ELEC 447: BIOMEDICAL SIGNAL PROCESSING - PROJECT REPORT

Determination of Sleep and Wakefulness with EEG Signals

Yamaç Ömür - 79458
Fatma Oğuz - 76747

1. INTRODUCTION

Signals received from a subject while sleeping, and sleep classification using these signals are very important to understand the health status of the individual and to understand whether the individual has a disease or illness. These signals can be used for purposes such as measuring the total hours of sleep, and determining the subject's sleep quality. Actigraphy and heart rate measurements are used in many of today's smart devices for these purposes. However, these measurements do not provide sufficiently detailed and accurate results. Electroencephalogram (EEG) is the signals obtained from measuring electrical activities in the brain. Since EEG signals are received directly from the brain, they provide more detailed and accurate results. In addition, EEG signals have a low amplitude ranging from 20-100 µV and are measured using electrodes placed on the individual's head [1]. These signals are frequently used to detect sleep-wake status and also to detect different stages of sleep such as REM and deep sleep.

In this project, we aimed to create a model that accurately classifies the sleep-wake state of the individual after analyzing EEG signals and applying different processes. In the project, EEG signal data were taken from PhysioNet's Sleep-EDF database [3], and different biomedical signal processing methods such as feature extraction, dimensionality reduction with PCA, classification with support vector machines (SVM), k-nearest neighbors (kNN), decision trees (DT) were applied to these data. In addition, the Leave-One-Out Cross Validation method was used in the training and testing phase, and then the performance of the model was analyzed by calculating different metrics such as accuracy and sensitivity.

2. METHODS

a. DATA PREPROCESSING

The data available in PhysioNet were obtained by recording the EEG signals of 153 individuals for 20 hours every day for 2 days. In this project, data from 10 individuals were selected to make the model more applicable and practical. In addition, the data in PhysioNet contained the signal records received from individuals divided into 30-second segments and the sampling frequency was 100 Hz. For this reason, we did not apply any extra preprocessing step and focused directly on the feature extraction and model development stages.

b. FEATURE EXTRACTION

In the feature extraction stage of the model, different features of the signal such as maximum, minimum, average, standard deviation, skewness, kurtosis, entropy and non-linear energy were extracted by using the feature table in [2]. In addition, Itakura Distance [4] values of the signals were also calculated in this project. In this method, auto regressive models were represented, a signal of the wakefulness state was taken as a reference and the Itakura Distance between it and other signals was calculated. Values lower than the recommended threshold value of 0.3 indicate alertness; Values higher than this value are classified as drowsiness. In wavelet decomposition, Daubechies 4 (db4) mother wavelet was used and then all the extracted features were combined in a feature matrix covering time and frequency domains.

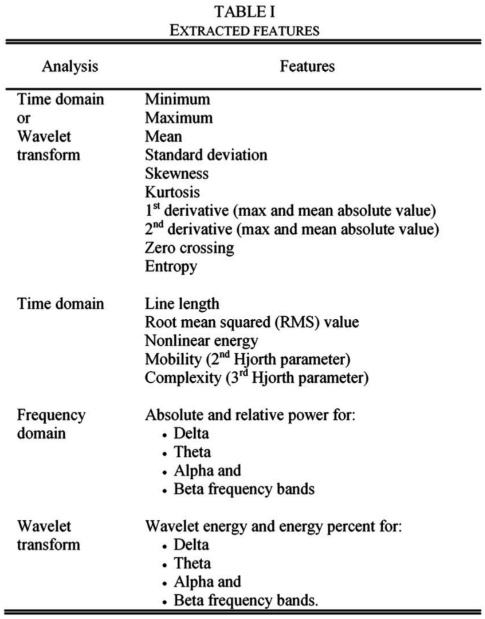


TABLE I
EXTRACTED FEATURES

Analysis	Features
Time domain or Wavelet transform	Minimum Maximum Mean Standard deviation Skewness Kurtosis 1st derivative (max and mean absolute value) 2nd derivative (max and mean absolute value) Zero crossing Entropy
Time domain	Line length Root mean squared (RMS) value Nonlinear energy Mobility (2nd Hjorth parameter) Complexity (3rd Hjorth parameter)
Frequency domain	Absolute and relative power for: • Delta • Theta • Alpha and • Beta frequency bands
Wavelet transform	Wavelet energy and energy percent for: • Delta • Theta • Alpha and • Beta frequency bands.

Figure 1: Feature Table from Article [2]

c. PRINCIPAL COMPONENT ANALYSIS (PCA)

When data sets are high-dimensional and the number of features extracted is quite large, data analysis can be challenging and can lead to computational complexity problems in the model. For this reason, Principal Component Analysis (PCA) was used in this project to eliminate these issues that may arise. Before applying PCA, a normalization process was applied, and then the connection between these features was understood by calculating the covariance matrix of these features. Afterwards, principal components were determined as a result of the analysis of eigenvectors and eigenvalues obtained from the covariance matrix. In addition to these processes, PCA was applied separately to the Fpz-Cz and Pz-Oz channels. As a result of this process, it was seen that for Fpz-Cz, 12 principal components were sufficient and these selected features were sufficient to explain a variance of 90% or more. When applied to the Pz-Oz channel, 11 principal components were found to be sufficient. Thanks to this dimension reduction, the risk of overfitting the model was reduced and it had positive effects such as a reduction in processing time, a more effective analysis, and so on.

d. MACHINE LEARNING

After reducing the number of columns to 12 in the Fpz-Cz and 11 in the Pz-Oz channels, we applied machine learning (ML) on the aggregated feature matrix containing all 10 patients. Using techniques like machine learning in sleep classification can be useful since it can notice a relationship between the features of the signal and the patient's sleep status. As discussed in lectures, supervised learning algorithms generally achieve better results than unsupervised learning algorithms in the context of biomedical signal processing applications. Since our dataset also contained the labels of the signals, we decided to assess the performance of three supervised learning algorithms on our classification task: Support Vector Machine, k-Nearest Neighbors, and Decision Tree. Each algorithm can be explained as follows:

SUPPORT VECTOR MACHINE:

The support vector machine (SVM) classification algorithm aims to find the optimal hyperplane that can maximize the distance between the data points of the label classes. This distance is also called a "margin". Furthermore, for the special case when the classes are not linearly separable as the below figure, engineers can also determine a "kernel function" for the SVM algorithm, that takes a data point and then generally maps it to another high dimensional space, where the classes are linearly separable. This algorithm can be implemented using the MATLAB function "fitcsvm". After trying a polynomial and linear order kernel, we set the 'KernelFunction' parameter of the function as 'linear', because it achieved the best results.

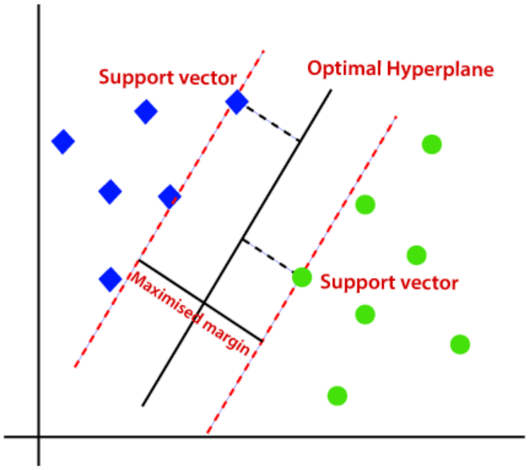


Figure 2: A visual representation of the SVM algorithm.

K-NEAREST NEIGHBORS:

The k-Nearest Neighbors (kNN) algorithm determines a test data's label by finding the training points that have the smallest distance to it, according to a distance metric such as Euclidean or Manhattan. These closest data points are called the test data point's neighbors. "K" is a hyperparameter that should be picked by us. After the top "k" closest neighbors of the data point are found, the algorithm uses majority voting to determine the class of the test data point. This algorithm can be implemented using the MATLAB function "fitcknn".

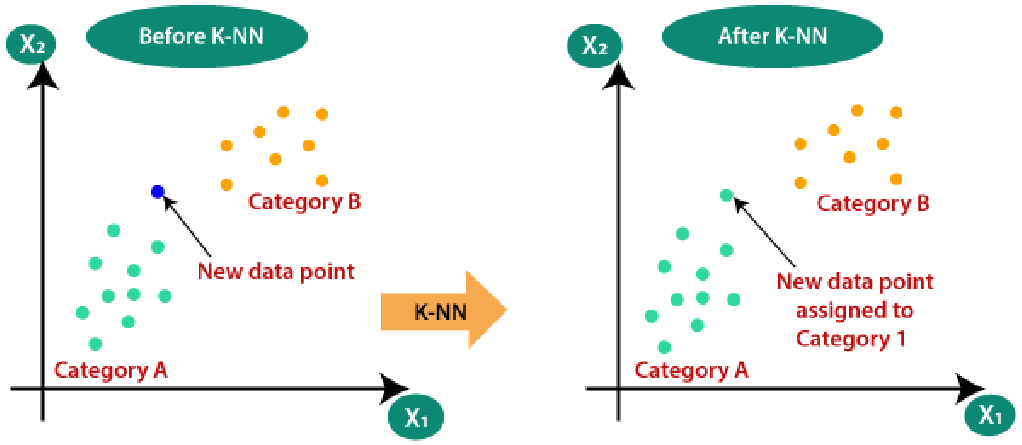
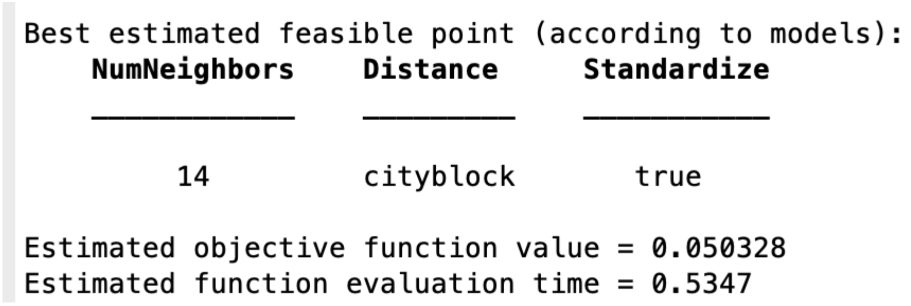


Figure 3: A visual representation of the kNN algorithm.

To find the best possible hyperparameters of the MATLAB function, we first ran the algorithm setting the 'OptimizeHyperparameters' parameter to 'auto', which ran the program by trying all possible hyperparameters and presenting the best ones. Our results were the following:



```
Best estimated feasible point (according to models):
    NumNeighbors    Distance     Standardize
    ____________    _________    ___________

         14         cityblock       true

Estimated objective function value = 0.050328
Estimated function evaluation time = 0.5347
```

Figure 4: The optimal parameters for the Pz-Oz channel data, using leave-one-out cross validation, with the first patient left out.

Although the above parameters were optimal only for the Pz-Oz channel data with the first patient left out for validation, we believed that the algorithm would take too much time if we had waited for it to optimize parameters for all evaluations across all channels. Therefore, we generalized the results and used the fitcknn function with the parameters NumNeighbors = 14, Distance = cityblock, and Standardize = true.

DECISION TREE:

Decision trees are tree-like classification algorithms that begin with a root node, further splitting into branches called "decision nodes". At every node, the algorithm determines a certain feature and certain threshold to split from. The attribute selection aims to minimize the impurity/error of a node, where metrics such as Information Gain or Gini Impurity can be used. This algorithm can be implemented using the MATLAB function "fitctree". We used the default parameters of the function, which uses Gini Impurity as the selection criterion.

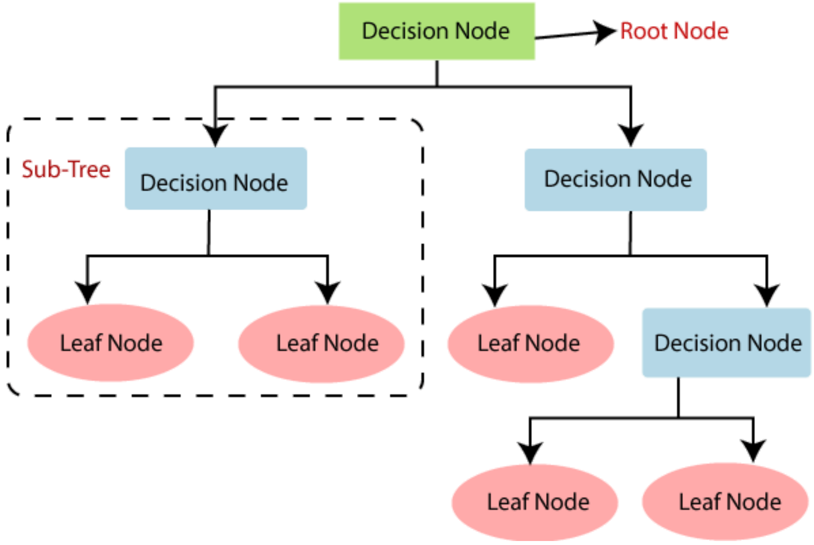


Figure 5: A visual representation of the decision tree algorithm.

3. RESULTS AND DISCUSSION

The training and validation of each ML algorithm was performed using leave-one-subject-out cross validation. We later extracted the average classification metric results of each algorithm for each left out subject. The calculations were done assuming that the positive class is "asleep", and the negative class is "awake". Our results were the following:

Pz-Oz Results	SVM	kNN	DT
Overall Accuracy	93.12%	89.04%	82.79%
Overall Precision	92.65%	84.38%	73.91%
Overall Sensitivity	90.54%	90.16%	86.57%
Overall Specificity	96.30%	90.82%	83.57%

Overall Area Under the ROC Curve	0.93	0.90	0.85

Table 1: The average classification results of each machine learning algorithm for the Pz-Oz channel data.

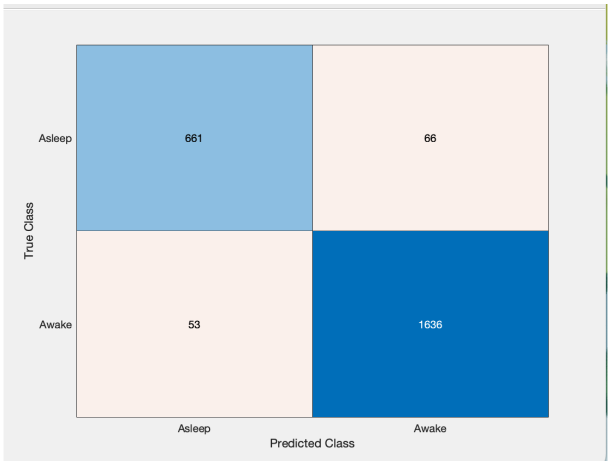
Fpz- Cz Results	SVM	kNN	DT
Overall Accuracy	86.60%	88.68%	83.21%
Overall Precision	84.24%	84.44%	74.67%
Overall Sensitivity	84.29%	84.94%	81.38%
Overall Specificity	90.18%	92.78%	86.55%
Overall Area Under the ROC Curve	0.87	0.89	0.84

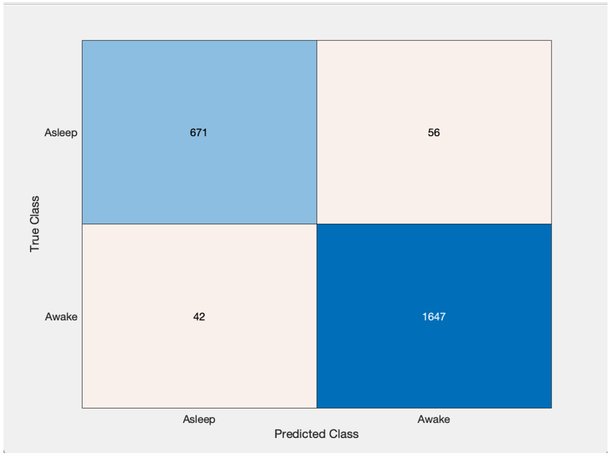
Table 2: The average classification results of each machine learning algorithm for the Fpz-Cz channel data.

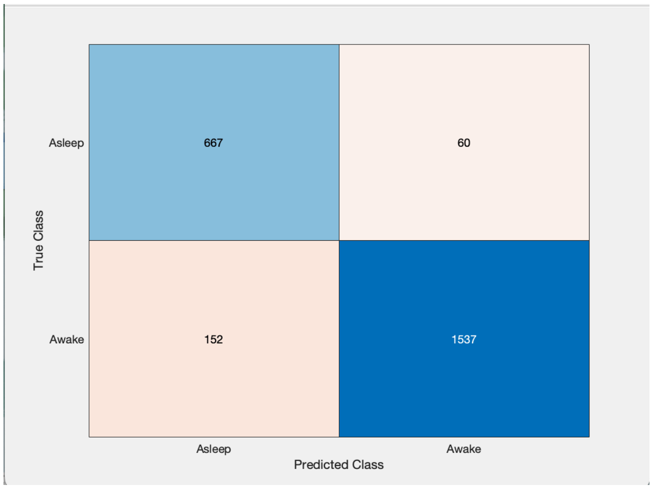
Table 1 and Table 2 show that all three algorithms have an overall high performance across all channels, since their overall area under the ROC curve is all above 0.80. In general, the worst performing algorithm was the Decision Tree. For the Pz-Oz channel, the best performance was achieved by the Support Vector Machine, whereas k-Nearest Neighbors outperformed the other algorithms for the Fpz - Cz channel. In addition, inter-subject variability was also prevalent in the individual subject results, where the third subject would generally receive an accuracy of below 80% whereas the other subjects were mostly receiving over 80% results. Hence, we were able to observe that the sleep status of a patient can be classified with great accuracy using EEG signals and machine learning methods.

Furthermore, we evaluated the confusion matrix of the classification models validated on the first subject for both Fpz-Cz and Pz-Oz channels:

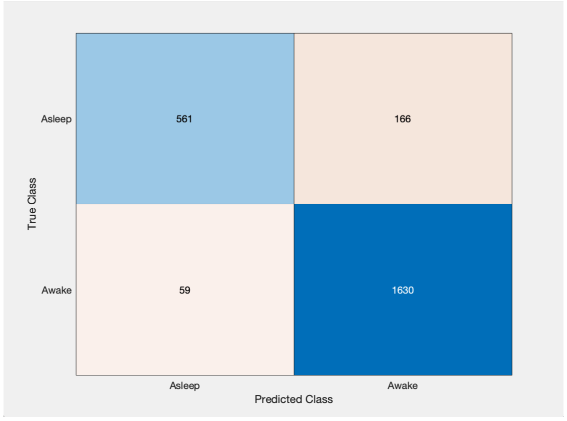
Pz-Oz Confusion Matrices (SVM, kNN, DT in order):

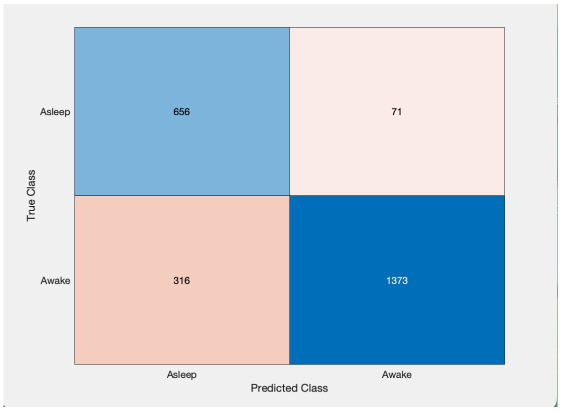


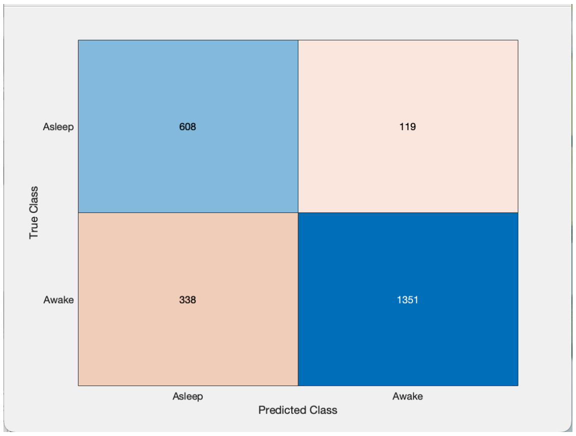




Fpz - Cz Confusion Matrices (SVM, kNN, DT in order):







REFERENCES

[1] B. Semiz, “Lecture 11: EEG,” ELEC447 Biomedical Signal Processing, Koç Üniversitesi, 2024

[2] V. Djordjevic, N. Reljin, V. Gerla, L. Lhotska, and V. Krajca, “Feature Extraction and Classification of EEG Sleep Recordings in Newborns,” in *Proc. 9th Int. Conf. Inf. Technol. Appl. Biomed. (ITAB)*, Larnaca, Cyprus, Nov 5-7, 2009. pp. 1-4, doi: 10.1109/ITAB.2009.5394439.

[3] B. Kemp, *Sleep-EDF Database Expanded*, version 1.0.0, PhysioNet, Published: Oct. 24, 2013. Accessed: Nov. 24, 2024. [Online]. Available: https://www.physionet.org/content/sleep-edfx/1.0.0/.

[4] E. Estrada, H. Nazeran, P. Nava, K. Behbehani, J. Burk, and E. Lucas, “EEG feature extraction for classification of sleep stages,” in *Proc. 26th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC),* San Francisco, CA, USA, Sep. 1-5, 2004. doi: 10.1109/IEMBS.2004.1403125.