American Journal of Applied Sciences

Original Research Paper

Random Forest Classification and Support Vector Machine for Detecting Epilepsyusing Electroencephalograph Records

Heri Kuswanto, Mutiah Salamah and Muhammad Idrus Fachruddin

Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Article history
Received: 14-01-2016
Revised: 26-04-2017
Accepted: 03-05-2017

Corresponding Author:
Heri Kuswanto
Department of Statistics,
Institut Teknologi Sepuluh
Nopember, Surabaya,
Indonesia
Email: heri_k@statistika.its.ac.id

Abstract: Complexity in data structure has led to the rapid development of computational statistics methods. Machine learning approaches have been introduced and applied to solve complex problems in many fields. This paper applies two common machine learning approaches, Random Forest (RF) and Support Vector Machine (SVM), in the detection of epilepsy. The diagnosis of epilepsy can usually only be made when a seizure is happening, which leads to some difficulties in the diagnostic process. The most recent way of diagnosing epilepsy is by using an Electroencephalograph (EEG) record. However, detecting epilepsy cases through EEG records takes a long time and may lead to misleading diagnostic results. The use of machine learning approaches is intended to generate fast and accurate classification results. As the EEG only generates a signal, direct analysis using RF or SVM cannot be carried out and the EEG record needs to be pre-processed. This paper uses Discrete Wavelet Transform and Line Length Features in the data pre-processing stage to decompose the signal by frequency and time. The classification results show that both RF and SVM perform very well and are able to classify cases of epilepsy accurately. The RF outperforms the SVM in the training dataset, while the SVM has a better performance in testing, with almost nom is classified cases. Several open problems relating to interpretation as well as parameter settings are described.

Keywords: Classification, Epilepsy, SVM, Random Forest

Introduction

The prevalence of epilepsy is still very high across the world. Epilepsy can be diagnosed during an epileptic attack. However, the diagnostic check conducted by a doctor is usually carried out during the post-attack period, without the doctor being able to look at the patient's condition directly while the attack is happening. This condition makes the diagnostic process difficult and may lead to wrong conclusions.

The most sophisticated way to recognize epilepsy is through an *Electroencephalograph* (EEG) record. An EEG records the electrical activity of neurons in the brain. Fluctuations in the electricity are measured from the voltage difference of electrodes connected to the brain network (Hughes, 2003). The diagnostic process using an EEG is carried out by monitoring patients continuously over several days. Nevertheless, most of the recorded data has to be observed and analyzed visually by an expert in order to detect epilepsy correctly. This detection procedure is considered to be inefficient, as it is time-consuming and costly. Therefore, tools or methods that lead to the fast and accurate diagnosis of epilepsy are required. In fact, numerous methods have been developed to detect cases of epilepsy. Recent works on EEG classification are, among others, those of Choe *et al*. (2010), who use statistical spectral feature extraction directly to classify the EEG signal, Anu and Thomas (2015), who apply k-NN to the spectral features of EEG records, and Al Ghayab *et al*. (2016), who perform the classification by simple random sampling and feature extraction. Most of the methods carry out feature extraction on the signals prior to classification and can thus be considered as two-step procedures. Although it has advantages, feature extraction may lead to some important information in the raw features being lost. Moreover, using extracted features to predict the class of new cases sometimes leads to poor performance, which means that the method per forms well in training data but fails to classify the testing data well. Hira and Gillies (2015) comprehensively discuss the weaknesses and advantages of feature selection and

Science Publications

extraction. This situation leads to the suggestion that the full features should be used in some cases.

This research applies two famous statistical methods in the classification of epilepsy using EEG signal data: Random Forest (RF) and Support Vector Machine (SVM), with no feature extraction. Both classification methods have been successfully applied in various cases. The RF applications can be seen in Diaz-Uriarte and de Andres (2006) for gene selection, Svetnik *et al.* (2003) for compound classification and Pal (2005) for remote sensing classification, among others, while Vapnik (1999; Drucker *et al.*, 1999; Squarcina *et al.*, 2015; Lian *et al.*, 2015; Huang and Zhou, 2015) have applied SVM to spam categorization, psychosis patients, influenza virus classification, etc.

Because of the nature of the EEG record, the RF and SVM methods cannot be directly applied to the signal generated from an EEG and the signal therefore has to be transformed. This research transforms the signal using Discrete Wavelet Transform (DWT) as the pre-processing method. DWT is a commonly recommended approach for analyzing data that have both time and frequency. By decomposing the signal into localized elements (both time and frequency), Mallat and Hwang (1992) argue that DWT is able to characterize the pattern well. The classification results generated from RF and SVM will be compared to find the best approach to diagnosing cases of epilepsy.

Materials and Methods

Data and Variables

The data used in this research come from a secondary dataset that has previously been analyzed by Lehnertz *et al.* (2002) and is published online at http://ntsa.upf.edu/downloads/andrzejak-rg-et-al-2001. The dataset has been used by Guo *et al.* (2010) to detect cases of epilepsy using neural networks. The variables used in this research are pre-processed using DWT. The number of predictor variables is determined by the number of levels used in the sub-band coding procedure. The level is chosen so that part of the signal is correlated with the frequency, which is represented by the wavelet coefficients. In this research, the decomposition levels are specified to be four, six and eight levels. Furthermore, the data obtained from those three levels is classified using RF and SVM. Subasi (2007) proved that Daubechies order four (db4) has a smoothing feature that fits the changes in the EEG very well. Therefore, the wavelet function chosen for this research is Daubechiesorder four (db4).

Signal Transformation and Wavelet Transform

Mathematical transformation is used to obtain further information from a signal through frequency analysis. One of the methods used to transform a signal from the time domain to the frequency domain is the Fourier Transform (FT). However, the FT can only be used when the signal is stationary and does not change over time. In fact, non stationary signals appear in many cases including EEG. Wavelet transform is a method used to transform a signal into the frequency as well as the time domain.

A wavelet transform signal is defined as:

$$\psi^{a,b}(t) = |a|^{-\frac{1}{2}}\psi\left(\frac{t-b}{a}\right) \qquad (1)$$

where, b is the position of the wavelet and a is the scale (Antonini and Barlaud, 1992). Wavelet transform is a popular method for transforming any data in the form of a signal, such as an acoustic signal from music. Research by Daubechies (1990) proved that the wavelet transform outperforms any other transformation approach such as a Short-time Fourier Transform. This research uses the Discrete Wavelet Transform (DWT) as the transformation approach. The DWT efficiently passes the signal through a low pass and a high pass filter, which is known as sub-band coding. The raw signal, denoted as $x[n]$, is processed through the high pass filter $g[n]$ and the low pass filter $h[n]$. After the filtering process, half of the sample is eliminated by applying the Nyquist rule (Akansu *et al.*, 2010). The rest of the signal is then divided into two outputs, which is known as the one-level decomposition process. This is mathematically written as:

$$y_{high}[k] = \sum_{n} x[n].g[2k-n] \qquad (2)$$

$$y_{low}[k] = \sum_{n} x[n].h[2k-n] \qquad (3)$$

where, $y_{high}[k]$ in (2) and $y_{low}[k]$ in (3) are the outputs of the high pass and low pass filters. The filtering process uses a linear convolution method, where $x[n]$ is a signal function for the input filtering, the functions $g[2k\text{-}n]$ in (2) and $h[2k\text{-}n]$ in (3) are the coefficients of the wavelet functions for the high pass and the low pass respectively, n is the signal length of the input and k is the length of the wavelet coefficient. The output of the high pass filter is Details (D) and the output of the low pass filter is Approximation (A). The one-level decomposition will generate as output D1 and A1. At level two, the output A1 is re-filtered to generate outputs D2 and A2. This process is repeated recursively up to the specified l-level. D1, D2,…,Dl and Al are the predictors in the classification process.

Line length is defined as the complexity measure or waveform fractal dimension line length, which is sensitive to the amplitude and signal frequency. It can be used to measure the pathology of the combination between the amplitude and frequency characteristics of the EEG signal. Esteller *et al.* (2001) define the line length as:

$$L = \frac{1}{N-1}\sum_{i=1}^{N-1} abs(x_{i+1} - x_i) \qquad (4)$$

where, x in Equation 4 is a signal, i represents the indices from the signal sample and N is the number of the signal used.

Random Forest

Random forest is a classification method consisting of independent classification trees (CART). The prediction of the classification is obtained by the majority voting of the classification trees that have been formed. Random forest is an extension of a collection of methods developed by Breiman (2001) and is used to improve the classification accuracy. Random forest differs from the bagging process in the sense that the bagging process uses a bootstrap to generate the classification tree in various versions and then combines these versions together to obtain the final prediction. In contrast, the randomization process in random forest to form the tree is carried out not only on the sample data but also on the predictor variables, leading to a collection of classification trees with different sizes and forms. The expected result is a collection of classification trees with very low correlation between the trees. This low correlation reduces the classification accuracy produced by random forest.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a very popular method in classification. SVM was first introduced by Vapnik (1999) in the Annual Workshop on Computational Learning Theory. SVM is one of the methods that has been developed to solve problems of classification and prediction that cannot be solved by the classical approach. SVM was developed using the principle of the linear classifier. However, most cases do not satisfy the linearity assumption and hence SVM has been developed to meet the nonlinear case by introducing the kernel concept. The study by Hsu *et al.* (2003) showed that classification using SVM will yield an accurate mapping.

The idea of SVM is to find the optimum hyperplane on the input space. The function of the hyperplane is used as a separator of two classes on the input space. The classes are usually denoted by -1 and +1. Fig. 1 illustrates the hyperplaneon SVM. The pattern on class -1 is shown by rectangles, while the pattern on class +1 is shown by circles.

Figure 1a shows the separator lines between the two classes (discriminant boundaries). The best line is the one with the maximum margin hyperplane. The margin is the distance between the hyperplane and the closest pattern in each class. The closest pattern is called the support vector. In Fig. 1b, the circles show the support vector for each class. Moreover, the bold line is the best hyperplane as it is located in the middle of the classes. The process of finding the position of the optimum hyperplane is at the core of SVM.

Evaluation of Classification Accuracy

Classification accuracy is used to assess how well the model represents the true process. One of the ways of measuring classification inaccuracy is the *Apparent Error Rate* (APER); the total accuracy rate is (1-APER). Table 1 shows how to calculate the APER.

Table 1. Cross-tabulation of classification results

	Prediction		
Observation	1	2	Total
1	n_{11}	n_{12}	n_1
2	n_{21}	n_{22}	n_2
Total	$n_{.1}$	$n_{.2}$	N

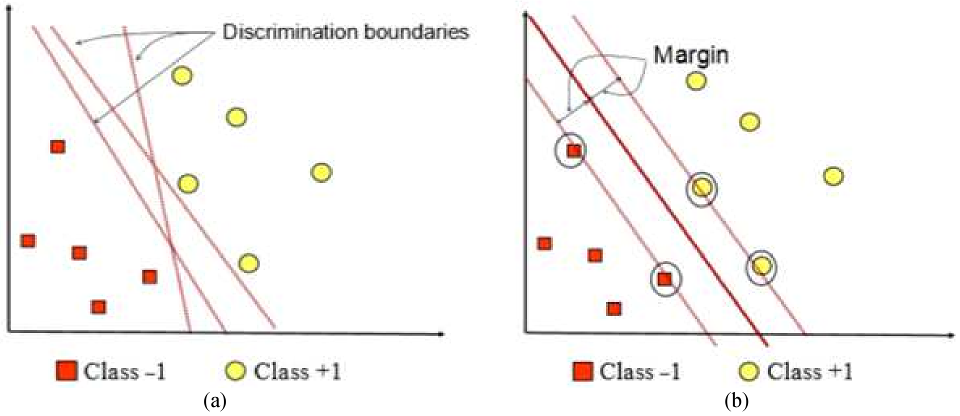

Fig. 1. Illustration of hyperplane on SVM (Hsu *et al.*, 2003)

where, n_{11} and n_{22} are the numbers of correct predictions in the corresponding class, while n_{12} and n_{21} are the numbers of incorrect predictions. Efron (1986) defines the APER value as:

$$APER = \frac{n_{12} + n_{21}}{N} \times 100\%$$

Results

This section describes the simple statistics for the pre-processed data using DWT and Line Length Feature (LLF) extraction. The data used in this research are the EEG records of 500 patients, some of whom are "normal" patients and some of whom are epileptic.

Figure 2 depicts the percentages for the data used in this research: 80% of the patients (400 patients) are "normal" and the rest have been detected to beepileptic. The dataset has been further pre-processed with DWT as well as Line Length Feature (LLF) extraction to extract sub signals with a vector size of 1 x (k+1), where k is the specified level of decomposition.

Classification using Random Forest

The first step to be conducted in RF analysis is to determine the number of predictor variables that will be randomly selected during the splitter selection in the classification tree. Following Breiman, the number of selected predictor variables is $\sqrt{p}$ where p is the total number of variables. The number p for the four-level decomposition is five, so that the number of selected control variables (predictors) is $\sqrt{5} = 2.24$ (which we round to two variables), the number of predictor variables for the six-level decomposition is $\sqrt{7} = 2.64$ (rounded up to three variables) while the number of predictors for the eight-level decomposition will be exactly 3. Each setting is run with numbers of trees equal to 100, 500 and 1,000. Tables 2-4 below display the classification results from analyzing the dataset with these three different decomposition levels consecutively.

Table 2 shows good classification accuracy in almost all settings. The combinations of data composition 75%:25%, 85%:15% and 95%:5% consistently have 100% accuracy on both test and training data. This result is observed for all combinations of the number of trees (K) and hence it is concluded that the number of trees does not significantly influence the classification accuracy. The data composition with the lowest accuracy (although it is still high enough) is the combination of 80%:20%, which gives 100% for training data and 95.83% for the test data. Similar results are obtained for the settings specified in Table 3 and 4.

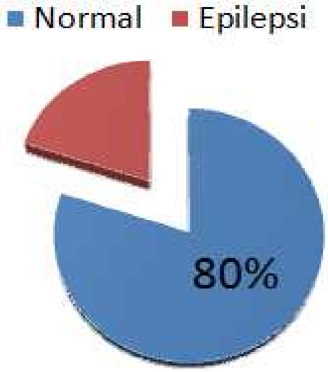

Fig. 2. Percentage of data in each category

Table 2. Comparison of classification accuracy using random forest for four-level decomposition data

	Accuracy (1-APER) (in %)					
Data combination	K = 100		K = 500		K = 1000	
Train text	Data train	Data test	Data train	Data test	Data test	Data test
75%:25%	100	100.00	100	100.00	100	100.00
80%:20%	100	95.83	100	95.83	100	95.83
85%:15%	100	100.00	100	100.00	100	100.00
90%:10%	100	97.96	100	95.83	100	95.83
95%:5%	100	100.00	100	100.00	100	100.00

Table 3. Comparison of classification accuracy using random forest for six-level decomposition data

	Accuracy (1-APER) (in %)					
Data combination	K = 100		K = 500		K = 1000	
Train text	Data train	Data test	Data train	Data test	Data train	Data test
75%:25%	100	99.19	100	99.19	100	99.19
80%:20%	100	97.96	100	97.96	100	97.96
85%:15%	100	97.26	100	97.26	100	97.26
90%:10%	100	97.96	100	97.96	100	97.96
95%:5%	100	100.00	100	100.00	100	100.00

Table 4. Comparison of classification accuracy using Random Forest for eight-level decomposition data

Data combination	Accuracy (1-APER) (in %)					
	K = 100		K = 500		K = 1000	
Train: Test	Data train	Data test	Data train	Data test	Data train	Data test
75%:25%	100	98.37	100	98.37	100	98.37
80%:20%	100	95.83	100	96.91	100	95.83
85%:15%	100	98.65	100	98.65	100	98.65
90%:10%	100	100	100	100	100	100
95%:5%	100	100	100	100	100	100

Classification using Support Vector Machine

The first step in SVM analysis is to determine the best C and γ parameters to use in the kernel function, that is, those that generate the lowest error model. The choice of C and γ will influence the accuracy of the constructed model. Among several approaches to determining the parameter combination are the uniform design and the tune approaches, which are available in R software. Huang and Zhou (2015) found that the optimum C in SVM should be located within the range of 10^{-2} to 10^{4}, while the parameter γ will be found within the range of $10^{-2}/\rho$ and $1.9/\rho$, where ρ is assumed to be 0.5. This research thus uses the tune approach for estimating the best combination of C and γ, which gives us 91 and 0.302 with an error of 0.008.

Table 5 above shows the comparison of the classification accuracy for analyzing the four-level decomposition data using SVM. The highest accuracy is obtained with the setting of 80% data as training data and 20% as testing data, where the accuracy is 99.75%. Meanwhile, setting 95% as training data yields the lowest accuracy. Tables 6 and 7 perform the accuracy results for six- and eight-level decomposition.

The results for the classification using SVM with six- and eight-level decomposition, as shown in Table 6 and Table 7, are similar to those using the dataset with the four-level decomposition. The accuracy for all settings is more than 90%. The values presented in Table 6 are obtained with the combination of C = 20 and γ = 0.011, while Table 7 uses the parameters C = 10,000 and γ = 0.001.

Table 5. Comparison of classification accuracy using SVM for four-level decomposition data

Data combination	Accuracy (1-APER) (in %)	
Train: Test	Data train	Data test
75%:25%	99.73	99.19
80%:20%	99.75	96.90
85%:15%	99.53	100.00
90%:10%	99.55	97.96
95%:5%	99.36	100.00

Table 6. Comparison of classification accuracy using SVM for six-level decomposition data

Data combination	Accuracy (1- APER) (in %)	
Train: Test	Data train	Data test
75%:25%	98.37	97.54
80%:20%	98.73	97.96
85%:15%	98.81	95.83
90%:10%	98.65	95.83
95%:5%	98.50	100.00

Table 7. Comparison of classification accuracy using SVM for eight-level decomposition data

Data combination	Accuracy (1- APER) (in %)	
Train: Test	Data train	Data test
75%:25%	99.73	99.19
80%:20%	99.75	98.99
85%:15%	99.76	100.00
90%:10%	99.77	100.00
95%:5%	99.79	100.00

Discussion

This section discusses the results presented in the previous section. One of the focuses of the discussion is the process of setting up the proportion of data used for the training and testing samples. All the settings showed that the data (testing and training) combinations lead to high classification accuracy. There is no suggestion that using any specific combination will lead to a more accurate result. This shows that both methods are very suitable for classifying whether the EEG records for a particular patient show that the patient suffers from epilepsy.

Another issue is the setting of the tree number (K) in the Random Forest classification as reported in Tables 2, 3 and 4. The tables reveal that the accuracy is only slightly different for the different settings of K. Meanwhile, the classification using SVM requires the parameters C and $\gamma$to be set. This research uses previously published works to determine the domain for both parameters. Indeed, we obtain convergent results with a high degree of classification accuracy.

Both RF and SVM have a high accuracy level, which means that both procedures are very capable of detecting cases of epilepsy. There is no reason to decide that one method outperforms the other in this case. This research shows that applying RF and SVM is not a simple matter, although both methods are capable of generating high classification accuracy. To implement the SVM, several parameters, such as C and γ, have to be specified and no formal procedure has been developed to estimate the optimum value of those parameters. Moreover, the exact range of the values is also unknown, which may lead to under- or over-estimation. A similar problem arises with the application of the RF method. The optimum value for K can only be found after comparing the RF accuracy for several K settings, which seems to be a trivial matter. Developing a statistical procedure to deal with the optimum parameter setting in SVM and RF is still an interesting and challenging problem and will be our future research topic.

Besides the problem of parameter specification, both methods suffer from alack of interpretation. As can be seen from this research, RF and SVM are purely machine learning approaches and they lack information about the specification or selection of the variables. To detect cases of epilepsy, we need to know if there is any clear and specific pattern that can explain the differences between the EEG records of epileptic and "normal" patients. Developing the RF and SVM methods to make them capable of conducting feature selection is also a subject for future research.

Conclusion

This paper investigates the performance of two common statistical methods to detect cases of epilepsy. The nature of the raw dataset generated from an electroencephalograph requires transformation and the DWT has been proven to be an effective procedure to decompose the signal into a time domain dataset. Several settings of the dataset are examined and the results show that both RF and SVM perform very well, with an accuracy level approaching 100% for both training and testing data. This confirms that the methods are capable of being used to detect, from EEG record data, whether or not a patient has epilepsy.

Acknowledgement

The authors gratefully acknowledge the financial support from the Ministry of Research, Technology and Higher Education Indonesia through Research Grant for International Publication and Research Collaboration scheme.

Author's Contributions

Heri Kuswanto: Cordinate the whole activities of the research, write the paper programming code validation.

Mutiah Salamah: Date management and cleaning, run the data analysis.

Muhammad Idrus Fachruddin: Date collection, help to analyze the data.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Akansu, A.N., W.A. Serdijn and I.W. Selesnick, 2010. Wavelet transforms in signal processing: A review of emerging applications. Phys. Commun.

Al Ghayab, H.R., H. Li, S. Abdulla, M. Diykh and X. Wan, 2016. Classification of epileptic EEG signals based on simple random sampling and sequential feature selection. Brain Informat., 3: 85-91. DOI: 10.1007/s40708-016-0039-1

Antonini, M. and M. Barlaud, 1992. Image coding using wavelet transform. J. IEEE Image Process., 1: 202-205. DOI: 10.1109/83.136597

Anu, N.S. and P. Thomas, 2015. An improved method for classification of epileptic EEG signals based on spectral features using k-NN. SSRG Int. J. Electron. Commun. Eng., 2: 35-38.

Breiman, L., 2001. Random forests. Machine Learn., 45: 5-32. DOI: 10.1023/A:1010933404324

Choe, S.H., Y. Chung and S. Kim, 2010. Statistical spectral feature extraction for classification of epileptic EEG signals. Proceedings of the 9th International Conference on Machine Learning and Cybernetics, Jul. 11-14, IEEE Xplore Press, Qingdao, pp: 1-6. DOI: 10.1109/ICMLC.2010.5580709

Daubechies, I., 1990. The wavelet transform time-frequency localization and signal analysis. J. IEEE, 36: 961-1005. DOI: 10.1109/18.57199

Diaz-Uriarte, R. and S.A. de Andres, 2006. Gene selection and classification of microarray data using random forest. Bioinformatics. DOI: 10.1186/1471-2105-7-3

Drucker, H., D. Wu and V. Vapnik, 1999. Support vector machines for spam categorization. IEEE Tran. Neural Netw., 10: 1048-1055. DOI: 10.1109/72.788645

Efron, B., 1986. How biased is the apparent error rate of a prediction rule? J. Am. Stat. Assoc., 81: 461-470. DOI: 10.1080/01621459.1986.10478291

Esteller, R., J. Echauz, T. Tcheng, B. Litt and B. Pless, 2001. Line length: An efficient feature for seizure onset detection. Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Oct. 25-28, IEEE Xplore Press, pp: 1707-1710. DOI: 10.1109/IEMBS.2001.1020545

Guo, L., D. Rivero, J. Dorado, J.R. Rabunal and A. Pajoz, 2010. Automatic epileptic seizure detection in

EEGs based on line length feature and artificial neural networks. J. Neurosci. Meth., 191: 101-109. DOI: 10.1016/j.jneumeth.2010.05.020

Hira, Z.M. and D.F Gillies, 2015. A review of feature selection and feature extraction methods applied on microarray data. Adv. Bioinform. DOI: 10.1155/2015/198363

Hsu, C.W., C.C. Chang and C.J. Lin, 2003. a practical guide to support vector classification. England University of Southampton.

Hughes, J.R. 2003. Practical Guide to Epilepsy. 1st Edn., Butterworth-Heinemann, Burlington, ISBN-10: 0750646217, pp: 335.

Huang, R. and Y. Zhou, 2015. Disease classification and biomarker discovery using ECG data. Biomed Res. Int. DOI: 10.1155/2015/680381

Lehnertz, K., R.G. Andrzejak, T. Kreuz, F. Mormann and C. Rieke *et al*., 2002. Analysis of EEG in epilepsy. Modelling Biomedical Signals.

Lian, W., J. Fang, C. Li, X. Pang and A.L. Liu *et al*., 2015. Discovery of influenza A virus neuraminidase inhibitors using support vector machine and naïve Bayesian models. Mol. Divers, 20: 439-451. DOI: 10.1007/s11030-015-9641-z

Mallat, S. and W.L. Hwang, 1992. Singularity detection and processing with wavelets. IEEE Tran. Inform. Theory, 38: 617-643. DOI: 10.1109/18.119727

Pal, M., 2005. Random forest classifier for remote sensing classification. Int. J. Remote Sens., 26: 217-222. DOI: 10.1080/01431160412331269698

Squarcina, L., U. Castellani, M. Bellani, C. Perlini and A. Lasalvia *et al*., 2015. Classification of first-episode psychosis in a large cohort of patients using support vector machine and multiple kernel learning techniques. Neuroimage, 145: 238-245. PMID: 26690803

Subasi, A., 2007. EEG signal classification using wavelet feature extraction and mixture of expert model. Expert Syst. Applic., 32: 1084-1093. DOI: 10.1016/j.eswa.2006.02.005

Svetnik, V., A. Liaw, C. Tong, C. Culberson and R.P. Sheridan *et al*., 2003. Random forest: A classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci., 43: 1947-1958. DOI: 10.1021/ci034160g

Vapnik, V., 1999. An overview of statistical learning theory. IEEE Tran. Neural Netw., 10: 988-1000. DOI: 10.1109/72.788640