A Deep Learning Framework for Inter-Patient ECG Classification

Manh-Hung Nguyen†, Vu-Hoang-Tran†, Thanh-Hai Nguyen†, Thanh-Nghia Nguyen†
†Faculty of Electrical-Electronics Engineering, HCMC University of Technology and Education, Vietnam

Summary
The robust automatic ECG classification systems has attracted researchers in recent years due to saving time and minimizing errors for heart clinical predictions. Inter-patient Electrocardiography (ECG) classification has been studied extensively and provided promising results, but remains a difficult task due to the diversity among patients. Moreover, conventional methods relied on feature selection frameworks to select suitable feature-sets for inter-patient ECG classification task. However, the hand-craft features are specifically designed for different purposes and may not characterize the original signal in an optimal way. In this paper, a proposed deep learning framework is applied to obtain learned patient-invariant features for ECG classification. In particular, two constraints are embedded in a unified Siamese structure to handle inter-patient diversity and ECG classification simultaneously. The first one indicates that if two ECG signals are in the same class, the extracted features should be similar even these signals are from two different patients. The second one is that the learned features should support well for the classification task. Experimental results have shown that the accuracy is improved significantly even in inter-patient datasets. Moreover, t-SNE visualization proves that the proposed framework can learn the discriminative features.

Key words:
ECG classification, Deep learning, Feature Learning, Siamese network.

1. Introduction

In order to monitor and diagnose cardiovascular diseases, electrocardiography is the most popular and broadly used method due to its simplicity, the cheapness of process, and non-invasive property [1]. Electrocardiogram (ECG) signal records the cardiac electrical activity and gives the important information about heart's condition necessitated for the treatment of heart patients. Manual analyzing long-term ECG records are extremely time-consuming and difficult because of the morphological variation of ECG signal. Besides, human errors might be caused by fatigue when analyzing the ECG signal over a long period of time. Therefore, improving and developing the automatic ECG analysis systems are necessary and they have been got more attention in the recent decade.

The common problem is that all automatic ECG analysis systems have the large variation in morphological characteristics of ECG signals of different patients [2]. In particular, with the same cardiovascular condition, but the ECG signals measured in two different patients may be completely different. This makes most of previous studies focus on patient-specific designs [1-3] or patient adaptation techniques [4-8] which have ability to adjust or improve the classifier corresponding to patient's ECG signal characteristics. Although these patient-specific methods still exist a lot of disadvantages such as: (1) it is time consuming to rebuild the classifier for each patient; (2) need the help of experts to label data from 2 to 5 minutes per patient; (3) the ECG of a healthy person with no history of cardiac arrhythmias might only contain normal beats but no abnormal beats for training, there are very few studies that deal with or successfully construct the generic (inter-patient) ECG classification system due to the high inter-patient variations of ECG.

Therefore, in this research, we would like to build an inter-patient ECG classification system which use a Siamese based deep network to solve the mentioned problems. The major contributions of the proposed method can be summarized as follows:

- A Siamese architecture [9] is introduced to guide the network to extract the patient-invariant features, which is useful for building a generic ECG classification system. As far as our knowledge, this is the first work proposed to use Siamese architecture to handle the inter-patient variation problem and build a robust generic ECG classification system.
- By introducing the Siamese structure into the learning network, the variety of the training set may be expanded and thus it may reduce the required amount of labeled data.
- A multi-task structure, which jointly considers classifying and clustering, is designed in our framework to extract the features. Therefore, the learned features are not only patient-invariant but also useful for classifying diseases.

The rest of this paper is organized as follows: the related works for inter-patient ECG classification are discussed in Section 2. In Section 3, the proposed network for ECG signal classification is described. The experimental results and discussions are given in Section 4. Sections 5 concludes this paper.

Manuscript received January 5, 2019
Manuscript revised January 20, 2019

2. Related Works

In order to handle the high inter-patient variation problem and build a robust inter-patient ECG classification system, extracting and selecting the appropriate features, which can characterize pathological signs, will be the key to success. Therefore, in this section, we would like to divide the previous inter-patient methods into two categories: (a) hand-craft based methods and (b) deep learning methods. The results on each method are reviewed in section 2.1 and section 2.2.

2.1 Hand-craft Based Methods

Similar to other classification tasks, the previous automatic ECG classification methods combined different classification models such as cluster analysis [10], artificial neural network [11], naïve Bayes classifier [12], Support Vector Machine (SVM) [13], decision tree [14], K-Nearest Neighbors (k-NN) [15] with some hand-crafted feature extraction methods such as S-transform [16], Discrete Wavelet Transform [17-19], Continuous Wavelet Transform [18], Discrete Cosine Transform [18], temporal vector-cardiogram [19], normalized RR-interval morphological features [20], Pan-Tompkins algorithm [21] to find optimal algorithms. In addition, to enhance the discriminability of feature representation, the feature fusion strategy [22] was applied to capture critical attributes from multiple aspects and hence it could boost the higher performance. However, combining several features may significantly increase the computational complexity.

In order to decrease the size of the feature vector, reduce computational cost and increase the generality of the method, feature selection and feature dimension reduction are employed. Feature selection methods, such as Genetic Algorithms (GA) [23], feature selection driven [24], and Binary PSO (BPSO) [25] decrease the size of the feature vector by selecting the useful features from the original set of features. To handle the diversity among patients, G. Garcia et al. [26] introduced a PSO framework for inter-patient ECG heartbeat experiment. On the other hands, feature dimension reduction methods such as Principal Component Analysis (PCA) [27], Linear Discriminant Analysis (LDA) [28], Independent Component Analysis (ICA) [29] generate the new compact low-dimensional features from the high-dimensional ones.

2.2 Deep Learning Methods

The hand-craft-based methods might achieve good performance in some specific datasets. However, the hand-craft features are specifically designed for different purposes and may not characterize the original signal in an optimal way. Therefore, to improve feature flexibility and generalization, researchers have recently come to reply on off-the-shelf deep learning methods for feature extraction and ECG signal classification simultaneously [30]. In supervised deep learning methods, with the supervised labels, discriminative deep learning networks would be designed to automatically learn appropriate features in different levels for the ECG signal classification task. However, the biggest weakness of supervised deep learning is the need for large amounts of labeled data. Collecting and labeling such large amounts of labeled data are extremely expensive and time-consuming.

Therefore, in [31], unsupervised deep learning is combined with supervised deep learning to decrease the demand on labeled training data. The idea is that, the unsupervised networks such as auto-encoder [32], Restricted Boltzmann Machine (RBM) [33], etc. are firstly employed to learn representations from unlabeled data. Then the fine-tuning process is used to improve the model performance by tuning the parameters of all layers using labeled data. The results have shown that automatic ECG signal classification systems founded on deep learning features are able to achieve better performance comparing to hand-crafted features.

Besides, in order to extract the more robust features, the traditional Siamese structure [9] has been applied for ECG signals but for different purposes. In [33], this structure was used to build an authentication apparatus based on ECG signal. In [34], it was used to build an arousal recognition from ECG. To estimate arousal level, it requires the user-specific template, which represents the user's neutral affective state, to compare with the current user's ECG. These framework uses Siamese structure as a comparison method for different purposes instead of handling the inter-patient variation problem to build the robust ECG classification system. With the best of our knowledge, very rare research in this field discusses how to design a deep network that can solve the common problem faced in the automatic ECG analysis systems and the diversity of ECG signals of different patients.

3. Proposed Method

To compensate for the diversity of ECG signals of different patients, the deep learning-based framework has been proposed. The ECG beats are firstly extracted from the ECG signals; then they are fed into a deep Siamese network for the classification task. In detail, properties of heart-beat-signals in section 3.1 are introduced and the framework is in section 3.2.

3.1 ECG Signal and Heart Beat Detection

Electrocardiography is used to record the heart operation process using electronic leads. In particular, an ECG signal involves several ECG-beats, therefore, in order to analyze the ECG signal, the ECG-beats should be extracted in advance. As shown in Fig. 1(a), a typical ECG-beat includes: one P-wave, one QRS complex, and one T-wave. The easiest way to extract the ECG-beats is to firstly detect R-peaks in the ECG signal, then to select N points before and after each peak to form an ECG-beat. In our experiment,

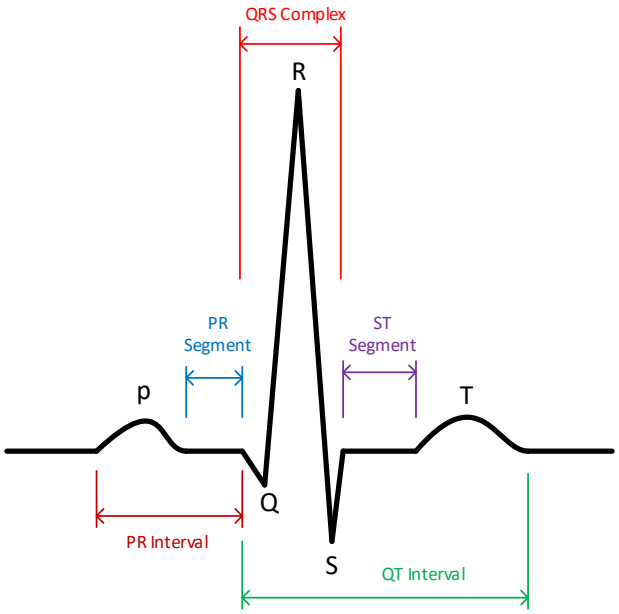

Fig. 1(a) The typical ECG-beat

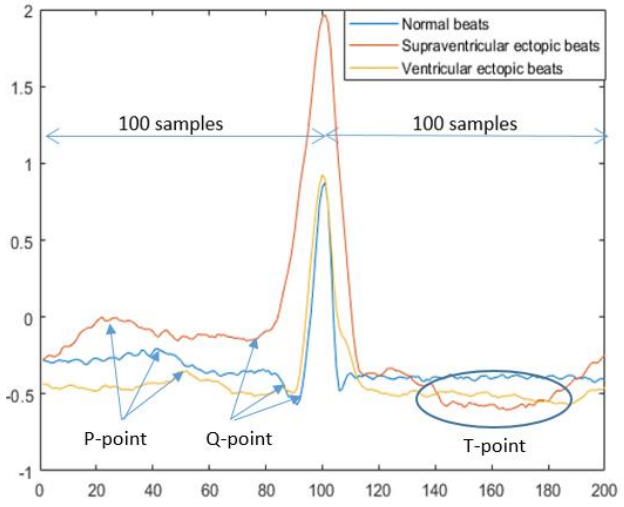

Fig. 1(b) Particular ECG-beat for Normal beats (N), Supraventricular ectopic beats (S), and Ventricular ectopic beats (V)

the MIT dataset are acquired at 360 Hz, hence 200 sampled points is enough to cover a heartbeat [35]. Therefore, we select N = 100 as in Fig.1b. Given the ECG-beats, some physical attributes in the time domain such as heartbeat interval, duration factors (QRS, QT, and PR), amplitude factors (QRS, ST), and combined factors (Q/R ratio, S/R ratio) [36, 37] could be considered to recognize their status. However, as presented in Fig. 1(b), a particular ECG beat is affected seriously by noise or patient motion during data collection process and it is intractable to extract the attributes in the time domain. For instance, the Q point of the Supraventricular ectopic beat does not follow the standard of a formal beat in Fig.1a; and the T point may not be recognized in the example. Therefore, in order to remove noise, it is better to represent the heart-beat signal in the frequency domain [35]. Unlike conventional methods where many feature extraction methods and preprocessing methods are often applied, in this paper, a deep network is aimed to learn the useful features of ECG signals. Therefore, a 4-level discrete wavelet transform (DWT) is simply utilized to represent an ECG-beat in the frequency domain without any pre-processing. In addition, a PCA algorithm [27] is applied to remove unnecessary information in the frequency domain. Following the experiment in this paper, with 40 dimensions, the compressed data could recover 99% of the original information. Therefore, the PCA is applied to project the ECG data into the 40-dimension vectors which are connected into the inputs of the proposed network.

3.2 The Siamese-based Multi-Task Network

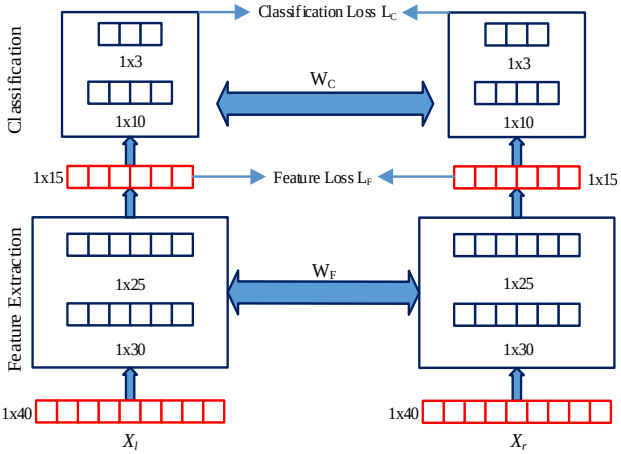

Fig. 2 The proposed Siamese-based multi-task network for heart beat classification

In the feature extraction part, we aim to overcome the diversity among patients. In [26], the authors have pointed out that with the same diseases, the extracted features of ECG signals collected from two different patients may follow different distributions as shown in Fig. 3(a). With these kinds of features, no matter how good the classifier is,

the performance is difficult to improve. Therefore, in order to achieve a robust automatic ECG analysis system, extracting the good features should be a critical issue. In this paper, a feature extraction function $G_{W_F}(X)$, which is firstly introduced, can be applied to extract the patient-invariant features as shown in Fig. 3(b). It means that the similar diseases could be grouped together and the different diseases are separated in the learned feature domain.

In this research, the proposed framework is introduced for inter-patient ECG classification and it includes two main parts: feature extraction and classification. The first one is employed to learn robust features, whereas the second is to learn a model for classification task. Unlike conventional methods, where feature extraction and classification are designed independently, this paper represents the combination of these tasks into a unified framework and training them in an end-to-end fashion. In particular, Fig. 2 introduces the overall structure of the proposed framework using the network for heart beat classification.

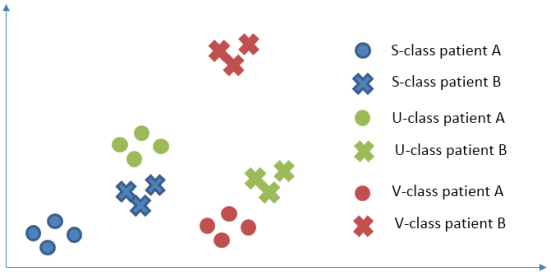

Fig. 3(a) Diversity among patient

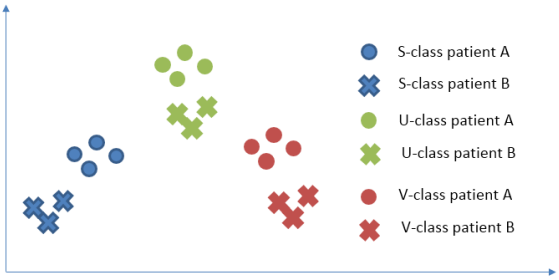

Fig. 3(b) Expected feature distribution

To realize the idea, the Siamese architecture [25] to learn the patient-invariant features is proposed as shown in Fig. 2. In particular, the Siamese architecture has two input sources X_l and X_r, which are randomly sampled from the training dataset $X \in R^{40}$ with the set of 40-dimensional vectors, in which the vectors are achieved using the DWT and PCA on training ECG beats as mentioned in the sub-section II.A. Moreover, each input is projected into a 15-dimensional feature vector using a neural network (NN), which contains three fully connected layers of size 30, 25 and 15, each layer is followed by one sigmoid function. To identically ensure the feature extraction function G_{W_F} of both input sources, two feature extraction NNs are forced to share their weights W_F. Therefore, the distance between two extracted features $G_{W_F}(X_l)$ and $G_{W_F}(X_r)$ can be measured as follows:

$$D_{W_F} = \left\| G_{W_F}(X_l) - G_{W_F}(X_r) \right\|_2 \quad (1)$$

As shown in Fig. 3(b), in order to achieve the patient-invariant features, the parameters W_F are determined such that the distance D is small, if X_l and X_r belong to the same class, $Y_l = Y_r$. In contrast, it should be large, if the classes X_l and X_r are different, (that is $Y_l \neq Y_r$). The idea can be realized by the contrastive loss L_F. In detail, given a set of training data pairs $X_l, X_r\}$ and their corresponding labels $X_l, X_r\}$, the goal is to learn the feature extraction parameters W_F in order to minimize the contrastive loss function L_F defined in the following equation:

$$L_F(W_F \mid \{X_l, X_r, Y_l, Y_r\} = \frac{1}{2N}\sum_{n=1}^{N}(S^n).(D_{W_F}^n)^2 + (1-S^n).\{\max(0, m - D_{W_F}^n)\}^2 \quad (2)$$

where N is the number of training pairs and $D_{W_F}^n$ is the feature distance of n^{th} training pair $X_l, X_r\}$. Moreover, m is the margin value which is chosen to be constant and greater than 0. The margin helps to control the minimum distance among classes and its visual meaning is expressed in Fig. 4. Finally, S^n is the similar indicator of the n^{th} pair and it is described as follow:

$$S^n = \begin{cases} 1 & if \quad Y_l^n = Y \\ 0 & if \quad Y_l^n \neq Y_r^n \end{cases} \quad (3)$$

In order to obtain the category of a given heart-beat, a classification model on the top of feature extraction network is built as shown in Fig. 2. The input of the network model is the extracted contrastive features. Therefore, in the feature domain, data is ideally grouped into distinct categories and then a simple model is created for the classification task. In this work, the model is defined as a one-hidden-layer neural network with the size of 10 nodes, followed by a sigmoid function and moreover, its output has three nodes indicating the probabilities of three classes: normal beat (N), supraventricular ectopic beat (S), and ventricular ectopic beat (V) given a heart-beat. It is denoted that the network parameters for the classification function $G_{W_C}(.)$ as W_C. Given the training set $X, Y\}$, the optimizing parameters $W_F, W_C\}$ is focused to minimize the multi-class cross entropy loss function L_C which is expressed by the following equation:

$$L_C(W_F, W_C | \{W, Y\} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{3} y_i^y \log \hat{y}_i^y \quad (4)$$

where y_i^n and $\hat{y}_i^n$ are the label and predicted probabilities of the i^{th} class of the given n^{th} sample X^n; y_n can be directly obtained from Y}; and N is the number of training samples. Therefore, $\hat{y}_i^n$ can be estimated by the following equation:

$$\hat{y}^n = \sigma(G_{W_C}(G_{W_F}(X^n))) \quad (5)$$

In Eq. (5), σ(.) denotes the sigmoid function.

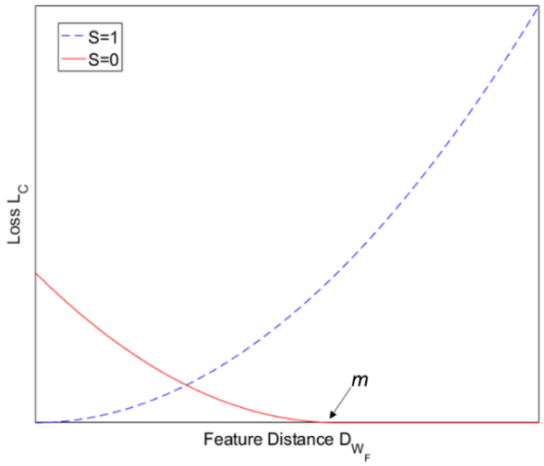

Fig. 4 Visual meaning of contrastive loss L_C.

It is denoted that the classification losses for the left half and right half of the proposed network are L_C^l and L_C^r correspondingly and the multi-task objective function for the whole network as shown in Fig. 2 can finally be expressed as follows:

$$L = (L_C^l + L_C^l) + \lambda_1 L_F + \lambda_2 . \|W\|_2^2 \quad (6)$$

In Eq. (6), to avoid overfitting problem, we introduce L_2-norm $\|W\|_2^2$ as a regularization term, where W denotes all of the weights of the network. The hyper parameter λ_1 and λ_2 are used to control the balance between the terms of the final loss function L. The whole network is employed to train data using the back-propagation method [38] to determine the feature extraction parameters (W_F), and the classification parameters (W_C) simultaneously. It should be noted that WF and WC are shared between the left-half and right-half of the framework. And the framework shown in Fig. 2 is used in training phase, for testing phase, only a half of the framework is used.

4. Experimental Results and Discussion

4.1 Datasets and Evaluation Methods

To demonstrate benefits of the proposed method, a popular dataset, named MIT-BIH [39], is used. The dataset consists of records from 48 patients and each record was acquired at the 360 Hz sampling frequency over 30 minutes. In addition, the paper aims to show the patient diversity, so training and testing data should be collected from patients. Therefore, based on the settings in [26], records from 22 patients to form the training data and other 22 patients to form the testing data are selected in this research. It should be noted that four records related to patients wearing an electronic pacemaker are ignored. The details of the dataset are given in Table I and particularly three categories to test the system include: Normal beats (N), Supraventricular ectopic beats (S), and Ventricular ectopic beats (V). It means that the training sample pairs used in the network are to train. If there are M samples in training set, one can achieve C_2^M training sample pairs. This expands the variety of the training dataset in this research and thus it may reduce the required amount of labeled data.

Table 1: Training set and testing set for intra experiment

	Patient ID	**#samples in each class**		
		N	S	V
Training Set	101, 106, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223, 230	45543	782	3469
Testing Set	100, 103, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 219, 221, 222, 228, 231, 232, 233, 234	44049	1808	3143

Table 2: Confusion matrix for a three-class classifier

		Ground Truth		
		N	S	V
Prediction	n	*nN*	*nS*	*nV*
	s	*sN*	*sS*	*sV*
	v	*vN*	*vS*	*vV*

To evaluate the proposed method, four metrics are utilized as follows: Sensitivity (Se), Positive Predictive (Pp), False Positive Rate (FPR), and Accuracy (ACC). In particular, Se is to measure the proportion of actual classes for correctly recognizing; Pp shows how many percent of predictions one class is correct; ACC is the percentage of beats classified correctly over all categories. Therefore, there metrics are expected to get the high values for a good classifier. In contrast, FPR is to calculate the probability of false alarm within a class which should be small for a good classifier. Given a confusion matrix for three classes as

shown in Table II, the evaluation metrics are defined by equations (7) - (16).

$$Se_N = \frac{nN}{nN + sN + vN} \quad (7)$$

$$Pp_N = \frac{nN}{nN + nS + nV} \quad (8)$$

$$FPR_N = \frac{nS + nV}{nS + nV + sS + vV + sV + vS} \quad (9)$$

$$Se_S = \frac{sS}{sS + nS + vS} \quad (10)$$

$$Pp_S = \frac{sS}{sS + sN + sV} \quad (11)$$

$$FPR_S = \frac{sN + sV}{sN + sV + nN + vV + vN + Nv} \quad (12)$$

$$Se_V = \frac{vV}{vV + sV + nV} \quad (13)$$

$$Pp_S = \frac{sS}{vV + vS + vN} \quad (14)$$

$$FPR_V = \frac{vS + vN}{vS + vN + sS + nN + sN + nS} \quad (15)$$

$$ACC = \frac{nN + sS + vV}{nN + nS + nV + sN + sS + sV + vN + vS + vV} \quad (16)$$

The proposed method is evaluated based on three following aspects: the influence of hyper-parameters; the system performance; and the discrimination of learned features. Firstly, the effect of the hyper-parameters is discussed: λ_1, λ_2 and the margin m on the final result in Section 4. 2 are chosen to be suitable for experiments in this research. Therefore, in Section 4. 3, the results with other methods based on the selected hyper-parameters are compared together. The learned features are visualized by t-SNE [39] to understand the capabilities of the proposed framework in the intuitive way in Section 4. 4. In Section 4.5 we justify the need of using DWT and PCA in our proposed method. Finally, the computational complexity and the robustness in parameter setting are discussed in Section 4.6.

4.2 Hyper-parameters Influence

To evaluate the effect of hyper parameter m on accuracy, firstly $\lambda_1 = 0.001$, $\lambda_2 = 0.001$ are fixed and the m is varied using Eq. (2). Therefore, the change of performance on training dataset is represented in Table 3. The result points out that the system tends to ignore the class *S* when $m = 0$. Setting $m = 0$, it means that the minimum distance among the classes is equal to zero. This eliminates the thrust between the different classes and thus it reduces the ability to separate the classes. Therefore, with these cases, it is too difficult to separate and to account for only a small proportion of the training data and the model is allowed to ignore them. In practical experiment, it is the class S with 45543 samples in the class N, 3469 samples in the class V and only 782 samples (1.57% of the total training data) during training process. Therefore, it could lead to the model directly ignoring samples in the class S and only focusing on two other classes. In contrast, the higher m will make the thrust stronger to separate the different classes, but at the same time it will make the model more sensitive as shown in Table 3. To get the balance in the system performance, $m = 0.5$ is chosen so that the accuracy is highest.

Table 3: Effect of the parameter m where $\lambda_1 = 0.001$, $\lambda_2 = 0.001$ on training dataset (unit is %)

m		0	0.1	0.5	1	5
ACC		95.3	93.92	**95.77**	93.74	93.53
CLASS N	Se	98.35	96.93	**97.95**	95.16	95.26
	Pp	96.71	96.91	**97.78**	98.18	97.86
	FPR	35.8	33.1	**23.32**	18.94	22.28
CLASS S	Se	0	48.34	24.64	**55.88**	51.53
	Pp	0	31.06	32.74	**33.03**	34.68
	FPR	0	1.71	1.08	1.81	1.55
CLASS V	Se	76.74	64.6	**88.1**	83.66	80.25
	Pp	76.52	74.13	**83.48**	67.07	64.74
	FPR	1.76	1.69	**1.23**	3.08	3.27

Table 4: Effect of the parameter λ_1 where $m = 0.5$, $\lambda_2 = 0.001$ on training dataset (unit is %)

λ_1		0	0.001	0.005	0.01	0.05
ACC		95.5	95.77	97.07	**97.42**	96.35
CLASS N	Se	98.58	97.95	99.03	**98.96**	99.03
	Pp	96.67	97.78	97.99	**98.59**	97.2
	FPR	29.51	23.32	19.37	**14.57**	25.07
CLASS S	Se	38.21	24.64	51.31	**60.72**	43.1
	Pp	59.08	32.74	**69.95**	53.96	72.25
	FPR	0.66	1.08	**0.48**	0.73	0.45
CLASS V	Se	78.08	88.1	86.26	**85.51**	83.21
	Pp	88.33	83.48	91.18	**91.84**	90.72
	FPR	0.88	1.23	0.66	**0.61**	0.7

When the parameter m is selected, the value λ_1 is found to be suitable using Eq. (6) and λ_1 is to control the contribution of contractive loss L_F in the total loss of the system. If λ_1 is small, for instance, $\lambda_1 = 0$ as described in the first column of the Table 4, the proposed framework becomes the traditional NN-based classifier with the raw feature input. Without the feature constraint and Siamese structure, the model tends to ignore the class S due to the lacking of training data. In addition, the performance on the class V is not good because of the diversity of ECG signals. With the bigger λ_1, the model focuses more on learning the invariant features and then it reduces the attention on classification task. Hence, the overall accuracy tends to reduce, when λ_1 increases as shown in Table 4. For that

reason, the trade-off λ_1 should be chosen to balance these two tasks. Based on the results as shown in Table 4, λ_1 = 0.01 is chosen in this experiment.

Table 5: Effect of the parameter λ_2 where m = 0.5, λ_2 = 0.01 on training dataset (unit is %)

λ_2		0	0.001	0.005	0.01	0.05
ACC		**97.53**	97.42	97.14	95.31	91.46
CLASS N	Se	99	98.96	**99.09**	97.62	91.46
	Pp	98.41	**98.59**	97.96	97.93	100
	FPR	16.04	**14.57**	19.48	22.93	0
CLASS S	Se	58.46	**60.72**	43.26	29.2	0
	Pp	67.14	53.96	**71.87**	32.23	0
	FPR	0.53	0.73	**0.45**	1.08	1.57
CLASS V	Se	88.78	85.51	**92.08**	84.41	0
	Pp	**92.82**	91.84	92.13	75.12	0
	FPR	**0.54**	0.61	0.59	1.83	6.97

Table 6: Effect of the parameter λ_2 where m = 0.5, λ_2 = 0.01 on testing dataset (unit is %)

λ_2		0	0.001	0.005	0.01	0.05
ACC		96.79	**96.82**	96.3	95.27	89.9
CLASS N	Se	99.05	99.06	**99.17**	96.37	89.9
	Pp	97.46	**97.57**	96.87	99.25	100
	FPR	19.76	**19.07**	23.1	9.06	0
CLASS S	Se	61.37	**64.34**	54.96	0	0
	Pp	83.46	82.25	**86.45**	0	0
	FPR	0.64	0.69	**0.53**	3.69	3.69
CLASS V	Se	**93.41**	90.09	94.35	81.61	0
	Pp	**95.1**	94.62	93.95	94.3	0
	FPR	**0.34**	0.37	0.41	0.39	6.41

In this paper, the suitable m and λ_1, have been already selected and the effect of λ_2 on the system performance needs to be explored. Moreover, the parameter λ_2 is used to control the flexibility of the proposed model and particularly it can be varied from 0 to 0.05 and the report of the results is described in Table 5 and Table 6 for training and testing datasets respectively. Usually, the lower λ_2 might allow for the more flexible model, therefore it might handle the bias in data and make the model to achieve the good result in training data. In return, the risk of overfitting might increase due to the noise in training data and the model may result in reducing performance with testing data. However, in the experiments, the model works well with the low λ_2 (0.001 and 0.005) for both training and testing datasets as shown in Table 5 and 6. In this case, the overfitting problem does not seem to be happening, because the PCA is applied to compress the data into a minimum size so that the noise and redundancy can be dropped before training. With the higher λ_2 the regularization term comes in to shrink the learned weights towards zero, and hence it discourages learning a more complex model to avoid the risk of overfitting. However, if the λ_2 parameter is too high, for example, λ_2 = 0.05, the model seems to try to be as simple as possible. Therefore, it leads to ignore the classes occupying small amounts in training datasets such as the classes S and V, as shown in the final column of Table VI. Supported by the results in Table 6, λ_2 = 0.001 is chosen in this experiment because it not only provides a high accuracy but also has the ability to adapt with the bias problem.

4.3 Performance Comparison

Based on the selected hyper-parameters, the proposed method is compared with state-of-art methods within three following approaches: hand-craft feature based methods [19, 20]; feature selection-based methods [23, 26]; and deep learning-based method [30]. Finally, the results are presented in Table 7.

Table 7: Comparison results – unit is %

Method		**Hand-craft**		**Feature Selection**			**Deep Learning**		
		[20]	[19]	[23]	PSO on VCG [26]	PSO [26]	DRBM [30] (training)	DRBM [30] (testing)	Proposed method (testing)
ACC		90.8	91.2	93	78	92.4	98.85	70.91	**96.82**
Class N	Se	91.6	95	95	79.1	94	98.74	93.36	**99.06**
	Pp	**99.3**	96.5	98	96.3	98	98.11	72.14	97.57
	FPR	-	27.9	-	27	**17.4**	0.97	80.02	19.07
Class S	Se	**81.4**	29.6	77	31.2	62	98.14	3.47	64.34
	Pp	31.6	26.4	39	8.4	53	99.18	11.89	**82.25**
	FPR	-	3.1	-	13	2.1	0.4	4.45	**0.69**
Class V	Se	86.2	85.1	81	89.5	87.3	99.66	30.79	**90.09**
	Pp	73.7	66.3	87	46.1	59.4	99.28	87.66	**94.62**
	FPR	-	3.0	-	7.2	4.1	0.37	1.22	**0.37**

Supported by the results in Table 7, it means that the simply applying deep learning does not always get the good result, especially in the inter-patient case as discussed in this paper. In particular, the Restricted Boltzmann Machine (DRBM) [30] can learn the good features that perform well on training dataset, where the accuracy is up to 98.85%. However, when applied to testing data, they do not perform well, the ACC drops dramatically to 70.91% due to the inter-patient problem. In comparison, the accuracy is 93% given by the feature selection-based method [23] or 91.2 % given by the hand-craft feature-based method in [19], the conventional deep learning framework's result may not be a promising result due to the lack of training data. This is a big weakness of deep learning based methods, if the training data is not large and generalized enough, the trained model is also not generalized. Therefore, in order to apply deep learning for ECG classification, the framework should be designed carefully, this is one of the major contributions in this paper.

Unlike the DRBM method that relies only on data to approximate a classification model, the proposed framework is designed to handle the practical challenges for ECG classification such as inter-patient problem, bias and lack of training data issues. Therefore, we achieve not only the higher ACC but also better results in each particular class. In overall, the accuracy of 96.82% is achieved on the inter-patient dataset as shown in Table 7. In addition to the inter-patient problem, the dataset has serious bias problem,

where the class N covers 91.5% of training data while the classes S and V only occupy 1.5% and 7%, respectively. Even so, in the class V, the proposed method has the smallest FPR and the highest Se and Pp metrics compared to others. With the most challenging case, the class S in this method could also achieve 64.34% on sensitivity and 82.25% on the positive rate, while the best reports of the previous works are just 81% in [20] and 53% in [26], correspondingly.

4.4 Feature Visualization

In order to inspect the learned features for the proposed network, t-SNE [40] is used for feature visualization and the learned features are analysed following three modes: input features, learned features without contrastive loss LF, and learned features with contrastive loss LF; with two aspects: the centrality of features extracted from the same class, and the separation of features extracted from different classes. Due to the imbalance in the data, to easily see the difference, two challenging cases of the classes S and V are investigated.

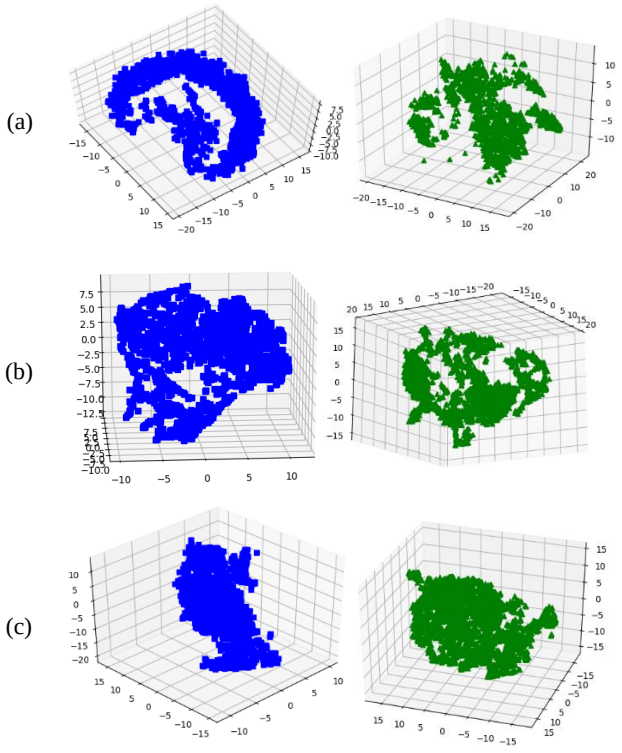

Fig. 5. The t-SNE visualizations of features extracted from the same class. (Left) Class S. (Right) Class V. (a) Input features, (b) Learned features without L_F, (c) Learned features with L_F.

As shown in Fig. 5(a), without learning procedure, the input features are quite dispersed even they are extracted from the same class due to the inter-patient problem. This phenomenon explains why the previous methods failed to achieve the good results with these two classes. By simply applying the traditional NN-based classification, the learned features seem to have only a little improvement as shown in Fig. 5(b). The feature distribution still shows the complex feature variations, therefore the complexity of learned classification model needs to be increased and it also means that its generality will be reduced. This results in the model performing well in training data, but it is not in the testing data. With the help of contrastive loss LF, the compact features can be applied to learn as shown in Fig. 5(c), thereby it may reduce the workload of the classification model, make it more general and boost the performance in both training and testing datasets.

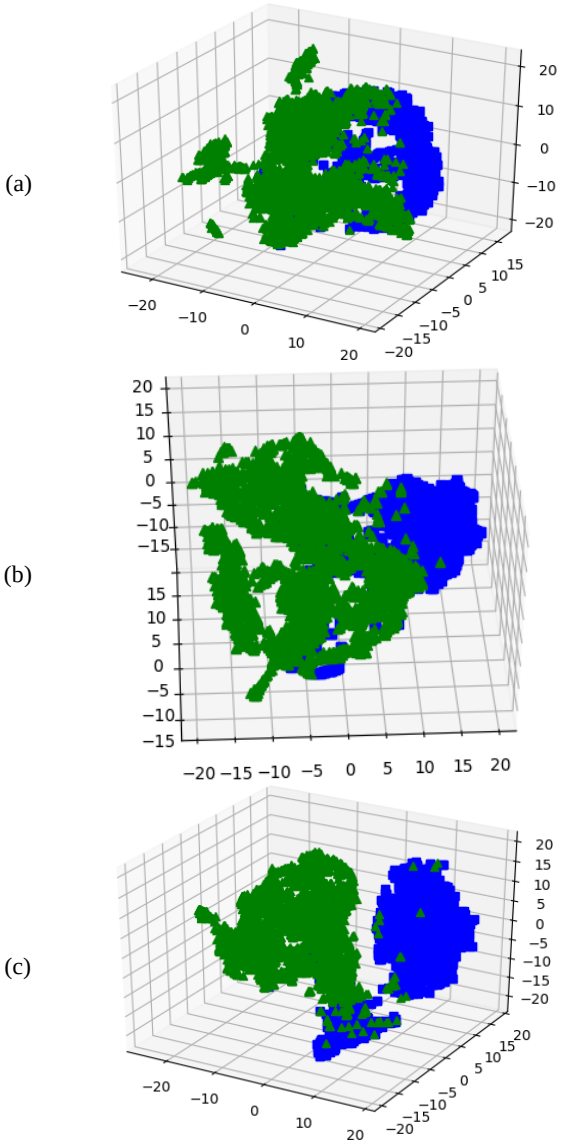

Fig. 6. The t-SNE visualizations of features extracted from the different classes. (Blue) Class S. (Green) Class V. (a) Input features, (b) learned features without L_F, (c) learned features with L_F..

To evaluate the separation of features extracted from different classes, the extracted features from the classes S and V are visualized as shown in Fig. 6. Without training process or with the traditional NN training, the extracted features from the classes S and V tend to be tied together with the complex variations. This makes the classification even more difficult. By proposing the joint objective function covering contrastive loss and classification loss, the proposed deep network could learn discriminative features as shown in Fig. 6(c). In particular, ignoring a small amount of confused samples, the learned features of the same class certainly form a cluster and the learned features of different classes are far apart. This explains why the performance in this experiment is significantly improved as compared to the previous works. Besides, it proves that the proposed framework can allow to learn the patient-invariant features for ECG classification.

4.5 The Contribution of DWT and PCA

In this section, we evaluated the effect of DWT and PCA on the ECG classification system by various experiments. In the first experiment, we use directly the original ECG signals as the inputs of our network. For the second experiment, we apply Fast Fourier Transform (FFT) to convert the data to frequency domain instead of DWT; here we use a similar PCA for dimensional reduction to evaluate the effect of frequency transform methods. Third, we use our proposed method without using PCA to evaluate the effect of dimensional reduction process. The training and testing set are similar to the setting in the Table I. Among many hyper-parameter options, the highest accuracy is reported in Table 8.

Table 8: The performance among various features

METHOD		**Original ECG signal**	**FFT with PCA**	**DWT without PCA**	**DWT with PCA**
ACC		85.99	87.8	**97.18**	96.82
CLASS N	Se	85.5	88.88	**99.37**	99.06
	Pp	97.33	93.41	**97.55**	97.57
	FPR	1.19	3.41	**16.7**	19.07
CLASS S	Se	91.51	94.75	**67.49**	64.34
	Pp	89.35	86.62	**90.8**	82.25
	FPR	15.8	21.2	**0.28**	0.69
CLASS V	Se	42.53	23.71	89.04	**90.09**
	Pp	30.96	52.57	95.54	94.62
	FPR	7.29	1.64	**0.22**	0.37

The results point out that with the original signals in time domain which are affected seriously by noise and patient motion, the overall accuracy is only 85.99% (reduced nearly 11% compared to our method). When FFT is applied to transform the original signals into frequency domain, the noises are removed so the performance is increased nearly 2%. However, by using FFT, we will lose all the temporal information, which is also important for ECG classification purpose, the improvement is still limited. With DWT, both temporal and frequency information are captured so the performance is significantly improved (increased around 9% compared to FFT). Without using PCA to reduce feature dimension, the performance is slightly better compared to PCA version. Because without feature reduction, the classifier has more information to learn and make the decision. However, more features mean that more parameters need to be learned and more time need to be used for training and testing. In detail, without using PCA the training time is 3824,16 seconds; with the help of PCA, the training time reduced significantly to 310.22 seconds.

4.6 The Computational Complexity and The Robustness Against Noise Setting

We implemented our framework in a PC equipped an i7-5820k CPU and 16GB RAM using Tensorflow library [41]. It takes around 0.0252 seconds to process 49,000 ECG samples in the testing set. It means that the processing time is very fast and our framework can be used for the real-time application.

To evaluate the robustness of the proposed technique against the noise and variations of the parameters of the proposed classifier, we evaluate the accuracy variation under various network settings. Except the input and output layers which have the fix number of nodes ($n_{feature}$ and n_{class}), we randomly change the number of nodes in the other four hidden layers. Denote $n_{i|i=1\sim4}$ is the number of nodes in i^{th} hidden layer, the selection for n_i will be randomly but need to satisfy the constraint $n_{feature} > n_1 > n_2 > n_3 > n_4 > n_{class}$. The Fig. 8 demonstrates the histogram of accuracy over 200 randomly settings. The results point out that in most cases, the accuracy is 96% which similar to our original setting. In overall, 75% of settings achieve the accuracy higher than 95%. It means that our proposed method is quite robust to the change of number of parameters.

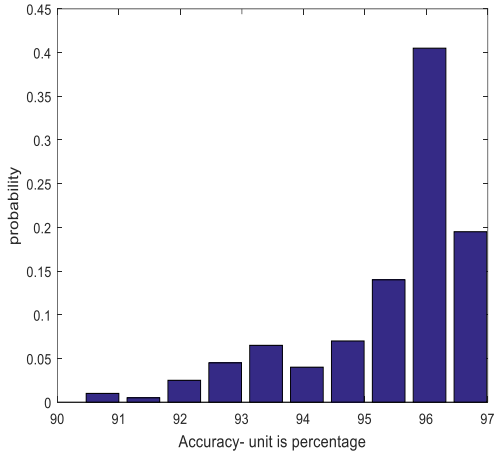

Fig. 7. Histogram of the system accuracy over various settings.

5. Conclusion

In this paper, the multi-task based deep learning framework was introduced for ECG heartbeat classification. In particular, the framework was designed to address the practical challenges in ECG heartbeat classification such as inter-patient ECG heartbeat, bias and lack of training data issues. In the proposed network, the Siamese architecture and contrastive loss were introduced to learn patient-invariant features. Therefore, the classification network was mounted on the top of the Siamese architecture to classify the ECG heartbeats. By considering both classification and clustering tasks in the end-to-end training process, the promising results are achieved while reducing the demand for the amount of training data. In addition, the feature visualization proves that the learned features from the proposed framework are not only patient-invariant but also useful for classifying diseases.

Acknowledgment

The authors would like to acknowledge the support of Ministry of Education and Training, Vietnam with Grand No. B2017.SPK.03 and HCMC University of Technology and Education, Vietnam.

References

[1] H. Atoui, J. Fayn, and P. Rubel, "A neural network approach for patient-specific 12-lead ECG synthesis in patient monitoring environments," in Computers in Cardiology, pp. 161-164, 2004.

[2] R. E. Gregg, S. H. Zhou, J. M. Lindauer, E. D. Helfenbein, and D. Q. Feild, "Limitations on the Re-use of patient specific coefficients for 12-lead ECG reconstruction," in Computers in Cardiology, pp. 209-212, 2008.

[3] Ki Moo Lim, Jae Won Jeon, Min-Soo Gyeong, Seung Bae Hong, Byung-Hoon Ko, Samsung Electronics, South Korea Yongin, Sang-Kon Bae, et al., "Patient-Specific Identification of Optimal Ubiquitous Electrocardiogram (U-ECG) Placement Using a Three-Dimensional Model of Cardiac Electrophysiology," IEEE Transactions on Biomedical Engineering, Vol. 60, pp. 245-249, 2013.

[4] Serkan Kiranyaz, Turker Ince, and Moncef Gabbouj, "Personalized Monitoring and Advance Warning System for Cardiac Arrhythmias," Scientific Reports, Vol. 7, pp. 9270–9278, 2017.

[5] Mariano Llamedo and Juan Pablo Martinez, "An Automatic Patient-Adapted ECG Heartbeat Classifier Allowing Expert Assistance," IEEE Transactions on Biomedical Engineering, Vol. 59, pp. 2312-2320, 2012.

[6] Serkan Kiranyaz, Turker Ince, Ridha Hamila, and Moncef Gabbouj, "Convolutional Neural Networks for patient-specific ECG classification," in 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2608-2611, 2015.

[7] Serkan Kiranyaz, Turker Ince, and Moncef Gabbouj, "Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks," IEEE Transactions on Biomedical Engineering, Vol. 63, pp. 664-675, 2016.

[8] Udita Dev Roy, Santanu Ghorai, and Anirban Mukherjee, "Kernel-based feature extraction for patient-adaptive ECG beat classification," in International Conference on Systems in Medicine and Biology (ICSMB), pp. 144-147, 2016.

[9] Yann LeCun, "Learning invariant feature hierarchies," in Computer Vision - ECCV 2012. Workshops and Demonstrations, pp. 496–505, 2012.

[10] Cheng Wen, Teng-Chiao Lin, Kuang-Chiung Chang, and Chih-Hung Huang, "Classification of ECG complexes using self-organizing CMAC," Measurement, Vol. 42, pp. 399-407, 2009.

[11] Marina Ronzhina, Tomas Potocnak, Oto Janousek, Jana Kolarova, Marie Novakova, and Ivo Provaznik, "Spectral and higher-order statistical analysis of the ECG: Application to the study of ischemia in rabbit isolated hearts," in Computing in Cardiology, pp. 645-648, 2012.

[12] Omid Sayadi, Mohammad B. Shamsollahi, and Gari D. Clifford, "Robust Detection of Premature Ventricular Contractions Using a Wave-Based Bayesian Framework," IEEE Transactions on Biomedical Engineering, Vol. 57, pp. 353-362, 2010.

[13] Yakoub Bazi, Naif Alajlan, Haikel AlHichri, and Salim Malek, "Domain adaptation methods for ECG classification," in International Conference on Computer Medical Applications (ICCMA), pp. 1-4, 2013.

[14] Venkatachalam Mahesh, Arumugam Kandaswamy, Chandran Vimal, and Balakrishnan Sathish, "ECG arrhythmia classification based on logistic model tree," Journal of Biomedical Science and Engineering, Vol. 2, pp. 405-411, 2009.

[15] Saeed Karimifard and Alireza Ahmadian, "A robust method for diagnosis of morphological arrhythmias based on Hermitian model of higher-order statistics," Biomedical engineering online, Vol. 10, pp. 10-22, 2011.

[16] Samit Ari, Manab Kumar Das, and Anil Chacko, "ECG signal enhancement using S-Transform," Computers in Biology and Medicine, Vol. 43, pp. 649-660, 2013.

[17] Minhas FU and Arif M, "Robust electrocardiogram (ECG) beat classification using discrete wavelet transform," Physiological Measurement, Vol. 29, pp. 555-70, May 2008.

[18] Hamid Khorrami and Majid Moavenian, "A comparative study of DWT, CWT and DCT transformations in ECG arrhythmias classification," Expert Systems with Applications, Vol. 37, pp. 5751-5757, 2010.

[19] Gabriel Garcia, Gladston Moreira, Eduardo Luz, and David Menotti, "Improving automatic cardiac arrhythmia classification: Joining temporal-VCG, complex networks and SVM classifier," in International Joint Conference on Neural Networks (IJCNN), pp. 3896-3900, 2016.

[20] Chun-Cheng Lin and Chun-Min Yang, "Heartbeat Classification Using Normalized RR Intervals and Morphological Features," Mathematical Problems in Engineering, Vol. 2014, p. 11, 2014.

[21] R. Acharya, A. Kumar, P. S. Bhat, C. M. Lim, S. S. lyengar, N. Kannathal, and S. M. Krishnan, "Classification of cardiac abnormalities using heart rate signals," Medical and Biological Engineering and Computing, Vol. 42, pp. 288–293, 2004.

[22] Manab Kumar Das and Samit Ari, "ECG Beats Classification Using Mixture of Features," International Scholarly Research Notices, Vol. 2014, pp. 1-12, 2014.
[23] R. Poli, S. Cagnoni, and G. Valli, "Genetic design of optimum linear and nonlinear QRS detectors," IEEE Transactions on Biomedical Engineering, Vol. 42, pp. 1137-1141, 1995.
[24] Llamedo M and Martinez JP, "Heartbeat classification using feature selection driven by database generalization criteria," IEEE Transactions on Biomedical Engineering, Vol. 58, pp. 616-25, 2011.
[25] Liam Cervante, Bing Xue, Mengjie Zhang, and Lin Shang, "Binary particle swarm optimisation for feature selection: A filter based approach," in IEEE Congress on Evolutionary Computation, pp. 1-8, 2012.
[26] Gabriel Garcia, Gladston Moreira, David Menotti, and Eduardo José da Silva Luz, "Inter-Patient ECG Heartbeat Classification with Temporal VCG Optimized by PSO," Scientific Reports, Vol. 7, pp. 10543–10554, 2017.
[27] Emina Alickovic and Abdulhamit Subasi, "Effect of Multiscale PCA De-noising in ECG Beat Classification for Diagnosis of Cardiovascular Diseases," Circuits, Systems, and Signal Processing, Vol. 34, pp. 513-533, 2015.
[28] Jeen-Shing Wang, Wei-Chun Chiang, Ya-Ting C. Yang, and Yu-Liang Hsu, "An Effective ECG Arrhythmia Classification Algorithm," in Bio-Inspired Computing and Applications, Berlin, Heidelberg, pp. 545-550, 2012.
[29] Mohammad Sarfraz, Ateeq Ahmed Khan, and Francis F. Li, "Using independent component analysis to obtain feature space for reliable ECG Arrhythmia classification," in IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 62-67, 2014.
[30] Yan Yan, Xingbin Qin, and Lei Wang, "ECG Annotation and Diagnosis Classification Techniques," in Health Informatics Data Analysis: Methods and Examples, Cham, pp. 129-154, 2017.
[31] Hinton GE and Salakhutdinov RR, "Reducing the dimensionality of data with neural networks," Science, Vol. 313, pp. 504-511, 2006.
[32] Asja Fischer and Christian Igel, "An Introduction to Restricted Boltzmann Machines," in Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Berlin, Heidelberg, pp. 14-36, 2012.
[33] Chisung BAE, Jin Woo Shin, Sung-Soo Ahn, and Sang Joon Kim, "Electrocardiogram (ecg) signal based authentication apparatus and method," Samsung Electronics Co LtdKorea Advanced Institute of Science and Technology (KAIST), pp. 1-19, 2016.
[34] Patane A and Kwiatkowska MZ, "Calibrating the classifier: siamese neural network architecture for end-to-end arousal recognition from ECG," presented at the AffecTech - Personal Technologies For Affective Health, 2018.
[35] Roshan Joy Martisa, U. Rajendra Acharyaab, and Lim Choo Min, "ECG beat classification using PCA, LDA, ICA and Discrete Wavelet Transform," Biomedical Signal Processing and Control, Vol. 8, pp. 437-448, 2013.
[36] Philip de Chazal, M. O'Dwyer, and R.B. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," IEEE Transactions on Biomedical Engineering, Vol. 51, pp. 1196-1206, 2004.
[37] Giovanni Bortolan, Christian Brohet, and Sergio Fusaro, "Possibilities of using neural networks for ECG classification," Journal of Electrocardiology, Vol. 29, pp. 10-16, 1996.
[38] Michael A. Nielsen, "Neural Networks and Deep Learning," Determination Press, 2015.
[39] G.B. Moody and R.G. Mark, "The impact of the mit-bih arrhythmia database," IEEE Engineering in Medicine and Biology Magazine Vol. 20, pp. 45–50, 1996.
[40] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," Journal of Machine Learning Research, Vol. 9, pp. 2579–2605, 2008.
[41] Aurélien Géron, Hands-On Machine Learning with Scikit-Learn and TensorFlow: O'Reilly Media, 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2017.

Manh-Hung Nguyen received the B.S., M.S. degree in electrical engineering from National University of Technology and Education, Ho Chi Minh City, Vietnam, in 2009, 2011, respectively; and the Ph.D. degree in electrical engineering from National Kaohsiung University of Applied Sciences, Taiwan, in 2016. His research interests are in machine learning and data analysis. (E-mail: hungnm@ hcmute.edu.vn)

Vu-Hoang Tran received the B.S. degree in Electrical Engineering from Ho Chi Minh City University of Technology and Education, Vietnam in 2012, the Master degree from National Kaohsiung University of Applied Sciences, Taiwan, in 2015, and the Ph.D. degree from National Chung Cheng University, Taiwan, in 2018. His research interests are in image processing, pattern recognition, computer vision, machine learning, and transfer learning. (E-mail: hoangtv@hcmute.edu.vn)

Thanh-Hai Nguyen received his M.Eng. in Electrical - Electronics Engineering from HCM City University of Technology in 2002 and his PhD in Electronics – Telecommunication Engineering from university of Technology, Sydney, Australia in 2010. His research interests include smart wheelchair, artificial intelligent, Bio-signal and image processing, healthcare devices. (E-mail: nthai@hcmute.edu.vn)

Thanh-Nghia Nguyen received the BE. and MEng. degrees in Electrical-Electronics Engineering from the HCMC University of Technology and Education, Viet Nam, in 2007 and 2012, respectively. Currently, he is a Ph.D. student with Electronics Engineering at the HCMC University of Technology and Education. His research interests include biomedical signal processing, biomedical image processing and artifact intelligent. (E-mail: nghiant@hcmute.edu.vn)