Multi Domain Feature Fusion from a combination of Empirical Wavelet Transform and Boruta feature selection for Improved Sleep Stage Classification

Aashish Khilnani
Banaras Hindu University
aashishkhilnani@bhu.ac.in

Jyoti Singh Kirar
Jawaharlal Nehru University
kirarjyoti@gmail.com

***Abstract*—The potential of machine learning can be significant in revolutionizing clinical studies on human sleep. Enabling a machine to diagnose the sleep stages could alleviate the workload for clinicians. Here, we revisit the task of classifying sleep stages by applying machine learning algorithms. We have proposed a combination of Empirical wavelet transformation (EWT) based feature extraction and Boruta feature selection approach for automated classification of sleep stages. Various multi-domain features were extracted from intrinsic mode functions(IMF) of EWT. We have used two publicly available datasets in this research work. The proposed work performed better as compared to existing works in terms of classification accuracy.**

***Keywords*—Sleep staging, Empirical Wavelet Transform, Boruta Algorithm, EEG signals**

I. Introduction

Everyone sleeps around one-third of their life on average. However, not everyone can have the same quality of sleep. Machinery, stressful life, and psychological and neurological disorders can disturb sleep quality. Poor sleep can lead to drowsiness, fatigue, irritation and impact cognitive actions. To diagnose sleep, it is vital to study the stages of sleep and score the quality of sleep.

Sleep scoring, also known as sleep staging, involves using a polysomnograph (PSG) to record data from various sources, including electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), electrocardiogram (ECG), oxygen saturation, and airflow. Sleep experts manually annotate epochs, typically lasting 30 seconds, to classify sleep stages based on established guidelines like Rechtschaffen and Kales (R&K) and the American Academy of Sleep Medicine (AASM).

The R&K classification categorizes stages into Wakefulness, Non-Rapid Eye Movement (NREM) with four sub-stages (S1, S2, S3, and S4), and Rapid Eye Movement (REM). AASM has updated guidelines, defining wakefulness (W), three intermediate stages (N1, N2, N3), and REM. A full sleep cycle typically lasts 90 to 110 minutes, and this cycle repeats throughout the night [12]. As the night progresses, we spend less time in deeper sleep stages and more time in REM sleep, experiencing this cycle four or five times on an average night.

Sleep scoring is a complex and time-consuming process if done manually by humans as per standards. Moreover, manual annotation is prone to human error and subjective bias. Various machine learning and deep learning models are present to automate sleep classification, but still, humans are considered the standard for sleep scoring. The contributions of our study are as follows: We extracted various features from the intrinsic mode function (IMF) of the Empirical Wavelet Transform (EWT). This gave us improved performance. We selected the important features needed for the classification. We discussed the relevance of channel and ML classification algorithms to be used for sleep scoring

The whole research article is organised as follows. Section 2 shares insights into the related works in the sleep stage classification domain. Section 3 discusses the proposed approach. A brief introduction about dataset used is given in section 4. Section 5 shares the evaluations of the results. conclusion along with future directions is given in Section 6.

II. Related Work

Multiple combinations of features and classifiers have been used to detect sleep stages accurately [16]. [2] reviewed suitable methods for pre-processing, feature extraction and selection for sleep scoring and gave an accuracy value of 88.7% with a random forest classifier. [12] extracted features from the signal's decomposed frequency bands, and the best features were selected using the minimum redundancy maximum relevance (mRMR) algorithm. [19] proposed using bubble entropy and dispersion entropy-based sleep staging combined with multivariate fixed boundary empirical wavelet transform (MFBEWT) technique on the multiple channel EEG data. [14] proposed the range entropy and [4] investigated the importance of Range Entropy in sleep staging. [20] developed a python package YASA that aims to be easily scalable and flexible.

III. Methodology

In this work, we proposed a 3 step multi domain feature approach for automated sleep stage classification. In the first step, we decomposed the signal into 3 intrinsic mode functions (IMF) using EWT. In the second step, we extracted various

979-8-3503-6486-6/24/$31.00 ©2024 IEEE

2024 IEEE Region 10 Symposium (TENSYMP) | 979-8-3503-6486-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/TENSYMP61132.2024.10752243

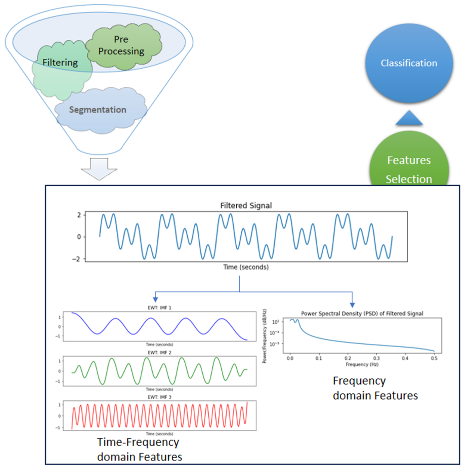

Fig. 1. Proposed sleep staging

multi-domain features from these IMFs. In this step, 16 time-frequency domain features from each IMF are extracted. We also extracted the Higuchi fractal dimension directly from signal and 17 frequency domain features using power spectral density (PSD) as shown in table I. Thus a total of 66 features per epoch, per channel were used in this study. In step 3, we performed Boruta feature selection to measure the importance of each feature. Fig. 1 depicts the process followed in this paper.

A. Emperical Wavelet Transformation

Due to the distinct frequency ranges of various components in EEG signals, researchers in composite algorithms have increasingly explored the application of decomposition techniques, such as Empirical Mode Decomposition (EMD) and Empirical Wavelet Transform (EWT) [3] etc. The EWT technique is employed to derive diverse components of the Fourier spectrum from the EEG signal by creating an adaptive wavelet filter bank. An overview of the EWT is given below:

1) Employ the FFT algorithm to acquire the spectrum of the Signal.
2) Identify all local maxima within the spectrum.
3) Determine boundaries based on the distances between local maxima.
4) Partition the spectrum in accordance with the established boundaries.
5) Formulate empirical wavelets and conduct signal decomposition into distinct components.

Each of the decomposed component is called Intrinsic Mode Function (IMF). We have decomposed our signal into 3 IMFs and extracted features from each IMF.

B. Feature Extraction

Feature extraction involves pulling out a group of traits to create a meaningful and condensed data representation. EEG signals are non-stationary, and thus, there are various methods available to study the time domain, frequency domain and

TABLE I
FEATURES EXTRACTED

Feature Extracted from each signal: Higuchi fractal dimension	
Features extracted from each decomposed IMF	
Mean	Skewness
Variance	Kurtosis
Range	Zero Crossing Count
Root Mean Square	Hjroth mobility
Standard Deviation	Hjroth complexity
IQR	Petrosian Dimension
Range Entropy	Permutation Entropy
Sample Entropy	Spectral Entropy
Frequency Domain Features	
Sum	Mean
Variance	Skewness
Kurtosis	Shannon Entropy
Median	Absolute spectral power
Relative spectral power in:	Spectral power ratio of:
Delta Band (0.4-4 Hz)	Alpha / Theta
Theta band (4-8 Hz)	Delta / Beta
Alpha band (8-12 Hz)	Delta / Sigma
Sigma band (12-16 Hz)	Delta / Theta
Beta band (16-30 Hz)	

time-frequency domain independently. We calculated 66 multi-domain features (extracted from time and frequency domain) per 30 sec epoch.

1) Time Frequency Domain: EEG signals exhibit multiple frequency modulations. Breaking down these signals into Intrinsic Mode Functions (IMF) gives a detailed understanding of fundamental characteristics. We used EWT to decompose into 3 IMFs. The features we extracted are:

- Statistical Parameters: Commonly employed time domain features for EEG signals include the 1st to 4th-order moments, which correspond to the mean, standard deviation, skewness, and kurtosis, respectively. Along with these, we also extracted the root mean square, Peak-to-peak range, variance, and IQR range of the signal distribution.
- Hjroth Parameters: Hjroth mobility H_m and Hjroth complicity H_c [7] of a time series $e(t)$ tell the proportion of standard deviation (σ) of the power spectrum and change in the frequency, respectively. They are formulated as:

$$H_m = \frac{\sigma(e'(t))}{\sigma(e(t))}; H_c = \frac{H_m(e'(t))}{H_m(e(t))}$$

- Zero Crossing Count (ZCC) The Zero Crossing Count (ZCC) tells us how many times a signal crosses a mean value of the signal. Since different sleep stages exhibit distinct characteristics in the time domain, ZCR can vary between stages.

2) Non Linear Features: The presence of nonlinear characteristics such as the Persian fractal dimension and the Higuchi fractal dimension contributes significantly to understanding the irregularity within signals. A diminished fractal dimension, signifying high irregularity, is correlated with elevated neuronal activity and is presumed to be connected with the N1 or REM sleep stage. The two non-linear features used in study are:

- Petrosian Fractal Dimension (PFD): It indicates signal complexity based on chaotic dynamics. In a time series comprising N data points, the [15] PFD is rapidly computed by converting the time series into binary sequences. In this binary representation, 1 denotes a positive difference between successive data points, while -1 indicates a negative difference. The PFD estimation can subsequently be calculated using the following formula:

$$\text{PFD} = \frac{\log_{10} N}{\log_{10} N + \log_{10} \frac{N}{N+0.4*N_\sigma}}$$

  where N_σ is the number of changes in the sign of the binary sequence.
- The Higuchi fractal dimension (HFD) [6]: Typically, the autocorrelation demonstrated by a signal is perceived as fractal in nature. Certain segments of the signal exhibit resemblances to the signal itself, and this resemblance recurs recursively. HFD serves as a quantitative gauge of this resemblance. It was directly derived from the signal to quantify its nonlinear characteristics, reflecting the complexity of the waveform in the time domain.

3) Entropy based features: Entropy serves as a metric for assessing randomness and acts as an indicator of system complexity. High entropy in an epoch indicates a probable association with N1 or REM stages. Shannon's information entropy is computed using the expression $-\sum_j p_j \log p_j$, where p_j represents the probability distribution of the observed data. In EEG analysis, numerous derived variations of information entropy are employed. We considered the entropies: Permutation entropy [1], Range entropy [14], Spectral Entropy, Sample entropy and Shannon entropy.

4) Frequency based Features: We used the Welch Power Spectral Density (PSD) to analyse the frequency characteristics of signals. This helped us to understand how the energy in the signal is distributed across different frequency bands which are the delta (0.4-4.0 Hz), theta (4.0-8.0 Hz), alpha (8.0-12.0 Hz), sigma (12-16 Hz) and beta (16.0-30.0 Hz). We also measured the total power of the wide-range signal and calculated power ratios, including $\frac{\delta}{\theta}$, $\frac{\delta}{\sigma}$, $\frac{\delta}{\beta}$, and $\frac{\alpha}{\theta}$ along with other statistical features.

C. Feature selection with Boruta Algorithm

Feature selection aims to achieve a minimal and optimal set of features, essentially seeking the smallest yet most effective feature subset. We used the Boruta algorithm [9] [10] in this study, which is a wrapper-based method centred around a Random Forest (RF) classifier. The Boruta algorithm extends feature information by creating shadow features through duplication and attribute shuffling. Using a random forest (RF) classifier, it calculates the mean decrease impurity (MDI) matrix to determine the importance of shadow features.

The Boruta algorithm follows a systematic process outlined as follows [11]:

1) Initiate the feature set by including all original features.
2) Generate shadow features by randomly permuting the values of the original features.
3) Train a machine learning model utilizing both the original and shadow features.
4) Assess the significance of each original feature by comparing it with its corresponding shadow features. Various metrics such as Gini impurity, information gain, and feature weights can be employed to quantify relevance.
5) Determine the statistical significance of a feature by comparing its importance against a predetermined threshold. A feature is considered significant if its importance surpasses the cumulative importance of its shadow counterparts.
6) Remove unnecessary features from the feature list.
7) Iterate through steps 2 to 6 until all features are either accepted or rejected.

Shadow features are generated by shuffling the values of the original features while preserving the target variable. Boruta evaluates the statistical significance of each feature by comparing its importance with that of its corresponding shadow feature. Features confirmed by Boruta at the end of the algorithm are considered suitable for the prediction task, while those rejected are deemed unimportant.

Boruta offers several advantages. It exhibits versatility by accommodating various types of machine learning models, encompassing both tree-based and non-tree-based algorithms. It effectively handles interactions and redundancy among features. Moreover, Boruta demonstrates resilience to noise, yielding stable outcomes even in the presence of noisy datasets. It also introduces randomness to identify truly essential features based on statistical significance. However, it's crucial to acknowledge that Boruta may not unfailingly pinpoint the optimal subset of features. Its performance can be impacted by the selection of the importance threshold and the quality of the feature importance metric employed. Furthermore, Boruta's computational demands can escalate, especially when dealing with datasets containing numerous features.

IV. Dataset and Experimental Setup

NTX Challenge Dataset: NeuroTechX conducted the Hackathon'23 across the globe in Dec 2023. Under the category of data challenge, EEG data of sleep stages was released on https://www.codabench.org/competitions/1777/ [22]. The data was collected with IDUN Guardian Earbuds, with a sampling frequency of 250 Hz, with two tip-shaped dry electrodes, one for each ear. The reference electrode is in the left ear canal, and the measuring electrode is in the right ear canal. Data from 4 subjects was collected, each with one night of recorded sleep. Per the competition rules, training was to be done on subjects two and three, validation on subject four and testing on subject 1. The labels were binary encoded, indicating the presence of sleep neuromarkers: spindles, K-complexes, Rapid Eye Movements, Sleep Onset, Sleep Off-set, arousal, and Microsleep. Thus, for each epoch of 30 sec, we have 14 binary labels, making the challenge a multiple class Machine learning Challenge with class imbalance.

Sleep EDF (Expanded) Dataset: We also implemented the proposed model on the publicly available sleep EDF expanded database on Physionet [8]. The sleep cassette has 153 PSG recordings. This study used the two EEG signals at Fpz and Pz of about 10 h each, sampled at 100 Hz. Subsequently, each 30-second epoch is categorized into sleep stages (W, N1, N2, N3, N4, REM) based on the R&K manual. For alignment with AASM recommendations, classes N3 and N4 are consolidated into a single stage labelled N3.

The two datasets used in this study have different structures. The first dataset by NTX has multiple class binary encoded labels indicating the presence or absence of sleep neuromarkers: sleep spindles, K complexes, REM, Sleep onset, sleep offset, arousal and Microsleep. On the other hand, the labels of the second dataset are a single 1D array stating the sleep states of the epochs. [17] Fig. 2 depicts the presence of various sleep neuromarkers among the states of sleep. The first stage of sleep N1 is a transitional phase between wakefulness and sleep, the period when sleep onsets. During this period, there is a slowdown in respiration and heartbeat, muscle tension and core body temperature. Sleep spindles - a rapid burst of high-frequency EEG signals, and K-complex - a very high amplitude pattern of EEG signals is found in stage N2. Arousals may occur within stages N1, N2, N3, or R if there is an abrupt shift in EEG frequency, encompassing alpha, theta, or frequencies more than 16 Hz for a minimum duration of 3 seconds. Additionally, at least 10 seconds of stable sleep should be preceding the observed abrupt change. [21]. Microsleeps are brief movements in which the brain falls asleep only to snap back awake. In our dataset, the presence of microsleep marks the offset of sleep.

A. Pre-Processing

The datasets present challenges with class imbalances. The difference between the number of epocs corresponding the the Wake state and deep sleep state varies greatly. Thus the initial 30 mins of Wake state was removed from the datasets. The datasets were high-pass filtered with a cutoff frequency of 30 Hz. For the first dataset, Data of subjects 2 and 3 were concatenated together to shape (2240 epochs, 2 channels, 7500 time points). The binary encoded labels were of shape (2240, 14). For the second dataset, we segmented the data into epochs of shape (836 epochs, 2 channels, 3000 time points). The corresponding number of epochs of W, N1, N2, N3 and REM are 183, 58, 250, 220 and 125 respectively. Both datasets have two signals. So we experimented thrice, once with each channel and then with a combination of both channels.

B. Classification

To evaluate the selected features' relevance, we split the dataset into train data and test data at a ratio of 80:20. We used random forest, SVC, logistic regression, decision tree, and Gaussian NB classification models. First, all the features were selected, and then the best 30 features chosen with high importance scores were used with the algorithms.

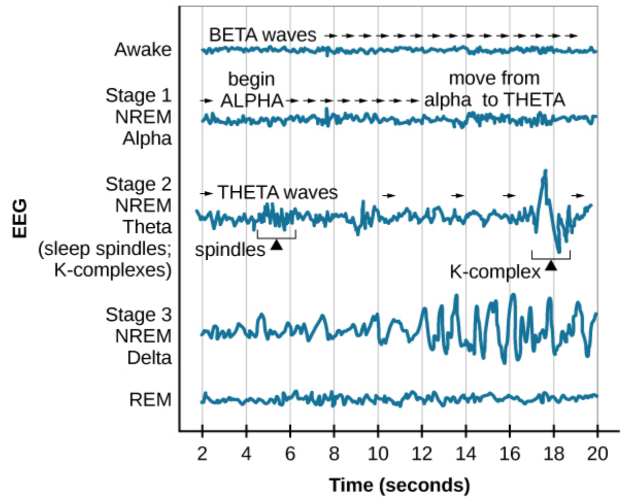

Fig. 2. Presence of sleep markers in sleep states (Adapted from [17])

V. Evaluations

A. Importance of Features

Fig. 3 illustrates the importance score of all features. The suffix after each feature denotes the Intrinsic Mode Function (IMF) from which the feature was extracted. Across all experiment runs, the importance order remained consistent. Features extracted from the third IMFs were relatively more important than those from the second and first IMFs. The blue boxplots represent the Z scores of the shadow attribute, specifically depicting the values for Shadow Minimal, Shadow Mean, and Shadow Maximum. Features coloured in red were deemed irrelevant due to poor importance scores, including skewness, kurtosis, Hjroth Complexity, and mean. [4] claimed the importance of Range entropy in sleep staging. It has a good importance score, but it increases the time taken to run by around 4 times. Permutation Entropy and Petrosian Dimension extracted from the third IMF emerged as the features with the highest importance, followed by the relative spectral power in the delta and alpha bands. We can see that the importance gradually becomes constant after 20 features, and then suddenly dips after 30 features. Thus 20 best features with the highest importance score in fig. 3 are selected.

B. Channel Relevance

All tests were performed in triplicate, first with Fpz-Cz, then Pz-Oz, and finally with both channels for dataset 2. Extracting features from both channels yielded better results compared to individual runs. Notably, the frontal area channel showed superior performance, as depicted in fig. 4 and fig. 5. The reason for the better performance of Fpz-Cz is its ability to capture a wide range of frequencies, including delta activity, K-complexes, and lower-frequency sleep spindles, which are vital for sleep staging. Whereas the Pz-Oz channel focuses on Theta activity and higher-frequency sleep spindles. [13]

C. Classifiers

For the first dataset, only Random forest, KNN, and decision tree were used due to its nature of binary multiclass labels. For

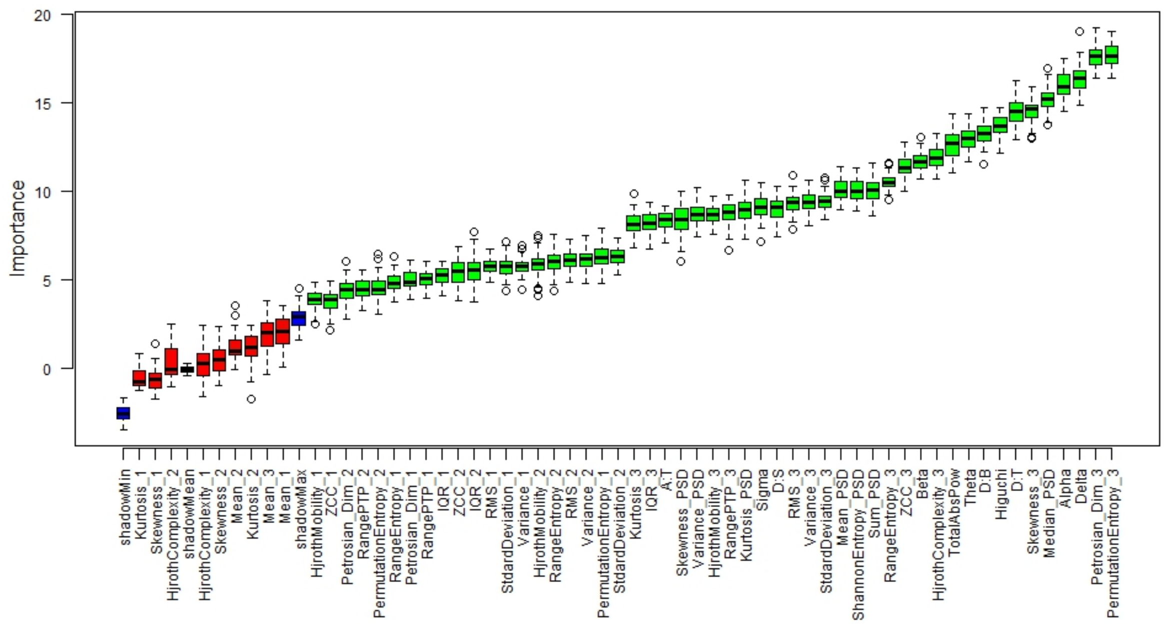

Fig. 3. Importance score of the features selected in study

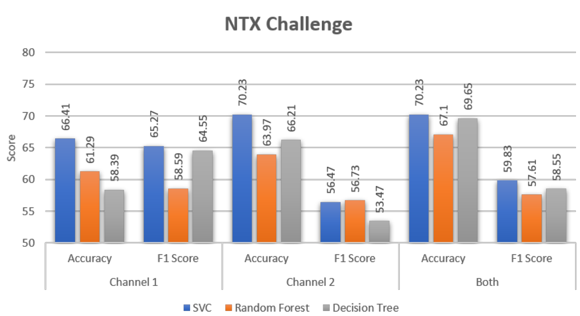

Fig. 4. Accuracy and F1 score of NTX Challenge with various classifiers

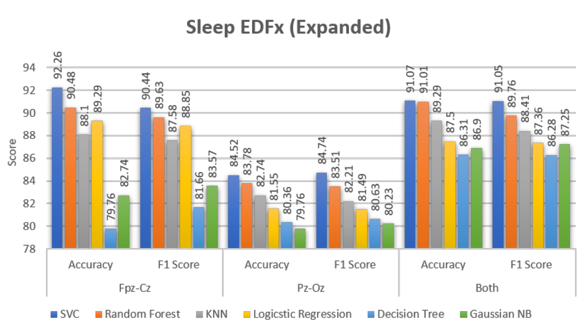

Fig. 5. Accuracy and F1 score of Sleep EDFx with various classifiers

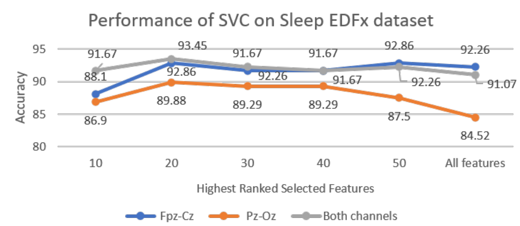

Fig. 6. Accuracy of SVC on sleep EDFx dataset with various selected features

the dataset of sleep EDFx, The extracted features were trained on machine learning models, including Random Forest(RF), Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), Logistic Regression(LR), Decision Tree(DT), and Gaussian Naive Bayes(GNB). Support Vector Classifier outperformed all others, as shown in fig. 4 and fig. 5, illustrating the performance of classifiers on both datasets.

D. Optimal solution

We selected only the best 20 features from fig. 3 and tested the accuracy score for all the datasets. Since SVC performed best, as in fig. 4 and fig. 5, we performed the final run only with SVC. Fig. 6 and fig. 7 show the SVC's accuracy change when the best features were selected for both datasets. We can see that the accuracy in all the cases peaked at 20 features. Although the change in accuracy is minor, the computational complexity is reduced by one-third. Thus our feature selection helped to reduce the computational complexity, while maintaining the accuracy.

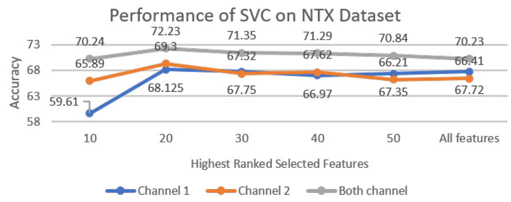

Fig. 7. Accuracy of SVC on NTX dataset with various selected features

TABLE II
COMPARISION OF PROPOSED STUDY WITH EXISTING SLEEP SCORING STUDIES (WITH THE SINGLE CHANNEL) ON THE EXPANDED DATASET OF SLEEP EDFX

Study	Fpz-Cz		Pz-Oz	
	Accuracy	F1 Score	Accuracy	F1 Score
SleepEEGNet [13]	84.3	79.7	77.6	70
DeepSleepNet [18]	82	79.6	79.8	73.1
1D-CNN-HMM [23]	83.9	76.9	83.2	74.7
LDFASSC [5]	88.5	81	87.6	75.2
Proposed Approach	**92.86**	**90.44**	**89.88**	**84.74**

E. Comparision with the existing work

The dataset provided by NTX is novel, and no one has worked on it. Table II shows the comparison of the proposed model with existing sleep scoring (with the single channel) on the EDFx dataset.

VI. CONCLUSION AND FUTURE WORK

This study focuses on classifying sleep neuromarkers like spindles, K-complexes, REM, arousal, and microsleeps in the first dataset and identifying sleep stages in the second dataset. Feature extraction was performed on signals decomposed using Empirical Wavelet Transform. The Boruta algorithm helped pick out the most important features. Although keeping these features only increased accuracy by about 2 per cent on average, it significantly simplified the computational process. Using the features, we trained different machine learning models, and the Support Vector Classifier showed the most promising results. With the enhanced performance of our approach, we anticipate its potential application in diagnosing and treating sleep disorders. With the 20 best-performing features, we got the accuracy/F1 score of 69.33 % / 56.47% with the first dataset and 91.07% / 90.44% with the sleep EDFx dataset using just one EEG channel.

We experience a cycle of sleep stages approximately 4-5 times during the night. Consequently, the distribution of sleep stages is time-dependent. Including time as a feature in future research may enhance the model's accuracy. One limitation is that this study used a relatively small dataset. When dealing with a larger dataset, the differences in the training data for each sleep stage might impact the accuracy differently.

REFERENCES

[1] Christoph Bandt and Bernd Pompe. Permutation entropy: a natural complexity measure for time series. *Physical review letters*, 88(17):174102, 2002.

[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] Jerome Gilles. Empirical wavelet transform. *IEEE transactions on signal processing*, 61(16):3999–4010, 2013.

[4] Mohammad Gafer Hadra, Iman Abuel Maaly, and Ibrahim Dweib. Range entropy as a discriminant feature for eeg-based alertness states identification. In *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 395–400. IEEE, 2021.

[5] Xia Heng, Miao Wang, Zhongmin Wang, Jie Zhang, Lang He, and Lin Fan. Leveraging discriminative features for automatic sleep stage classification based on raw single-channel eeg. *Biomedical Signal Processing and Control*, 88:105631, 2024.

[6] Tomoyuki Higuchi. Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena*, 31(2):277–283, 1988.

[7] Bo Hjorth. Eeg analysis based on time domain properties. *Electroencephalography and clinical neurophysiology*, 29(3):306–310, 1970.

[8] Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

[9] Miron B Kursa, Aleksander Jankowski, and Witold R Rudnicki. Boruta–a system for feature selection. *Fundamenta Informaticae*, 101(4):271–285, 2010.

[10] Miron B Kursa and Witold R Rudnicki. Feature selection with the boruta package. *Journal of statistical software*, 36:1–13, 2010.

[11] G Manikandan, B Pragadeesh, V Manojkumar, AL Karthikeyan, R Manikandan, and Amir H Gandomi. Classification models combined with boruta feature selection for heart disease prediction. *Informatics in Medicine Unlocked*, 44:101442, 2024.

[12] Pejman Memar and Farhad Faradji. A novel multi-class eeg-based sleep stage classification system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(1):84–95, 2017.

[13] Sajad Mousavi, Fatemeh Afghah, and U Rajendra Acharya. Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PloS one*, 14(5):e0216456, 2019.

[14] Amir Omidvarnia, Mostefa Mesbah, Mangor Pedersen, and Graeme Jackson. Range entropy: A bridge between signal complexity and self-similarity. *Entropy*, 20(12):962, 2018.

[15] Arthur Petrosian. Kolmogorov complexity of finite sequences and recognition of different preictal eeg patterns. In *Proceedings eighth IEEE symposium on computer-based medical systems*, pages 212–217. IEEE, 1995.

[16] Rym Nihel Sekkal, Fethi Bereksi-Reguig, Daniel Ruiz-Fernandez, Nabil Dib, and Samira Sekkal. Automatic sleep stage classification: From classical machine learning methods to deep learning. *Biomedical Signal Processing and Control*, 77:103751, 2022.

[17] Rose M Spielman, William Jenkins, and Marilyn Lovett. Psychology 2e. 2020.

[18] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

[19] Rajesh Kumar Tripathy, Samit Kumar Ghosh, Pranjali Gajbhiye, and U Rajendra Acharya. Development of automated sleep stage classification system using multivariate projection-based fixed boundary empirical wavelet transform and entropy features extracted from multichannel eeg signals. *Entropy*, 22(10):1141, 2020.

[20] Raphael Vallat and Matthew P Walker. An open-source, high-performance tool for automated sleep staging. *eLife*, 10:e70092, oct 2021.

[21] Anna Elisabetta Vaudano, Nicoletta Azzi, and Irene Trippi. Normal sleep eeg. *Clinical Electroencephalography*, pages 153–175, 2019.

[22] Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543, 2022.

[23] Bufang Yang, Xilin Zhu, Yitian Liu, and Hongxing Liu. A single-channel eeg based automatic sleep stage classification method leveraging deep one-dimensional convolutional neural network and hidden markov model. *Biomedical Signal Processing and Control*, 68:102581, 2021.