International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

Sleep stage classification via ensemble and conventional machine learning methods using single channel EEG signals

Hamza Osman İlhan[1], Gökhan Bilgin*[1]

Accepted : 26/10/2017 Published: 30/12/2017

Abstract: Sleep-stages play important roles in the diagnosis of the sleep disorders and the sleep-related illnesses. In this sense, accurate identification of the sleep-stages is a necessity for more robust and efficient diagnosis systems. Several traditional machine-learning and pattern recognition algorithms are deployed on the modern computer aided diagnosis systems. However, current results are not as satisfactory as expected. In the last two decade, a new concept has emerged with 'ensemble learning' title. It has attracted the attention of many researchers from various disciplines. In this study, several ensemble-learning methods are utilized and inspected on EEG signals for sleep-stage classification. Conventional machine-learning methods are also performed in same testing phase to report comparative results. Additionally, methods are evaluated in two different scenarios; subject specific and independent. Study proves that combination of DTs and SVMs in bagging theorem surpasses all of the conventional methods used in the experiments. Moreover, test trials reveal that both conventional and ensemble models need to be improved for subject independent scenario which is more essential case in the development of patient independent computer based diagnosis systems.

Keywords: *Sleep-stage classification, EEG, machine-learning, ensemble-learning, PhysioNet*

1. Introduction

Nowadays, inventive researches are being carried out to develop new methods for the identification and treatment of sleep disorders such as narcolepsy, idiopathic hypersomnia and sleep apnea. Problems related with sleep adversely affect physical and social quality of life of a person [1]. Besides the sleep disorders, sleep-related illnesses, including diabetes, cardiovascular diseases, obesity, etc. are other focuses of the researches [2, 3, and 4]. For this reason, the accurate identification of sleep-stages is an important subject in computer aided diagnosis that may lead to more precise diagnoses. A handbook about the determination and scoring of the human sleep stages has been published by twelve researchers, under the editorship of Rechtschaffen and Kales [5, 6]. According to this manual, the duration of the sleep for a healthy person can be divided into two main stages; rapid eye movement (REM) and non-rapid eye movement (NREM) stages. The NREM stage also consists of four sub-periods (NREM I, NREM II, NREM III, and NREM IV) that have discriminative amplitudes of certain frequencies. All stages are defined according to Polysomnography (PSG) results of the patients. According to sleep staging method developed by the American Academy of Sleep Medicine (AASM), NREM III and IV stages are defined in single stage, known as slow wave sleep (SWS) or deep sleep [7]. Polysomnography (PSG) is a "gold standard" method for clinical diagnosis; sleep medicine industry and sleep-stage classification studies. PSG contains crucial physiologic signals, including electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), pulse oximetry (SpO_2), and electrocardiography (ECG). Analysis of PSG requires the participation of an expert in a specific sleep centre during recording, but this is a relatively expensive and time-consuming procedure for both the patients and the experts on sleeping. Hence, automatic sleep-staging has become an important challenge for researchers in different disciplines [8, 9, and 10].

In literature, several methods have been studied on the classification of sleep-stages. The frequency-domain analysis methods [11, 12, and 13], wavelet transform [10, 14] and fuzzy logic [15] are examples of some methods with agreement rates ranging from 60% to 80%. Virkkala *et al.* have classified the sleep stages using only EOG signals with the agreement of 72% [16]. Mendez *et al.* have utilized a Hidden Markov Model (HMM) with spectral features of heart rate variability to classify NREM and REM and the classification accuracy is measured around 80% in both training and test sets [17]. Liang *et al.* have presented a rule-based sleep-stage classification method using features of temporal and spectral analyses of the EEG, EOG, and EMG signals with an agreement rate of 86.68% [18]. Different types of feature selection methods, including the multiple iterative, suitable linear and non-linear methods have been proposed for classification of sleep stages by Zoubek *et al.* [19], and accuracies of wakefulness (W), NREM I, II, SWS, and REM are obtained as 84.57%, 64.56%, 85.55%, 92.90% and 72.81%, respectively. In [20], the energy features of single-channel EEG signals are utilized for classification of sleep-stages using neural classifiers and EEG epochs were classified as wakefulness, NREM I, II, SWS or REM, and the overall accuracy is 81.8%. In another study, Koley and Dey [21] applied a Support Vector Machine (SVM) based ensemble method on their data set to classify it into five stages as similar to [20] with using different feature extraction methods. Furthermore, a type of SVM based recursive feature elimination algorithm is applied on 39 extracted features in order to enhance their result. It is reported in the study that 85% and 87% agreements were obtained with training and independent testing data sets respectively.

[1] *Dpt. Of Computer Eng. Yildiz Technical University, Istanbul – 34220, TURKEY*
* *Corresponding Author: Email: gbilgin@yildiz.edu.tr*

The goal of this study is to evaluate and compare the latest and conventional learning methods on sleep-stage classification. The best classifier for EEG signals based automatic sleep-staging system may be assessed according to the obtained results of this study. For this purpose, several well-known conventional methods (i.e., Support Vector Machines (SVMs), Naive Bayes (NB), Linear Discriminant Classifier (LDC), K-Nearest Neighbour (KNN) and Decision Tree (DT)) and some of their combination in ensemble learning (Bagging and Adaboost) are selected as classifier. It is aimed to demonstrate the effectiveness of ensemble combinations on results with comparative tables.

This paper consists of six sections. The details of the Sleep-EDF database are presented in Section 2. Signal pre-processing, feature extraction, and classification methods are described in Section 3. Performance metrics and the results are placed in Section 4. Discussions about results are given in Section 5, and the study ends up with conclusion and future works in Section 6.

2. Materials

PSG is a multi-parametric test that is used to identify illnesses caused by sleep disorders. It is also effectively used to derive the characteristic schema of the sleep. Several PSG records which are obtained from PhysioNet open database have been used in this study [22]. The PhysioNet [23] is a well-known biomedical data source, which is frequently used in many studies [18, 19, and 20]. PSG data sets contain several signals from various sensors. It is obtained from the records of eight Caucasian male and female volunteers aged from 21 to 35 years. The records are separated into two groups according to obtaining procedures. The first four patients with "sc" initial letters are combined into Group I. The other group contains the rest of four patients, which are designated with "st" initial letters. Group I records were acquired over 24 hour period from healthy patients in normal daily life with a modified cassette tape recorder. Group II records were obtained from patients having mild difficulty falling asleep and otherwise healthy in a hospital setting a 12-hour night period. None of the patients in both groups have been given medication for any illnesses or disorders.

Group II records are more challenging according to sleep disorder reports of the patients. Group I have more clear signals because of the modified analog cassette recorder. Furthermore, Group I recordings have more samples than Group II, which provides more efficient classification performance. Despite the differences, all PSG records commonly contain two EEG channels (Fpz-Cz and Pz-Oz) and a single EOG channel with a sampling frequency of 100 Hz. Additionally, both groups have sub-mental EMG signals with different sampling frequencies. Moreover, Group I recordings include additional signals, oral nasal airflow and rectal body temperature, sampled at a 1-Hz frequency.

The records are scored by using Rechtschaffen and Kales (R&K) rules with 30-second intervals, which are called epochs. Each epoch has one sleep-stage label stored in a hypnogram. According to R&K standards, sleep-stages are divided into six stages, namely W (Wakefulness), REM (Rapid Eye Movement), NREM I, II, III, and IV (Non-Rapid Eye Movement). In some studies NREM III and IV stages are combined into a single stage (SWS - Slow-wave Sleep) to increase classification ratio [19, 20], but in our study all stages will be separately taken into account in order to show the efficiency of ensemble methods. In the proposed study, sleep-stage classification is performed based on only EEG signals. The representation of the EEG channels is referred as a montage, and different montages are available in practical sessions at hospitals [24]. The sleep EDF data set includes two channels recorded under the sequential montage. Because the Fpz-Cz channel is more distinctive than others according to some recent research papers [20], we also utilize and focus on Fpz-Cz channel in our study.

3. Methodology

Fpz-Cz channel of EEG recordings is selected as input for the evaluations of the classification methods. Fig. 1 indicates the flow diagram and steps of this study.

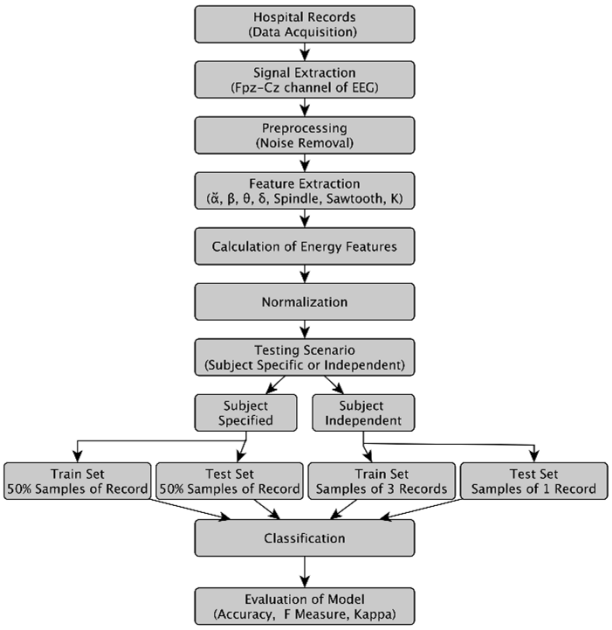

Fig. 1. The flow chart of the methodology

Methodology will be investigated in four sub-sections: (a) signal preprocessing, (b) feature inference and extraction, (c) formation of training and testing sets, and (d) classification. Performance metrics will be explained as a final step.

3.1. Signal Pre-processing

Biomedical signals can be easily affected by artificial or natural non-controllable factors hence signal pre-processing step is a necessity and inevitable process in order to isolate raw signals.

A small part of samples are labeled as 'undefined' in the records. Therefore, these samples are accepted as noise and removed. Additionally, Butterworth band-pass filters with 0.2 Hz and 40 Hz cut off frequencies are implemented on the records as noise-removal process. Signals over 40 Hz and below 0.2 Hz frequencies are mostly EEG irrelevant signals. Sample distributions with corresponding stages after de-noising processes are presented in Table 1.

Table 1. Number of samples after signal preprocessing step

	Data Set	WAKE	NREM I	NREM II	NREM III	NREM IV	REM	Total
Group I	sc4002e0	1884	59	373	94	203	215	2828
	sc4012e0	1823	92	660	80	16	176	2847
	sc4102e0	1908	117	607	25	0	199	2856
	sc4112e0	2103	18	396	90	21	151	2779
Group II	st7022j0	75	74	353	127	157	159	945
	st7052j0	128	121	396	53	127	226	1051
	st7121j0	70	34	452	120	83	267	1026
	st7132j0	60	89	384	83	20	216	852

3.2. Feature Inference and Extraction

Feature extraction methods directly state the success of the used classification algorithm in next step; hence feature extraction is

crucial step for any signal classification as in biomedical field. Features should express the original signal as much as possible. Moreover, they must be discriminative and informative in order to increase classification results.

Generally in the literature, feature extraction methods can be categorized into three sections: time, frequency, and spatial domain based techniques. Additionally, combined time-frequency based techniques are also available such as short time Fourier transforms (STFTs) and wavelet transforms [25]. In this study, frequency domain based feature extraction methods are selected.

The EEG signals can be represented in frequency domain with seven characteristic waves, namely alpha ($\boldsymbol{\alpha}$), beta ($\boldsymbol{\beta}$), theta ($\boldsymbol{\theta}$), delta ($\boldsymbol{\delta}$), spindle, saw-tooth, and K-complex. The 10th-order infinite impulse response (IIR) Butterworth filters are designed with relevant cut off frequencies, and applied on signals after preprocessing step in order to obtain these waves. The names of the waves and the corresponding spectral-band frequencies are presented in Table 2.

Table 2. Cut-off frequencies of Butterworth filters

Characteristic Wave	Spectral Band Freq.
Alpha	8 - 13 Hz.
Beta	12 - 30 Hz.
Theta	4 - 8 Hz.
Delta	0.5 - 2 Hz.
Spindle	12- 14 Hz.
Saw-tooth	2 - 6 Hz.
K Complex	1 Hz.

Subsequently, energy values of characteristic waves are calculated with (1);

$$Energy_k = \sum_{n=1}^{N} x_n^2 \tag{1}$$

where N denotes the total number of samples in one epoch, which is 3000 by taking 30-second intervals at the sampling frequency of 100 Hz. Here, x_n representsthe n_{th}sample in corresponding characteristic wave.The sum of the energy values of relevant waves are assigned as discriminative features of the signal. However, the distributions of feature characteristics vary with different ranges; hence these features need to be normalized in order to use them together. Additionally, the normalization procedure provides more accurate assessment for the subject independent scenario. In that case, all features are normalized into the [0-1] range before the classification step as follows:

$$Normalized_{e_i^k} = \frac{e_i^k - E_{min}^k}{E_{max}^k - E_{min}^k} \tag{2}$$

where e_i^k shows the normalized i_{th} energy value for the corresponding k characteristic wave. E_{min}^k and E_{max}^k are minimum and maximum values within the k_{th} characteristic wave.

As a summary, normalized energy values of extracted characteristic waves of Fpz-Cz EEG channel signals are selected as feature sets which will be divided into training and testing set in next section.

3.3. Training and Testing set formations

Two testing approaches are mainly presented in literature; subject independent and subject specific [26]. Classification methods are tested under both scenarios in this study. There are differences in sample selection step between the scenarios. Training samples are selected from one patient's records with a split ratio in the subject specific strategy, and the testing is performed on the rest of data belongs to same patient. On the other hand, training set is formed by entire records of all patients except one in the subject independent strategy. Isolated patient is reserved for testing within the same group. First strategy gives more theoretical information about the success of the model in terms of the machine learning concept, and other scenario is related with more practical experiments in order to apply the methods on unseen samples. Automatic sleep-staging system deals with more practical problems which is more likely to be encountered in hospitals, clinics and institutes.

A well-known method named as *k-fold* cross-validation is implemented for the subject specified strategy. The *k* value represents the number of partitions. Samples are divided into *k* equal sizes for every class. Number of *k-1* defines the size of training set and the rest of data is assigned to testing set. The minimum *k* can be two which indicates that training and testing sets are formed with equal number of samples. The *k* also represents the rotation number which indicates the number of repetitions with different samples but same size in training set. *k* is chosen 2 in this study. Additionally, testing process is repeated 10 times for strengthen the results. Final decision is made by majority voting technique which is based on calculation of average score of all results.

Another cross validation method, *leave-one-out* cross-validation, is used in order to arrange the sample proportions for the subject independent strategy. All the records of three patients are selected as training set and the remaining is considered as testing set within the same group.

3.4. Classification Methods and Parameter Settings

Several well-known classification methods are utilized in this study. Selected methods are separated into two titles: a.) Conventional machine-learning and b.) Ensemble-learning methods. The brief descriptions of the utilized methods and parameter settings are explained in following sub-sections. As a preliminary work, each method are tested with their different parameter settings in order to find the model's best accurate results and corresponding settings. Parameter setting tests are performed on the same dataset (50% of data set assigned as training, another as testing) at once. Afterwards, all methods with defined parameters are evaluated in experimental section with abovementioned formation of data set. Same as parameter setting tests, comparative tests are also performed on same testing data set at once for all methods in experimental tests.

3.4.1. Conventional Machine Learning Methods

Many algorithms are developed with the fast advance of the machine-learning. Majority of these algorithms are highly utilized on biomedical data sets to derive more meaningful information and classify with better accuracy. This study contains several familiar machine-learning algorithms (Support Vector Machines (SVMs), Decision Tree (DT), K-Nearest Neighbor (KNN), Naive Bayes (NB), Artificial Neural Network (ANN) and Linear Discriminative Classifier (LDC)) to evaluate the methods on the abovementioned sleep-stage data sets.

3.4.1.1. Support Vector Machine (SVMs)

SVM is one of the prominent classification algorithm which can be used large-scale data sets and provides more efficient results than statistical and neural classifiers. In SVM, higher classification

accuracies can be achieved by even small size train sets with the help of well-fitted cost function in kernel space as well [27]. In this section, SVM terminology and its usage in the sleep stage classification are briefly explained.

SVM uses the core idea of kernel based learning. Kernel based learning aims to separate data in high dimensional feature space by mapping data points with a kernel function. SVM creates a decision surface between the samples of different classes by finding the optimal hyperplane that is closest to the deciding training samples (support vectors). That way an optimal classification can be achieved for linearly separable classes. In case of linearly inseparable situations, kernel versions of SVM are defined. The main purpose of kernel approach in SVM is to transform the data to a higher dimensional space ($\Theta : R^n \rightarrow R^h$, $h > n$) where binary classification can be achieved linearly again [28]. Kernel functions are mainly used to define cost function, and the response of the cost function defines the weight and bias values in the learning model. Fig. 2 graphically demonstrates an example of binary classification problem. Data samples are classified into two classes -1 and +1 with a formed hyperplane by SVM model. The cost function is set o to define the margin distances between the support vectors.

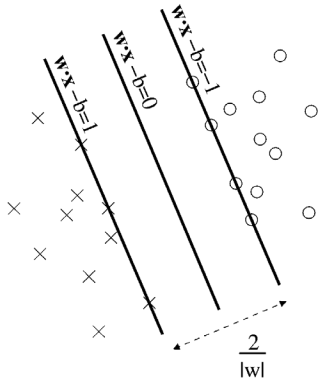

Fig. 2. SVM methodology on the binary classification.

The minimum response/value of the cost function provides the best position of the hyperplane. The penalty parameter is another variable used in the calculation of the cost function. Model flexibility in the formulation can be adjusted with penalty parameter given by the user. The large values of penalty parameter make the model stricter, and it ends with more misclassification errors. On the contrary, with the small values, the model becomes loose and, therefore classifies some outliers as well [29].

SVM maintains a binary classification of two-class datasets. In order to use SVM in multiclass structures, “one against one” or “one against all” are the most popular strategies in literature. Each strategy has own advantages and disadvantages mentioned in [30]. In order to define well-fitted settings of SVM on sleep stage classification problem, different penalty (1, 10, 100, 250, 500), kernels (radial-basis, Polynomial, quadratic, linear) and its parameters are tested at the initial part of study and registered in Table 3. According to the accuracies of parameter testing, two different SVM model with polynomial and RBF kernels are included in the study. Polynomial kernel selected as 3th degree of equation and RBF kernel fixed with sigma ‘1’. Penalty parameters are defined as 25 and 100 respectively. Additionally, one against one strategy is used for evaluation between classes owing to 6 classes’ presence in sleep stage datasets.

Table 3. SVM parameter test results

	Penalty Parameters						
Kernels	1	10	**25**	50	**100**	200	500
Linear	64.95	67.17	67.62	68.12	67.86	67.79	67.80
Quadratic	71.19	74.82	75.57	75.90	76.58	75.87	76.83
RBF (σ) 0.1	61.12	60.17	61.05	60.7	61.5	59.95	60.40
RBF (σ) 0.5	72.53	74.96	76.14	75.77	76.17	75.32	75.96
RBF (σ) **1**	**69.58**	**75.46**	**76.62**	**77.11**	**77.23**	**76.79**	**77.05**
RBF (σ) 5	58.21	66.13	67.37	68.73	69.81	71.34	72.71
RBF (σ) 9	50.8	61.49	63.79	65.71	66.74	68.75	69.94
Poly (d) 2	71.19	74.82	75.57	75.91	76.58	75.87	75.83
Poly (d) **3**	**75.49**	**76.58**	**77.57**	**76.01**	**76.68**	**75.21**	**75.30**

3.4.1.2. Artificial Neural Network (MLP)

The idea of Artificial Neural Network in machine learning is same as in biological concept of central nervous systems. In biological meaning, neural synapses are connected with each other and transmit each sense to the brain to feel and understand the sense. This transmission can be increased by the power and variety of the sense. Moreover, some senses can be transmitted by some predefined path in order to react quickly to the sense. Similar to this definition, each sample is considered as sense to be classified in machine learning terminology, and parameters (bias, weight) are the impact factors of the samples showing the importance of senses. Each neuron has interconnections to other neurons to provide the transmission of the information. It is maintained by different mathematical functions owing to distributions of the samples (similar to different senses uses various transmission path) between layers which is the group of neurons. Number of the neurons and layers defines the complexity of the network.

MLP (Multi-Layer Perceptron) is the advanced version of ANN. Minimum two layers connected with two functions should be utilized. Different parameters and functions are tested at initial studies in order to define best settings of MLP network for sleep stage classification. According to results, MLP network as in Fig. 3 is considered for experimental tests with hyperbolic tangent activation function in the hidden layer. Weight and Bias are fixed with 0.8 and 1, respectively. Total 100 neurons were utilized in the hidden layer.

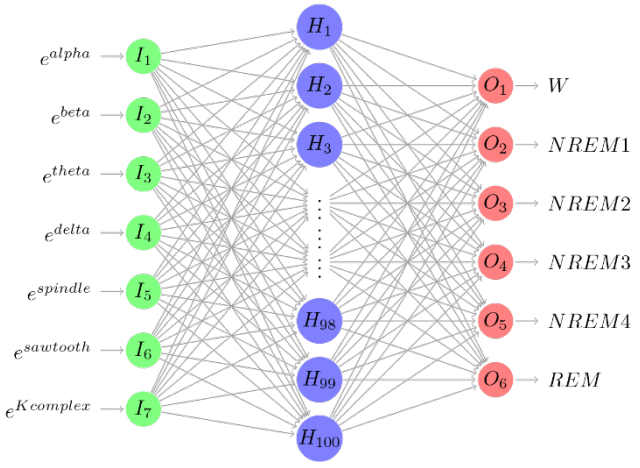

Fig. 3. MLP network for Sleep Stage Classification

3.4.1.3. Naïve Bayes Classifier (NBC)

Naive Bayes is a kind of probabilistic approaches in machine learning concept using modified Bayes theorem. Generally, in probabilistic classification, it is maintained based on the sample distributions, and samples aren’t strictly assigned into the classes. Models give the probabilities of samples over set of classes instead

of single class. Bayes theorem uses all probabilities of the features, but Naive Bayes assumes that features are independent each other. In this sense, algorithm can be performed with less computational cost rather than regular probabilistic methods. Naive Bayes result will provide us to see the probabilistic classifier success on Sleep Stage Classification problem within this study.

3.4.1.4. Linear Discriminative Classifier (LDC)

Linear classification is major issue in the machine learning literature. In this study, linear discriminative classifier realizes simple classification using only covariance matrices. Obtained model forms a multivariate normal density to each group derived from training set and estimates testing samples' labels with calculated covariance with estimations [31]. Basic linear classification is tested in order to demonstrate effects of a simple linear model on the sleep stage classification besides complex methods.

3.4.1.5. K-Nearest Neighbour (KNN)

KNN is a benchmark method in many classification problems in the literature due to the high accuracy results and easy to implement. As a short explanation of the KNN, samples are classified based on predefined *K* labels of the nearest neighbors. In the testing stage of the algorithm, new samples from test set are assigned to the classes according to the closest *K* number of *K* samples' class label in training set with majority voting technique. The *K* value is the key determinant parameter in the definition of class labels. In this study, *K* value is set to 5 according to preliminary studies on parameter selection of *K*. Distance metric determined as "*Euclidean*" algorithm within tested other distance metrics (Euclidean, Cityblock, Chebychev, Minkowski, Mahalanobis, and Cosine). Table 4 shows the accuracies of other *K* values with different distance metrics.

Table 4. Accuracy Results of KNN

		K					
		3	**5**	7	9	11	13
METRICS	**Euclidean**	**82.69**	**83.29**	**83.27**	**83.15**	**82.79**	**82.62**
	Cityblock	82.33	82.89	82.81	82.76	82.68	82.46
	Chebychev	81.00	81.39	81.39	81.11	80.99	80.76
	Minkowski	81.85	82.4	82.52	82.34	82.12	81.85
	Mahalanobis	81.76	82.38	82.43	82.45	82.21	81.97
	Cosine	79.94	80.67	80.75	80.55	80.48	80.21

3.4.1.6. Decision Tree (DT)

Decision Tree is known as rule based machine-learning method. Basically, it works based on tree terminology. The path from root to leaf presents classification rules. The roots represent the most informative features and the leaves indicate the labels. Information gain (IG) is the rule defining criteria. The most widely used methods are entropy, twoing, and Gini to calculate the IG.

Decision Tree is easy to implement similar to KNN. Additionally, interpretation of the classification is much easier than other methods and, it can be useful for some regression problems. However, DT produces low performance on large scale data sets with few training samples compare to SVM [32]. Furthermore, the pruning process is another obstacle point to avoid from over-fitting. According the results of preliminary studies on parameter settings, DT model was modified with pruning functionality and Gini's Diversity Index for IG.

3.4.2. Ensemble Learning Methods

Ensemble learning methods are evolved from the principles of conventional machine learning concepts. The key point of the ensemble learning relies on the proper combination of several machine learning algorithms. Not only one method as in conventional methods, many learners contribute to decision step of classification in ensemble methods, therefore it provides higher success. Machine learning classifiers such as decision trees, Bayes classifiers, KNN, etc. is called base learners or weak classifier in ensemble models. Three ensemble models, which have different base learner combinations and/or sample selection strategies, are implemented in this study. Majority voting is used to define final decision of base learners.

3.4.2.1. Random Forest (RF) - (DT + Bagging)

Random Forest is combination of multiple decision trees with bagging sample selection strategy. Bagging is shortened form of bootstrap aggregation, which is a way for improving the quality of estimates by the aid of well-formed train samples. It is also cited as re-sampling. The main strategy underlying the bagging is to distort the data set by re-sampling, and to train weak learners using re-sampled training sets. The distortion of the samples is carried out with a voting process of weight parameters. The weights of the samples are defined equally in bagging, therefore, train sets are generated by random selection. As a result of bagging, different samples are selected in train set iteratively. Process helps to enhance the diversity of the samples' distribution. The average of the each decision of base learners determines the final decision. More information about RF can be found in [33].

RF is commonly used by many studies in literature because of fast computation time, high accuracy, easy to handle with noise and over fitting problems. Various number of decision tree combination from 10 to 1000 is tested over sleep stage dataset in order to define best parameter settings. According to results, 200 decision tree combination is dedicated to use in experimental tests.

3.4.2.2. Adaptive Boosting - (DT + Adaboost)

Boosting is another technique similar to bootstrap. The difference between boosting and bootstrap is at the re-sampling step. Bootstrap ignores the weight values of the samples and it re-samples randomly, however boosting technique defines different weights for each samples after first iteration. At the end of the first step, the probabilities of misclassified samples are boosted for the second step, and subsequent classifiers are trained. Likewise, other steps are sustained with different weight parameters defined by technique. Readers are referred to an essential guide [34] for boosting theorem in literature.

Adaboost is abbreviation of adaptive boosting which mainly outperforms other regular boosting techniques and, more robust for over-fitting problem. However, it is still easily affected by noise in data and outliers. In this study, the same ensemble model structure in RF strategy is used to assess the effects of Adaboost re-sampling over the sleep stage classification (200 DTs combination).

3.4.2.3. Random Subspace (RSS) - (KNN + Bagging)

RSS is a generalized form of the RF algorithm. RF is composed of decision tree ensembles whereas RSS can be derived from any other classifiers. In this study, KNN classifiers are used in RSS as base learners. The identical number of base learners similar to other ensemble models are utilized in order to demonstrate the effect of re-sampling on regular KNN methods in terms of Ensemble concept.

3.4.2.4. Ensemble SVM - (SVM + Bagging)

SVM is already explained in previous sub-section, but regular SVM uses random sample selection within the concept of binary classification. However, this study aim to present comparative results, hence, regular SVM is modified with bagging process to indicate the effect of ensemble theory. Polynomial kernel SVM is only adapted with Bagging re-sampling and combination theory. Same parameters are arranged for base learners in ensemble SVM model (25 for penalty parameter and polynomial kernel having 3th degree of equation). More details can be found in [35].

4. Evaluation Metrics and Testing Results

Kappa, Accuracy, F-measure, sensitivity (recall) and precision values are considered as performance measurements in this study. Brief information about evaluators is provided in the following sub-sections.

4.1. Evaluation Metrics

Generally, performance metrics is derived from confusion matrix which is an essential table to summarize all classification results under four notations as in Fig. 4. True Positive (TP) shows the relevant samples classified correctly in desired class by model, whereas wrong grouped samples are gathered under False Positive (FP), in other words Type I Error. False Negative (FN), indicates the samples misclassified in desired class which is also referred to as Type II error as well. The last notation is True Negative (TN) which is about true classification for undesired samples in undesired classes. The higher classification performance can be gathered with higher scores in TP, TN and lower samples in FN, FP together.

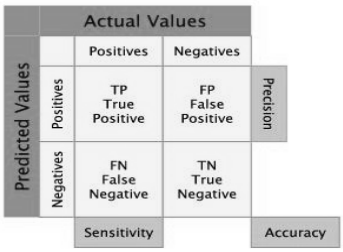

Fig. 4. Notations in the confusion matrix form

Accuracy is the key benchmark metric for any classification. It signifies the percentage of the correctly classified samples within all testing set by using (3). An accuracy of 100% shows the given samples in test set is all correctly classified. However, higher accuracies does not mean the success of the model entirely. In the terminology in literature, accuracy paradox [36] tells that all distribution of the confusion matrix is important to evaluate the model success, however accuracy only indicates the true classified samples. Other metrics are also given in the studies in order to prove complete model success.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

Sensitivity (SE) and precision (PR) are accepted as other performance metrics in order to evaluate the model in terms of Type I and Type II errors. Sensitivity and precision can be calculated by (4) and (5), respectively.

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

F-Measure (F1-Score) can be derived from sensitivity and precision measures as in (6). It reaches to the best value at 1 and worst score at 0. F-Measure values are more reliable than accuracy rates due to inclusion of the FP and FN in the results.

$$F\text{-}Measure = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{6}$$

Cohen's Kappa (κ) is another performance metric, commonly used in many statistical problems [37]. It mainly assesses the inter-rater agreement which covers the similarity of the raters to each other. κ statistics reveals more informative results due to taking into account the prior probabilities than other metrics. Also in some cases having similar accuracy values but different confusion matrix, κ gives more reliable information about the success of the learning model. It also evaluates the raters. κ score is calculated with using (7).

$$K_j = \frac{P_A - P_{C_j}}{1 - P_{C_j}} \tag{7}$$

where P_A represents the proportion of observed values and $P_C$demonstrates proportion of real values derivedfrom confusion matrices. P_A and P_C will be generated from (8) and (9) in multi-class models with two raters.

$$P_A = \frac{P_{11} + P_{22} + P_{33} + \cdots + P_{66}}{P_{NN}} \tag{8}$$

$$P_{C_j} = \left(\frac{P_{PC_j} * P_{AC_j}}{P_{NN^2}}\right) + \left(\left[1 - \frac{P_{PC_j}}{P_{NN}}\right] * \left[1 - \frac{P_{AC_j}}{P_{NN}}\right]\right) \tag{9}$$

j is the total number of classes and P_{PC_j} and P_{AC_j} are abbreviated as predicted and actual values for j_{th} class respectively. In sleep stage classification case, six stages are defined under R&K rules. Two raters are considered as actual hypnogram and predicted results. The Kappa schema for the sleep-stage classification can be seen on Table 5. Kappa score for each class is calculated based on this schema with referred formulas (7, 8, and 9).

Table 5. Confusion matrix schema used in calculation of κ

		Predicted Values by Classifier						
		W	NREM1	NREM2	NREM3	NREM4	REM	Total
Actual Values	W	P_{11}	P_{21}	P_{31}	P_{41}	P_{51}	P_{61}	P_{PC1}
	NREM1	P_{12}	P_{22}	P_{32}	P_{42}	P_{52}	P_{62}	P_{PC2}
	NREM2	P_{13}	P_{23}	P_{33}	P_{43}	P_{53}	P_{63}	P_{PC3}
	NREM3	P_{14}	P_{24}	P_{34}	P_{44}	P_{54}	P_{64}	P_{PC4}
	NREM4	P_{15}	P_{25}	P_{35}	P_{45}	P_{55}	P_{65}	P_{PC5}
	REM	P_{16}	P_{26}	P_{36}	P_{46}	P_{56}	P_{66}	P_{PC6}
	Total	P_{AC1}	P_{AC2}	P_{AC3}	P_{AC4}	P_{AC5}	P_{AC6}	P_{NN}

Table 6. Individual and average accuracies for subject specified scenario

	Data sets	Methods																	
		Conventional Machine Learning										Ensemble Learning							
		NB		KNN		LDC		NN		DT		SVM		KNNs+Bagging		DTs+Adaboost		DTs+Bagging		SVMs+Bagging	
		Acc	Std	Acc	Std	Acc	Std	Acc	Std	Acc	Std	Acc	Std	Acc	Std	Acc	Std	Acc	Std	Acc	Std
Group I	sc4002e0	85.52	0.65	90.61	0.42	83.68	2.21	91.65	0.55	89.12	0.78	85.31	0.85	71.43	0.79	86.93	0.49	92.11	0.35	93.02	0.27
	sc4012e0	71.99	1.34	86.93	0.62	65.21	2.11	84.21	1.82	84.65	0.87	82.16	1.27	68.99	0.95	80.74	1.83	88.43	0.99	89.38	0.39
	sc4102e0	79.57	1.34	88.62	0.66	77.35	1.24	86.31	1.17	85.78	1.05	86.72	0.86	72.41	0.45	80.37	1.36	89.53	0.68	91.99	0.17
	sc4112e0	88.96	0.84	94.86	0.32	92.22	1.02	94.33	1.66	93.62	0.52	93.02	1.17	84.68	0.62	91.86	2.27	95.33	0.31	96.85	0.35
	Average	81.51	1.04	90.26	0.5	79.62	1.65	89.12	1.3	88.29	0.8	86.8	1.04	74.38	0.7	84.98	1.49	91.35	0.58	**92.81**	**0.30**
Group II	st7022j0	70.81	1.38	70.69	1.05	71.57	1.95	67.31	4.54	69.17	2.02	63.31	2.39	55.26	1.66	65.98	0.76	75.44	1.39	76.06	1.67
	st7052j0	75.56	1.28	77.91	1.65	77.28	3.37	57.28	2.31	79.07	0.96	74.61	3.62	60.44	0.89	68.97	3.19	85.99	1.1	86.81	1.24
	st7121j0	75.91	1.06	75.78	1.65	67.47	2.03	75.3	2.55	74.87	1.37	70.12	1.33	61.04	1.58	72.32	0.66	81.06	0.72	81.57	1.12
	st7132j0	70.94	1.47	74.79	1.37	69.55	2.27	72.65	1.49	72.51	1.91	69.2	1.67	64.2	2	69.38	1.08	76.91	1.68	77.10	1.23
	Average	73.3	1.3	74.79	1.43	71.47	2.4	68.14	2.72	73.91	1.56	69.31	2.25	60.24	1.53	69.16	1.42	79.85	1.22	**80.39**	**1.31**
	Total Avg.	77.4	1.17	82.52	0.97	75.54	2.02	78.63	2.01	81.1	1.18	78.06	1.65	67.31	1.12	77.07	1.46	85.6	0.9	**86.60**	**0.80**

General κ scores of the models are derived by using (10). It calculates the averages of each kappa scores corresponding to classes. κ_{avg} scores will be present in comparison tables.

$$\kappa_{avg} = \frac{1}{6}\sum_{j=1}^{6} \kappa_j \tag{10}$$

4.2. The Subject Specific Scenario

The subject specific scenario contains the analyses of the methods on a certain record. Both training and testing samples are selected from the specific record. Models are concurrently trained with predefined number of samples located in selected record with the cross-validation technique. The rest of the samples in the same record are allocated for testing set. Individual and averaged accuracies of methods with relevant group divisions are presented in Table 6. Tests are repeated 10 times to strengthen and generalize the results. Standard deviations (*Std*) occur between repetitive tests because of the sample rotation in testing set with the theory of cross validation. *Std* values are also noted in the Table 6 to show the consistency of the corresponding methods. It is certainly more preferred to have minimum deviation between all tests, hence, methods will be evaluated in this respect as well.

Table 7. The κ results for subject specified scenario

	Methods		Group I	Group II	Averaged
Machine Learning	NB		0.66	0.64	0.65
	KNN		*0.80*	*0.65*	*0.73*
	LDC		0.65	0.62	0.64
	NN		0.78	0.54	0.66
	DT		0.76	0.65	0.70
	SVM		0.74	0.59	0.67
Ensemble Learning	Adaboost	DTs	0.69	0.56	0.62
	Bagging	KNNs	0.38	0.42	0.40
		DTs	0.82	0.72	0.77
		SVMs	**0.83**	**0.76**	**0.79**

Accuracy scores can give an idea about the performance of methods in general, but scores is more stronger and trustable criteria that contains inter-rater comparison as well. In this sense, averaged scores of groups are given in Table 7. Best numerical results for each learning concepts are separately signified with bold and italic numbers. Bold style is used for ensemble, and italic is assigned for conventional methods best case registration.
Additionally, F-measure scores of each methods are demonstrated as bar charts in Fig. 5 and 6 for Group I and II respectively. In this way, more individual and meaningful inferences of each method can be derived from visual demonstrations. Explications about tables and figures will be made in Discussion section.

4.3. The Subject Independent Scenario

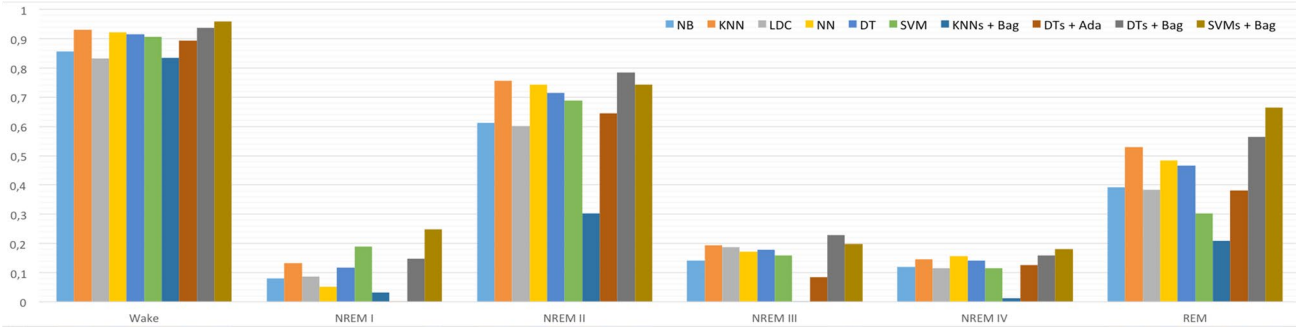

Fig. 5. F-Measure rates for Group I sleep stages (Subject Specified Scenario)

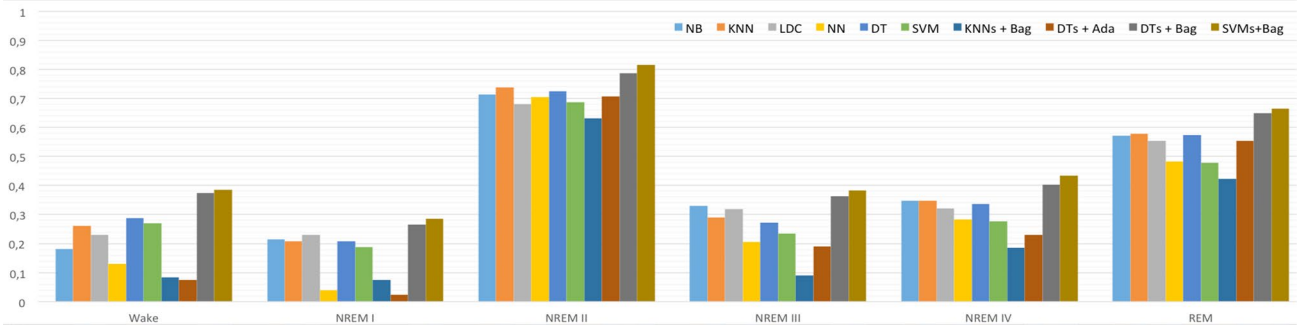

Fig. 6. F-Measure rates for Group II sleep stages (Subject Specified Scenario)

Table 8. Individual and average accuracies and κ scores for subject independent scenario

Data sets		Methods																			
		Conventional Machine Learning												Ensemble Learning							
		NB		KNN		LDC		NN		DT		SVM		KNNs+Bagging		DTs+Adaboost		DTs+Bagging		SVMs+Bagging	
		Acc.	κ	Acc.	κ	Acc.	κ	Acc.	κ	Acc.	κ	Acc.	κ	Acc.	κ	Acc.	κ	Acc.	κ	Acc.	κ
Group I	sc4002e0	70.16	0.33	74.62	0.34	61.00	0.22	**77.37**	0.49	68.03	0.34	67.86	0.31	70.91	0.19	65.87	0.08	75.02	0.37	76.32	0.34
	sc4012e0	67.90	0.47	74.75	0.36	40.50	0.18	67.97	0.26	70.14	0.35	67.40	0.35	67.54	0.18	65.30	0.05	**75.98**	0.36	75.02	0.38
	sc4102e0	69.47	0.40	**77.80**	**0.55**	55.36	0.37	75.07	**0.56**	64.36	0.33	59.98	0.34	71.04	0.19	70.44	0.23	70.60	0.39	74.05	0.41
	sc4112e0	70.64	0.37	70.60	0.38	58.08	0.33	**85.86**	**0.64**	77.83	0.48	75.85	0.48	70.47	0.36	75.28	0.46	79.67	**0.64**	78.17	**0.67**
	Average	69.54	0.39	74.44	0.41	53.73	0.28	**76.57**	0.49	70.09	0.38	67.77	0.36	69.99	0.23	69.22	0.20	75.31	0.44	75.89	0.45
Group II	st7022j0	55.24	0.41	44.97	0.20	40.21	0.25	51.01	0.34	47.09	0.30	41.59	0.25	47.31	0.16	50.47	0.11	62.15	0.39	**63.60**	0.41
	st7052j0	23.50	0.03	27.52	0.10	23.60	0.03	25.40	0.06	31.78	0.15	23.12	0.11	29.52	0.06	28.49	0.07	30.30	0.13	**32.38**	0.15
	st7121j0	64.91	**0.57**	51.01	0.28	47.66	0.33	64.32	0.44	59.16	0.40	39.08	0.18	44.89	0.22	50.14	0.27	62.59	0.49	**70.57**	**0.50**
	st7132j0	39.55	0.23	48.96	0.22	37.32	0.23	48.24	0.27	42.72	0.22	24.53	0.14	47.27	0.20	37.62	0.12	53.28	0.28	**54.78**	0.30
	Average	45.80	0.31	43.11	0.20	37.20	0.21	47.24	0.28	45.19	0.27	32.08	0.15	42.25	0.16	41.68	0.14	52.08	0.32	**55.33**	0.34
Total Avg.		57.67	0.35	58.78	0.30	45.47	0.24	61.91	0.38	57.64	0.32	49.93	0.26	56.12	0.19	55.45	0.17	63.70	0.38	**65.61**	0.40

In the subject independent scenario, models are trained with three records whereas other record in the same group is reserved for the testing. This selection method is referred as *'leave-one-out* cross-validation' in literature. The goal of this scenario is to evaluate the success of the model on classification of unseen samples, which is likely to be encountered in clinics and hospitals. Results are presented in similar forms as in the subject specific tests. Only one difference can be seen that standard deviation does not occur for this scenario, because there is no sample rotation in testing set. Individual and averaged accuracies with scores are recorded in Table 8. It can be derived from the table that the subject independent scenario is obviously more challenging than the subject specific scenario because of the relatively low accuracy and κ scores with the same configuration of the handled methods.

F-measure scores are presented in Fig. 7 and 8 for Group I and Group II respectively. Figures indicate individual performances of classifiers over each sleep stages. In other words, methods can be analyzed with more detailed based on sleep stages. The main challenge is the individual differences of the patients in this scenario. Additionally, different artifacts can be occurred while recording the signals with several kind of noises. All features are normalized at the pre-processing step in order to scale the signals in a standard form and overcome outlier problems. Otherwise, some records can be incompatible or inconsistent with each other. However, remaining outliers induce under-training in model learning significantly, and as a result of that evaluation metrics produce relatively lower results than the subject specified scenario. On the other hand, the subject independent tests are more important than the subject specified tests, because the key idea behind the subject independent scenario is to provide the results in a patient-free system. System can be trained by previously retrieved healthy and unhealthy records to build a model, then the diagnosis of unknown case can be made based on predefined criteria. In this sense, the subject independent tests are more beneficial in current computer aided diagnosis systems which mainly aim to give diagnose directly. However, the subject specific tests depend on the long term diagnoses of specific patients. The variations in the conditions of patient during the treatment can give a trace about the diagnosis in the specific scenario.

5. Discussions

Ensemble SVM with a bagging resampling idea surpasses over all other methods in overall accuracy according to Table 6. Another ensemble method, Random Forest (DTs combination with

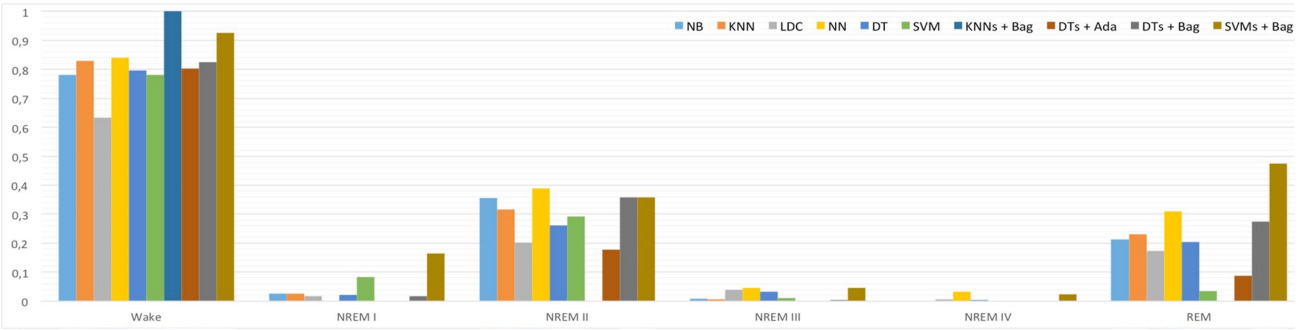

Fig. 7. F-Measure rates for Group I sleep stages (Subject Independent Scenario)

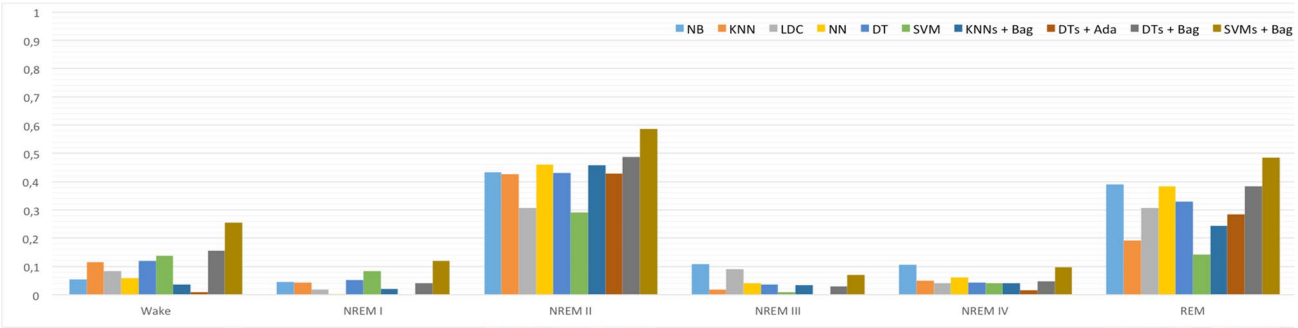

Fig. 8. F-Measure rates for Group II sleep stages (Subject Independent Scenario)

Bagging), is measured as second successful method. However, same combination with Adaboost resampling strategy doesn't result in as successful as in Bagging version. Similar to Adaboost and DT combination, also KNN with Bagging ensemble method is not successful and meaningful combination for subject specified sleep stage classification as it can be seen on Table 6 with worst accuracy results of all.

KNN is evaluated as the best accurate classifier within conventional machine learning methods with 82% accuracy. It is also graded as third rank classifier in overall accuracy. However, it is not consistent like RF and ensemble SVM. Several subjects are classified with low accuracies and KNN is evaluated as fourth rank or more in another cases. On the contrary, RF and ensemble SVM are always steady during all subjects' classification. Ensemble SVM classified all individual subjects with the best accuracies and RF comes second. Besides to the best results, another ensemble method, Random Sub Space having KNN classifier with bagging resampling, is also steady in misclassification of all subjects. In both state; successive and failure, results prove that ensemble methods act more stable which makes the algorithms more reliable for implementation on real system design. Controversially, regular methods result with different rankings on each datasets. For example, NB gives better results on some subjects within Group II, whereas KNN or LDC come up in classification of another subjects within same group. This is suspicious aspects of regular methods in sleep stage classification usage.

Another consistency criteria, Standard Deviation (*std*), also emphasize the importance of ensemble SVM or any other ensemble method within this study. Classifiers with minimum standard deviation is more preferable in practical usage. In that meaning, ensemble SVM with the 0.80 deviation is highlighted within all other tested methods. Similar to accuracy sorting of methods, best three methods are same in consistency based on standard deviation; SVM (1st), RF (2nd) and KNN (3th) in subject specified scenario.

As a kernel based method, regular SVM (using random sample selection strategy), is evaluated with less classification accuracies contrary to ensemble version. Similarly, also DT classification results indicates the importance of having several learner instead of using only one learner. When compare to ensemble combination of DTs, accuracies of each subjects classification in DT stay far behind from multi DT combination. It can be derived that the rule-based or even a kernel based method is not useful alone for sleep stage classification based on single EEG channel. When several of them combined with a bagging re-sampling strategy, results are increased noticeably in terms of subject specified scenario.

In some cases, accuracy is not enough criteria for evaluation. Principal problem in accuracy formula is the ignorance of Type I and Type II error in confusion matrixes. In order to resolve that problem, summarized scores of each methods are also presented besides the accuracy table to verify the results. In the theory, formula grades the results as 'accidental (by chance)' or 'not accidental'. Lower results than 0.5 score is submit as 'accidental', and it is advised that method should be avoided for application. This study presents that the classification of KNN combination with bagging re-sampling is directly an accidentally resulted process for subject specified scenario. It is an important criteria for medical science, because accidentally results shouldn't be taken into account in human life. In that respect, ensemble KNN is entirely useless for sleep stage identification. Other methods have different scores which are all over 0.5, but it is better to use the one which nearest to 1. In that meaning, as in accuracy results, ensemble SVM or DT should be utilized for sleep stage classification in subject specified scenario as well.

Fig. 5 and Figure 6 are derived to show methods success on each individual sleep stage classification. Figures are graphically demonstration of F-Measure scores. F-Measure score gives an average of SE and PR rates which solve the Type I and II error in confusion matrix forms. Methods except ensemble version of KNN are mostly gave successful results on Wake, NREM II and REM stages in Group I samples. Figures also prove that ensemble KNN is entirely worthless method in the subject specified scenario, because it is resulted under 10% F-measure score in mostly NREM stages. Additionally, Adaboost in terms of re-sampling method is useless when compare to Bagging on DT combination. NN and LDC classify the stages with the worst scores in terms of regular machine learning concept.

Group II records are more challenging because of the including patients having some mild difficulties in sleeping. This effects the results with less success of methods in classification when compare to Group I. NREM III is the deepest point of sleep, and more distinctive. As a result, all methods give the most successful F-Measure scores at that point. Still, Ensemble KNN with bagging and DT with Adaboost are finalized with the worst scores. Specifically, NN is the worst method within regular machine learning algorithms for Group II.

Another information can be depicted from figures that REM and NREM II stages are more clear and distinctive stages to classify in any condition. Wake is easily classified stage for heathy records but hard to analyze in disorder cases according to tested methods.

Some studies approved combination of NREM I into NREM II and NREM III into NREM IV as one class named NREM I and SWS. In these studies, classification will be made into 4 class. However in this study, all stages are individually observed to emphasize ensemble methods. Obviously, combined approach will increase the success rates.

For an overall classification results without group division in subject specific scenario, Ensemble SVM with bagging resampling idea is the most successful method in terms of accuracy rate of 86.60%. The second promising method is another ensemble method; DTs with bagging. Third one and also the most accurate method within conventional machine learning algorithms is KNN with 82.52% accuracy.

Another tested scenario, subject independent, is considered as a special case for more practical usage be-cause tests are performed on unseen records such as in hospitals. Subject specified tests generally resulted with relatively high accuracies, but it is not at satisfactory level in the subject independent case. Diversity of the records and artifacts during recordings are the main reason of inefficiency. Artifacts and individual differences are tried to be eliminated by normalization and noise removal procedures but some samples still remained as outliers for the models.

Tests are performed with *leave-one-out* cross validation method which focuses on one subject's entire samples in each testing step. The rest of three records are used for training set for learning process of method. As presented in Table 8, similar to subject specified scenario, Group II classification accuracies are lower than Group I and the best obtained accuracies are unstable. Not only specific algorithm, different algorithms concluded with best accuracies, so it is better to analyze results independently based on records instead of using overall success. In that meaning, the maximum accuracy results of each records are separately written with bold numbers in Table 8. Group II records are classified with more consistent and stable than Group I.

NN is the most accurate method for Group I with an overall score

of 76.57%. It achieved its maximum accuracy with the last record in Group I whereas others failed. Additionally, NN is also prosperous for the first record. Ensemble DT with bagging resampling and regular KNN have more power on second and thirty records respectively.

Group II classification results are more determined. SVM is graded as the best model with 53.77 %, and DT with Bagging comes behind with little difference. As an overall conclusion for Group II, all methods give unsuccessful accuracy in subject independent scenario. Additionally, κ scores also promote that outcome.

The κ scores are provided in detailed form rather than summarized table as in subject specified scenario, because each score of the method is important in terms of tested subject record. κ scores over than 0.5 means not accidental results. It gives more meaning to corresponding accuracy rates, otherwise, calculated accuracies are not important because method classified the samples by chance.

Only two records accuracies are acceptable in both groups according to κ scores. Rest of them is accidentally classified. Even if the accuracy rate is high but κ score below than 0.5, it is defined as unsuccessful classification such as in subject 2 with 75.98 % accuracy but 0.36 score in ensemble DT method. The worst classification result obtained by all methods on subject 2 in Group II. As an overall result, the highest score is obtained by Ensemble SVM, but it is not acceptable and needs to be improved because it is lower than 0.5.

Similar to subject specified scenario, Fig 6 and 7 are prepared for subject independent case. As it can be seen on both figures, success of the methods is low in terms of individual stage classification as well. Only the KNN with bagging resampling has an impact on Wake stages, but in other stages, similar to subject specified scenario, KNN with bagging is still useless. Mostly methods classified NREM I and REM stages in subject independent scenario.

It can be derived from all results of subject independent scenario that the tested methods are insufficient and accidental because of the aforementioned complexities and differences.

6. Conclusions

In this study, sleep stages are classified based on single-channel EEG signals. Several prominent ensemble and conventional machine learning methods are tested on a well-known dataset. Comparative results provide to define best method which can be used in an automatic sleep staging system in the future. Furthermore, detailed figures and explanations shed light on the compelling side of stages.

Dataset is divided into two parts as Group I and II according to patient status. Group I records are taken during daily life with an analog modified cassette, thus records within the Group I has more samples and more clear than Group II. On the other hand, Group II records are obtained in hospital during night period with digital recorders which can be affected by other devices in terms of external factors. Effects can be described as artifacts and noises. Additionally, Group II have some mild difficulties in sleeping which causes more complexity in EEG signals.

At the first step, preprocessing is performed on EEG signals in order to remove outliers and noises. In the feature extraction phase, some frequency based characteristic waves are obtained from signals and, subsequently, a set of energy features are derived from these waves as representative features. Normalization process is applied on energy features to scale the ranges of various records of subjects.

In the subject specific scenario, obtained results are in efficient level with 92.81% and 80.39% of averaged accuracy rates for Group I and II, respectively. The highest individual accuracies of each record vary between 76% and 97%. The highest accuracy rates are obtained by multiple SVM combination with bagging resampling in terms of ensemble learning concept. Another ensemble method, DT combination with bagging, is the second prominent classifier. In addition, KNN resulted as the best method within conventional methods whereas stated in 3rd place in overall success list.

Contrary to the subject specified scenario, the subject independent tests are resulted with lower success because of the different patients data used for another patient's sleep stage prediction. Differences in each metabolism and device settings affect the prediction results. As a consequence, the averaged results stay behind the regular specified scenario with an accuracy of 75.89% and 53.77% for Group I and II. Ensemble SVM is still the most robust classifier for Group II whereas algorithms are resulted with various rates for Group I. The highest individual accuracies can be seen between 32% and 86% κ scores are more reliable criteria instead of accuracy rates. κ scores prove that the extra processes still need to be applied to eliminate outliers and noises in order to increase the classification even if it has remarkable accuracy rates.

As a conclusion, the subject independent scenario is not an easy task for computer aided diagnosis. Mainly ensemble SVM and partly some other methods provide better results, but practical problems needs more generalized solution which can be implemented in any condition. Ensemble SVM surpasses other methods in terms of classification accuracy for Group II, but not robust enough according to κ scores. Furthermore, results are inconsistent in Group I. Several methods are finalized with the best accuracies on different recordings instead of only one as in Group II. In order to constitute an automatic sleep staging system, a combined method of those should be optimized or an artificial decision system performed to decide the classification method according to each recording characteristics, for example 3rd recording should be utilized with KNN whereas NN should be used especially for 4th records.

This study is aim to be a source for sleep stage classification based on both ensemble and conventional machine learning algorithms by using single channel EEG. As a future work, some extra preprocessing and feature extraction techniques methods will be investigated especially for Group II data sets and subject independent scenario. Additionally, different machine learning ensembles will be tested for further improvement on the sleep stage classification in order to compose more acceptable solution on a computer aided diagnosis system.

7. Acknowledgment

This research has been supported by Yildiz Technical University, Scientific Research Projects Coordination Department, and Project Number: 2014-04-01-KAP01.

References

[1] M. A. Reimer, W. Flemons,"Quality of life in sleep disorders," Sleep Med Rev, vol.7, no. 4, pp. 335 – 349, Aug. 2003.

[2] F. Zizi, G. Jean-Louis, C. Brown, G. Ogedegbe, C. Boutin-Foster, S. McFarlane, "Sleep duration and the risk of diabetes mellitus: epidemiologic evidence and pathophysiologic insights," Curr. Diab. Rep., vol. 10, no. 1, pp. 43–47, Jan. 2010.

[3] Almazaydeh, L , Elleithy, K , Faezipour, M . "A highly Reliable and Fully Automated Classification System for Sleep Apnea Detection". International Journal of Intelligent Systems and Applications in

Engineering, vol. 4, no. 3,pp. 66-70, 2016.

[4] R. Grunstein, I. Wilcox, T. Yang, Y. Gould, J. Hedner, "Snoring and sleep apnea in men: association with central obesity and hypertension," Int J Obes Relat Metab Disord, vol. 17, no. 9,pp. 533–540, 1993.

[5] A. Rechtschaffen, A. Kales, "A Manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," Public Health Service, U.S. Government Printing Office, Washington, DC, 1968.

[6] T. Hori,Y. Sugita,E. Koga,S. Shirakawa,K. Inoue,S. Uchida, H. Kuwahara, M. Kousaka, T. Kobayashi, Y. Tsuji,M. Terashima, K. Fukuda,N. Fukuda, "Proposed supplements and amendments to the Rechtschaffen & Kales (1968) standard," Psychiatry Clin Neurosci, vol. 55, No. 3, pp. 305 – 310, Jun. 2001.

[7] C. Iber, S. Ancoli-Israel, A. Chesson, S. Quan, "The AASM Manuel for Scoring of Sleep and Associated Events-Rules: Terminology and Technical Specification," American Academy of Sleep Medicine, 2007.

[8] P. Achermann, R. Hartmann, A. Gunzinger, W. Guggenbhl, A.Borbly, "Correlation dimension of the human sleep electroencephalogram: Cyclic changes in the course of the night," Eur J Neurosci, vol. 6, no. 3,pp. 497 – 500, Mar. 1994.

[9] R. Agarwal, J. Gotman, "Computer-assisted sleep staging," IEEE Trans Biomed Eng, vol. 48, no. 12, pp. 1412–1423, Dec. 2001.

[10] E. Oropesa, H. L. Cycon, M. Jobert, "Sleep stage classification using wavelet transform and neural network," International Computer Science Institute, Mar. 1999.

[11] C. Robert, C. Guilpin, A. Limoge, "Review of neural network applications in sleep research," J Neurosci Methods, vol. 79, no.2, pp. 187 – 193, Feb. 1998.

[12] T. Shimada, T. Shiina, Y. Saito, "Sleep stage diagnosis system with neural network analysis," in *Proc.* 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Hong Kong, China, 1998, vol. 4, pp. 2074–2077.

[13] J. Principe, S. Gala, T. Chang, "Sleep staging automaton based on the theory of evidence," IEEE Trans Biomed Eng, vol. 36, no. 5, pp. 503–509, Aug. 2002.

[14] W.-Y. Hsu, C.-C. Lin, M.-S. Ju, Y.N. Sun, "Wavelet-based fractal features with active segment selection: Application to single-trial EEG data," J Neurosci Meth, vol. 163, no. 1, pp. 145–160, Jun. 2007.

[15] H. G. Jo, J. Y. Park, C. K. Lee, S. K. An, S. K. Yoo, "Genetic fuzzy classifier for sleep stage identification," Comput Biol Med, vol.40, no. 7, pp. 629 – 634, Jul. 2010.

[16] J. Virkkala, J. Hasan, A. Varri, S.L. Himanen, K. Muller, "Automatic sleep stage classification using two-channel electrooculography," J Neurosci Methods, vol.166, no. 1, pp. 109-115, Oct. 2007.

[17] M. O. Mendez, M. Matteucci, V. Castronovo, L. Ferini Strambi, S. Cerutti, A. M. Bianchi, "Sleep staging from heart rate variability: time-varying spectral features and Hidden Markov Models," Int J Biomed Eng Technol, vol. 3, no. 3, pp. 246 – 263, 2010

[18] S.F. Liang, C.E. Kuo, Y.H. Hu, Y.S. Cheng, "A rule-based automatic sleep staging method," J Neurosci Methods, vol. 205, no. 1, pp.169 – 176, Mar. 2012.

[19] L. Zoubek, S. Charbonnier, S. Lesecq, A. Buguet, F. Chapotot, "Feature selection for sleep/wake stages classification using data driven methods", Biomed Signal Process Control, vol. 2, no. 2, pp. 171 – 179, Jul 2007.

[20] Y.L. Hsu, Y.T. Yang, J.S. Wang, C.Y. Hsu, "Automatic sleep stage recurrent neural classifier using energy features of EEG signals," Neurocomputing, vol. 104, pp. 105 – 114, Mar. 2013.

[21] B. Koley, D. Dey, "An ensemble system for automatic sleep stage classification using single channel EEG signal", Comput Biol Med, vol. 42, no. 12, pp. 1186 – 1195, Dec. 2012.

[22] B. Kemp, J. Olivan, "European data format plus (edf+), an edf alike standard format for the exchange of physiological data,"Clin Neurophysiol, vol. 114, no. 9, pp. 1755– 1761, Sep. 2003.

[23] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdor , P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.K. Peng, H. E. Stanley, "PhysioBank, Physio Toolkit, and Phys-ioNet: Components of a new research resource for complex physiologic signals," Circulation, vol. 101, no. 23, pp. 215– 220, Jun. 2000.

[24] S. Ng, P. Raveendran, "Comparison of different montages on to EEG classification," in P*roc.* 3rd Kuala Lumpur International Conference on Biomedical Engineering, Kuala Lumpur, Malaysia, 2006, vol. 15, Springer Berlin Heidelberg, pp. 365–368.

[25] T. Kalayci, O. Ozdamar, "Wavelet preprocessing for automated neural network detection of EEG spikes,"IEEE Eng Med Biol Mag, vol. 14, no. 2,pp. 160–166, Apr. 1995.

[26] M. Xiao, H. Yan, J. Song, Y. Yang, X. Yang, "Sleep stages classification based on heart rate variability and random forest," Biomed Signal Process Cont., vol. 8, no. 6, pp.624–633, Nov. 2013.

[27] B. Scholkopf, A. J. Smola, "Learning with Kernels: Support Vector Machines," Regularization, Optimization, and Beyond, MIT Press, 2001.

[28] G. Camps-Valls, L. Bruzzone, "Kernel-based methods for hyperspectral image classification," IEEE Trans Geosci Remote Sens,vol. 43, no. 6, pp.1351–1362, May. 2005.

[29] C. Burges, "A tutorial on support vector machines for pattern recognition," Data Min Knowl Discov, vol. 2, no. 2, pp.121–167, Jun. 1998.

[30] J. Milgram, M. Cheriet, R. Sabourin, "One against one or one against all: Which one is better for handwriting recognition with SVMs?," in *Proc.* Tenth International Workshop on Frontiers in Hand-writing Recognition, La Baule, France, 2006.

[31] S. Motamedi-Fakhr, M. Moshrefi-Torbati, M. Hill, C. M. Hill, P. R. White, "Signal processing techniques applied to human sleep EEG signals; A review," Biomed Signal Process Control, vol. 10, pp. 21 – 33, Mar. 2014.

[32] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, "Classification and Regression Trees," Wadsworth International Group, Belmont, CA, 1984.

[33] L. Breiman, "Random forests," Mach Learn, vol. 45, no. 1, pp. 5– 32, Oct. 2001.

[34] R. Rojas, "Adaboost and the super bowl of classifiers a tutorial introduction to adaptive boosting", Freie University, Berlin, Tech. Rep, 2009.

[35] H.C. Kim, S. Pang, H.M. Je, D. Kim, S.Y. Bang, "Support vector machine ensemble with bagging", in Pattern Recognition with Support Vector Machines. Lecture Notes in Computer Science, vol. 2388, pp. 397–408, Springer, Berlin, Heidelberg.

[36] X. Zue, I. Davidson, "Knowledge discovery and data mining," In-formation Science Reference, 2007.

[37] J. Cohen, "A Coefficient of Agreement for Nominal Scales," Educational and Psychological Measurement, vol. 20, no. 1, pp.37-46, 1960.