EEE-Based Automatic Classification of Insomnia Using Machine Learning Algorithm

First Name
Department of Electrical and Electronic Engineering,
Khulna University of Engineering & Technology

Second Name
Department of Electrical and Electronic Engineering,
Khulna University of Engineering & Technology

Second Name
Department of Electrical and Electronic Engineering,
Khulna University of Engineering & Technology

***Abstract*—This electronic document is a "live" template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. **CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract*.** (*Abstract*)

Keywords—component, formatting, style, styling, insert *(key words)*

INTRODUCTION

Sleep is still one of the greatest mysteries of human life. It is a natural physiological process that is essential to both physical and mental health, affecting cognitive function, emotional regulation, and overall quality of life. Insomnia is a type of sleep disorder and refers to a condition in which an individual has problems with initiating or maintaining sleep or gets non-restorative sleep even when he/she has the opportunity to sleep adequately and it is among a status of sleep disorders, which is highly common during pregnancy. Insomnia has a great impact not only on individual health but also on society through reduced productivity and increased healthcare costs. Global prevalence studies indicate that a substantial portion of the population experiences insomnia, with estimates ranging from 10% to 30% [1]. Accurate and early diagnosis of insomnia is crucial for effective intervention and treatment. However, traditional methods for diagnosing insomnia primarily rely on subjective assessments, including patient self-reports, sleep diaries, and standardized questionnaires[2]. Electroencephalography (EEG) is a neurophysiological method that detects brain generated electrical activity via sensors positioned on the scalp. It has been widely used in the field of sleep research in order to identify the various stages of sleep and detect sleep abnormalities[3].

Numerous studies have employed EEG to investigate the neurophysiological characteristics of insomnia, aiming to identify objective markers of the disorder. Findings from this body of research have consistently pointed towards alterations in the power of specific EEG frequency bands in individuals with insomnia. In EEE signal analysis, the frequency bands (alpha, beta, gamma, delta, and theta) play an important role in neurological and psychological research. Alpha band (8–12Hz) activity is often associated with relaxed wakefulness and tends to dominate when an individual is calm, awake, and alert but not engaged. Beta band (13-30 Hz): Indicates higher-frequency activity as it relates to active thinking, problem solving, and concentration. The gamma band (30-100 Hz) is associated with higher cognitive function, e.g., attention and information processing. Theta band (4-8 Hz) is associated with light sleep, relaxed drowsiness, deep relaxation, and meditation. The delta band (1-4 Hz) is the lowest-frequency band associated with deep sleep stages, particularly slow-wave sleep [4].

I. LITERATURE REVIEW

Over the years, many researchers have made significant contributions to this field. Shahin et al. proposed a two-stage method for the detection of insomnia by analyzing EEG data from 115 participants. The initial stage employed deep neural networks (DNNs) to automate the assessment of sleep stages and an epoch-level insomnia classifier, which extracted 57 temporal and spectral features from two EEG channels. Support vector machines were employed to derive and classify subject-level features in the second stage, resulting in a maximum F1 score of 0.88, 84% sensitivity, and 91% specificity[5]. A 1D-convolutional neural network (CNN) was employed by Yildirim, O. et al. to classify sleep stages using EEG and electrooculogram (EOG) signals. The model obtained accuracies ranging from 91.00% to 98.06% for a variety of sleep stages [6]. Rahman, M.M. et al. analyzed single-channel EOG data with discrete wavelet transform (DWT) and statistical features. Their proposed system achieved an accuracy of 91.7% [7]. Kuo, C.-E., et al. proposed a short-time insomnia detection system based on sleep EOG signals using refined composite multiscale entropy (RCMSE) analysis. Though their method achieved an accuracy of 89.31%, it is effective only for short-term insomnia detection [8]. Sharma et al. proposed an automated insomnia detection system using single-channel EEG signals and optimal bi-orthogonal wavelet decomposition. Their model achieved 95.6% accuracy and a Cohen's Kappa of 0.91 using ensemble bagged decision trees on the CAP sleep database, demonstrating its efficiency for at-home and lab-based monitoring applications [9]. Hanif, U., et al. used polysomnographic and clinical variables to classify insomnia using machine learning models. The study reported a balanced accuracy of 71% with logistic regression and identified key features such as depression, age, sex, and EEG power in insomnia detection, highlighting the potential of using machine learning for screening insomnia. The dataset's imbalance between insomnia and healthy subjects reduced the model's predictive power [10]

II. MATERIALS AND METHODS

A. OVERVIEW

This study utilizes the EEG signals from 22 patients. The data undergoes preprocessing steps including downsampling, bandpass filtering, artifact removal, segmentation, trimming, and labeling. Feature extraction was performed using Fast Fourier Transform (FFT) and power band analysis. Recursive Feature Elimination (RFE) is employed to select features, and four machine learning models—Random Forest, Decision Tree, KNN, and XGBoost—are employed for classification.

(A figure Will be added At last)

B. DATASET DESCRIPTION

The dataset employed in the current study was taken from the cross-sectional study on the 'Sleep EEG spectral analysis in psychophysiological insomnia and normal sleep subjects' conducted at the Sleep Disorders Research Center (SDRC) at Kermanshah, Iran. The dataset consists of polysomnography (PSG) recordings from 22 participants, 11 diagnosed with psychophysiological insomnia and 11 normal sleepers matched for age and BMI. The sample consisted of 14 female and 8 male subjects, ages 18–63. The dataset contains signals from 24 electrodes including 14 Electroencephalogram (EEG) channels (C4A1, C3A2, F3, F4, C3, C4, A1, A2, O1, O2, F3A2, F4A1, O1A2, O2A1), 6 electrooculogram (EOG) channels (EOG1, EOG2, EOG1A1, EOG2A1, EOG1A2, EOG2A2), 3 Electromyogram (EMG) channels (EMG, EMG1, EMG2), as well as ECG channel with sampling frequency of 256 Hz[11].

Table 01: Dataset Description

Classes	Male	Female	Samples
Normal	6	5	11
Insomnia	2	9	11
Total	8	14	22

C. DATA PREPROCESSING

The preprocessing pipeline in our investigation comprises six significant steps: downsampling, bandpass filtering, data segmentation, trimming, labeling, and artifact removal. The algorithm 1 was employed to execute the preprocessing stages.

```
Algorithm 1: EEE Preprocessing
Input: Raw EEG data $X_{raw}$(256 Hz)
Output: Processed EEG data $X_{processed}$ (128 Hz)
  1. Downsampling:
       $X_{down}$=Downsample($X_{raw}$,128 Hz)
  2. Bandpass Filtering:
       $X_{filtered}$= Bandpass Filter ($X_{down}$, 0.5 Hz, 35 Hz)
  3. Artifact Removal:
       $X_{clean}$= ApplyICA ($X_{filtered}$)
  4. Data Segmentation:
       $X_{seg}$= Segment( $X_{clean}$, 30 sec)
  5. Trimming:
       Datastart = 120 no. epoch
       Datalength = 840-120 = 720 epoch
       $X_{processed}$=Trim($X_{seg}$,120,720)
  6. Labeling: $y = \begin{cases} 1, & insomnia \\ 0, & normal \end{cases}$
    Result: $X_{processed}$, y
```

1. Downsampling:

The original sampling frequency of brain activity (the EEG signals) was 256 Hz. We down-sampled the data to 128 Hz to mitigate computational demands and storage without loss of fundamental signal features. Downsampling was performed as a means to avoid aliasing and ensure that the EEG records were accurate at the same time and enable more efficient analysis[12].

2. Bandpass Filtering:

The frequency component of relevance was extracted from the EEG data using a bandpass filter. The filter's lower cutoff frequency was 0.5 Hz, while its upper cutoff frequency was 35 Hz. This filtration procedure was crucial for the removal of high-frequency noise and low-frequency drifts, thus helping to ensure the data used to detect insomnia predominantly reflected pertinent brain activity.

3. Artifact Removal:

EEG recordings are frequently contaminated by non-cerebral artifacts originating from eye movements, muscular activity, and external electromagnetic interference. To eliminate these artifacts, we used Independent Component Analysis (ICA). ICA is a statistical technique used to separate a multivariate signal into additive, statistically independent components[13].

4. Segmentation

The data was divided into multiple 30-second epochs, with each epoch containing the same number of samples as the sampling frequency.

5. Trimming

(if possible, try to mention, why does it improve quality) The first and last portion of each subject were removed using the trimming procedure to enhance the quality of segmented data. The initial point of data segmentation was designated as "datastart".

$$\text{datastart} = \frac{60\times 60\times 1}{30} = 120$$

And the effective data length as:

$$\text{datalength} = \frac{60\times 60\times 7}{30} - 120 = 720$$

This trimming eliminated the initial one hour of data, which is not as significant as the data from the subsequent six hours.[12].

6. Labeling

In this study, patients are labeled based on their condition. Insomnia patients were labeled as 1 while normal patients, without insomnia, were labeled as 0. This labeling system was used to distinguish between the two groups for analysis.

D. Feature Extractions

Feature extraction is an essential component of biological signal analysis. The extraction process aims to classify data with the fewest and most precise parameters feasible. The preprocessed EEG signals are subjected to feature analysis.

The frequency domain features of EEG data were analyzed in this study using the Fast Fourier Transform (FFT). Later, we decomposed the data into several power bands.

a) Fast Fourier Transform:

The Fourier transform (FFT) is a computationally efficient algorithm that is employed to convert any signal from the time domain to the frequency domain. It decomposes complex signals into their constituent frequencies and amplitudes. This transformation yields a frequency spectrum that shows the signal's energy at various frequencies. The FFT is significantly faster than the direct Discrete Fourier Transform (DFT) calculation, making it suitable for EEG analysis [14].

To obtain the Fourier coefficient for a given signal x(n) in the frequency range [0, 2π] using the discrete Fourier transform algorithm. The discrete Fourier transform is defined as equation (1):

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-i2\pi k\frac{n}{N}}; 0 \leq k \leq N-1 \qquad (1)$$

b) Band Power

While the amplitude spectrums were obtained after performing the FFT. By this, spectral power was derived in five canonical EEG frequency bands: Delta (1-4 Hz), Theta (4-8 Hz), Alpha (8-13 Hz), Beta (13-30 Hz), Gamma (30-50 Hz)[4].

For each epoch, the power within these frequency bands was computed by summing the squared magnitudes of the FFT components falling within the corresponding frequency range. This process was repeated for all 14 EEG channels, generating a comprehensive set of frequency-domain features.

c) Feature Selection

A total of 70 features were generated for feature extraction. Feature reduction was deemed necessary for two main reasons: to eliminate highly correlated and redundant features, and to increase interpretability of the selected machine learning algorithm. Recursive feature elimination (RFE) was used for feature reduction in this study. RFE is a feature selection method that recursively trains a model and removes the least important features, as determined by the model. This process continues until all the requisite number of features is reduced. The number of features was reduced from 70 to 30 [10], [15].

E. Classification

Four machine learning models were employed to classify insomnia: K-Nearest Neighbor, Decision Tree, Random Forest (RF), and XGBoost. To obtain predictions for the entire dataset, all models were trained and evaluated using 5-fold cross-validation.

1. Random Forest:

The Random Forest algorithm is an ensemble learning technique that incorporates the outputs of numerous decision trees to generate a final prediction. This approach improves predictive accuracy and generalization by reducing overfitting[16]. In this work, we utilized a Random Forest Classifier and applied GridSearchCV for hyperparameter tuning. Here, RF was used because it effectively handles high-dimensional data and provide robustness against noise, leading more reliable prediction.

2. Decision Tree:

A decision tree is a supervised learning system that use a tree-like framework to illustrate decisions and their possible outcomes. Each internal node signifies an evaluation of an attribute, each branch denotes the result of the evaluation, and each leaf node indicates a class label (for classification) or a predicted value (for regression). This model was employed to estimate the value of a target variable by generating straightforward decision rules from the data features[17], [18].

3. K-Nearest Neighbor:

KNN is a non-parametric, instance-based learning algorithm employed for classification and regression. It is more computationally efficient and less complex than other classifiers [19]. We employed K-NN with k=5 because it yielded favorable outcomes for classifying insomniac and normal subjects through EEG signals.

4. XGBoost:

XGBoost is an efficient and scalable gradient boosting algorithm known for its superior performance in machine learning tasks, particularly with structured data[20]. We utilized the XGBoost classifier in the study for its high accuracy and speed in model training and prediction.

III. Result and Discussion

The study used several machine learning models to analyze frequency band power features extracted from EEG signals. After the feature extraction process, 70 features were initially extracted, then reduced to 30 through Recursive Feature Elimination (REF).

The selected 30 features using REF are listed in Table 02

C3A2 Theta, C3A2 Alpha, C3A2 Beta, F3 Delta, F3 Theta, F3 Alpha, F3 Beta, F4 Delta, F4 Theta, F4 Alpha, F4 Beta, C4 Theta, A1 Theta, A1 Alpha, A2 Alpha, A2 Beta, O1 Alpha, O1 Beta, O2 Alpha, O2 Beta, F3A2 Theta, F3A2 Alpha, F3A2 Beta, F4A1 Theta, F4A1 Alpha, F4A1 Beta, O1A2 Theta, O1A2 Alpha, O1A2 Beta, O2A1 Theta

Table 02: 30 selected features using RFE

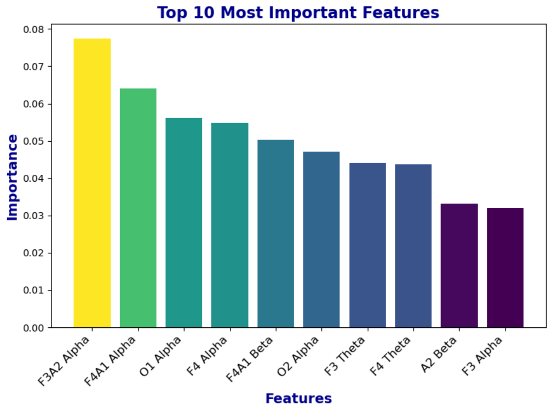


Figure 0*: Top 10 most important features

These selected features were subsequently fed into several machine learning models: Random Forest (RF), Decision Tree (DT), K-Nearest Neighbor (KNN) with k=5, and XGBoost with 5-fold cross-validation applied to ensure the reliability of the results.

Model	Accuracy	Precision	Recall	F-1 Score
Random Forest	0.9939	0.9931	0.9911	0.9931
Decision Tree	0.9695	0.9695	0.9695	0.9695
KNN	0.97643	0.9767	0.9764	0.9764
XGBoost	**0.9941**	**0.99411**	**0.9941**	**0.9941**

Table 03: Comparison table of classifiers

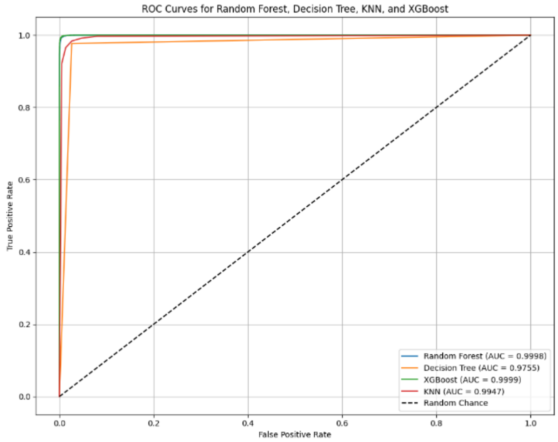


Figure 0*: ROC curve for the classifiers

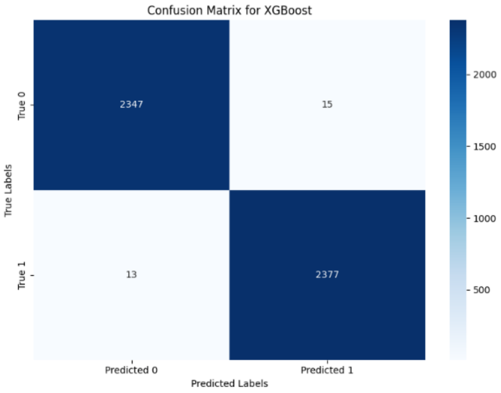


Figure 0*: Confusion matrix for XGBoost Classifier

IV. CONCLUSION

In this paper, we proposed a method for automatically detecting insomnia in patients using 14-channel EEG signals. We first preprocess the signals by downsampling, filtering, artifact removal, segmentation, and trimming. After preprocessing Fast Fourier Transform was performed, and the feature was extracted based on frequency band powers. EEG signals. Recursive Feature Elimination (REF) was used to decrease the original 70 features that were retrieved from the feature extraction method to 30 features. After that we performed several classifies with 5-fold cross validation and get the highest accuracy of 99.31% with F-1 score of 99.37% using Random Forest. These results demonstrate the effectiveness of our approach in accurately detecting insomnia using EEG data, showing promising potential for clinical applications. (Read some conclusion of other papers and try to write it again)

V. REFERENCES

[1] S. Bhaskar, D. Hemavathy, and S. Prasad, “Prevalence of chronic insomnia in adult patients and its correlation with medical comorbidities.,” *J Family Med Prim Care*, vol. 5, no. 4, pp. 780–784, Dec. 2016, doi: 10.4103/2249-4863.201153.

[2] C.-Y. Yang, N. Premakumara, H. Samani, and C. Premachandra, “Single Channel EEG Based Insomnia Identification Without Sleep Stage Annotations.” 2024. [Online]. Available: https://arxiv.org/abs/2402.06251

[3] R. Largo, M. Lopes, K. Spruyt, C. Guilleminault, Y. Wang, and A. Rosa, “Visual and automatic classification of the cyclic alternating pattern in electroencephalography during sleep,” *Brazilian Journal of Medical and Biological Research*, vol. 52, p. e8059, Feb. 2019, doi: 10.1590/1414-431X20188059.

[4] A. Ameera, A. Saidatul, and Z. Ibrahim, “Analysis of EEG Spectrum Bands Using Power Spectral Density for Pleasure and Displeasure State,” *IOP Conference Series: Materials Science and Engineering*, vol. 557, no. 1, p. 012030, Jun. 2019, doi: 10.1088/1757-899X/557/1/012030.

[5] M. Shahin, L. Mulaffer, T. Penzel, and B. Ahmed, “A Two Stage Approach for the Automatic Detection of Insomnia,” in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 466–469. doi: 10.1109/EMBC.2018.8512360.

[6] O. Yildirim, U. B. Baloglu, and U. R. Acharya, “A Deep Learning Model for Automated Sleep Stages Classification Using PSG Signals.,” *Int J Environ Res Public Health*, vol. 16, no. 4, Feb. 2019, doi: 10.3390/ijerph16040599.

[7] M. M. Rahman, M. I. H. Bhuiyan, and A. R. Hassan, “Sleep stage classification using single-channel EOG,” *Computers in Biology and Medicine*, vol. 102, pp. 211–220, 2018, doi: https://doi.org/10.1016/j.compbiomed.2018.08.022.

[8] C.-E. Kuo and G.-T. Chen, “A Short-Time Insomnia Detection System Based on Sleep EOG With RCMSE Analysis,” *IEEE Access*, vol. 8, pp. 69763–69773, 2020, doi: 10.1109/ACCESS.2020.2986397.

[9] M. Sharma, V. Patel, and U. R. Acharya, “Automated identification of insomnia using optimal bi-orthogonal wavelet transform technique with single-channel EEG signals,” *Knowledge-Based Systems*, vol. 224, p. 107078, 2021, doi: https://doi.org/10.1016/j.knosys.2021.107078.

[10] U. Hanif, U. Gimenez, A. Cairns, D. Lewin, N. Ashraf, and E. Mignot, “Automatic Detection of Chronic Insomnia from Polysomnographic and Clinical Variables Using Machine Learning.,” *Annu Int Conf IEEE Eng Med Biol Soc*, vol. 2023, pp. 1–5, Jul. 2023, doi: 10.1109/EMBC40787.2023.10340587.

[11] M. Rezaei, H. Mohammadi, and H. Khazaie, “EEG/EOG/EMG data from a cross sectional study on psychophysiological insomnia and normal sleep subjects,” *Data in Brief*, vol. 15, pp. 314–319, Dec. 2017, doi: 10.1016/j.dib.2017.09.033.

[12] M. R. Khan, A. A. Tania, and M. Ahmad, “A comparative study of time–frequency features based spatio-temporal analysis with varying multiscale kernels for emotion recognition from EEG,” *Biomedical Signal Processing and Control*, vol. 107, p. 107826, Sep. 2025, doi: 10.1016/j.bspc.2025.107826.

[13] A. Mayeli, V. Zotev, H. Refai, and J. Bodurka, “An automatic ICA-based method for removing artifacts from EEG data acquired during fMRI in real time,” in *2015 41st Annual Northeast Biomedical Engineering Conference (NEBEC)*, Apr. 2015, pp. 1–2. doi: 10.1109/NEBEC.2015.7117056.

[14] J. Nie, H. Shu, and F. Wu, “An epilepsy classification based on FFT and fully convolutional neural network nested LSTM,” *Frontiers in Neuroscience*, vol. Volume 18-2024, 2024, doi: 10.3389/fnins.2024.1436619.

[15] B. Akkaya, *The Effect of Recursive Feature Elimination with Cross-Validation Method on Classification Performance with Different Sizes of Datasets*. 2021.

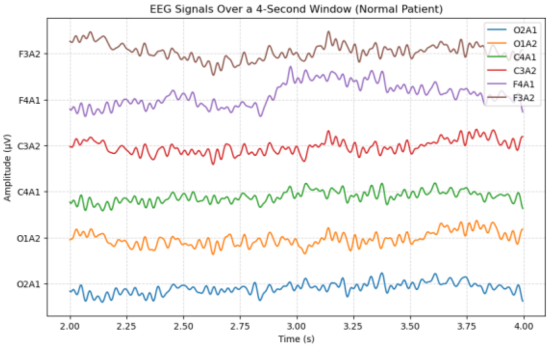
[16] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

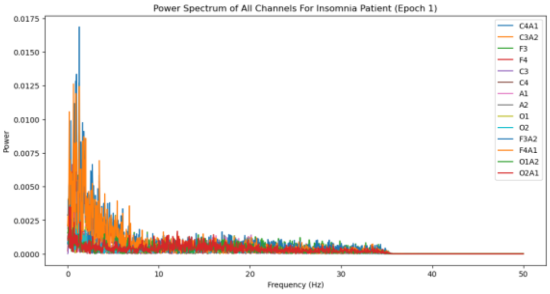
[17] J. R. Quinlan, “Induction of Decision Trees,” *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1023/A:1022643204877.

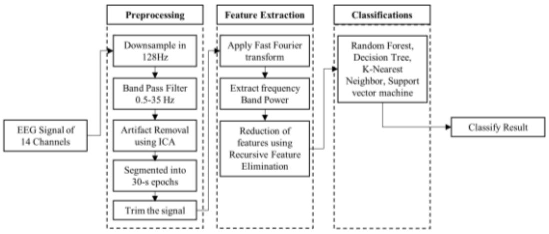
[18] H. Blockeel, L. Devos, B. Frénay, G. Nanfack, and S. Nijssen, “Decision trees: from efficient prediction to responsible AI,” *Frontiers in Artificial Intelligence*, vol. Volume 6-2023, 2023, doi: 10.3389/frai.2023.1124553.

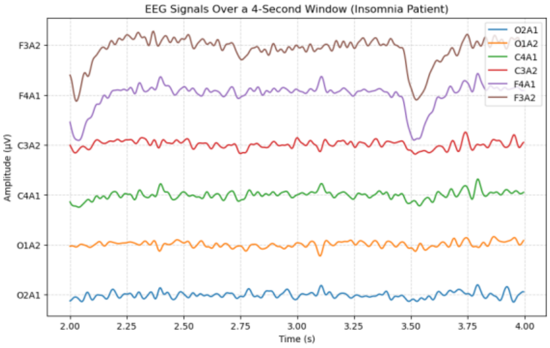
[19] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.

[20] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.






Preprocessing
Feature Extraction
Classifications
EEG Signal of 14 Channels
Downsample in 128Hz
Band Pass Filter 0.5-35 Hz
Artifact Removal using ICA
Segmented into 30-s epochs
Trim the signal
Apply Fast Fourier transform
Extract frequency Band Power
Reduction of features using Recursive Feature Elimination
Random Forest, Decision Tree, K-Nearest Neighbor, Support vector machine
Classify Result


EEG Signals Over a 4-Second Window (Insomnia Patient)
Amplitude (µV)
F3A2
F4A1
C3A2
C4A1
O1A2
O2A1
Time (s)
O2A1
O1A2
C4A1
C3A2
F4A1
F3A2