Multi-Layer Perceptron for Sleep Stage Classification

Intan Nurma Yulita, Rudi Rosadi, Sri Purwani, Mira Suryani
Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Jl. Raya Bandung – Sumedang Km. 21, Sumedang 45363, Indonesia
intan.nurma@unpad.ac.id

Abstract. Sleep apnea is a sleep disorder that causes decreasing or even stopping of breathing during sleep. One way to detect whether a person has the disorder or not, then it can be done by conducting a sleep test (polysomnogram). Polysomnogram provides overall body activity during sleep. Polysomnogram records every process of breath changes, muscle tension, brain waves, eye movements that occur in sleep from awake to the patient has dreams and finally wakes up. Once polysomnogram is obtained, then the doctor will check it. One of the targets of the analysis conducted is sleep stage classification. It takes a long time if done manually. Therefore, it needs an application that automatically to make classification efficiently. It is the main reason for this research that must be done. Specifically, this research applies Multi-Layer Perceptron (MLP) to classify the sleep stage. The results show that MLP has a higher performance than Naïve Bayes, Bayesian Networks, K-Nearest Neighbours, and Decision Tree.

1. Introduction
The sleep test is a standard test in medicine, as well as radiology, heart recording or blood tests. These test used to provide a diagnosis of disease. The difference is, if all the medical examinations during this time were done when the patient awake, the sleep test done to see the patient's body functions while sleeping. Today's health concept not only recognizes nutrition and exercise as a source of health, but healthy sleep is equally important. Lack of sleep will positively affect the physical body. Just like nutrition, sleep is also important to rest the body. When sleeping, the body will begin in the process of cell renewal and improve muscle. It becomes the time for the brain to synthesize all the information it has received throughout the day. Lack of sleep next will cause fatigue and stress. Also, this sleep deprivation can cause other organ disorders.

Through this sleep test, analysis can not only identify the type of sleep disorder but evaluate the overall body. However, doctors take more than three days to analyze the results. So this research needs to be done to shorten the analysis time and reduce human error rate [1]. Error human rate is an error that occurs in a person due to continued and extended work. The solution offered is a system that automatically checks the sleep recording. Specifically, this research focuses on classifying the sleep stage on sleep recording.

Research related to sleep stage classification has been done. LE Calvet et al. [2] using Naïve Bayes as one component of the proposed method for classification. The result of the combination with Muller C-elements succeeds in becoming an efficient classifier with only a minimum amount of information. If LE Calvet et al. focuses on good results through the development of classifier, then it is different with the research done by C Fu et al. which also focus on feature extraction [3]. The extraction feature
applied is Dynamic time warping while the classifier is Bayesian. Unfortunately, the classification is only done to classify the two stages of awake and sleep.

Sleep stage classification also requires machine generated feature extraction to produce the better performance. Several studies have implemented this mechanism through Deep Belief Networks and Convolutional Neural Networks. The mechanism is performed before the classification stage. To perform this mechanism without separately, the classifier used has the same function as the machine-generated feature extraction. Therefore, the classifier used in this research is Multi-Layer Perceptron (MLP). MLP consists of several layers. This layer is indirectly able to act as a machine-generated feature extraction.

On the other hand, another important thing in the sleep stage classification is the type of signal used. In a study conducted by AR Hassan and MIH Bhuiyan, sleep staging in mice was performed using single channel electroencephalogram (EEG) [4], while V Gao, F Turek, and M Vitaterna combined Electroencephalogram (EEG) and electromyogram (EMG) [5]. In the present study, however, the data used were based on a study by Martin Langkvist et al. using three types of data: Electroencephalogram (EEG), electromyogram (EMG), and Electrooculography (EOG) [6]. So the focus on research is MLP which aims to feature learning and classification of EEG, EMG, and EOG for sleep stage classification. The method is based on non-sequence classification, so that it is different with the concept of Martin Langkvist et al. that apply sequence classification.

2. Study Literature
MLP is a type of feed forward artificial neural network. MLP is a well-known method of stability, ease of use, and a reasonably small structure in solving some problems when compared to other structures [7]. This network will study the relationship between input and output factors by replacing the weight and bias values of the network [8]. The network architecture of MLP consists of an input layer, hidden layer, and output layer. Each layer consists of a processed unit called a neuron [9].

Each neuron will receive an input which will be multiplied by a specific weight and then added with a bias value. Input will be computed using certain mathematical equations. The results of the computation process will be activated using a non-linear activation function. So the output of the hidden layer is as follows:

$$h_j = f \left( \sum_{i=1}^{l} w_{ij} x_i + b_j \right); \text{for } 1 \leq j \leq J$$

(1)

Where $h_j$ is the output of the hidden layer. $x_i$ and $w_{ij}$ represent the inputs and weights of the input in the hidden unit on the first layer. $b_j$ is the bias of hidden units and $f(\cdot)$ is a transfer function. There are several types of activation functions used such as sigmoid function, tanh function, and relu function.

The output of the predicted result on the $k^{th}$ node in the output layer, in this case symbolized $y_k$, can be written as follows

$$y_k = \sum_{j=1}^{l} w_{jk} h_j; \text{for } 1 \leq k \leq K$$

(2)

where $w_{jk}$ represents the weight of the hidden layer at the output layer. Weight changes that occur during the learning process obtained from the multiplication of input, error and learning rate. $k$ represents the number of outputs of neurons. By combining (1) and (2), the MLP can be written as follows:
\[ y_k = \sum_{j=1}^{J} w_{jk} f \left( \sum_{i=1}^{I} w_{ij} x_i + b_j \right) \]  

(3)

\[ y_k = \sum_{j=1}^{J} w_{jk} f \left( \sum_{i=1}^{I} w_{ij} x_i + b_j \right) \]  

(3)

for 1 \leq j \leq J, 1 \leq k \leq K

3. **Methodology**

Implementation of the proposed methods was tested using the dataset to the architecture of the MLP. In particular, the architecture consists of many layers so that learning features can be done before classification.

3.1 **Dataset**

The dataset in this study came from nine patients with sleep apnea at St Vincent University Hospital. Table 1 describes the characteristics of the nine patients. The dataset can be downloaded at https://physionet.org/pn3/ucddb/.

| No | Gender | BMI | SEF | ESC | Time |
|----|--------|-----|-----|-----|------|
| 1  | M      | 33.9| 84  | 16  | 6.2  |
| 2  | M      | 31.8| 81  | 13  | 7.3  |
| 3  | M      | 32.4| 63  | 19  | 6.9  |
| 4  | M      | 30.2| 89  | 3   | 6.7  |
| 5  | M      | 25.1| 90  | 15  | 6.8  |
| 6  | F      | 28.4| 64  | 1   | 6.4  |
| 7  | M      | 31.3| 80  | 19  | 7.7  |
| 8  | M      | 39.3| 92  | 2   | 7.6  |
| 9  | M      | 28.6| 60  | 8   | 7.5  |

Where

- BMI (Body Mass Index) stands for Body Mass Index, which is a measure used to evaluate between height and weight of a person.
- ECS (Epworth Sleepiness Scale) is a scale that represents the sleepiness in the daytime.
- Time is the duration of the test.
- SEF (Sleep Efficiency) is the comparison between the amount of time a person is asleep compared to the amount of time during being in bed.

3.2 **Architecture**

Sleep stage classification performed in this study consists of several stages, as shown in Figure 1.
- **Pre-processing**
  Pre-processing was done choose to choose the type of signal needed. In this study, processing only selects EEG, EMG, and EOG to be tested for sleep stage classification.

- **Hand-crafted feature extraction**
  Three types of signals were extracted to obtain the characteristics of the signals. The results obtained 28 features. The types of features were described in Table 2.

| No | Signals | Features                                      |
|----|---------|-----------------------------------------------|
| 1  | EEG     | Delta, Theta, Alpha, Beta, Gamma by using Fast Fourier Transform (FFT) |
| 2  | EEG     | Delta, Theta, Alpha, Beta, Gamma by using Fast Fourier Transform (FFT) |
| 3  | EOG     | Delta, Theta, Alpha, Beta, Gamma by using Fast Fourier Transform (FFT) |
| 4  | EMG     | The median of the absolute value |
| 5  | EOG     | Eye correlation coefficient |
| 6  | EEG     | Kurtosis |
| 7  | EOG     | Kurtosis |
| 8  | EMG     | Kurtosis |
| 9  | EOG     | Standard Deviation |
| 10 | EEG     | Entropy |
| 11 | EOG     | Entropy |
| 12 | EMG     | Entropy |
| 13 | EEG     | Spectral mean |
| 14 | EOG     | Spectral mean |
| 15 | EMG     | Spectral mean |
| 16 | EEG     | Fractal Exponent |

- **Imbalance class handling**
Based on the dataset of only from nine patients, the processed data had an imbalanced class. It means that some sleep stages had more data proportion than any other sleep stage. If this was left, then the classification tent to the dominant class. The process of handling in the study was done through under-sampling, i.e. the amount of data from the dominant sleep stages was discarded so that it had the same number of minor sleep stages.

- Normalization
  At this stage, some features were normalized by rebooting on a particular range.

- Multi-Layer Perceptron
  This study focused on the learning rate and the number of hidden layers used to build models on MLP. Learning rate is a parameter of learning speed to obtain the optimal model, while hidden layers aim as machine-generated feature extraction or feature learning. Through the machine-generated feature extraction or feature learning contained in the MLP, the classification process was expected to be maximal because the final layer processes the optimal attributes in differentiating between sleep stages. The iterations used in the MLP training were 500 times.

- Evaluation
  The evaluation was based on 10-cross-validation where the measurement parameters were accuracy, precision, recall, and F-measure.

4. Experiments
The experiment aims to test the effect of the number of hidden layer and learning rate on MLP. The optimal conditions of the two experiments were then compared with shallow classifiers. Shallow classifier used included Naïve Bayes, Bayesian Networks, Decision Tree, and K-Nearest Neighbour (K = 5).

4.1 Effect Analysis of Hidden Layer Amount
The analysis was done by comparing the four experiments by changing the value of the number of hidden layers, among others:

- The first experiment used the number of layers as much as the number of sleep stages (5)
- The second experiment used the number of layers as much as half of the number of attributes and sleep stages (17)
- The third experiment used the number of layers as many as the number of attributes (28)
- The second experiment used the second number of layers as many as the number of attributes and sleep stages (33).

| Number of hidden layers | Accuracy | Precision | Recall | F-Measure |
|-------------------------|----------|-----------|--------|-----------|
| 5                       | 70.7     | 70.6      | 70.7   | 70.5      |
| 17                      | 77.3     | 77.1      | 77.3   | 77        |
| 28                      | 79.5     | 79.4      | 79.5   | 79.4      |
| 33                      | 79.3     | 79.3      | 79.3   | 79.3      |

Table 3 shows the experimental results by evaluating each of these parameter values against accuracy, precision, recall, and F-Measure. The table describes that five hidden layers had performance above 70%. This performance increased with the added hidden layers, as the second experiment shows. The
addition of 12 layers could improve performance above 7%. If the number of layers was set to same as the number of features, then the performance increased over 2% so that the final performance was above 79%. It indicates that the number of layers makes feature learning better.

The fourth experiment shows that the addition of more hidden layers was not able to improve performance better. It means that 28 hidden layers were the most optimal in the research. Even using 33 hidden layers, it had a slightly smaller performance than the 28 hidden layers.

4.2 Effect Analysis Learning Rate

This experiment was to evaluate the influence of learning rate by doing three experiments. These three experiments were done with different learning rate sizes of 0.1, 0.3, and 0.5, but the number of hidden layers was the same. The number of hidden layers used was 28, as was the best condition in Table 3. Learning rate plays a vital role in the development model of MLP so that the exact parameter value is probably very influential to find the optimal model. Table 4 shows the results of experiments on the learning rate.

| Size of learning rate | Accuracy | Precision | Recall | F-Measure |
|-----------------------|----------|-----------|--------|-----------|
| 0.1                   | 79.4     | 79.3      | 79.4   | 79.3      |
| 0.3                   | 79.5     | 79.4      | 79.5   | 79.4      |
| 0.5                   | 78.1     | 78.1      | 78.1   | 78        |

The best performance was obtained when the learning rate is 0.3. If the value of learning was reduced to 0.1, then the performance was decreased by 0.1%. It shows that a small learning rate hinders the learning process of finding the optimal models. However, the learning rate was too large, as shown by the learning rate = 0.5, it further aggravated the performance of classification. It means that the learning rate was too large, it is too much to make repairs that were not needed.

4.3 Comparison of Multi-Layer Perceptron with Shallow Classifier

This study compared MLP to shallow classifiers to see the performance of MLP success, as shown in Table 5. Shallow selected classifiers included:

- Naïve Bayes (NB) is a simple probabilistic classifier that applies Bayes Theorem with the assumption of independent independence [10].
- Bayesian Networks (BN) is a graphical probabilistic model representing a set of random variables and their conditional dependencies through Directed Acyclic Graph [11].
- Decision Tree (DT) is a classification algorithm formed from three types of vertices namely the root, intermediate, and leaf node [12].
- K-Nearest Neighbour is a method to classify objects based on learning data closest to the object [13]. In the study, the number of neighbours used was five.

| Methods | Accuracy | Precision | Recall | F-Measure |
|---------|----------|-----------|--------|-----------|
| NB      | 60.6     | 62.8      | 60.6   | 59.2      |
| BN      | 68.6     | 69.7      | 68.6   | 68.3      |
| DT      | 73.5     | 73.5      | 73.5   | 73.5      |
| KNN     | 66.7     | 67.5      | 66.7   | 66.5      |
Based on the test results in Table 5, MLP had the highest performance. It shows that feature learning helped MLP to find more optimal classification models. The importance of feature learning was increasingly visible when compared with Naïve Bayes. Naïve Bayes ignored the possible dependence between features, so there was not much exploration of the features performed. It made the Naïve Bayes had the lowest performance in this study. Elimination of independent properties through network structures, as performed by Bayesian Networks shows a significant performance increase of 9% for F-Measure.

On the other side, the construction of the dependence between attributes in the form of trees by Decision Tree showed a higher performance than Bayesian Networks. However, if the dependence between these features were presented in the form of dependence between objects, then its performance would degenerate considerably. It is as shown by KNN's performance, but its value was better than Naïve Bayes which had no dependency at all.

5. Conclusion
Based on the experiments, the MLP in this study achieved the optimal conditions on the use of hidden layers as much as the number of features of the dataset. The use of fewer hidden layers gave worse performance due to lack of learning feature abilities. However, too many hidden layers were not always followed by performance improvements. MLP in this study also depended on the learning rate as shown in Table 4. Learning rates are too small to slow down. If MLP was compared with shallow classifiers, the method got the better result. It can be concluded that the mechanism of MLP can handle the complex problem in sleep data. But the performance is still low. So that, it is still needed the superior method, for example Deep Learning or sequence classifier.

References:
[1] S Gupta P Kumar NC Karmakar SK Palei 2013 Quantification of human Error rate in underground coal mines - A fuzzy mapping and rough set-based approach IEEE Int. Conf. Industrial Eng. Eng. Manag. IEEEM 140-144
[2] LE Calvet JS Friedman D Querlioz P Bessière J Droulez 2016 Sleep stage classification with stochastic Bayesian inference 2016 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH) 117-122
[3] C Fu et al A 2017 Bayesian approach for sleep and wake classification based on dynamic time warping method Multimedia Tools, and Applications 76(17) 17765-17784
[4] AR Hassan MIH Bhuiyan 2016 Computer-aided sleep staging using complete ensemble empirical mode decomposition with adaptive noise and bootstrap aggregating Biomedical Signal Processing and Control 24 1-10
[5] V Gao F Turek M Vitaterna 2016 Multiple classifier systems for automatic sleep scoring in mice Journal of neuroscience methods 264 33-39
[6] M Längkvist L Karlsson and ALoutfi 2014 A review of unsupervised feature learning and deep learning for time-series modeling Pattern Recognition Letters 42 11-24
[7] H Hong et al 2015 Spatial prediction of landslide hazard at the Yihuang area (China): a comparative study on the predictive ability of backpropagation multi-layer perceptron neural networks and radial basis function neural networks, Cartography-Maps Connecting the World Springer International Publishing 175-188
[8] S Mirjalili SM Mirjalili A Lewis2014 Let a biogeography-based optimizer train your multi-layer perceptron, Information Sciences 269 188-209
[9] S Mirjalili 2015 How effective is the Grey Wolf optimizer in training multi-layer perceptrons
[10] TR Patil SS Sherekar 2013 Performance analysis of Naive Bayes and J48 classification algorithm for data classification *International Journal of Computer Science and Applications* 6(2) 256-261

[11] B Cai et al 2014 Multi-source information fusion based fault diagnosis of ground-source heat pump using Bayesian network *Applied energy* 114 1-9

[12] B Pradhan 2013 A comparative study on the predictive ability of the decision tree support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS *Computers & Geosciences* 51 350-365

[13] Y Park et al 2015 Reversed CF: A fast collaborative filtering algorithm using a k-nearest neighbor graph *Expert Systems with Applications* 42(8) (2015) 4022-4028

[14] M Davis H Putnam 1960 A computing procedure for quantification theory *Journal of ACM* 7(3) 201-215