DEEP NEURAL NETWORKS VERSUS SUPPORT VECTOR MACHINES FOR ECG ARRHYTHMIA CLASSIFICATION

Sean shensheng Xu, Man-Wai Mak and Chi-Chung Cheung

Department of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hong Kong SAR of China

ABSTRACT

Heart arrhythmia is a heart disease that threatens the health of many people. As Electrocardiography (ECG) is an efficient measurement of heart arrhythmia, lots of research efforts have been spent on the identification of heart arrhythmia by classifying ECG signals for health care. Among them, support vector machines (SVMs) and artificial neural networks (ANNs) are the most popular. However, most of the previous studies reported the performance of either the SVMs or the ANNs without in-depth comparisons between these two methods. Also, a large number of features can be extracted from ECG signals, and some may be more relevant to heart arrhythmia than the others. This paper is to enhance the performance of heart arrhythmia classification by selecting relevant features from ECG signals, applying dimension reduction on the feature vectors, and applying deep neural networks (DNNs) for classification. A holistic comparison among DNNs, SVMs, and ANNs will be provided. Experimental results suggest that DNNs outperform both SVMs and ANNs, provided that relevant features have been selected.

Index Terms— Heart arrhythmia classification; ECG, SVM, deep neural networks; Fisher discriminant ratio.

1. INTRODUCTION

Heart arrhythmia is a kind of heart diseases in which the patients suffer from irregular heartbeat. It is one of the main heart diseases that threaten the health of human, especially the elderly. Some types of heart arrhythmia such as atrial fibrillation, ventricular escape and ventricular fibrillation may even cause strokes and cardiac arrest. A recent report published by American Heart Association (AHA) suggests that more than four million Americans have recurrent arrhythmias [1].

Heart arrhythmia can be detected by using electrocardiography (ECG), which records the electrical activities of the heart of a patient for a period using two electrodes attached to the skin. Because ECG signals reflect the physiological conditions of the heart, they are commonly used by medical doctors to diagnose heart arrhythmia. Thus, being able to identify the dangerous types of heart arrhythmia from ECG signals is an important skill of medical professionals. However, interpreting the ECG waveforms manually is tedious and time-consuming. With the increasing use of personal portable devices to acquire ECG data, the amount of data will easily overload the medical professionals. As a result, the development of automatic techniques for identifying abnormal conditions from daily recorded ECG data is of fundamentally importance. Moreover, timely first-aid procedures can be applied if such abnormal conditions can be detected automatically by health monitoring equipment. In this regard, machine learning will play an important role [2].

Support vector machines (SVMs) and artificial neural networks (ANNs)[1] are the two commonly used classifiers for identifying heart arrhythmia [3–5]. For example, Kohli *et al.* [6] used a one-versus-rest SVM as the classifier to predict heart arrhythmia and achieved good performance on the UCI benchmark dataset [7] (the best classification accuracy on their test data is over 70%). In [8], Khare *et al.* proposed a hybrid approach combining rank correlation [9] and principal component analysis (PCA) [10] for feature extraction and SVMs for classification. They demonstrated that the hybrid approach achieves much better performance than the predictor proposed by Kohli *et al.* [6] on the same dataset. However, the hyper-parameters of the heart arrhythmia classifiers in these works were optimized based on the test data. Thus, the claimed accuracy in these studies may be over-estimated. In [11], ANNs were applied to the same arrhythmia dataset. The authors showed that the best performance of the ANNs is close to that of the SVMs. Unfortunately, they did not specify the network structures and parameter settings in their paper, causing difficulty in comparing the capability of ANNs and SVMs in predicting heart arrhythmia.

Previous work either optimized the hyper-parameters of the feature extractors and classifiers using test data (e.g., [6, 8]) or provided a single random split of the benchmark dataset into a training set and a test set (e.g., [6, 8, 11]). These experimental settings make comparison of methods difficult. In this paper, we perform 10-fold cross validations on the

This work was in part supported by Hong Kong Innovation and Technology Commission, Grant No. UIT/133.

[1] In this paper, ANNs refer to the feedforward networks with a shallow architecture and DNNs refer to the networks with a deep structure.

dataset and repeat the cross-validation runs a number of times, each with a different random split of the dataset. Therefore, unlike previous work, we report not only the classification accuracy but also its range in these repeated runs. We have also investigated feature pre-processing methods, including Fisher discriminant ratio (FDR) [12] and PCA, and various classification methods, including SVMs and deep neural networks (DNNs). More importantly, we attempt to investigate which feature pre-processing methods are appropriate for which classification methods. Performance evaluations on the UCI benchmark dataset suggests that feature selection together with DNNs achieve the best performance.

The paper is organized as follows. Section 2 briefly introduces the feature selection method called Fisher discriminant ratio (FDR) and describes how to use SVMs and DNNs for constructing classifiers to classify the selected features. Section 3 outlines the experimental protocol and compares the performance of DNNs against ANNs and SVMs based on the full feature set and the selected feature set. The name of the most important features will also be identified. Finally, Section 4 concludes our findings.

2. METHODOLOGY

2.1. Feature Pre-Processing

It is not uncommon for biological data to contain missing values and heart arrhythmia data derived from ECG signals are of no exception. For example, in the UCI benchmark dataset, there are 408 missing entries, which account for about 0.33% of the total number of entries. In this work, we filled these missing entries with the average value of the corresponding features. Another characteristic of heart arrhythmia data is the high dimensionality of the feature vectors. For example, in the UCI dataset, the dimension is 279 but the number of feature vectors is only 452. To address this problem, we used Fisher discriminant ratio (FDR) [12] to select relevant features and PCA to reduce the dimension of feature vectors.

FDR is a simple and effective measure of features for classification problems. For the two-class problem, FDR of the j-th feature is defined as:

$$\mathrm{FDR}(j) = \frac{\left[\mu_j^{(1)} - \mu_j^{(2)}\right]^2}{\left[\sigma_j^{(1)}\right]^2 + \left[\sigma_j^{(2)}\right]^2}, \tag{1}$$

where $\mu_j^{(1)}$, $\mu_j^{(2)}$, $\sigma_j^{(1)}$ and $\sigma_j^{(2)}$ represent the class-conditional means and standard derivations of the j-th feature, respectively. In Eq. 1, the superscript represents the class labels. For multi-class problems, we may estimate the average FDR values across all class pairs.

A high FDR implies that the corresponding feature produces large separation between different classes. Therefore, its classification capability is stronger, and it should be selected for classification. In practice, the FDR of individual features can be computed independently and ranked in descending order. We retained the features with FDR scores larger than a predefined threshold (0.001 in this work). FDR can remove all insignificant features from the data set. Performance evaluations show that dropping some irrelevant features by FDR helps the training of SVMs and boost the performance of DNNs.

2.2. Classification

We investigated two popular classifiers (SVMs and DNNs) for the classification of heart arrhythmia. We also investigated which of the feature pre-processing approaches is the best for these two types of classifiers.

To apply SVMs for K-class classification, we constructed K one-versus-rest RBF-SVM [10, 13], one for each class. Specifically, the k-th SVM is trained to discriminate between the feature vectors of the k-th class and those of the other classes. During recognition, given an unknown vector $\mathbf{x}$, its class label is predicted according to the maximal output:

$$l(\mathbf{x}) = \underset{k\in\{1,\ldots,K\}}{\arg\max}\ h^k(\mathbf{x}), \tag{2}$$

where

$$h^k(\mathbf{x}) = \sum_{i\in\mathcal{S}_k} \alpha_i^k y_i^k K(\mathbf{x},\mathbf{x}_i) + b^k \tag{3}$$

is the output of the k-th SVM. In Eq. 3, $\mathcal{S}_k$ is the set of support vector indexes corresponding to the k-th SVM, $y_i^k \in \{-1,+1\}$ are the target output of the k-th SVM, α_i^k's are the Lagrange multipliers, b^k's are bias terms, and $K(\cdot,\cdot)$ is a kernel function. In this work, the radial basis function (RBF) kernel was used.

To apply DNNs for K-class classification, we trained a DNN with several hidden layers comprising sigmoid nonlinearity and a soft-max output layer comprising K outputs nodes. We applied the greedy layer-wise training [14] to pre-train the hidden layers which are formed by stacking a number of restricted Boltzmann machines (RBMs) [15,16]. Then, we fine-tuned the whole network (including the soft-max output layer) with backpropagation. The pre-training step is very important for arrhythmia classification because the number of training vectors is typically small for this task. The architecture of the DNN with stacked RBMs is shown in Figure 1.

3. EXPERIMENTS AND RESULTS

3.1. Data Set and Evaluation Protocol

The UCI cardiac arrhythmia [17] was used in the experiments. In the dataset, one of the classes is named “Normal”. It contains 245 samples. The remaining fifteen classes represent different kinds of heart arrhythmia. Since the numbers of samples in these classes are highly imbalance, we combined them into one class, called “Abnormal”. Thus, the “Abnormal” class contains 206 samples.

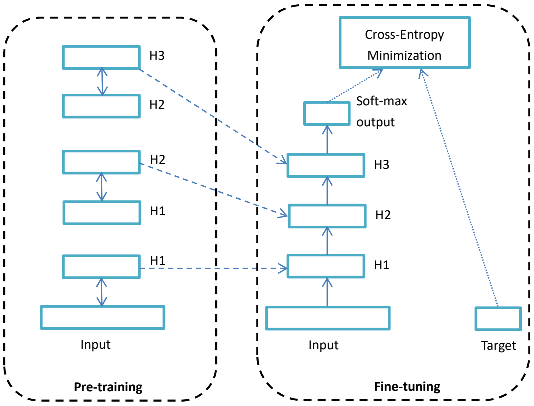

Fig. 1: DNN with stacked RBMs

To rigorously estimate the accuracy of different classifiers, 10-fold cross validation was performed. For each configuration of feature pre-processing and classification, the corresponding 10-fold cross-validation was repeated 10 times, each with a random reshuffling of the samples in the dataset. Then, the average accuracy and the range of accuracy were obtained from the results of the 10 repetitions. The DNNs program is based on G. E. Hinton's Matlab code [18].

3.2. Selected Features

Table 1 shows the top-10 features selected by FDR, i.e., features with the top-10 FDR scores. As can be seen from the table, these features can be divided into five types, each having features obtained from different channels. The five types include QRSTA, QRS duration, Amplitude of T, Average width of R, and the number of intrinsic deflections. These are the features that were found important by medical professionals [19]. Therefore, our feature selection method agrees well with the diagnostic criteria of medical doctors.

3.3. Performance of SVM Classifiers

For the RBF-SVMs, the hyper-parameters (RBF width and penalty factor) were further optimized based on the training data in each fold. Specifically, for each fold of the 10-fold cross-validation, we applied an inner 5-fold cross validation on the training split to optimize the hyper-parameter of the RBF-SVMs. The optimal RBF-SVMs were then tested on the remaining data in the test split. In other words, we further partition the training split of each fold into 5 portions in the inner 5-fold cross validation. We set the base 2 logarithm of the RBF width and penalty factor to values from -16 to 16 for the inner cross validation.

Table 2 shows the performance of the SVM classifiers with different feature pre-processing methods. For FDR, the cut-off threshold for feature selection is 0.001, which results in 236 selected features. For PCA, we kept 95% of the variance after projection, which results in 89-dimensional projected vectors. The results show that FDR is the best pre-processing method for SVMs and PCA degrades the performance. This is reasonable because SVMs are known to be able to handle high dimensional data and PCA will inevitably remove some useful information features. On the other hand, feature selection is able to keep the relevant features.

Table 2: The average accuracy (across ten 10-fold cross-validations) of SVM classifiers with different feature pre-processing methods

Feature Pre-Processing	Feature Dimension	Classification Acc. (average)
Nil (All features)	279	77.77%
FDR	236	78.23%
PCA	89	76.97%

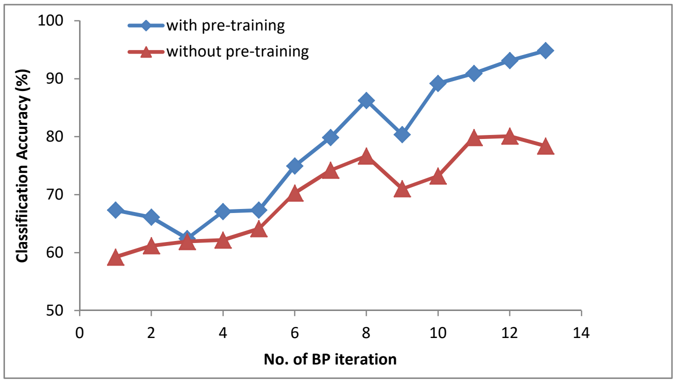

Fig. 2: Classification accuracy of the DNN with or without pre-training. The results were based on one of the 10 runs of the 10-fold cross-validation.

3.4. Performance of DNN Classifiers

Figure 2 show the effect of applying pre-training on a DNN with three hidden layers. For the network without pre-training, the backpropagation algorithm was applied to a DNN whose weights were initialized with small random values. On the other hand, 5 epochs of contrastive divergence (CD-1) [16] were applied to pre-train the network when pre-training was applied. The result clearly shows that pre-training can help the backpropagation to find a better solution.

Figure 3 shows the effect of increasing the number hidden nodes (in all hidden layers) on the classification accuracy. It shows that peak performance (80.64%) is achieved when the number of hidden nodes is 25, with the second best (80.04%) occurs at 20 nodes. Therefore, we used 25 hidden nodes per layer in the rest of our experiments.

To optimize the network structure, we fixed the number of hidden nodes per layer to 25 and varied numbers of hidden layers. According to Table 3, the performance becomes worse if the number of hidden layers is more than four because of the small number of training samples in this dataset.

Table 1: The top-10 features selected by FDR

Rank	Feature ID	FDR Score	Feature Information
1	199	0.237	QRSTA from channel AVR
2	5	0.230	Average QRS (msec.)
3	167	0.204	Amplitude of T wave measured in millivolts from channel DI
4	169	0.200	QRSTA from channel DI
5	197	0.183	Amplitude of T wave measured in millivolts from channel AVR
6	277	0.173	Amplitude of T wave measured in millivolts from channel V6
7	91	0.155	Average width of R wave in msec. from channel V1
8	279	0.139	QRSTA from channel V6
9	179	0.125	QRSTA from channel DII
10	93	0.122	Number of intrinsic deflections from channel V1

Table 3: Performance comparisons of DNN with different numbers of hidden layers

Feature Pre-Processing	Feature Dimension	Network Structure	Classification Accuracy (average)
Nil (All Features)	279	[25 25] [25 25 25] [25 25 25 25] [25 25 25 25 25]	79.00% 79.18% 79.29% 78.25%
FDR	236	[25 25] [25 25 25] [25 25 25 25] [25 25 25 25 25]	79.23% 80.64% 79.91% 79.54%
PCA	89	[25 25] [25 25 25] [25 25 25 25] [25 25 25 25 25]	74.89% 73.65% 73.50% 71.11%

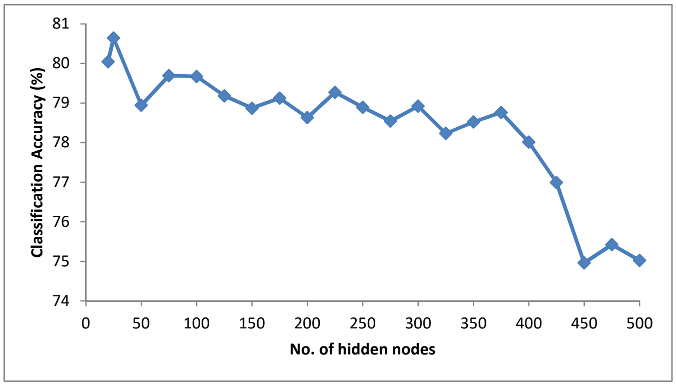

Fig. 3: The effect of increasing the hidden nodes on the DNN

Table 4 shows the performance of DNNs with different feature pre-processing methods. From the table, DNNs with FDR outperform DNNs with PCA and DNNs without any feature pre-processing (i.e., using the full features). The results also show that PCA does not work well with DNNs.

Figure 4 shows the range and rough distributions of the classification accuracies across the 10 runs of 10-fold cross-validation for different feature pre-processing methods combined with different classification methods. In the figure, the central mark inside each box indicates the median accuracy, and the bottom and top edges of each box indicate the 25th and 75th percentiles, respectively. The horizontal dashes represent the lowest and highest accuracies. The results in Figure 4 clearly show that FDR can improve the performance of DNN and SVM. However, PCA degrades their performance. Moreover, the performance of DNN is better than SVM, except when PCA is applied.

Table 4: The average accuracy (across ten 10-fold cross-validations) of DNN classifiers with different feature pre-processing methods

Feature Pre-Processing	Feature Dimension	Classification Acc. (average)
Nil (All features)	279	79.18%
FDR	236	80.64%
PCA	89	73.65%

A reason for the poor performance of PCA is that it is a linear transformation method that reduces the dimensionality of data while retaining most of the variance. Therefore, PCA is not suitable when the data lie on a nonlinear manifold of the feature space. Table 2, Table 4 and Figure 4 suggest that PCA is not an appropriate pre-processing method for this dataset, regardless of the classification methods used. Intuitively, when the data dimension is high and the amount of

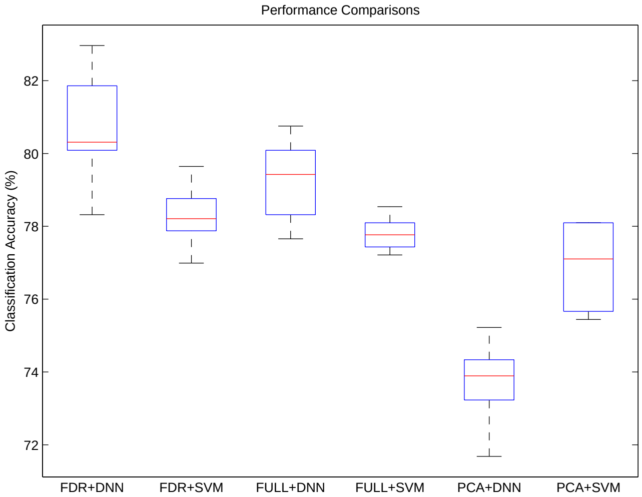

Fig. 4: The distribution of classification acc. of different algorithms

training data is small (the so-called small sample-size problem), PCA should be able to reduce the dimension so that the overfitting problem can be avoided. However, our results suggest that PCA is not necessary and that overfitting does not occur in our DNNs even for such a small dataset. This is mainly because we pre-trained [14, 15] our DNNs before applying backpropagation with early stopping (20 epoches). The pre-training step provides the necessary regularization to the networks [20] and the early stopping strategy avoids overfitting.

3.5. Comparing with Other Studies

Because there is no standard protocol for this dataset, different studies used different evaluation protocols, causing difficulty in comparing performance across studies. For examples, in [8], 30% of the data were used for training and the remaining 70% were used for testing, whereas in [11], various percentages of splitting were tried and the best result was obtained from the split where 90% of the data were used for training and the remaining 10% were used for testing. Also, these studies optimized the hyper-parameters (such as the number of hidden nodes and parameters of RBF kernels) of the classifiers based on the test set, which may give overoptimistic performance. Nevertheless, we attempt to compare our classifiers with [11] and [6] whose evaluation protocols are closest to ours.

Two-class Case: As [11] reported the best performance of its ANN, for fair comparisons, we compare its accuracy with the highest achievable accuracy of our DNNs. The results are shown in Table 5, which show that the performance of DNNs is comparable with that of the ANN in [11]. When relevant features have been selected, the DNN slightly outperforms the ANN in [11].

Table 5: The best accuracy (across ten 10-fold cross-validations) achieved by the DNN classifiers with different feature pre-processing methods

Feature Pre-Processing with ANNs/DNNs	**Feature Dimension**	**Classification Acc. (best)**
ANNs only [11]	279	82.22%
DNNs only	279	81.42%
FDR with DNNs	236	82.96%
PCA with DNNs	89	75.22%

Table 6: The class labels and number of samples in each class after the data preparation steps

Class ID	**Class Label**	**No. of Samples**
01	Normal	237
02	Ischemic changes	36
04	Old Inferior Myocardial Infarction	14
06	Sinus bradycardy	24
10	Right bundle branch block	48
16	Others	18

Multi-class Case: we have also compared the performance of our heart arrhythmia classifiers with those in [21] under the multi-class scenarios. We generally followed the evaluation protocol and data preparation procedures in [21] to make performance comparisons meaningful. Specifically, we followed [21] to remove the features whose values are all zeros across all samples and to remove the samples that contain missing values. After this data preparation steps, 377 samples remain. These samples are distributed into 6 classes as shown in Table 6. By dropping Classes 04, 06 and 16, which contain a small number of samples only, we reduce the 6-class problem to a 3-class one. Similar to [21], we selected half of the samples for training and remaining half for testing. However, unlike [21], we repeated the division of data 100 times, each with a different training and test sets, to obtain the average accuracy.

In [19], PCA was applied to reduce the dimension of feature vectors. In this work, we not only applied PCA to reduce dimension but also used FDR to select relevant features. Although FDR is originally designed for binary classification problems, it can be easily adopted to the multi-class scenarios by noting that each SVM in the one-versus-rest SVM classifier is a binary classifier. Therefore, for a K-class problem, there will be K sets of FDR-selected features, one set for each SVM. While this strategy works very well for one-versus-rest SVM classifiers, it is not applicable to the DNN classifiers. Therefore, we did not use DNNs for comparison.

Table 7 shows the performance of the classifiers in this paper and the best arrhythmia classifier in [21] under the 6-class and 3-class scenarios. Two conclusions can be drawn

Table 7: Performance of the best SVM classifier in [21] and the SVM classifiers in this paper.

Feature Pre-Processing	Feature Dimension	Classification Accuracy
Nil [21]	166	75.0%
Nil	245	77.77%
FDR	236	78.23%
PCA	80	76.97%

(a) 6-class Case

Feature Pre-Processing	Feature Dimension	Classification Accuracy
Nil [21]	166	78.13%
PCA [21]	70	83.71%
Nil	245	86.15%
FDR	236	86.26%
PCA	77	85.04%

(b) 3-class Case

from Table 7. First, FDR not only reduces the feature dimension but also helps the SVM classifier to achieve better performance. Second, our classifier outperforms the best classifier in [21].

4. CONCLUSION

In this paper, SVMs and DNNs were applied to classify heart arrhythmia. Results show that for classifying normal against abnormal heart arrhythmia, the best combination of feature pre-processing and classification is FDR and DNNs. For multi-class classification, FDR can be easily adopted to one-vs-rest SVMs. Results also show that pre-training DNNs is an essential step for training DNN classifiers, especially when the number of training samples is very limited.

5. REFERENCES

[1] American Heart Association, "Arrhythmia.," `http://www.heart.org/Arrhythmia`.

[2] S. Hijazi, A. Page, B. Kantarci, and T. Soyata, "Machine learning in cardiac health monitoring and decision support," *IEEE Computer Magazine*, vol. 49, no. 11, pp. 38–48, Nov. 2016.

[3] A. H. Khandoker, M. Palaniswami, and C. K. Karmakar, "Support vector machines for automated recognition of obstructive sleep apnea syndrome from ECG recordings," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 1, pp. 37–48, Nov. 2009.

[4] F. Melgani and Y. Bazi, "Classification of electrocardiogram signals with support vector machines and particle swarm optimization," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 5, pp. 667–677, Sep. 2008.

[5] S. Osowski and T. H. Linh, "ECG beat recognition using fuzzy hybrid neural network," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 11, pp. 1265–1271, Nov. 2001.

[6] N. Kohli, N. K. Verma, and A. Roy, "SVM based methods for arrhythmia classification in ECG," in *International Conference on Computer and Communication Technology (ICCCT)*, 2010, pp. 486–490.

[7] School of Information University of California, Irvine and Computer Science, "UCI machine learning repository," `http://archive.ics.uci.edu/ml`.

[8] S. Khare, A. Bhandari, S. Singh, and A. Arora, "ECG arrhythmia classification using Spearman rank correlation and support vector machine," in *Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS)*, 2011, pp. 591–598.

[9] W. Wayne, *Applied Nonparametric Statistics*, Boston: PWS-Kent, 2nd edition, 1990.

[10] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[11] S. M. Jadhav, S. L. Nalbalwar, and A. A. Ghatol, "ECG arrhythmia classification using modular neural network model," in *IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 2010, pp. 62–66.

[12] P. Pavlidis, J. Weston, J. Cai, and W.N. Grundy, "Gene functional classification from heterogeneous data," in *International Conference on Computational Biology*, 2001, pp. 62–66.

[13] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[14] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in Neural Information Processing Systems*, vol. 19, pp. 153, 2007.

[15] G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[16] G. E. Hinton, "A practical guide to training restricted boltzmann machines," 2010, UTML Tech Report, Univ. Toronto.

[17] UCI Machine Learning Repository, "Cardiac arrhythmia database," `http://archive.ics.uci.edu/ml/datasets/Arrhythmia`.

[18] G. E. Hinton, "Training a deep autoencoder or a classifier on MNIST digits," `http://www.cs.toronto.edu/~hinton/MatlabForSciencePaper.html`.

[19] M. Zijlmans, D. Flanagan, and J. Gotman, "Heart rate changes and ECG abnormalities during epileptic seizures: prevalence and definition of an objective clinical sign," *Epilepsia*, vol. 43, no. 8, pp. 847–854, Aug. 2002.

[20] D. Erhan, Y. Bengio, A. Courville, P. A. Manzagol, P. Vincent, , and S. Bengio, "Why does unsupervised pre-training help deep learning?," *The Journal of Machine Learning Research*, vol. 11, pp. 625–660, Feb. 2010.

[21] N. Kohli and N. K. Verma, "Arrhythmia classification using SVM with selected features," *International Journal of Engineering, Science and Technology*, vol. 3, no. 8, pp. 122–131, 2011.