Classification of Electroencephlography (EEG) Alcoholic and Control Subjects using Machine Learning Ensemble Methods

Article · January 2015

6 authors, including:

Lal Hussain
University of Azad Jammu and Kashmir

Wajid Aziz
University of Azad Jammu and Kashmir

Anees Qumar Abbasi
Karlsruhe Institute of Technology

Mohsin Manshad Abbasi
University of Azad Jammu and Kashmir

Classification of Electroencephlography (EEG) Alcoholic and Control Subjects using Machine Learning Ensemble Methods

Lal Hussain[1], Wajid Aziz[2], Amjad Saeed Khan [3], Anees Qammar Abbasi[2] , Syed Zaki Hassan[2] and Mohin Manshad Abbasi[2]
[1,2] Computer Science and Information Technology
University of Azad Jammu and Kashmir
Muzaffarabad, Pakistan
[3] Department of Computing, Lancaster University, UK
lall_hussain2008@live.com

***Abstract*—Electroencephalography (EEG) is a method to measure the electrical activity of brain signals by electrodes attached to the scalp at multiple locations. In this study we used the EEG signals of Alcoholic and Control subjects were obtained from Machine Learning repository according to the 10-20 International System. There were 29 subjects from control group and 29 from alcoholic comprising of electrodes from different brain lobes such as Central Lobe (C3, C4), Frontal (F3, F4, F7, F8), Occipital Lobe (O1, O2), Parietal Lobe (P3, P4), Temporal Lobe (T7) and Front Polar (Fp1, Fp2). In the first phase, the nonlinear complexity based features are computed such as Approximate Entropy, Sample Entropy and Wavelet Entropy. These features for each electrode are passed as input to the Machine Learning classifiers such as Multilayer Perceptron (MPL), K-nearest Neighbor (KNN) and LIB Support Vector Machine (SVM) to classify for alcoholic and control group. The highest accuracy was obtained using MLP of 98.22 % at electrode C3. Moreover, above 90% accuracy was obtained using MPL at electrode C3, C4, F7; KNN at electrodes C3, C4, F7 and LIB SVM at electrodes C3, C4, F7 where KNN gives highest accuracy of 97.67% at electrode C4 and LIB SVM an accuracy of 94.67 % at electrode F7. Secondly, the ensemble methods are employed such as Minimum, Maximum, Sum, Average, Product, Majority Vote, Bayes, Decision Template and Dempster Shefer Fusion. Using ensemble methods most electrodes depicted higher accuracy than individual classifier such as Electrode F4, Fp2, O1, O2 and T7. While electrode C3 the ensemble methods Moving Average, DT, DFT gives highest accuracy of 98.22% and at electrodes C4 Moving Average, Sum, Average, DT, DFT provided an accuracy of 97.11%.**

Keywords—*Electroencephalogram(EEG); Multiscale Sample Entropy (MSE); Multiscale Approx. Entropy (MAE); Multiscale Wavelet Entropy (MWE), Multilayer Percerptron (MLP); K-Nearest Neighbor (KNN); Support Vector Machine (SVM); Ensemble Methods (EM); Decision Template (DT) and Dempster Shefer Fusion (DSF).*

I. INTRODUCTION

Electroencephalographic (EEG) measurements are commonly used in medical and research areas. EEG is a representative signal containing information about the condition of human brain. Electroencephalography is a medical imaging technique that reads scalp electrical activity generated by brain structures. It is defined as electrical activity of an alternating type recorded from the scalp surface after being picked up by metal electrodes and conductive media [1].

By placing electrodes on the scalp it is possible to record the summed electrical activity of the cortex using a methodology known as Electroencephalography (EEG) [2].EEG records average neuronal activity from the cerebral cortex and can detect changes in activity over large areas but with low sensitivity for sub-cortical activity. EEG recordings are sensitive enough to detect tiny electrical impulses lasting only a few milliseconds. Most EEG devices have good temporal resolution, but low spatial resolution. EEG unity between pairs of scalp locations can provide important information about brain state. Current enter through skull, skin and other layers between electrode and neuronal layers. Weak electrical

Signals detected by the scalp electrodes are massively amplified, and then displayed on paper or stored to computer memory [3].

The EEG is a measure of the increasing examine of neurons in various parts of the brain. It contains the info about changes in electrical potential of the brain obtained from a set of recording electrodes [1].EEG signals are the reflection of the electricity activities of cerebral tissues and brain function status. The EEGs are useful for diagnosis and treatment of mental and brain diseases and abnormalities.

Recent progress in the theory of non-linear dynamics has provided new methods for the study of the EEG [5]. Non-linearity in the brain is introduced even at the cellular level, since the dynamical behaviour of individual neurons is governed by threshold and saturation phenomena. Moreover, the hypothesis of an entirely stochastic brain can be rejected due to its ability to perform sophisticated cognitive tasks. Considering this, non-linear dynamical analysis techniques may be a better approach than traditional linear methods to obtain a better understanding of abnormal dynamics in EEG signals [6,7]. In this study we are considering two groups alcoholic and control different brain lobes are considered C4,F3,F4 and O1 etc.

Considering the chaotic and non-stationary nature of EEG data, approximate entropy (ApEn) (e.g. Pincus et al., 1991) has been applied, instead of spectral entropy (Inouye et al., 1991), to measure the complexity of EEG series. The more regular the EEG is, the smaller the ApEn will be. The exact value of the ApEn will depend on three parameters: N (number of samples), m (embedding dimension) and r (noise threshold). The ApEn specifies a noise threshold, and so may be better than spectral entropy in the quantification of complexity of EEG recording (Bruhn et al., 2000, 2001). The disadvantage of ApEn is that it is heavily dependent on the record length, and is often lower than expected for short records. Another disadvantage is that ApEn lacks relative consistency (Richman and Moorman, 2000). To overcome the disadvantages of ApEn, a sample entropy (SampEn) was proposed to replace ApEn by excluding self-matches (Richman and Moorman, 2000), so reducing the computing time by one-half in comparison with ApEn. Another advantage of SampEn is that it is largely independent of record length and displays relative consistency. Further details of SampEn are described elsewhere (Richman and Moorman, 2000; Lake et al., 2002).

The complexity of an EEG series can also be quantified by using symbolic dynamic. A new permutation method was proposed by (Bandt and Pompe, 2002) to map a continuous time series onto a symbolic sequence; the statistics of the symbolic sequences was called permutation entropy. Permutation entropy quantifies the diversity of possible orderings of the values a random or deterministic system can take, as Shannon entropy quantifies the diversity of values.

The complexity of the EEG time series Alcoholic and control investigated. The relationships between these states and the complexities of the EEG are assessed. Alcoholic subjects are addicted persons and control subjects are normal.

In this study three classifiers are used such as SVM, MLP and KNN. Given a set of training data, each labled as belonging to one of two classes, an SVM training algorithm builds a model that assigns new examples into one classes or the other. An SVM is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. An MLP is a network of simple neuron called Perceptron. The basic concept of a single Perceptron was introduced by Rosenblatt in 1958. A multilayer Perceptron (MLP) is a Feed Forward model that plot sets of input data onto a set of appropriate outputs. An MLP is a directed graph consists of multiple layers of nodes that are fully connected to the next one layer. Except for the input nodes, each node is a neuron with a nonlinear activation function. k-Nearest Neighbors algorithm (k-NN) is a non-parametric method used for classification. The input consists of the k closest training examples in the feature space.

An ensemble consists of a set of individually trained classifiers whose predictions are combined when classifying new instance. Previous research has shown that an ensemble is often more accurate than any of the single classifiers in the ensemble. In this study we combine these three classifiers (SVM, MLP, KNN) with nine different combining ensemble methods.

II. PROPOSED METHODS

A. *EEG Recordings*

In the present study, the datasets for EEG motor movement tasks comprising of baseline eye open and eye close were taken from publically available database of Physionet. EEG signals are extracted from electrodes – F3, F4, F7, F8, Fp1, Fp2, O1, O2, P3, P4, C3 and C4 and T7 complying with the international 10-20 system and sampled at 160 HZ of one to two minute recording.

This data arises from a large study to examine EEG correlates of genetic predisposition to alcoholism. It contains measurements from 64 electrodes placed on subject's scalps. There were 29 subjects from alcoholic group and 29 from that of control group.

B. *Features Extraction*

Before classification we extracted the features of EEG signals. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. These features are used to understand the nature of signals, for example in investigating a certain brain disorder. On the basis of these features signals are categorized into two groups alcoholic and controlled. We consider three entropies as features.

Different approaches for extraction of quantitative features from the EEG signals were proposed more than 70 years ago, where these methods are used to explore the information from EEG. In this study, the sample entropy, approximate entropy and wavelet entropy is used as feature extraction methods.

Non-linear analysis metrics are valuable in the assessment of physiological time series, because ‘‘hidden information’’ related to underlying mechanisms can be sometimes obtained [8,9].

1) Approximate Entropy

Approximate entropy (ApEn) is the most popular non-linear method that has been applied to physiological time series to measure the regularity index ApEn presents some shortcomings, such as bias, relative inconsistency and dependence on the sample length. Approximate entropy (ApEn) is used as a measure of complexity that is applicable to noisy and medium-sized datasets presented by pincus [10]. ApEn determines the conditional probability of similarity between data sets of segments of the same duration. The higher the probability, the smaller the ApEn value, indicates less irregularity of data.

$$ApEn(S_{N,M,r}) = \ln\left[\frac{C_m(r)}{C_m + 1(r)}\right]$$

2) Sample Entropy(SpEn)

Entropies are basic invariants for dynamical systems. A modification of ApEn, named sample entropy (SampEn), which overcomes the deficiencies of approximate entropy. Sample entropy analyzes a time series for similar datasets and assigns a non-negative number to the sequence, with larger values corresponding to more irregularity in the data. It is used for assessing the complexity of a signal. Unlike ApEn, SampEn shows good results such as data length independence and trouble-free implementation.

$$SpEn(S_{N,M,r}) = -\ln\left[\frac{C_m(r)}{C_m + 1(r)}\right]$$

3) Wavelet Entropy (Wentropy)

Wavelet entropy measures built on wavelet analysis can signify the complexity of unsteady signal or system in both time domain and frequency domain.

C. Classification

Classification may refer to categorization, the process in which ideas and objects are recognized, differentiated, and understood. Classifying future or unknown objects, this is used. This model estimates the accuracy of the model. The known label of test sample is compared with the classified result from the model. Test set is independent of training set e.g.: if some tuples with certain data is given in the training dataset, in which these tuples are distributed among different classes, then this dataset is used to further determine the class of new tuple arrived for classification.

1) Support Vector Machine

A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification. , a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. The hyperplanes in the higher-dimensional space are defined as the set of points whose dot product with a vector in that space is constant. Support vector machine have recently gained performance in the field of machine learning and classification [11]

The operation of the SVM is based on finding the hyper plane that gives the largest minimum distance to the training examples. This distance receives the important name of margin within SVM's theory. The optimal separating hyper plane maximizes the margin of the training data.

SVM achieves great generalization performance. SVM is based on the concept of decision planes that define decision boundaries between a set of objects having different class memberships An SVM also uses a discriminate hyperplane to identify classes. However, concerning SVM, the selected hyperplane is the one that maximizes the margins.

Support vector machine is basically a two category classifier so that transformed data always separated by a hyper plane depends on nonlinear training data to higher dimension. Patterns are transformed according to their suitable kernel function. Support Vector Machines, only learn the way of discriminating the classes or the class membership in order to classify a feature vector directly [12] [13]

Its objective is to find a separate hyper plan with the largest margin while training an SVM so the classifier has a greatest generalization performance [14]. The small number of support vector and low error rate can be arise by using a kernel function which is capable of separating a data and therefore much important. In optimization process we will use quadratic discreminant analysis with minimum optimization [15][16] .

2) Multilayer Perceptron (MPL)

Multilayer Perceptron are a popular form of feed forward artificial neural network with many successful applications in data classification. The supervised learning process of MLP with input data x and target t, require the use of an objective function in order to assess the predicted output value.

MPL consist of multiple layers of simple, two state sigmoid processing elements (nodes) and neurons that interact using weighted connections [17]. The Perceptron (threshold unit) is an algorithm for supervised classification of an input into one of several possible non binary outputs. A single layer Perceptron network can be used for classification of linearly separable problems. For classification problem linear non separated groups of points there we used Multilayer Perceptron network with one or more hidden layers[18].The aim of MLP network is to classify inputs in appropriate class from classes. The logistic sigmoid or hyperbolic tangent is common choices for the activation function. We can use the same activation function on all layers.

K Nearest Neighbor (KNN)

The simplest classification algorithm is KNN which is based on an assumption that samples closer in instance space have same class values. The nearest neighbor classifier is an example of instance based learning approach. The training examples are stored and the distance function is used to determine which member of training set is closest to an unknown test instance. Once the nearest training instance has been located, its class is predicted for the test instance. The only remaining problem is defining the test function. From the features the KNN classifier can classify an EEG signal into the class which most of its neighbor belongs to KNN find its K nearest neighbor with respect to suitable distance function and classify new data object. The k-nearest-neighbor classifier is one of the most basic classifiers for pattern recognition or data classification. This method is based on the concept that data instances of the same class should be closer in the feature space. As a result, for a given data point x of unknown class, we can simply compute the distance between x and all the data points in the training data, and assign the class determined by the K nearest points of x. KNN is a simple algorithm that store all available cases and classify new cases based on a similarity measure. KNN require an integer, a set of labeled examples (training data) and a metric which is used to measure closeness. It is a non-parametric method and due to its effeteness and easy to implementation properties tried for many applications. Also the effect of noise in classification is reduced is reduced by the larger values of k.

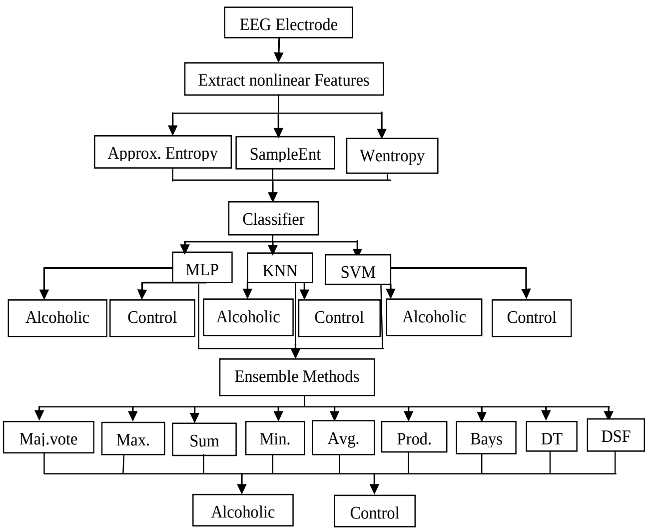

Fig.1: Block Diagram for Classification of EEG Alcoholic and Control Subjects using Ensemble Methods

D. *Ensemble Methods*

An ensemble consists of a set of individually trained classifiers (such as neural networks) whose predictions are combined when classifying novel instances. Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking the weight of their prediction.

1) *Minimum*

Find the minimum score of each class between the classifiers and assign the Input pattern to the class with the maximum score among the maximum scores.

2) *Maximum*

Find the maximum score of each class between the classifiers and assign the Input pattern to the class with the maximum score among the maximum scores.

3) *Product*

Multiplies the score provide by each base classifiers and assigns the class label with maximum score to given input pattern.

4) *Sum*

Adds the score provide by each base classifier and assigns the class label with maximum score to given input pattern.

5) *Average*

Finds the mean of scores of each class between the classifiers and assigns the input patterns to the class with the maximum score among the means.

6) *Majority Vote*

Is the assuming that higher rank values mean more confidence of the classifier. Correct class appears at the high rank but not at the top.

7) *Decision Template*

DP is the matrix of outputs of the classifiers in an ensemble. Each cell i,j represent that the patterns come from which class. The DT captures the most typical DP for each class and classifies new patterns by comparing the DP with the DT.

8) *Bayes*

Compute the posterior probability for all values of C using the Bayes theorem.

9) *Dempster Shefer Fusion*

The Dempster-Shefer Fusion method uses decision profile to find the overall support for each class and subsequently labels the instance x in the class with the largest support.

III. RESULTS AND DISCUSSIONS

We examined the classification performance here using three classifiers such as MLP KNN SVM on electrodes C3, C4, F3, F4, F7, F8, Fp1, Fp2, O1, O2, P3, P4 and T7. Among these classifiers MLP gives the higher accuracy. Higher accuracy shows that MLP more accurately classify the alcoholic and controlled subjects. In this study electrode C3 gives better performance among all the electrodes. Brain contains different lobes such as Central, Frontal, Occipital, Parietal, and Temporal. From the results depicted in Tables it is evident that the Central lobe is most important part of the brain to distinguish between these two groups (alcoholic and controlled). C3 is helpful for the best separation of alcoholic and controlled groups so this the most important part of the brain for classification which distinguish the brain activity. Electrode C4 is better classified by KNN among these classifiers, Electrode F3 is better classified by KNN, Electrode F4 is better classified by MLP among these classifiers, Electrode F7 is better classified by LIB SVM, Electrode F8 is better classified by KNN among these classifiers.

TABLE I. SINGLE CLASSIFIER ACCURACY RESULTS

Electrodes	MLP	KNN	LIB SVM
C3	0.9822	0.9711	0.9322
C4	0.9656	0.9767	0.9656
F3	0.6333	0.7067	0.6367
F4	0.7700	0.6867	0.7533
F7	0.9289	0.9300	0.9467
F8	0.7422	0.7822	0.7522
Fp1	0.7589	0.7933	0.8322
Fp2	0.7615	0.7600	0.8178
O1	0.6556	0.7778	0.7478
O2	0.6856	0.6956	0.6567
P3	0.8589	0.8300	0.7356
P4	0.6700	0.7267	0.6422
T7	0.7333	0.7467	0.7756

Similarly Electrode Fp1, Fp2, O1, O2, P3, P4 and T7 are better classified by Classifier SVM, SVM, KNN, KNN, MLP, KNN and SVM respectively. Moreover, the classifier SVM, MLP and KNN gives more than 80 % performance on electrodes C3, C4 and F7. Classifier lib SVM gives more than 80% performance on electrode Fp1and Fp2; whereas MLP and KNN give more than 80% performance on electrode P3.

Combining the ideas of different experts to obtain an overall ensemble decision is rooted in our culture at least from the classical age of ancient Greece, and has been formalized during the Enlightenment with the Condorcet Jury Theorem [19] that proved the ensemble methods are superior to single classifiers.

A plethora of term other than ensemble has been used such as fusion, combination that indicate a set of learning machines work together to solve a problem[20][21][22]. First of all Multiple classifier system conference organized by Rolli, kittler, Windeat and other researchers of this area [23][24].

Allwein, Schapier and Singer interpreted the improved generalization capabilities of ensembles of learning machines in the framework of large margin classifier[25][26], Kleinberg in the context of Stochastic Discrimination Theory[27], and Breiman and Freidman in the light of bais analysis borrowed from classical statistics[28].

IV. CONCLUSION

In the present study, electroencephalography (EEG) background activity in the patients with alcoholic and control groups is examined using Machine Learning classifiers and Ensemble Methods. Electrodes wise information was extracted from Central, Frontal, Parietal, Occipital and Front Polar Brain Lobes. Complexity based nonlinear features are extracted and passed to the Machine Learning classifier to distinguish the alcoholic and control subjects. MLP gives highest performance using single classifier based classification while KNN and SVM provided second highest accuracy. However, Ensemble methods depicted more accuracy than individual classifier at most of the electrodes which confirm the performance using the combined classifiers.

E	MV	Max	Sum	Min	Avg	Pro	Bayes	DT	DSF
C3	0.9822	0.9389	0.9822	0.9389	0.9822	0.9211	0.9767	0.9822	0.9822
C4	0.9711	0.9656	0.9711	0.9656	0.9711	0.9656	0.9656	0.9711	0.9711
F3	0.6644	0.6489	0.6644	0.6444	0.6644	0.6478	0.6533	0.6478	0.6422
F4	0.7678	0.7267	0.7678	0.7200	0.7678	0.6744	0.7789	0.7900	0.7900
F7	0.9467	0.9400	0.9467	0.9289	0.9467	0.9122	0.9289	0.9467	0.9467
F8	0.7878	0.7689	0.7878	0.7689	0.7878	0.7011	0.7756	0.7811	0.7878
Fp1	0.8200	0.8000	0.8200	0.8056	0.8200	0.7444	0.8333	0.8200	0.8200
Fp2	0.8178	0.7926	0.8178	0.7926	0.8178	0.7022	0.8104	0.8252	0.8252
O1	0.7778	0.7044	0.7778	0.7056	0.7778	0.6256	0.7900	0.7956	0.7833
O2	0.6733	0.6300	0.6733	0.6367	0.6733	0.5622	0.6956	0.7178	0.7178
P3	0.8322	0.8267	0.8322	0.8211	0.8322	0.7600	0.8378	0.8378	0.8378
P4	0.7100	0.6467	0.7100	0.6700	0.7100	0.6189	0.6933	0.7044	0.7100
T7	0.8200	0.7322	0.8200	0.7100	0.8200	0.6156	0.7956	0.8089	0.8200

TABLE 2. ENSEMBLE CLASSIFIERS CLASSIFICATION ACCURACY RESULTS

REFERENCES

[1] Niedermeyer, E., & da Silva, F. L. (Eds.). (2005). Electroencephalography: basic principles, clinical applications, and related fields. Lippincott Williams & Wilkins.

[2] Fisch, B. J., & Spehlmann, R. (Eds.). (1999). Fisch and Spehlmann's EEG primer: basic principles of digital and analog EEG. Elsevier Health Sciences.

[3] Tyner, F. S., & Knott, J. R. (1983). Fundamentals of EEG Technology: Basic concepts and methods (Vol. 1). Lippincott Williams & Wilkins.

[4] Bronzino, J. D. (1995). Principles of electroencephalography. The Biomedical Engeneering Handbook, 201-212.

[5] Jeong, J. (2004). EEG dynamics in patients with Alzheimer's disease. Clinical neurophysiology, 115(7), 1490-1505.

[6] Kantz, H., & Schreiber, T. (2004). Nonlinear time series analysis (Vol. 7). Cambridge university press.

[7] Zhang, X. S., Roy, R. J., & Jensen, E. W. (2001). EEG complexity as a measure of depth of anesthesia for patients. Biomedical Engineering, IEEE Transactions on, 48(12), 1424-1433.

[8] Pincus, S. M., & Goldberger, A. L. (1994). Physiological time-series analysis: what does regularity quantify?. American Journal of Physiology, 266, H1643-H1643.

[9] Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. American Journal of Physiology-Heart and Circulatory Physiology, 278(6), H2039-H2049.

[10] Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. Proceedings of the National Academy of Sciences, 88(6), 2297-2301.

[11] Vapnik, V. (2000). The nature of statistical learning theory. springer.

[12] Jordan, A. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Advances in neural information processing systems, 14, 841.

[13] Rubinstein, Y. D., & Hastie, T. (1997, August). Discriminative vs Informative Learning. In KDD (Vol. 5, pp. 49-53).

[14] Kaper, M., & Ritter, H. (2004, September). Generalizing to new subjects in brain-computer interfacing. In Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE (Vol. 2, pp. 4363-4366). IEEE.

[15] Mak, G. (2000). The implementation of support vector machines using the sequential minimal optimization algorithm (Doctoral dissertation, McGill University).

[16] Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

[17] McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). Parallel distributed processing. Explorations in the microstructure of cognition, 2.

[18] Luger, G. F. (2005). Artificial intelligence: structures and strategies for complex problem solving. Pearson education.

[19] Lam, L. (2000). Classifier combinations: implementations and theoretical issues. In Multiple classifier systems (pp. 77-86). Springer Berlin Heidelberg.

[20] Lam, L., & Suen, C. Y. (1995). Optimal combinations of pattern classifiers.Pattern Recognition Letters, 16(9), 945-954.

[21] Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers.Pattern Analysis and Machine Intelligence, IEEE Transactions on, 20(3), 226-239.

[22] Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. Pattern Analysis and Machine Intelligence, IEEE Transactions on,22(1), 4-37.

[23] Mayr, T. R., Corstanje, R., Hannam, J. A., Zawadzka, J. E., Holden, A., & Jones, R. J. A. (2014). Multiple Classifier System.

[24] Kittler, J., & Alkoot, F. M. (2003). Sum versus vote fusion in multiple classifier systems. Pattern Analysis and Machine Intelligence, IEEE Transactions on,25(1), 110-115.

[25] Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of statistics, 1651-1686.

[26] Allwein, E. L., Schapire, R. E., & Singer, Y. (2001). Reducing multiclass to binary: A unifying approach for margin classifiers. The Journal of Machine Learning Research, 1, 113-141.

[27] Kleinberg, E. M. (2000). On the algorithmic implementation of stochastic discrimination. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(5), 473-490.

[28] Breiman, L. (1996). Bias, variance, and arcing classifiers.