Final report

Qizheng Xia

November 20, 2023

Abstract

Sleep stage is a growing field of interest in medical and neuroscience research. There is scientific consensus on the characteristics of sleep stages related to PSG data. Traditional methods are focused on serval different signal channels, such as Electroencephalogram (EEG), Electrooculography (EOG), Electromyography (EMG), which is time and cost-wasting. In this project, we propose to extract five kinds of features from four EEG channels. Then, we use four kinds of machine learning and deep learning methods to classify the sleep stages based on extracted features. Finally, we evaluate the performance by accuracy and confusion matrix. The best-proposed method reached up to an average of 74% correctly classified stages, and N3, R, and W stages up to over 80%. The developed method supports a good guide to use a few channels and features for sleep stage classification.

1 Introduction

The sleep stage is a complex physiological process that cycles through several stages throughout the night. The sleep stage is usually divided into two categories: rapid eye movement (REM) and non-rapid eye movement sleep (NREM). NREM is further classified as N1, N2, and N3[2]. These sleep stages can be observed and classified by experts based on polysomnography (PSG)[11] including Electroencephalogram (EEG), Electrooculography (EOG), Electromyography (EMG), and other standards. Understanding and studying sleep stages has multiple important applications and significance, such as evaluating sleep quality, and diagnosis of sleep disorders, and it is also relevant to some neurological illnesses, like depression[3]. However, to get the classification of the sleep stage, researchers usually need to observe lots of EEG, EOG, and EMG channels, which is time and cost-wasting. Therefore, the overall objective of this research is to establish machine learning methods to use only 4 EEG channels to get the classification of sleep stages. The work is expected to provide supplement, save cost, time, and effectiveness.

2 Methods

We proposed to use machine learning methods (ML) to analyze EEG signal data. The entire schematic for the proposed method is in Fig 1. To extract the feature from different EGG channels, a number of entropy and distance metrics were utilized. The result was evaluated by classification accuracy and confusion matrix. The experimental dataset employed for validation was introduced in Section 2.1, and method procedures were described in Sections 2.2-2.4.

2.1 Experimental Dataset

The data was collected by the Haaglanden Medisch Centrum in 2018[1, 4]. The whole-night polysomnographic was recorded in 256 Hz and contained 4 EEG channels, 2 EOG channels, 1 ECG channel, and

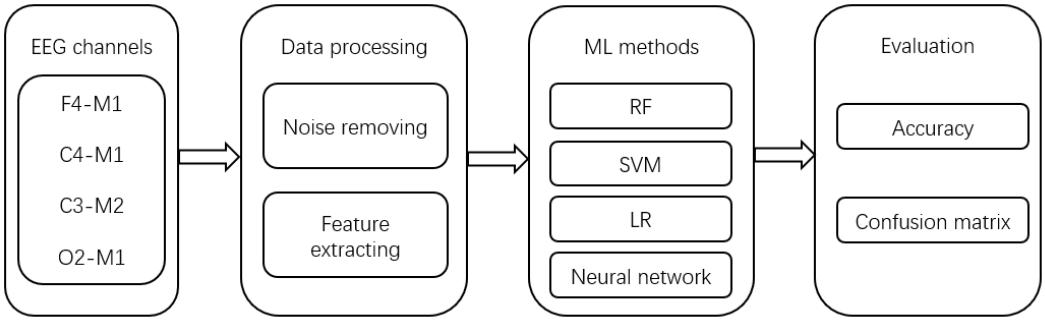

Figure 1: A schematic for the proposed method

1 EMG channel. The sleep stages were recorded in the epochs of 30 seconds of duration and scored manually using Rechtschaffen and Kales (R & K) criteria[12]. And in our project, we will use 5 people and all four EEG channels (F4-M1, C4-M1, O3-M2, C3-M2) to analyze. The detailed information about the sleep records used in this study is in Table 1.

Table 1: Detailed information about the sleep records

Sleep stage	W	R	N1	N2	N3	Total Samples
Count	917(19.76%)	841(18.12%)	486(10.54%)	1782(38.41%)	610(13.14%)	4639

2.2 Data preprocessing

In order to reduce the impact of noise and artifacts, the EEG signals were preprocessed. First of all, all EEG signals were filtered by bandpass with a cut-off frequency of 0.1 Hz and 100 Hz to remove the noise and artifacts. Then cleaning signals were resampled at 100 Hz. In Fig 2 and 3, I provided examples for one whole filtered EEG signal and resampled signal in 5 seconds separately.

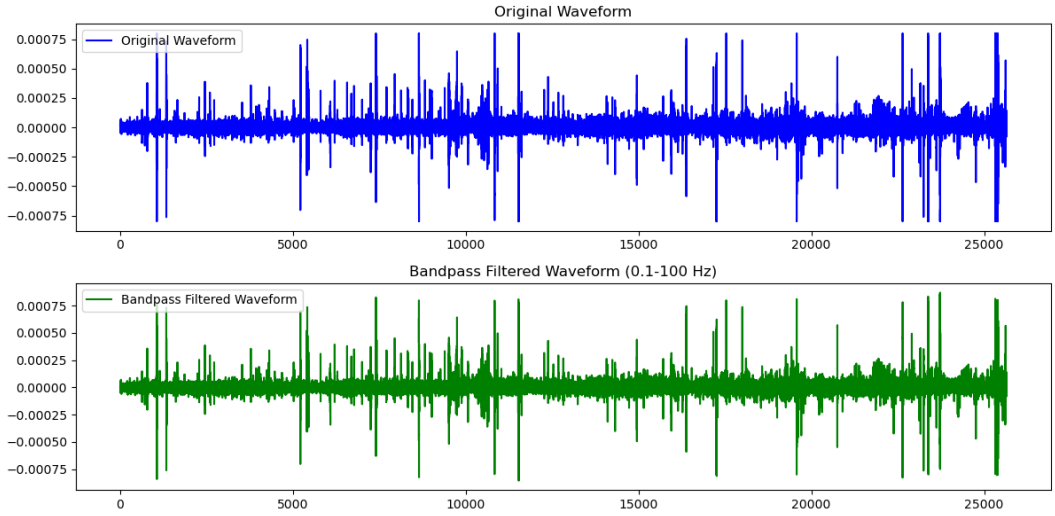

Figure 2: Original signal (up) v.s. Filtered signal(down)

From the plots, we could see that the EEG signal filtered by bandpass was as same as the original signal, which meant that there were few noises and artifacts recorded during sleep. The dataset was a high-quality set.

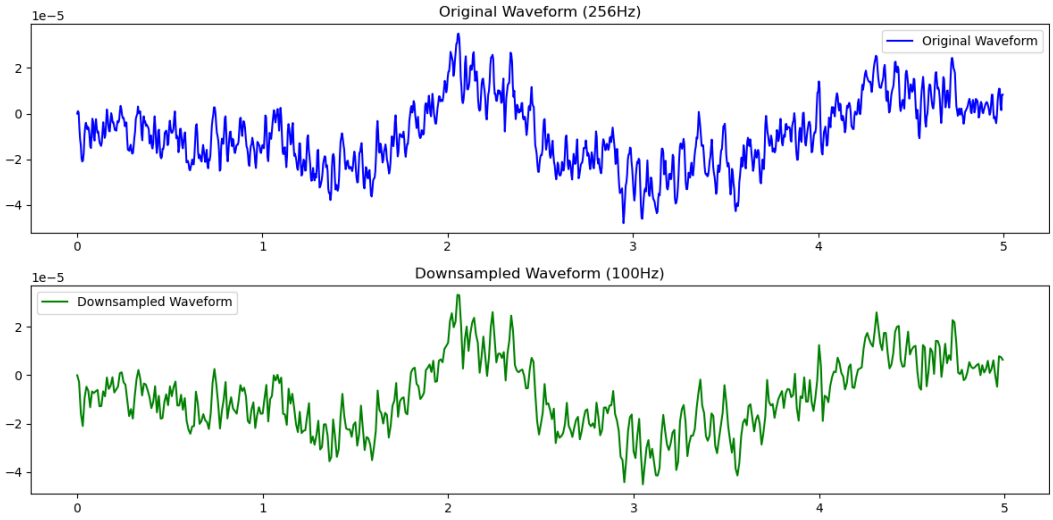

Figure 3: Original signal (up) v.s. resampled signal(down) in 5 seconds

From the step of resampling, we could find that the data were more sparse than the original signal, while the trend of the signal did not change. It met our expectations only diluted the data, but did not lose the information.

2.3 Feature Extraction

The next step was to extract features from signals. Given an input discrete time series $x \in \mathbb{R}^N$ in 30 seconds epoch, so $N = 3000$ $(30s \times 100Hz)$ A feature vector $u \in \mathbb{R}^q$, q is the number of features, and $q \ll N$.

2.3.1 Shannon entropy

Shannon entropy (SE) IS a measure to quantify the uncertainty of information or the average amount of information[6]. The formula of SE is

$$H(X) = -\sum_{i=1} P(x_i) \log_2 P(x_i)$$

where $p(x_i)$ is the probability of $p(x = x_i)$. In this work, we divided 16 equidistant intervals for all EEG data, and $p(x_i)$ is the percentage of each interval. SE was calculated for each channel independently $(u_1 - u_4)$.

In Fig 4, I gave an example boxplot of SE result of the EEG1 (F4-M1) channel. From the plot, we can see that the mean of SE of the wakefulness stage (W) is lowest among 5 sleep others, and the W and N2 stages are highest. It shows that SE is different among sleep stages, especially W and N2 stage.

2.3.2 Approximate Entropy

Approximate Entropy (ApEn) is applied to measure the complexity or unpredictability of time series and the frequency and randomness of patterns in time series[8]. Its formula is

$$\text{ApEn}(m, r, N) = \Phi^m(r) - \Phi^{m+1}(r)$$

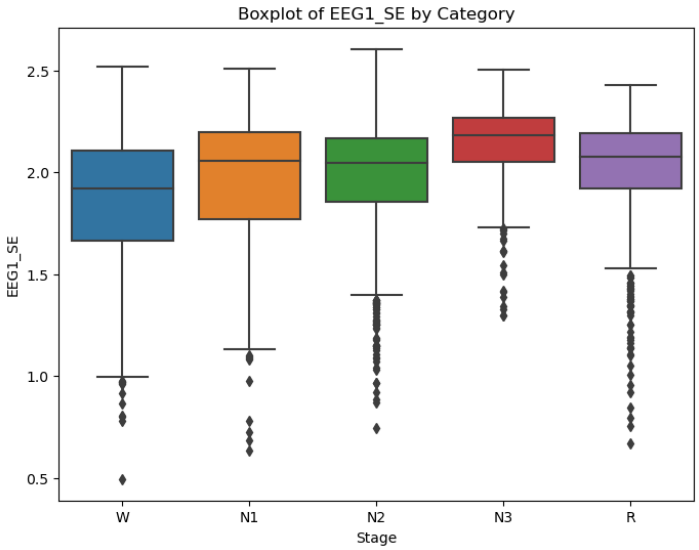

Figure 4: The boxplot of SE of EEG F4-M1 channel

where $\Phi^m(r) = \frac{1}{N-m+1}\sum_{i=1}^{N-m+1}\log C_i^m(r)$, m and r are parameters. We chose $r = 2$ and m being the range of all EEG data in this work. ApEn was computed for all four EEG channels ($u_5 - u_8$). In Fig 5, we can see that the mean of SE of the REM stage (R) is lower than others, and the W and N2 stages are highest. It shows different results with SE. We can conclude that we extract different information from EEG signals.

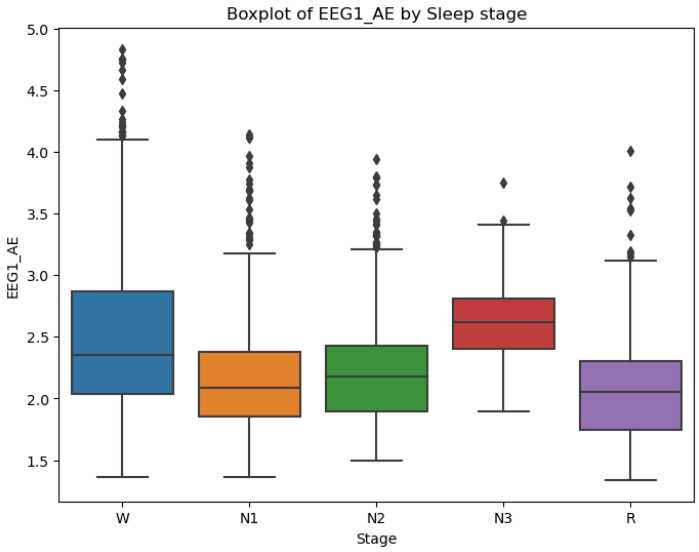

Figure 5: The boxplot of AE of EEG F4-M1 channel

2.3.3 Dispersion entropy

Dispersion entropy (DE) is a measure to quantify the regularity of time series[10]. To calculate DE, firstly we need to map x into z from 1 to c. We use $z_i^c = \text{round}(cf(x_i) + 0.5)$, where $f(x_i) \in [0, 1]$. For embedding vector $z_i^{m,c}$ with embedding dimension m and time delay d, we have $z_i^{m,c} = (z_i^c, z_{i+d}^c, ..., z_{i+(m-1)d}^c)$, and then map it to a dispersion pattern $\pi_{v_0 v_1 ... v_m}$, where $z_i^c = v_0$, $z_{i+d}^c = v_1$,

..., $z^c_{i+(m-1)d} = v_m$. The DE value is calculated as follows

$$\mathrm{DE}(x, m, c, d) = -\sum_{\pi=1}^{c^m} p(\pi_{v_0 v_1 \ldots v_m}) \ln p(\pi_{v_0 v_1 \ldots v_m})$$

where $p(\pi_{v_0 v_1 \ldots v_m}) = \frac{\#\{i | z_i^{m,c} \text{has type } \pi_{v_0 v_1 \ldots v_m}\}}{N-(m-1)d}$. In this work, we chose $d = 1$, $m = 4$, and $c = 6$. DE was calculated for each channel independently ($u_9 - u_{12}$).

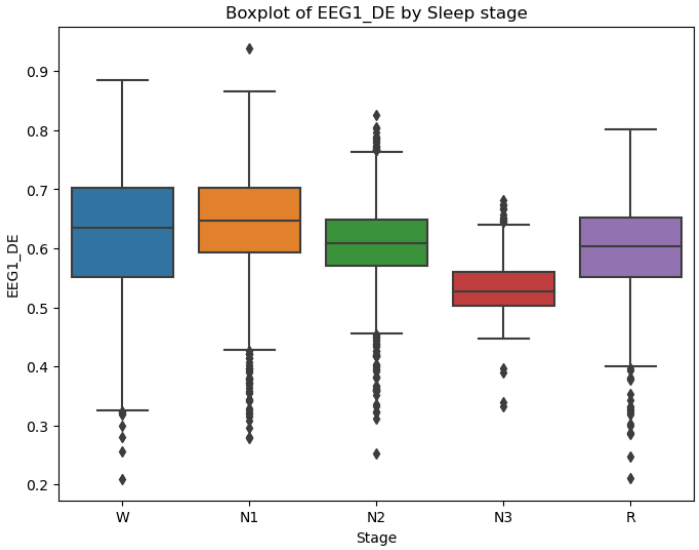

Figure 6: The boxplot of DE of EEG F4-M1 channel

In Fig 6, we can see that the mean of AE of W and N1 are highest, and the W and N3 stage has the lowest DE.

2.3.4 Hurst exponent

Hurst exponent (HE) is an indicator for describing the long-term memory of time series data[9]. If it is close to 0, it indicates that the time series has strong mean regression characteristics, meaning that future trends may be opposite to past trends; if it is close to 0.5, it indicates that the time series is a geometric Brownian motion with unpredictability, similar to a random walk; if it is close to 1, it indicates that the time series has persistence, meaning that future trends may continue past trends. It can be written as

$$\log[R(n)/S(n)] = H \log(n) + c$$

where $S(N) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$, and $R(n) = \max_{1 \le k \le n} \sum_{i=1}^{n}(x_i - \bar{x}) - \min_{1 \le k \le n} \sum_{i=1}^{n}(x_i - \bar{x}), n = 1, 2, ..., N$. HE is the estimated value of the slope, H. HE was computed for all four EEG channels ($u_{13} - u_{16}$).

In Fig 7, we can see that the mean of HE of N3 stage is lower than others, but the difference among other stages is not obvious. It indicates that the long-term memory of N3 performs more anti-persistent action from F4-M1 channel, while other stages perform just like random walks.

2.3.5 Dynamic Time Warping

Dynamic Time Warping (DTW) is a method to measure the similarity between two time series[7]. Given two series $x = (x_1, x_2, ...x_N)$ and $y = (y_1, y_2, ...y_N)$, the matrix $D(i, j)$ is the optimal path

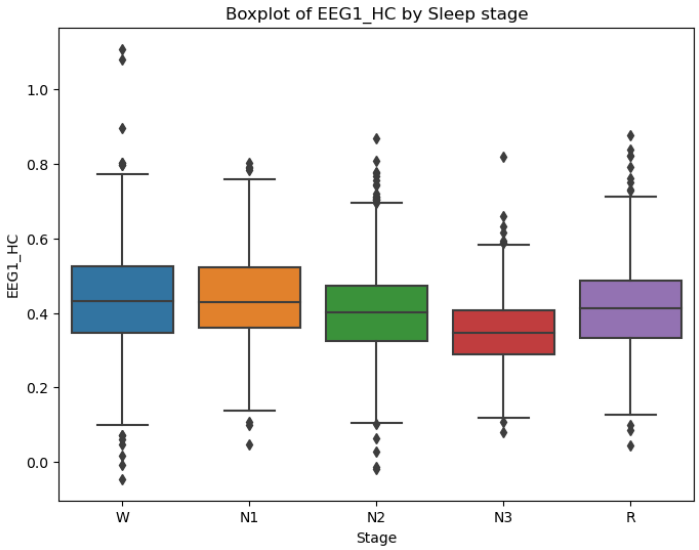

Figure 7: The boxplot of DE of EEG F4-M1 channel

between x_i and y_j. We have recurrence relation

$$D(i,j) = d(x_i, y_j) + \min(D(i-1,j), D(i,j-1), D(i-1,j-1))$$

where $d(x_i, y_j)$ IS the euclidean distance between x_i and y_j. The DTW can be got by

$$D = \sum_i \sum_j D(i,j)$$

DTW was calculated for all pairs of 4 EEG channels ($u_{17} - u_{22}$).

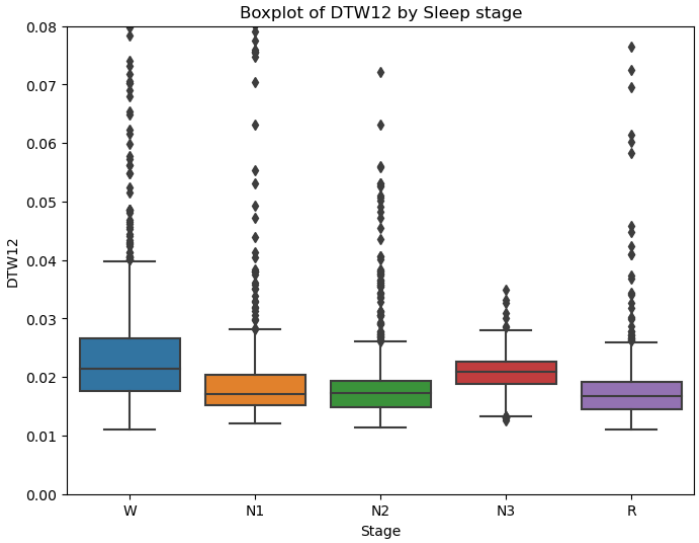

Figure 8: The boxplot of DTW between F4-M1 and C4-M1 channels

In Fig 8, we can see that the mean of DTW between F4-M1 and C4-M1 channels N3 and W stages is higher obviously. It shows that in N3 and W stages, the F4-M1 and C4-M1 channels perform similarly, but differently in other stages.

2.4 Supervised learning and Deep Learning Models

The potential features extracted by 22 different estimators are applied to several appropriate classifiers to classify the PSG dataset into sleep stages. Logistic regression, SVM, Random forest, and neural network (NN) are used as classifiers. we choose a linear kernel for SVM to classify the different sleep stages. In Random Forest, we use CV grid method to choose the best parameters, the number of estimators, and the maximum depth. NN has been used in different classification problems and shows excellent performance. There are 6 layers in our NN model: an input layer and a fully connected layer followed by a ReLu activation function and a batch normalization layer, and an output layer for classifying 5 sleep stages. All layers of the model and detailed parameter representations of these layers are provided in Table 2.

Table 2: Details of layers and parameters used in the proposed NN model

Num.	Layer Name	Input Size	Activation Function	Output Size
1	Input Layer	22	-	-
2	Fully Connected	22	ReLU	128
3	Batch Normalization	128	-	128
4	Fully Connected	128	ReLU	256
5	Batch Normalization	256	-	256
6	Output Layer	256	Softmax	5

3 Results and Discussion

3.1 Experiment

The experiment was conducted using PYTHON environment tools. We selected 80% data as the training dataset and 20 % as the testing dataset randomly. We observed that the dataset was imbalanced, and used SMOTE (Synthetic Minority Over-sampling Technique) to undersample and oversample for the training dataset. The distribution of sleep stages before and after SMOTE is in Table3

Table 3: Distribution of sleep stages in training dataset

Stage	Before	After
N1	391	700
N2	1407	1000
N3	478	700
R	682	682
W	753	753

3.2 Results

We use two methods to test the performance of different classifiers, accuracy and confusion matrix. Table 4 presents the accuracy values of 4 classifiers in the training and testing dataset. We can see that the training and testing accuracy in Logistic regression, SVM, and Neural network are close to each other. This indicates that there is no overfitting or underfitting in those models, and implies that those models have good generalization ability. The training accuracy in Random Forest is very high, but there is a big difference from test testing accuracy. It implies that the overfitting problem may exist in Random Forest.

Table 4: Model Accuracy Rate (%) in training and testing

Methods	Training	Testing
Logistic regression	59%	59%
SVM	58.19%	59.5%
Random Forest	99.94%	73.06%
Neural network	81.95%	74.03%

Fig 9 shows the performance graphs of the 4 proposed classifiers during the training dataset. The red bar represents the total accuracy, and the others represent the accuracy of different sleep stages. We can see that the total accuracy of Random Forest and Neural network is more than 70%, better than Logistic regression and SVM, which is just nearly 60%. The performance of classifying N3 stage is excellent in all methods, more than 80%, but 4 proposed classifiers all perform poorly in N1 stage. Among all proposed classifiers, we can see that Neural network performs the best. Except for N1, the

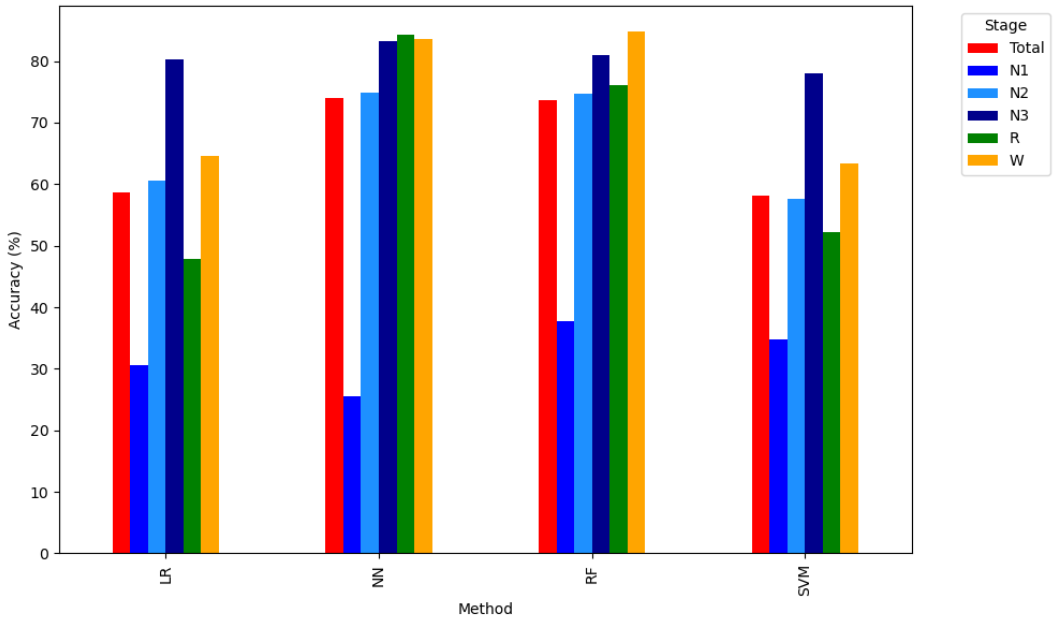

Figure 9: Performance graphs of the proposed classifiers

accuracy in the other stages is all above 70%, and the accuracy in N3, R, and W is more than 80%.

The detailed values of the evaluation for classification methods are given in Fig 10. For N1 stage, only about 30 samples are predicted correctly in all 4 proposed classifiers, while lots of samples are predicted as N2 or wake stage erroneously. For N3 stage, more than 100 samples are classified correctly out of 132. For R stage, 57 and 42 samples are predicted wrongly by logistic regression and SVM respectively out of 159 samples.

3.3 Discussion

In this project, we used 4639 sleep stage samples obtained from the whole night sleep dataset and 22 features extracted from singal data. We obtained the accuracy of 59%, 59.5%, 73.06%, and 74.03% using 4 EEG channels from Logistic regression, SVM, Random forest, and Neural network, respectively. The accuracy of N3, R, and W stage is higher than 80% in Neural network model, which is higher than Logistic regression, SVM, Random forest. All models cannot distinguish N1 from N2 and wake stages. We can see that the performance of Random forest and Neural network is better than

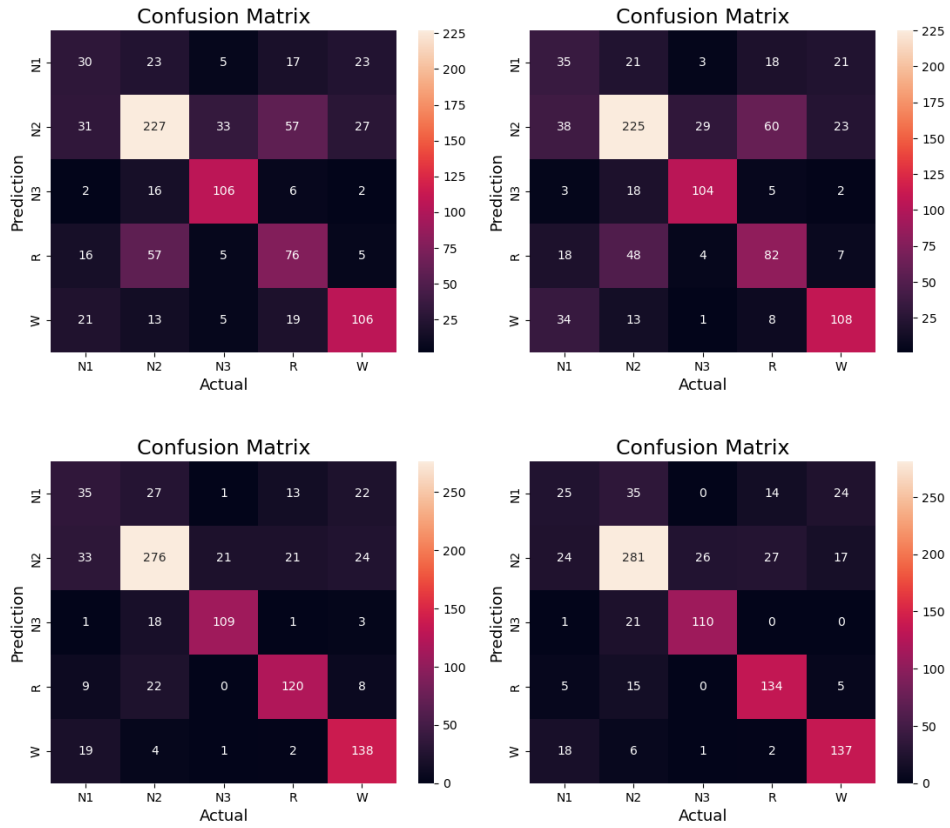

Figure 10: Confusion matrix for 4 classifiers: LR (left-top), SVM (right-top), RF (left-bottom), NN (right-bottom)

Logistic regression, and SVM. It may imply that the relationship between features and sleep stage is complicated but not just linear. The result of N1 stage is not very ideal, and one of the potential reason is that N1 and N2 are both light sleep stage[5], so their features represent the similar behaviours, while N3 belongs to deep sleep, so it can be classified clearly from otherstages.

4 Conclusion

This project tries to use different machine learning and deep learning methods to classify sleep stages, which has great medical and social significance. To reduce the cost and time consumption, we extract only 5 kinds of features from 4 EEG channels. In the proposed neural network model, we can get 74% total accuracy and over 80% accuracy in N3, R, and W stages. The results show that this method can learn useful features and achieve good classification performance. It is also a good beginning step for further research in sleep stage classification and guides the way forward to a better result.

In the future studies, we will try to extract other features from PSG dataset to improve the performance of classifiers, and focus on how to distinguish the difference between N1 and other stages. Moreover, improving the performance of classifiers by using different deep learning methods, such as CNN and LSTM can be a further goal.

References

[1] Diego Alvarez-Estevez and RM Rijsman. Haaglanden medisch centrum sleep staging database (version 1.0.1). *PhysioNet*, 2021.

[2] Reza Boostani, Foroozan Karimzadeh, and Mohammad Nami. A comparative review on sleep stage classification methods in patients and healthy individuals. *Computer methods and programs in biomedicine*, 140:77–91, 2017.

[3] J Christian Gillin, Wallace Duncan, Karen D Pettigrew, Bernard L Frankel, and Frederick Snyder. Successful separation of depressed, normal, and insomniac subjects by eeg sleep data. *Archives of general psychiatry*, 36(1):85–90, 1979.

[4] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[5] E Lamminmäki, A Saarinen, J Lötjönen, M Partinen, and I Korhonen. Differences in light sleep and deep sleep measured with ist vivago® wristcare. In *Ifmbe proceedings*, volume 11, page 1, 2005.

[6] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[7] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[8] Steven M Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301, 1991.

[9] Bo Qian and Khaled Rasheed. Hurst exponent and financial market predictability. In *IASTED conference on Financial Engineering and Applications*, pages 203–209. Proceedings of the IASTED International Conference Cambridge, MA, 2004.

[10] Mostafa Rostaghi and Hamed Azami. Dispersion entropy: A measure for time-series analysis. *IEEE Signal Processing Letters*, 23(5):610–614, 2016.

[11] Jessica Vensel Rundo and Ralph Downey III. Polysomnography. *Handbook of clinical neurology*, 160:381–392, 2019.

[12] Edward A Wolpert. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. *Archives of General Psychiatry*, 20(2):246–247, 1969.