed as the most suitable methodology. We should emphasize that the proposed CNN network is more appropriate to be utilized as a feature learning mechanism than a classifier, since the patient screening performance can be improved by combining the CNN with other conventional classifiers as compared to only an end-to-end CNN architecture. However, the end-to-end CNN shows a high PR among the other classification techniques and hence, the results are not completely uniform. Since, the CNN network shows the capability of learning features to screen patients, with further experiments using different loss functions on larger datasets there is a possibility to determine a proper architecture for an end-to-end CNN that outperforms in PR the other classification techniques.

The feature vectors that are extracted from the ECG signals corresponding to healthy individuals and the PAF patients are shown in Fig. 5 associated with four healthy and four patient individuals. Although our quantitative and numerical analysis of the classification performance using our learned features show the effectiveness of those features, one can also visually observe that there are noticeable differences and new feature patterns in the patients with PAF as compared to healthy individuals.

Associated with the 64-D feature space (the number of neurons that are selected in the fully connected layer of the CNN structure), some extracted features are fairly consistent between the patients and the healthy individuals while others can clearly discriminate the two groups. Moreover, there are certain features that are not consistent among the patients

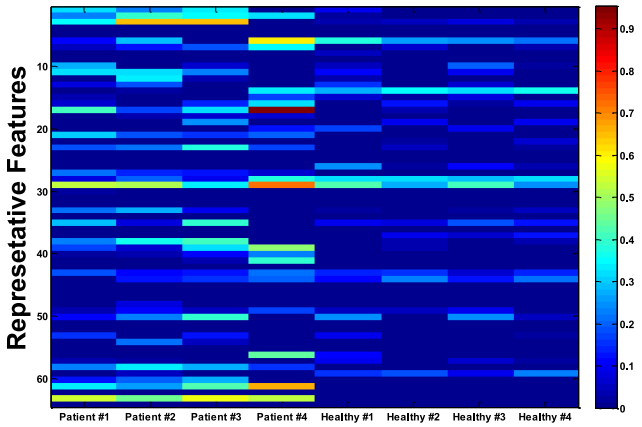

Fig. 5. Visualization of the learned feature vectors for healthy and PAF patient cases. The first four columns correspond to patients and the remaining four columns correspond to healthy individuals. The learned feature vectors can easily discriminate the patients from the healthy individuals.

and seem to encode certain specific information about each patient's ECG signal. It is interesting to note that we do not observe such patterns and phenomenon among the healthy individuals.

To further capture and realize what is being learned by the CNN network, in Fig. 6 we have shown the responses of all the 64 convolutional filters that have been learned from the ECG signals to screen the PAF patients. Most of the learned filter patterns approximately resemble the patterns associated with the ECG signals. It follows that the learned filters may not necessarily represent any characteristic features that are related to the PAF. However, these learned temporal patterns constitute as the building blocks that form discriminative representations for the PAF patients.

Although the learned filters depict different responses, there are some filters that have indeed very similar responses, indicating that there could be redundancy in these learned filters. Fig. 7 shows two examples of filters having similar responses. As expected, the similar filter responses are temporally shifted versions of one another. This occurs due to the fact that our CNN does not have the capability to directly model the dynamic characteristics of a time-series data, therefore it attempts to capture the underlying dynamics by learning temporally shifted filter patterns. By using a different network architecture, such as a recurrent or dynamic neural networks, and specifically long short-term memory networks, one can address this issue and could achieve fewer number of convolutional filters. This is left as a topic and problem for our future research.

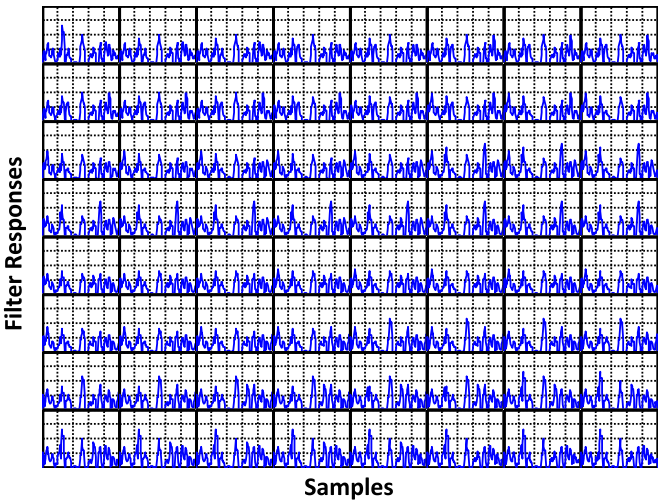

Fig. 6. Responses of the 64 filters corresponding to the first convolutional layer (each block corresponds to the time response of a single filter for the duration of 150 samples).

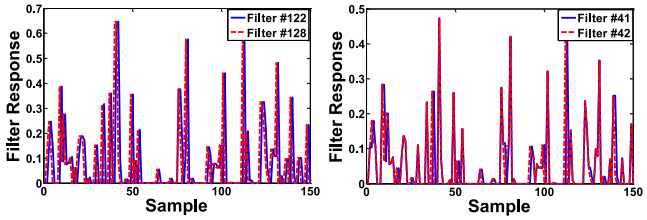

Fig. 7. Responses of the filters that have similar characteristics in the first convolutional layer.

From the computational time and required resources standpoint, the feature extraction that is utilized by our convolutional network is not a very resource demanding and time consuming scheme as compared to what we have experienced in other domains, such as image processing. This is due to the fact that the convolutional filters we have implemented are 1-D filters that are applied to 1-D signals, and hence, the computations are a lot lower as compared to multidimensional filters that are designed for multidimensional signal processing. For example, training the feature extractor network for image processing takes about 760 s per epoch and extracting the features from a 5 min ECG signal (38 400 samples) takes about 0.094 s on a CPU,[1] which is a reasonable metric for practical applications.

V. Conclusion

In this paper, a deep CNN architecture is proposed as a feature learning methodology to solve the PAF patient screening problem. The proposed CNN network is capable of operating with raw ECG time-series signals to extract and down-sample features through computing convolutions among the input vectors with their associated weights as well as determining the maximum outputs among the adjacent neurons. The CCRs under both training and testing ECG time-series data are obtained by performing extensive simulation case study scenarios.

Various machine learning classification methods, namely an end-to-end CNN, KNN, linear SVM, Gaussian kernel SVM, and MLP classifiers are also implemented to address the above problem. It is demonstrated through extensive simulation case studies that integration of our proposed CNN structure as a feature extractor with other conventional neural network-based classification methods will significantly improve the resulting classification performance when compared to only an end-to-end CNN network.

Experimental results confirm the effectiveness of the learned features for patient screening when compared to the hand-crafted features that are designed for PAF screening problem. Moreover, unlike the available conventional feature learning methods in the literature that are utilized for biomedical signal processing applications, our proposed CNN network does *neither need* biomedical domain knowledge and expertise *nor* a feature selection facility or mechanism. It must again be emphasized that in this paper the achieved classification results are significantly superior to the top score that was reported in the PAF screening challenge in 2001 as well as other available classification methods that are applied on the same data set.

[1]The CPU is an Intel Xeon E31230 running at 3.2 GHz. The training times per epoch and feature extraction are 13 and 4.2×10^{-4} s on an Nvidia TitanX GPU, respectively.

References

[1] G. Kaur, G. Singh, and V. Kumar, "A review on biometric recognition," *Int. J. Bio Sci. Bio Technol.*, vol. 6, no. 4, pp. 69–76, 2014.

[2] G. Y. Lip, C. M. Brechin, and D. A. Lane, "The global burden of atrial fibrillation and stroke: A systematic review of the epidemiology of atrial fibrillation in regions outside North America and Europe," *CHEST J.*, vol. 142, no. 6, pp. 1489–1498, 2012.

[3] G. Schreier, P. Kastner, and W. Marko, "An automatic ECG processing algorithm to identify patients prone to paroxysmal atrial fibrillation," in *Proc. Comput. Cardiol.*, Rotterdam, The Netherlands, 2001, pp. 133–135.

[4] W. Zong, R. Mukkamala, and R. G. Mark, "A methodology for predicting paroxysmal atrial fibrillation based on ECG arrhythmia feature analysis," in *Proc. Comput. Cardiol.*, Rotterdam, The Netherlands, 2001, pp. 125–128.

[5] P. de Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 7, pp. 1196–1206, Jul. 2004.

[6] P. de Chazal and R. B. Reilly, "A patient-adapting heartbeat classifier using ECG morphology and heartbeat interval features," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 12, pp. 2535–2543, Dec. 2006.

[7] A. H. Khandoker, M. Palaniswami, and C. K. Karmakar, "Support vector machines for automated recognition of obstructive sleep apnea syndrome from ECG recordings," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 1, pp. 37–48, Jan. 2009.

[8] K. A. Sidek, I. Khalil, and H. F. Jelinek, "ECG biometric with abnormal cardiac conditions in remote monitoring system," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 11, pp. 1498–1509, Nov. 2014.

[9] S. Kar and A. Routray, "Effect of sleep deprivation on functional connectivity of EEG channels," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 3, pp. 666–672, May 2013.

[10] L. Khadra, A. S. Al-Fahoum, and H. Al-Nashash, "Detection of life-threatening cardiac arrhythmias using the wavelet transformation," *Med. Biol. Eng. Comput.*, vol. 35, no. 6, pp. 626–632, 1997.

[11] O. T. Inan, L. Giovangrandi, and G. T. A. Kovacs, "Robust neural-network-based classification of premature ventricular contractions using wavelet transform and timing interval features," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 12, pp. 2507–2515, Dec. 2006.

[12] K.-I. Minami, H. Nakajima, and T. Toyoshima, "Real-time discrimination of ventricular tachyarrhythmia with Fourier-transform neural network," *IEEE Trans. Biomed. Eng.*, vol. 46, no. 2, pp. 179–185, Feb. 1999.

[13] H. Khorrami and M. Moavenian, "A comparative study of DWT, CWT and DCT transformations in ECG arrhythmias classification," *Expert Syst. Appl.*, vol. 37, no. 8, pp. 5751–5757, 2010.
[14] S. Asgari, A. Mehrnia, and M. Moussavi, "Automatic detection of atrial fibrillation using stationary wavelet transform and support vector machine," *Comput. Biol. Med.*, vol. 60, pp. 132–142, May 2015.
[15] X. Fu and L. Wang, "Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 33, no. 3, pp. 399–409, Jun. 2003.
[16] I.-S. Oh, J.-S. Lee, and B.-R. Moon, "Hybrid genetic algorithms for feature selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1424–1437, Nov. 2004.
[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 580–587.
[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[19] Z. Deng, M. Zhai, Y. Liu, S. Muralidharan, M. Javan Roshtkhari, and G. Mori, "Deep structured models for group activity recognition," in *Proc. British Mach. Vis. Conf. (BMVC)*, 2015, pp. 1–12.
[20] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. ACM 25th Int. Conf. Mach. Learn.*, Helsinki, Finland, 2008, pp. 160–167.
[21] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, "Time series classification using multi-channels deep convolutional neural networks," in *Web-Age Information Management*. Cham, Switzerland: Springer, 2014, pp. 298–310.
[22] H. Cecotti and A. Graser, "Convolutional neural networks for P300 detection with application to brain-computer interfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 433–445, Mar. 2011.
[23] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[24] T. N. Sainath, A.-R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Vancouver, BC, Canada, 2013, pp. 8614–8618.
[25] H. Li, D. Pan, and C. L. P. Chen, "Intelligent prognostics for battery health monitoring using the mean entropy and relevance vector machine," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 7, pp. 851–862, Jul. 2014.
[26] L. Wu, S. Liu, and Y. Yang, "A gray model with a time varying weighted generating operator," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 46, no. 3, pp. 427–433, Mar. 2016.
[27] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern Recognit. Lett.*, vol. 42, pp. 11–24, Jun. 2014.
[28] W. Wang, A.-H. Tan, and L.-N. Teow, "Semantic memory modeling and memory interaction in learning agents," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published, doi: 10.1109/TSMC.2016.2531683.
[29] B. Pourbabaee, M. Javan, and K. Khorasani, "Feature leaning with deep convolutional neural networks for screening patients with paroxysmal atrial fibrillation," in *Proc. IEEE World Congr. Comput. Intell.*, 2016, pp. 5057–5064.
[30] E. Ros, S. Mota, F. J. Fernández, F. J. Toro, and J. L. Bernier, "ECG characterization of paroxysmal atrial fibrillation: Parameter extraction and automatic diagnosis algorithm," *Comput. Biol. Med.*, vol. 34, no. 8, pp. 679–696, 2004.
[31] B. Pourbabaee and C. Lucas, "Automatic detection and prediction of paroxysmal atrial fibrillation based on analyzing ECG signal feature classification methods," in *Proc. Cairo Int. Biomed. Eng. Conf. (CIBEC)*, 2008, pp. 1–4.
[32] B. Pourbabaee and C. Lucas, "Paroxysmal atrial fibrillation diagnosis based on feature extraction and classification," in *Proc. IEEE Symp. Comput. Intell. Bioinformat. Comput. Biol. (CIBCB)*, Montreal, QC, Canada, 2010, pp. 1–8.
[33] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, 1980.
[34] I. Arel, D. C. Rose, and T. P. Karnowski, "Deep machine learning—A new frontier in artificial intelligence research [research frontier]," *IEEE Comput. Intell. Mag.*, vol. 5, no. 4, pp. 13–18, Nov. 2010.
[35] H. Cecotti and A. Graeser, "Convolutional neural network with embedded Fourier transform for EEG classification," in *Proc. IEEE 19th Int. Conf. Pattern Recognit.*, Tampa, FL, USA, 2008, pp. 1–4.
[36] G. Moody, A. Goldberger, S. McClennen, and S. Swiryn, "The PAF prediction challenge database," *Comput. Cardiol.*, vol. 28, pp. 113–116, Sep. 2001. [Online]. Available: http://physionet.org/physiobank/database/afpdb/

Bahareh Pourbabaee received the B.Sc. degree from Shahid Beheshti University, Tehran, Iran, in 2005, the M.Sc. degree from the University of Tehran, Tehran, in 2009, and the Ph.D. degree in electrical engineering from Concordia University, Montreal, QC, Canada, in 2016.

She is currently a System Design Engineer in CS Communication and Systems Canada, Montreal. Her current research interests include fault diagnosis, prognosis and health management, robust estimation theory, and neural network applications to pattern recognition and system identification problems.

Mehrsan Javan Roshtkhari received the B.Sc. and M.Sc. degrees from the University of Tehran, Tehran, Iran, in 2006 and 2009, respectively, and the Ph.D. degree in electrical engineering from McGill University, Montreal, QC, Canada, in 2014.

He is currently the Co-Founder and the Chief Technology Officer with SPORTLOGiQ, Montreal, the computer vision-based sport analytics platform that provides comprehensive games statistics to the professional teams using feeds from a single camera. He is also an Adjunct Faculty Member with the ECE Department, McGill University. He has published numerous research articles in the fields of computer vision and machine learning and holds several patents and patents pending applications. His passion is new technologies with a particular interest in intelligent systems and their positive impacts on our daily life.

Khashayar Khorasani (M'85) received the B.S., M.S., and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 1981, 1982, and 1985, respectively.

From 1985 to 1988, he was an Assistant Professor with the University of Michigan at Dearborn, Dearborn, MI, USA, and since 1988, he has been with Concordia University, Montreal, QC, Canada, where he is currently a Professor and Concordia University Tier I Research Chair with the Department of Electrical and Computer Engineering and Concordia Institute for Aerospace Design and Innovation. His current research interests include nonlinear and adaptive control, intelligent and autonomous control of networked unmanned systems, fault diagnosis, isolation and recovery, diagnosis, prognosis, and health management, satellites, unmanned vehicles, neural network applications to pattern recognition, robotics and control, adaptive structure neural networks, and modeling and control of flexible link/joint manipulators. He has authored/co-authored over 450 publications in the above areas.

Dr. Khorasani is currently serving as an Associate Editor for the IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS.