International Journal of
Intelligent Systems and
Applications in Engineering
ISSN:2147-6799 www.atscience.org/IJISAE
Advanced Technology and Science
Original Research Paper

Feature Selection using FFS and PCA in Biomedical Data Classification with AdaBoost-SVM

Rahime Ceylan[1], Mucahid Barstugan*[2]

Accepted : 06/01/2018 Published: 30/03/2018

Abstract: Recently, there has been an increasing trend to propose computer aided diagnosis systems for biomedical pattern recognition. A computer aided diagnosis method, which aims higher classification accuracy, is developed to classify the biomedical dataset. This process includes two types of machine learning algorithms: feature selection and classification. In this method, firstly, features were extracted from biomedical dataset, then the extracted features were classified by hybrid AdaBoost-Support Vector Machines (SVM) classifier structure. For feature selection, Forward Feature Selection (FFS) and Principal Component Analysis (PCA) algorithms were used, and the performance of the feature selection algorithms was tested by AdaBoost-SVM classifier. Following it, advantages and disadvantages of these algorithms were evaluated. Wisconsin Breast Cancer (WBC), Pima Diabetes (PD), Heart (Statlog) biomedical datasets were taken from UCI database and Electrocardiogram (ECG) signals were taken from Physionet ECG Database, and were used to test the proposed hybrid structure. The used two hybrid structures and other studies in the literature were compared with our findings. The obtained results show that the proposed hybrid structure has high classification accuracy for biomedical data classification.

Keywords: *AdaBoost, Biomedical Data Classification, Classification Performance, Feature Selection, Hybrid Structure, Machine Learning*

1. Introduction

For the last fifty years, researchers in the field of Biomedical Engineering have tried to improve a Computer Aided Diagnosis which generally use artificial intelligence techniques for detecting biomedical problems. In some situations, a biomedical problem can be an identification of an illness according to examination results, but sometimes it can be a signification of a signal. In the literature, lots of classification techniques have been proposed for the solution of both problems. Some of these techniques are algorithms of Decision Trees, Boosting, Artificial Neural Networks, and SVM. There are also some implementations done by ensemble classifiers (Naive-Bayes classifiers, AdaBoost, Bagging, Rotational Forest) to increase the classification accuracy. The main idea is to find the weak classifier which has the highest performance in lots of weak classifiers and to increase the weights of these weak classifiers in the ensemble. Some studies and obtained results in the literature are presented briefly as follows:

Yuan and Ma [1] proposed an AdaBoost-Genetic Algorithm system. The proposed algorithm was tested on benchmark datasets. The best performance on Breast Cancer dataset was obtained as 97.39%, and of Heart (Statlog) dataset as 83.09%. Dhakateet al. [2] introduced an ensemble feature selection approach to find the best first search feature selection algorithm to reduce the noise in the dataset. AdaBoost, Boosting and Bagging algorithms were used as ensemble classifiers, and the results were compared. The best performance for Breast Cancer on AdaBoost was obtained as 74.47%. Yunlong and Feng [3] designed an AdaBoost-kNN structure. In their system, some statistical regularity was obtained by AdaBoost. Then, kNN algorithm was run on feature space. The classification accuracy of proposed system on Breast Cancer was 96.44%; as 78.52% on Pima Diabetes. Chen et al. [4] modified the traditional AdaBoost method for One-Class Support Vector Machines. They used a binary class dataset from UCI benchmark and tested the proposed algorithm. The maximum classification performance on Breast Cancer was obtained as 97.03%. Lahiri and Biswas [5] proposed a new AdaBoost algorithm. In the proposed method, several learners were trained by ANNs on subsets of original feature spaces. With the proposed method, the classification performance on Breast Cancer dataset was obtained as 97.1%, and as 87.4% on Heart (Statlog) dataset. Huaxiang and Jing [6] proposed a fuzzy-boosting system. The C4.5 algorithm was used as a base classifier, and the proposed system obtained better results than AdaBoost and Bagging algorithm. The classification performance of proposed fuzzy-boosting system was obtained as 96.75% on Breast Cancer and as 77.32% on Pima Diabetes. Chen and Zhang [7] proposed a multiple Classifiers Ensemble based on Feature Selection (FSCE) in order to improve the classification performance. The proposed method was tested on UCI benchmark dataset [8], and the results were compared with AdaBoost algorithm. The best classification accuracy on Breast Cancer dataset was obtained as 97.13% by proposed FSCE method. Ghavidel et al. [9] proposed a new ensemble classifier generation method which aims to create more diverse base classifiers while making them more accurate. In their approach, training data for base classifiers were built by taking a bootstrap sample of the original training set. The proposed method was tested on 15 different UCI dataset. The best performance on Breast Cancer was obtained as 96.79%. Ham et al. [10] proposed a Boosted-PCA algorithm for efficient classification of two class dataset. In their proposed method, each principal component was

[1,2] *Electrical and Electronics Eng., Selcuk University, 42002, Selcuklu, Konya, Turkey*
* *Corresponding Author: Email: mubarstugan@selcuk.edu.tr*

treated as a weak classifier in AdaBoost algorithm to constitute a strong classifier for binary classification problems. The proposed algorithm was applied to UCI dataset and the obtained results were examined. The best classification accuracy was obtained as 97.37% on Breast Cancer, and as 69.22% on Pima Diabetes. Shu and Wang [11] proposed a new AdaBoost-AC (accelerated) method for classification. The algorithm was used to acquire the weights of the weak classifiers. The classification performance of the Breast Cancer dataset was obtained as 75.36% by the proposed method.

AdaBoost is mostly used for image classification in the literature. Also, there are many studies on biomedical data classification. In our study, we developed a hybrid algorithm for biomedical data classification. Firstly, we extracted features from dataset, and then classified by proposed AdaBoost-SVM algorithm. Two different feature extraction algorithms were used, and performance results obtained from AdaBoost SVM were compared. The aim of this study is to assess the performance of AdaBoost-SVM based on feature selection algorithms on the classification of the discontinuous dataset such as Breast Cancer, Pima Diabetes and classification of continuous dataset like ECG.

The paper has four parts. The first part reviews studies in the literature which are relevant to feature selection and classification algorithms. Then the used feature extraction methods and classification methods were presented. In the third part, the experimental results were summarized and discussed. The paper concludes with the discussion of obtained results and suggestions for further research.

2. Materials and Methods

The proposed system was formed by a feature selection algorithm and a classifier system which is implemented by AdaBoost and Support Vector Machines. The datasets firstly were processed by feature selection algorithms: FFS and PCA, then the classification process was done by AdaBoost-SVM hybrid classifier structure.

The used datasets were presented in Table 1.

Table 1. The features of the used dataset

Dataset	*Number of instances (patterns)*	*Number of features*	*Number of classes*
Wisconsin Breast Cancer	683	9	2
Pima Diabetes	786	8	2
Heart (Statlog)	270	13	2
ECG	1731	200	3

2.1. Feature Extraction Methods

2.1.1. Forward Feature Selection

Forward Feature Selection (FFS) process starts with obtaining all feature subsets which have only one attribute. One component subsets ($\{X_1\}$, $\{X_2\}$, … , $\{X_M\}$) are determined by One Is Out Cross Validation. Subset number M is the size of the input dataset. The most effective feature $X_{(1)}$ is selected with the feature selection process [12].

2.1.2. Principal Component Analysis

Principal Component Analysis is a method used to define instances in the dataset and to express the similarities and differences of the dataset. PCA is a strong method to use for analyzing the dataset [13] because it is hard to find instances of a dataset with high dimensions and in the situations where schematic representation is not possible.

2.2. Classification Methods

2.2.1. Ensemble Classifier AdaBoost

AdaBoost is an ensemble classifier method which creates a strong classifier by combining weak classifiers. In each iteration, the algorithm calls a simple learning algorithm, which was named as a base learner, and creates the classifier. Then, a weight coefficient is appointed to the classifier. The last classification result is obtained by weighted voting which is related to weight coefficients of weak classifiers. If the weak learner error is low, its weight is high in the last voting. The weak learners estimate a little better than random guessing, so there is a big flexibility in the weak learner set design [14].

The algorithm is as follows [15]:

AdaBoost Algorithm

Input: Training Data $(x_1, y_1), \dots, (x_m, y_m)$;

$x_i \in X$, $y_i \in Y = \{-1,+1\}$

$D_1(m) = 1/m$ Initialize the weights

Step 1: **for l=1, 2,…,t**

Weak classifiers φ_k in base learner are trained by weights

Weak classifier error is obtained:

$$\arg\min \varepsilon_t = \sum_{i=1}^{m} D_l(i)\,[y_l \neq \varphi_l(x_i)] \quad \text{.....(1)}$$

if $\varepsilon_t > 1/2$; Stop the iteration; end

Base learner h_t error a_t is computed:

$$a_t = \frac{1}{2}\log((1-\varepsilon_t)/\varepsilon_t) \quad \text{.....(2)}$$

Step 2: Weights are updated:

$$D_{t+1}\frac{D_t(i)\exp(-a_t y_t h_t(x_i))}{Z_t} \quad \text{.....(3)}$$

Z_t is the normalization factor

end for

Step 3: Ensemble classifier weighted voting output is computed:

$$H(x) = sign\left(\sum_{t=1}^{T}(a_t h_t(x))\right) \quad \text{.....(4)}$$

2.2.2. Support Vector Machines

Support Vector Machines (SVM) is a learning method which has high performance on various implementations. SVM is based on two main ideas. The first idea is to map feature vectors to high dimension space by a non-linear method and to use linear classifiers in this new space. The second idea is to find a hyper-plane which splits dataset with a big margin. This plane splits the dataset well as much as possible among infinite numbers of planes. Lots of planes have a similar performance on training dataset, but generalization performance on new dataset can differ significantly [16].

3. Experimental Results

In this study, AdaBoost-SVM ensemble classifier was presented to classify biomedical dataset. The most effective features of dataset were chosen, and dimension reduction was made; then the dataset were classified by SVM-based AdaBoost classifier structure. Two feature extraction algorithms, Forward Feature Selection and Principal Component Analysis, were applied to biomedical data. 10-folds cross-validation method was realized on all datasets and experiments. In stage of feature extraction, the features which are selected more than threshold value in 10-folds cross-validation are formed as the new dataset. The optimum threshold value is taken as 5 experimentally. By using this threshold value, 6 features are selected from Breast Cancer

and Pima Diabetes datasets. 1st,2nd,4th, 5th, 6th and 8th features of Breast Cancer dataset are selected. 1st,2nd,3rd, 5th, 6th and 7th features of Pima Diabetes dataset are selected. 10 features are selected from Heart (Statlog) dataset after experiments. 1st,2nd,3rd, 5th, 7th, 8th, 9th, 10th, 12th and 13th features of Heart (Statlog) dataset are selected. For selecting features of ECG dataset, the threshold values are taken as 1,2,4,7 and the features are selected according to these threshold values. 17 features are selected when the threshold value is taken as 7; 33 features are selected when the threshold value is taken as 4; 43 features are when the threshold value is taken as 2, and 70 features are selected when the threshold value is taken as 1. The selected features when the threshold value is taken as 2 can be seen on the used ECG wave in Fig. 1.

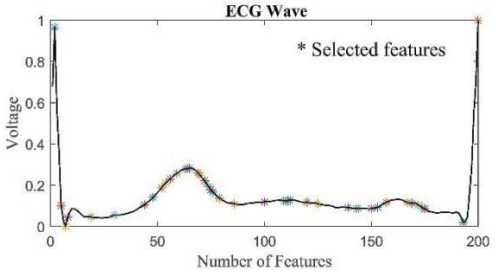

Fig. 1. Selected features in ECG signal (43 features were selected when threshold value is taken as 2)

The selected features were classified by AdaBoost-SVM structure. The performance of the classifier was evaluated by sensitivity, specificity and accuracy rates. There are some parameters in AdaBoost classifier algorithm such as base learner weight, coefficients, and weak learner error. In this study, the parameters for each of the features have been obtained during the training process. Therefore, during the test process, each of the features of test dataset was tested by its parameters which were obtained during the training process. The experiments were implemented on ASUS N550JK Intel (R) Core (TM) i7-4700HQ CPU @ 2.40 GHz notebook.

3.1. Testing Each of Features by Their Parameters

In this method, SVM-based AdaBoost classifier was trained by training set. SVM was used as weak learner in all base learners. During the training process, firstly, all features of the first pattern were classified by weak learners of the first base learner. All base learners have weak learners as much as the number of features. This process is implemented for all features. The 1st feature is classified by the 1st weak learner, and the 2nd feature is classified by the 2nd weak learner, and this pattern will continue. This process is repeated for all base learners. After the training process, weak learner errors and base learner weights are obtained, and the weak learner structures are held as a trained weak classifier to be used in test stage. In test stage, by using trained weak classifier, the pattern is classified. The classification result $h = \{-1, +1\}$ is multiplied with the weights of base learners (a). After classification process is completed for all base learners, a classification result called weight voting is obtained by Equation 4. The number of base learners and the number of weak classifiers were held the same. The training scheme of this method is presented in Fig. 2.

During the decision process, test dataset was given as input with parameters which were obtained after the training process. The weak learner structures are held as trained weak classifiers, and base learner weights are used to obtain the classification result. In all base learners, for the first pattern, classification results of their features are taken after classification with the weak learners. For *m* features, it is decided which class the features belongs to after classification. If the features of pattern are assigned to +1 class by more than the half of the feature number, the pattern was classified as +1, otherwise, the pattern is classified as -1. This process is repeated for all instances and base learners. After test stage, one classification result is obtained according to Equation 4. The scheme of test process was given in Fig. 3.

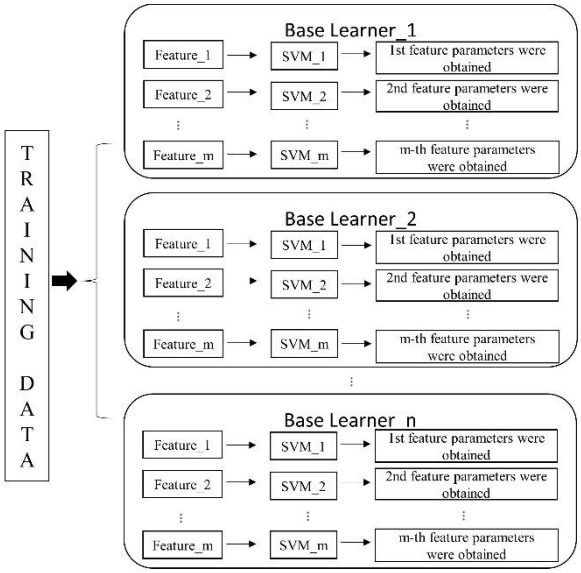

Fig. 2. Training process of classifier structure

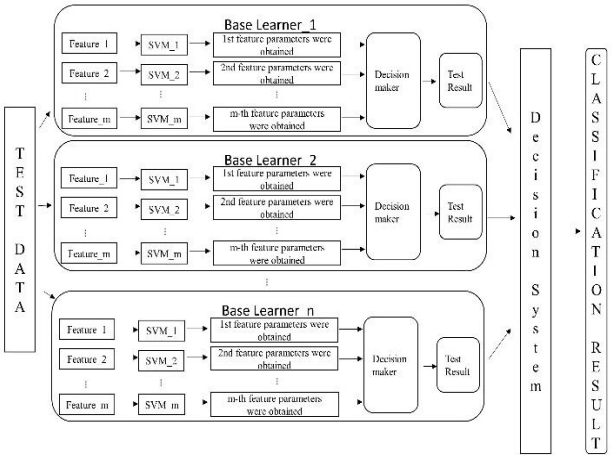

Fig. 3. Test process of classifier structure

3.2. Experiments on Discontinuous Dataset

In this study, two different data types, continuous time and discontinuous time, were used. Wisconsin Breast Cancer, Pima Diabetes and Heart (Statlog) datasets were taken from UCI database, and they were used as discontinuous data. These datasets were classified by FFS-AdaBoost-SVM and PCA-AdaBoost-SVM structures, and the performances of the structures were presented. According to experimental results, on FFS-AdaBoost-SVM structure, the optimum feature number was found as 4 for Breast Cancer, Pima Diabetes, and Heart (Statlog) datasets. The accuracy rates (ACC) on FFS-AdaBoost-SVM and PCA-AdaBoost-SVM structures for three datasets can be seen for different feature numbers of feature selection in Fig. 4.

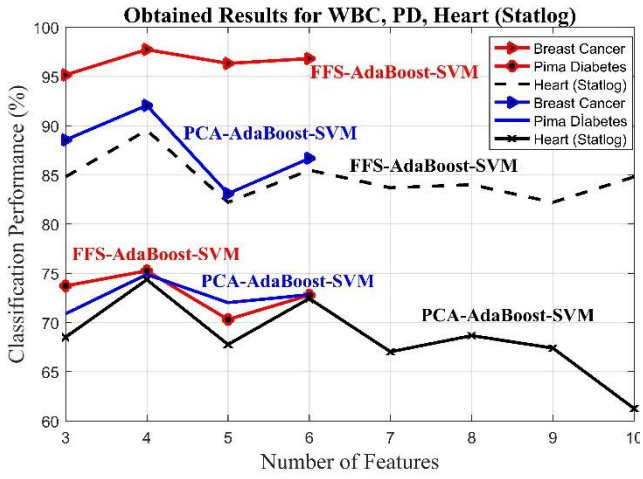

Fig. 4. Classification results of discontinuous dataset

It can be seen in Fig. 4 that the best performance was obtained in Breast Cancer dataset by 4 features, so the most effective 4 features of Breast Cancer dataset were selected by FFS algorithm, and these features were classified by AdaBoost-SVM structure. The accuracy rate was obtained as 97.74% for Breast Cancer dataset. As Fig. 4 shows, the best performance on Pima Diabetes data was obtained with 4 features using FFS algorithm, and the obtained features were classified by AdaBoost-SVM classifier. The optimum accuracy was obtained as 75.26% for Pima Diabetes data. FFS algorithm found the best performance with 4 features in Heart (Statlog) dataset, and these features were classified by AdaBoost-SVM classifier. The best accuracy was obtained as 89.54% for Heart (Statlog) dataset. Similarly, feature selection process was performed by PCA algorithm in Breast Cancer dataset, and the selected four features were classified by AdaBoost-SVM. The accuracy rate was found as 92.07% for Breast Cancer dataset. According to Fig. 4, the four features were selected on Pima Diabetes dataset by PCA algorithm, and the classification process was done by AdaBoost-SVM structure. The classification performance was obtained as 74.89% in Pima Diabetes dataset. As appropriate to the feature selection process in FFS algorithm, four features of Heart (Statlog) dataset were selected by PCA algorithm, and these 4 features were classified by AdaBoost-SVM. The optimum accuracy rate was obtained as 74.36% for Heart (Statlog) dataset. On both structures which were performed by FFS and PCA algorithms, the base learner number was held the same with the weak learner number. On Breast Cancer dataset, 5.7% higher performance was obtained in feature extraction process which was implemented by FFS according to PCA, and thus, the algorithm resulted in shorter time. On Pima Diabetes dataset, 0.4% higher performance was obtained by FFS algorithm according to PCA, but PCA algorithm resulted faster. On Heart (Statlog) dataset, FFS algorithm gave around 15% higher performance than PCA algorithm, but PCA algorithm resulted faster. The parameters of FFS-AdaBoost-SVM structure were used on PCA-AdaBoost-SVM structure. The number of base learners and weak classifiers for each dataset were held the same. The classification results on FFS-AdaBoost-SVM and PCA-AdaBoost-SVM for Wisconsin Breast Cancer, Pima Diabetes and Heart (Statlog) datasets were presented in Table 2.

As seen in Table 2(b), sensitivity (Sen) of Pima Diabetes dataset is inconsistent with specificity (Spe) and accuracy of the PCA-AdaBoost-SVM. Sensitivity shows the true positive (TP) performance of the classifier. The PCA-AdaBoost-SVM method could not able to classify positive values as positive. The reason could be that the principal components which represent the positive values were not defined well by the algorithm.

Table 2(a). The classification results on FFS-AdaBoost-SVM

Method	***FFS-AdaBoost-SVM***			
Dataset	***Sen***	***Spe***	***Acc***	***Time (sec)***
Breast Cancer	97.27	98.67	97.74	149.61
Pima Diabetes	80.02	72.38	75.26	410.2
Heart (Statlog)	80.18	95.6	89.54	87.6

Table 2(b). The classification results on PCA-AdaBoost-SVM

Method	**PCA-AdaBoost-SVM**			
Dataset	***Sen***	***Spe***	***Acc***	***Time (sec)***
Breast Cancer	98.91	69.48	92.07	304.2
Pima Diabetes	29.11	94.02	74.89	380.3
Heart (Statlog)	85.26	65.19	74.36	50.3

3.3. Experiments on Continuous Dataset

The Electrocardiogram (ECG) dataset was used as continuous time dataset in this experiment. ECG is a heart current graphic and a record of electrical activity in heart. The electrical signals in heart are being measured by surface electrodes and electronics devices. The obtained dataset is transformed to an ECG wave which has a characteristic pattern [17]. Three different ECG signal types (Right Bundle Branch Block, Left Bundle Branch Block and Normal Sinus Rhythm) were used. These datasets were taken from Physionet ECG Database [18]. Each of signals was sampled at 360Hz frequency with 11-bit resolution over a 10mV range. The presented classifier algorithms have binary structures, for that reason, multiple classification theories were used. In classification process, one- against-all (OAA) method was preferred. The method of OAA is: If the dataset is in $\Psi = \{\phi 1, \phi 2, \ldots, \}$ form, M is the number of different classes in a dataset. In OAA method, a binary classifier is processed to distinguish each class from the other classes in dataset. The aim of this binary classifier is to differ ($\Psi - \{\phi i\}$) from the other (ϕ_i, $i=1,2,...,M$) classes. Thereby, M numbers classifiers are being trained for every class [19].

First of all, dimension reduction process was done on ECG dataset by applying FFS and PCA feature extraction methods. The ECG dataset was divided to 5, 8, 10, 20, 25, 50 and feature selection process was done for all parts. Then, the selected features were classified by AdaBoost-SVM and the best results were obtained when the dataset was divided to 8. So, the dataset was divided to 8 and feature selection process was done on the divided parts. The scheme of the process was presented in Fig. 5.

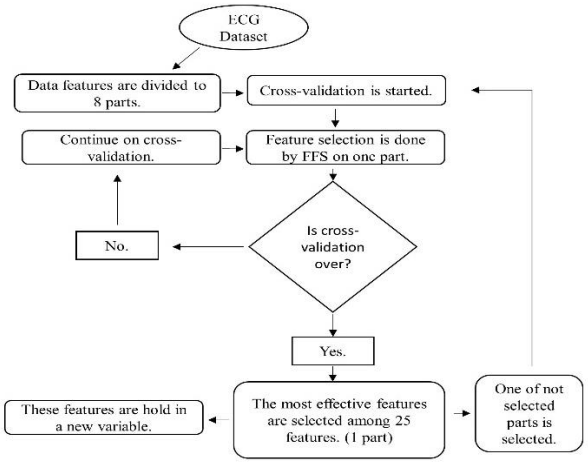

Fig. 5. Feature selection process on ECG

3.3.1. In taking number of the features and number of AdaBoost base learners as same

The new dataset was obtained after feature extraction (FFS or PCA) and it was presented as a new input to classifier structure. The number of base learners in classifier structure can be held the same with number of features, or it can be changed. The number of weak classifiers are also held the same with the number of features in each base learners. In Table 3, obtained results were presented while the number of features and the number of base learners were held the same.

For 17, 33, 43, 70 features, obtained results show that classifier structure with FFS has higher performance than the classifier structure with PCA. In all experiments, FFS has higher performance. PCA is better than FFS at only process time. The accuracy rates, due to different numbers of features on FFS-AdaBoost-SVM and PCA-AdaBoost-SVM structures, were presented in Fig. 6.

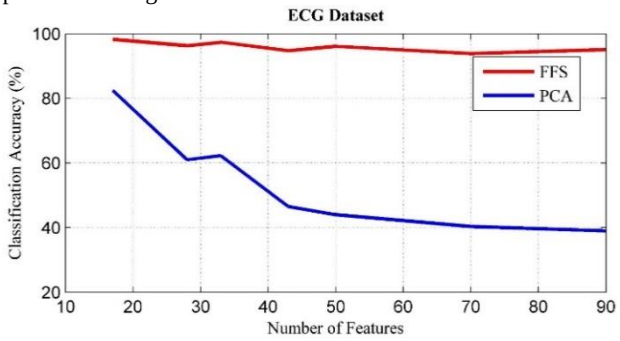

Fig. 6. FFS and PCA results according to numbers of features

Table 3(a). The obtained results for ECG dataset while the number of features and the number of the base learners were held same

Method	***PCA-AdaBoost-SVM***					
Number of Features	17			33		
Number of Base Learners	5	8	10	5	8	10
Sensitivity	76.32	80.28	71.42	75.76	71.64	74.63
Specificity	52.63	55.81	41.03	45.83	40.43	44.68
Accuracy	68.42	71.05	61.21	63.15	58.77	62.28
Time (sec)	110	165.1	201.4	173.5	270.2	359
Number of Features	43			70		
Number of Base Learners	5	8	10	5	8	10
Sensitivity	66.66	67.57	70.27	67.1	67.53	68.42
Specificity	33.33	35	40	34.21	35.14	36.84
Accuracy	54.39	56.14	59.65	56.14	57.02	57.89
Time (sec)	271.5	340.2	375.9	411.2	580.9	598.2

Table 3(b). The obtained results for ECG dataset while the number of features and the number of the base learners were held same

Method	***FFS-AdaBoost-SVM***			
Number of Feature & Base Learners	17	33	43	70
Sensitivity	98.68	98.67	96.05	96
Specificity	97.37	94.74	92.1	89.74
Accuracy	98.24	97.35	94.74	93.86
Time (sec)	222.25	290.49	339.59	679.09

Table 4(a). The obtained results for ECG dataset while the number of features and the number of base learners were held different

Method	***PCA-AdaBoost-SVM***			
Number of Feature & Base Learners	17	33	43	70
Sensitivity	91.17	74.63	61.54	56.45
Specificity	69.56	44.68	26.53	21.15
Accuracy	82.46	62.28	46.49	40.35
Time (sec)	120.79	390.3	540.79	1431.5

3.3.2. In taking number of the features and number of AdaBoost base learners as different

In this approach, the number of base learners were held different from the number of features while the number of weak learners in each base learners were held the same with the number of feature. The obtained results were presented in Table 4(a) and Table 4(b). And the graphical results for both structures were presented in Fig. 7.

Table 4(b). The obtained results for ECG dataset while the number of features and the number of base learners were held different

Method	***FFS-AdaBoost-SVM***					
Number of Features	17			33		
Number of Base Learners	5	8	10	5	8	10
Sensitivity	96.05	94.81	96.1	96.15	94.93	94.4
Specificity	92.11	91.89	94.59	97.22	97.14	97.29
Accuracy	94.74	93.86	95.61	96.49	95.61	97.37
Time (sec)	183.7	201.2	213.8	225.2	234.4	275.8
Number of Features	43			70		
Number of Base Learners	5	8	10	5	8	10
Sensitivity	97.37	96.1	98.66	97.33	94.81	97.37
Specificity	94.73	94.59	94.87	92.31	91.89	94.74
Accuracy	96.49	95.61	97.37	95.63	93.86	96.49
Time (sec)	238.7	279.6	325.1	316.7	370.2	433.3

As seen in Table 4(a), the highest classification rate was obtained as 97.37% with 10 base learners on FFS-AdaBoost-SVM structure at 33 and 43 features. The highest classification rate on PCA-AdaBoost-SVM structure was achieved as 71.05% with 8 base learners at 17 features as seen in Table 4(b). PCA is better than FFS only at some total training and test times as seen in Table 4.

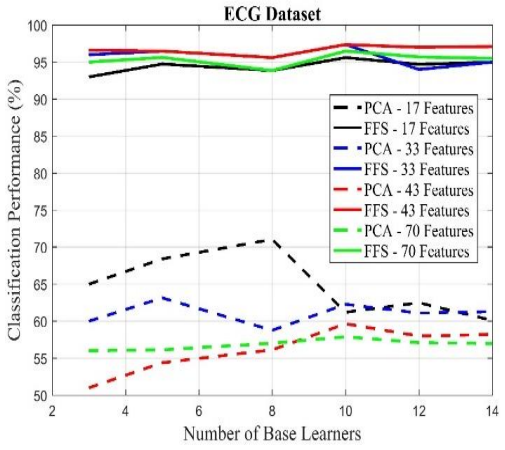

Fig. 7. PCA and FFS results on ECG according to the number of base learners for 17, 33, 43 and 70 features

4. Discussion and Conclusion

In this study, SVM-based AdaBoost ensemble classifier system was designed. Feature selection algorithms were added to this system to increase the classification performance. The dimension reduction was implemented on the biomedical dataset by FFS and PCA feature extraction algorithms. Thus, training and test times were reduced. Without feature extraction, total training and test time were 701,6 seconds in WBC dataset; 1285,2 seconds in PD dataset; 925,1 seconds in Heart (Statlog) dataset; 2804,7 seconds in ECG signal. The designed system was tested by continuous and discontinuous time dataset. The designed feature extraction-classifier structure was run by 30 times (10-folds CV). In experiments with the used discontinuous datasets, the number of base learners were held the same with the number of features. For 4 features, the best classification accuracy was obtained as 97.74% on WBC dataset; as 75.26% on Pima Diabetes dataset; as 89.54% on Heart (Statlog) dataset by FFS-AdaBoost-SVM structure. In experiments done by the used continuous dataset, the number of base learners were obtained by two different methods, and the system performance was observed. In the first method, the number of base learners were held the same with the number of features. In this method, the best classification accuracy was obtained by 17 features as 98.24% on FFS-AdaBoost-SVM structure as seen in Table 3(a).

In the second method, the number of base learners were investigated experimentally. The value of base learners' numbers was optimally found as 5, 8 and 10. 5, 8 and 10 base learners were tested by 17, 33, 43 and 70 features which were selected from ECG dataset. The best classification accuracy was obtained by 10 base learners with 33 and 43 features as 97.37% on FFS-AdaBoost-SVM structure as seen in Table 4(a).

The obtained results of this study were compared to the literature in Table 5. Table 5 shows that the proposed method is more effective than other studies to classify the used biomedical datasets. For Breast Cancer dataset, our study has the best classification accuracy according to Table 5, and the closest accuracy value was obtained as 97.35% in [10]. For Pima Diabetes dataset, the best classification accuracy is 77.34% in [3], and our study has lower classification accuracy than [3]. But our results are better than [10] for discontinuous dataset. For Heart (Statlog) dataset, our study has the best classification accuracy, and the closest accuracy value is 87.4% in [5]. For ECG dataset, our study has the highest classification accuracy as seen in Table 5. Also, it can be seen that in the Table 5, the studies in the literature did not use discontinuous and continuous dataset to test their algorithm. In this study, the proposed method was tested on both data types.

Table 5. Comparison with literature works

		Classification Accuracy (%)			
	Method	***Breast Cancer***	***Pima Diabetes***	***Heart (Statlog)***	***ECG***
This Study	FFS-Adaboost-SVM & PCA-AdaBoost-SVM	97.74	75.26	89.54	98.24
[2]	AdaBoost	74.47	Not used	Not used	Not used
[3]	AdaBoost- kNN	95.91	77.34	80	Not used
[4]	AdaBoost-SVM	97.13	Not used	72.24	Not used
[5]	AdaBoost-ANN	97.1	Not used	87.4	Not used
[6]	Fuzzy Boosting	96.75	77.32	82.28	Not used
[7]	Classifiers Ensemble based on Feature Selection	97.13	Not used	72.24	Not used
[9]	Ensemble Classifier Generation	96.79	Not used	Not used	Not used
[10]	AdaBoost-PCA	97.35	69.22	74.33	Not used
[20]	VQ feature extraction, Dictionary Learning	Not used	Not used	Not used	94.6
[21]	Parallel General Regression Neural Network	Not used	Not used	Not used	95

Acknowledgements

This work is supported by the Coordinatorship of Selcuk University's Scientific Research Projects.

References

[1] B. Yuan and X. Ma, "Sampling+ reweighting: boosting the performance of AdaBoost on imbalanced datasets," Neural Networks (IJCNN), IEEE International Joint Conference, 2012, pp.1-6.

[2] PP. Dhakate, K. Rajeswari and D. Abin, "An ensemble approach for cancerious dataset analysis using feature selection," Communication Technologies (GCCT), IEEE Global Conference, 2015, pp. 479-482.

[3] Y. Gao and F. Gao, "Edited AdaBoost by weighted kNN," Neurocomputing, vol. 73, no. 16, Oct. 2010, pp. 3079-3088.

[4] X.F. Chen, H.J. Xing and X.Z. Wang, "A modified AdaBoost method for one-class SVM and its application to novelty detection," Systems, Man, and Cybernetics (SMC), IEEE International Conference, 2011, pp. 3506-3511.

[5] A. Lahiri and PK. Biswas, "A scalable model for knowledge sharing based supervised learning using AdaBoost," Advances in Pattern Recognition (ICAPR), IEEE Eighth International Conference, 2015, pp. 1-6.

[6] H. Zhang and J. Lu, "Creating ensembles of classifiers via fuzzy clustering and deflection," Fuzzy sets and Systems, vol. 161, no. 13, 2010, pp. 1790-1802.

[7] B. Chen and H.X. Zhang, "An approach of multiple classifiers ensemble based on feature selection," Fuzzy Systems and Knowledge Discovery, IEEE FSKD'08 Fifth International Conference, 2008, pp. 390-394.

[8] UCI Machine Learning Repository, Access Date: January 2017.

[9] J. Ghavidel, S. Yazdani and M. Analoui, "A new ensemble classifier creation method by creating new training set for each base classifier," Information and Knowledge Technology (IKT), IEEE 5th Conference, 2013, pp. 290-294.

[10] SL. Ham and N. Kwak, "Boosted-pca for binary classification problems" Circuits and Systems (ISCAS), IEEE International Symposium, 2012, pp. 1219-1222.

[11] X. Shu and P. Wang, "An improved Adaboost algorithm based on uncertain functions," Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII), IEEE International Conference, 2015, pp 136-139.

[12] K. Deng, "OMEGA: On-Line Memory-Based General Purpose System Classifier," PhD thesis, Carnegie Mellon University, 1998.

[13] L. Smith, "A tutorial on Principal Component Analysis", 2002, pp. 001–027.

[14] B. Kégl, “Introduction to AdaBoost”, 2009, pp. 011-014.
[15] J. Sochman and J. Malas, “AdaBoost with totally corrective updates for fast face detection”, Automatic Face and Gesture Recognition, 2004, pp. 445-450.
[16] S.R. Kulkarni and G. Harman, “Statistical learning theory: a tutorial,” in Wiley Interdisciplinary Review: Computational Statistics, vol. 3, no. 6, 2011, pp. 543-556.
[17] S. Kutscher, “Algorithms for ECG Feature Extraction: an Overview”, 2013, pp. 001-008.
[18] PhysioNet, Access Date: January 2017.
[19] A. Beygelzimer, J. Langford and B. Zadrozny, “Weighted one-against-all”, American Association for Artificial Intelligence, 2005, pp. 720-725.
[20] Liu, T., et al.: “Dictionary learning for VQ feature extraction in ECG beats classification. Expert Systems with Applications”, Vol. 53, pp. 129-137 (2016)
[21] Li, P., et al.: “High-Performance Personalized Heartbeat Classification Model for Long-Term ECG Signal”, IEEE Transactions on Biomedical Engineering, Vol. 64, No.1, pp. 78-86 (2017)