Signals and Systems Project

Instructor: Prof. Hamid Aghajan

Sharif University of Technology

Epileptic Seizure Prediction Using
Spectral Entropy-Based Features of EEG

Authors:

Alireza Gargoori Motlagh — alireza.gargoorimotlagh@ee.sharif.edu
AmirReza Imani — amir.r.iman2012@gmail.com
S. Moienadin Makkiyan — moien.makkiyan@sharif.edu

Spring 2024

Preface

Notes on the project:

- Due date: 1403/04/10
- The project must be done individually. Each individual will present his results in an online session on 1403/04/11.
- Please submit your project report as a `.pdf` file. Include all outputs and final results in the report. Make sure to list the practice text questions and provide a concise explanation of your problem-solving approach in each section.
- Ensure that all codes are provided in a separate `.m/.py/.ipynb` file. If a code cannot be tested accurately upon submission, the reported results will be considered invalid, and no points will be awarded in such cases.
- You have the flexibility to utilize either `MATLAB` or `Python` for your project. However, please be aware that `MATLAB` is recommended since certain aspects of the project rely on `MATLAB` toolboxes.
- Ensure that you save all files, including your report, codes, helper functions, and any additional outputs, if required, in a compressed file format such as `.zip` or `.rar`. This compressed file should then be uploaded to the Coursework CW submission platform.
- Your file names must be in the following format:

  `Project_#StudentID.zip/.rar/.pdf/.m/.py/.ipynb`

- The details of the grading system of this project will be provided in the coming days. Generally, the project is worth a total of 1 point in total (phases 1 and 2 combined).
- In this project, it is essential to uphold the principles of academic integrity and refrain from any form of cheating or copying. Cheating undermines the learning process, diminishes personal growth, and compromises the trust placed in us as students/researchers/professionals. It is crucial to recognize that engaging in dishonest practices not only tarnishes our own reputation but also has serious consequences, both ethically and academically. We want to emphasize that if anyone is found to have cheated, their results will not be accepted in this project, and they will receive a zero mark.

Contents

1 Introduction

1.1 Background and Importance of EEG in Neurological Disorders

Electroencephalography (EEG) is a method of recording electrical activity of the brain using electrodes placed on the scalp. This technique captures voltage fluctuations resulting from ionic current flows within the neurons of the brain. EEG is widely used in clinical and research settings due to its non-invasive nature, high temporal resolution, and relatively low cost. It plays a crucial role in diagnosing and monitoring neurological disorders, including epilepsy, Alzheimer's disease, and other forms of dementia.

Epilepsy is a chronic neurological disorder characterized by recurrent, unprovoked seizures. These seizures are caused by abnormal electrical discharges in the brain. Predicting epileptic seizures can significantly improve the quality of life for patients by allowing timely interventions and reducing the risk of injury. This project focuses on the prediction of epileptic seizures using EEG data.

1.2 Understanding Epilepsy and Seizures

Epilepsy affects approximately 1% of the global population. It is a condition marked by sudden, excessive electrical discharges in a group of brain cells. Seizures can vary in their presentation, ranging from brief lapses of attention to severe convulsions. The unpredictable nature of seizures poses a significant challenge for patients and healthcare providers. Early prediction of seizures can enable patients to take preventive measures, potentially averting injuries and enhancing their overall well-being.

There is currently no definitive cure for epilepsy, and the exact mechanisms underlying seizure generation are not fully understood. However, EEG provides a window into the brain's electrical activity, offering valuable insights that can be used to predict seizures. By analyzing patterns in EEG signals, researchers can develop algorithms to forecast the onset of seizures, providing critical warnings to patients.

1.3 EEG Signal Characteristics and Frequency Bands

EEG signals are composed of multiple frequency bands, each associated with different types of brain activity:

- Delta (0.5-4 Hz): Deep sleep and slow-wave activity.
- Theta (4-8 Hz): Drowsiness, meditation, and early stages of sleep.
- Alpha (8-13 Hz): Relaxed, calm state, typically with closed eyes.
- Beta (13-30 Hz): Active thinking, problem-solving, and focus.
- Gamma (30-100 Hz): High-level cognitive functions and information processing.

Understanding these frequency bands is crucial for analyzing EEG data and extracting meaningful features for seizure prediction.

1.4 Goals of the Project

This project aims to develop a robust method for predicting epileptic seizures using spectral entropy-based features of EEG signals. The project is divided into two phases:

1.4.1 Phase 1: Theoretical Foundations and Initial Data Analysis

Build a strong theoretical foundation and perform initial preprocessing and analysis of EEG data.

- **Theoretical Foundations of EEG and Epilepsy:** Introduce the basics of EEG, its significance, and its role in monitoring brain activity. Discuss epilepsy and the importance of predicting seizures.
- **Understanding and Preparing the EEG Dataset:** Describe the dataset used, including details on subjects, channels, and recording conditions. Implement data loading, exploration, and preprocessing steps.

1.4.2 Phase 2: Advanced Analysis and Machine Learning

Extract advanced features, apply machine learning algorithms, and evaluate the performance of seizure prediction models.

- **Spectral Entropy and Feature Extraction:** Explain the concept of spectral entropy and its application in EEG analysis. Implement the calculation of Power Spectral Density (PSD) and Shannon entropy from EEG signals.
- **Advanced Feature Extraction and Selection:** Extract statistical and entropy-based features from EEG epochs. Apply statistical methods to select significant features for classification.
- **Machine Learning Classification:** Train and test Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) classifiers using the selected features. Evaluate the performance of these classifiers in predicting seizures.
- **Performance Evaluation and Comparison:** Define and calculate sensitivity, specificity, and latency metrics. Compare the performance of SVM and KNN classifiers.
- **Real-time Seizure Prediction:** Discuss the challenges and requirements for real-time seizure prediction. Implement a script to simulate real-time processing and prediction of EEG data.

By completing this project, you will gain a deep understanding of EEG signal processing, feature extraction, and machine learning classification for epileptic seizure prediction. You will apply theoretical concepts to practical implementations using MATLAB or Python, ultimately contributing to the development of predictive algorithms for epilepsy management.

This comprehensive approach ensures that you not only understand the theoretical underpinnings of EEG analysis but also acquire hands-on experience in data processing, feature extraction, and the application of machine learning techniques. This foundation will prepare you for further research and practical applications in the field of biomedical signal processing.

2 Electroencephalography (EEG)

2.1 What is EEG?

Electroencephalography (EEG) is a technique used for recording electrical activity in the brain. This method involves capturing brain signals through electrodes placed on the scalp, which detect changes in voltage caused by neuronal activity. The recorded signals are in the microVolt range, making them sensitive to small noises.

One significant advantage of EEG over other brain activity monitoring methods is its high temporal resolution, meaning it can capture rapid changes in brain activity. However, it has relatively low spatial accuracy compared to techniques like functional Magnetic Resonance Imaging (fMRI). Another benefit of EEG is the portability of the devices, which can be used outside of a laboratory setting. This makes it feasible to conduct long-term monitoring using EEG headsets that are capable of storing large amounts of data.

EEG employs various electrode placement systems, with the 10-20 system being the most widely recognized. The 10-20 system ensures standardized electrode placement, making it possible to reproduce and compare results across different studies. The system is named for the relative distances between electrodes, which are either 10% or 20% of the total front-back or right-left distance of the skull.

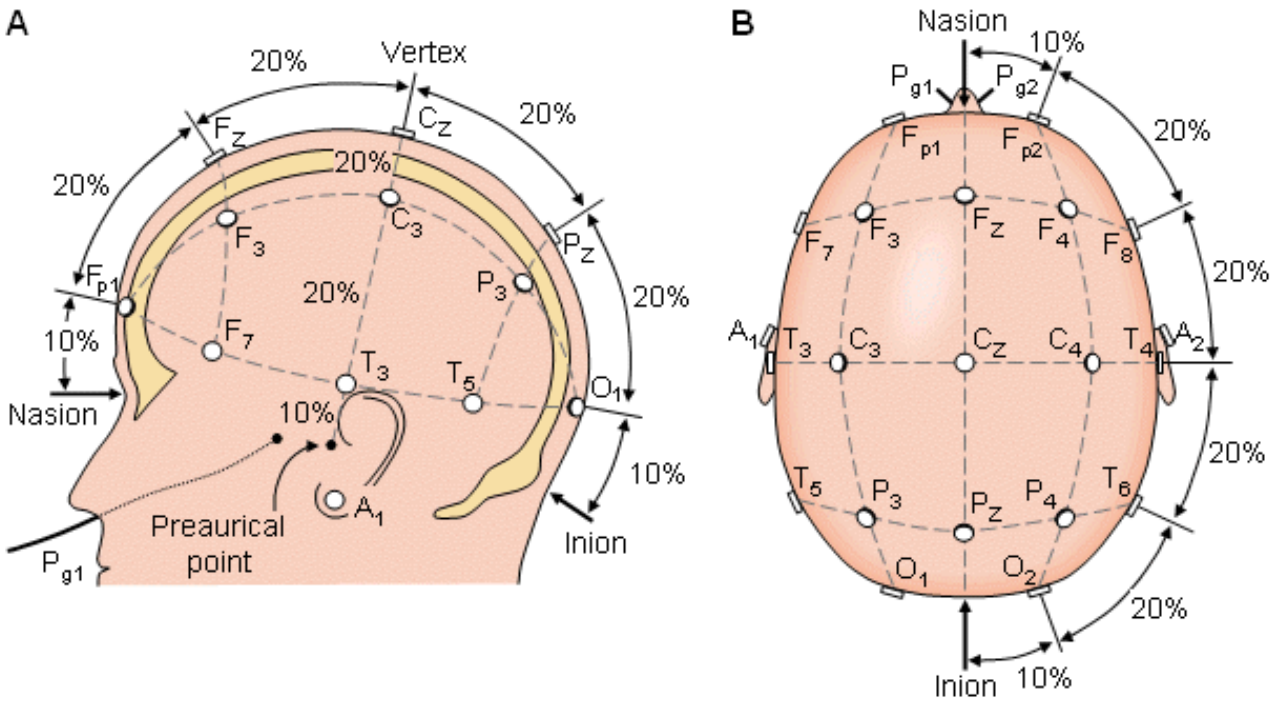

Figure 1: EEG 10-20 Electrode Placement System. (A) Side view of electrode placement showing distances relative to skull landmarks. (B) Top view of electrode placement showing a standard layout used in the 10-20 system.

Question: Based on the picture above, what does each electrode's name stand for? Explain the naming method used in the 10-20 EEG system.

2.2 Theoretical Background and Data Acquisition

Database: The EEG signals used in this study are from the CHB-MIT Scalp EEG Database, recorded at the Boston Children's Hospital. The dataset comprises recordings from children with intractable seizures, collected after several days of drug withdrawal. This setup ensures the capture of natural seizure activity without the interference of medication.

The database includes recordings from 23 cases, covering 22 different subjects (with one subject recorded twice over a two-year interval). The subjects range from 1.5 to 22 years old, including 5 males and 17 females. Due to the varying conditions under which EEG signals were recorded, the number of recording channels ranges from 23 to 38. All EEG signals were sampled at a rate of 256 Hz and with a 16-bit resolution. The International 10-20 system was used for electrode placement, ensuring consistent and reproducible data acquisition.

Data Summary:

- Subjects: 22 individuals, aged 1.5 to 22 years
- Gender Distribution: 5 males, 17 females
- Sampling Rate: 256 Hz
- Resolution: 16 bits
- Channels: 23 to 38 channels per recording
- Recording Conditions: Post-drug withdrawal to capture natural seizure activity

Table 1 in the referenced article provides detailed information about the subjects, including the number of seizures recorded for each patient. This information is crucial for understanding the variability in the dataset and designing algorithms that can generalize across different recording conditions and patient profiles.

Patient #	Sex	Age	No. of seizures
1	F	11	5
2	M	11	3
8	M	5.3	5
11	F	12	2
12	F	2	8
13	F	3	6
14	F	9	7
16	F	7	4
17	F	12	3
20	F	6	3
21	F	13	4
22	F	9	3

Table 1: Characteristics of the patients

Data Processing: The recorded EEG data from the CHB-MIT database is processed to extract meaningful features for seizure prediction. Key steps include:

- **Calculating Power Spectral Density (PSD):** PSD provides insight into how the power of the EEG signal is distributed across different frequencies. This is essential for identifying patterns associated with epileptic activity.
- **Calculating Shannon Entropy:** Entropy measures the randomness in the signal, with higher entropy indicating more disorder. By analyzing entropy, researchers can identify changes in brain activity that precede seizures.
- **Feature Extraction and Selection:** Features derived from the EEG signals are statistically analyzed to select the most significant ones for classification. This involves using statistical tests, such as t-tests, to identify features with high predictive power.

The ultimate goal is to use these features to train machine learning models, such as Support Vector Machines (SVM) and K-Nearest Neighbor (KNN) classifiers, to predict seizures with high accuracy and low latency.

This comprehensive approach ensures that the developed algorithm is robust and can effectively predict epileptic seizures, providing valuable warnings to patients and improving their quality of life.

2.3 Frequency Bands of EEG

EEG signals are divided into five distinct frequency bands, each associated with different types of brain activity. Understanding these frequency bands is crucial for analyzing EEG data and identifying patterns related to various neurological conditions.

Determine the activities each frequency band is associated with.

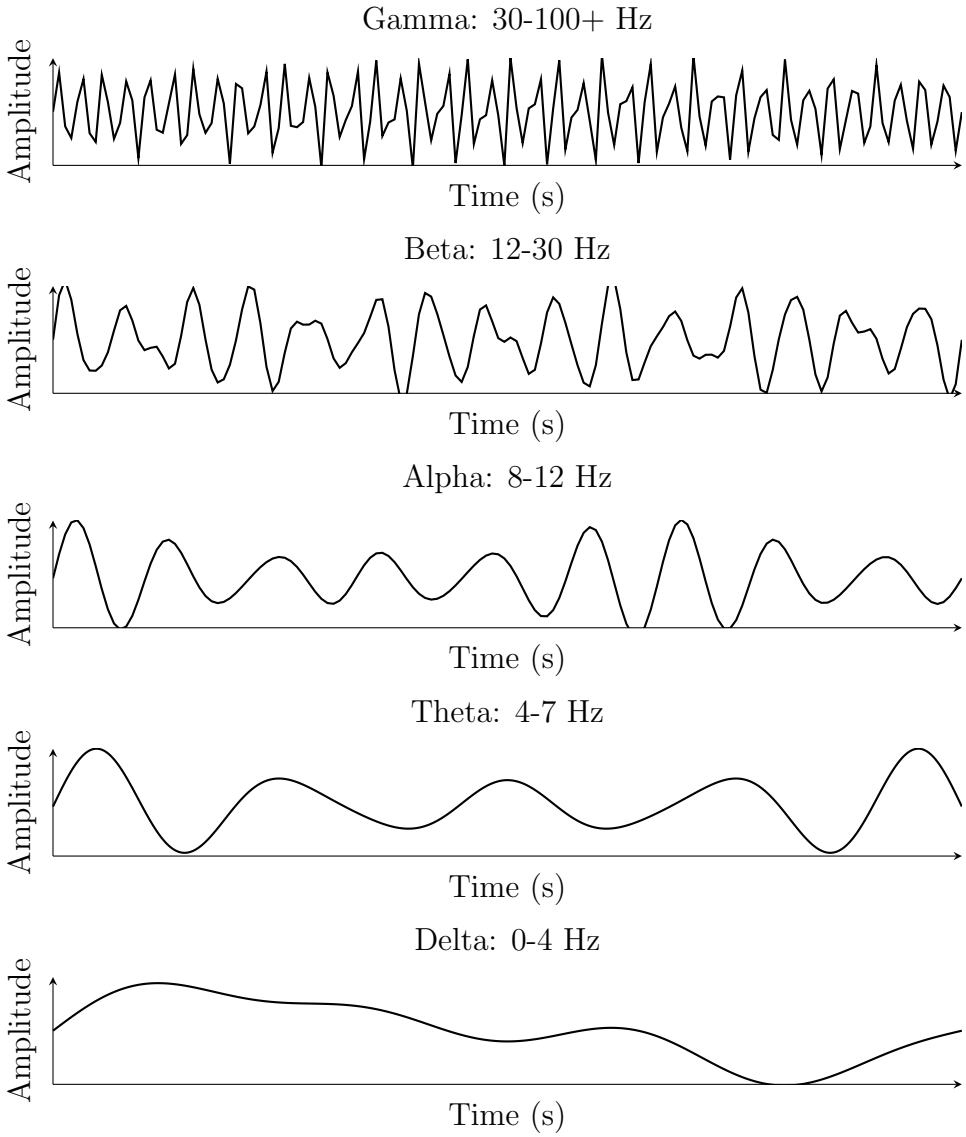

Figure 2: Comparison of EEG Bands

2.4 Sampling Frequency

The choice of sampling frequency for EEG signals is critical to ensure that the data accurately captures the relevant brain activities. According to the Nyquist criterion, the sampling frequency should be at least twice the highest frequency present in the signal to avoid aliasing. For EEG signals, which include frequencies up to 100 Hz, a common sampling frequency is 256 Hz. This rate ensures that all relevant frequency components are captured, providing sufficient resolution for both clinical and research purposes.

Based on frequency bands and Nyquist criterion, which sampling frequencies are preferred for EEG signals?

3 EEG Signal Processing

In this section, you will become familiar with the task and the structure of the data.

3.1 Task Definition

The primary objective of this project is to predict epileptic seizures using EEG data. The data is sourced from the CHB-MIT Scalp EEG Database, which includes recordings from children with intractable seizures. These recordings were made after several days of drug withdrawal to capture natural seizure activity. The data is structured in EDF (European Data Format) files, each containing one hour of continuous EEG data recorded at a sampling rate of 256 Hz and a resolution of 16 bits. The electrode placement follows the International 10-20 system, ensuring standardized and reproducible results.

Data Collection:

- EEG recordings capture electrical activity across 23 to 38 channels.
- Each recording session lasts one hour and can include multiple sessions per subject.
- Seizure events within these recordings are annotated with precise start and end times.

Objective:
The goal is to process these EEG recordings to extract features such as Power Spectral Density (PSD) and Shannon entropy, which can then be used to train machine learning models for seizure prediction.

3.2 Data Description

The dataset consists of EEG recordings from 5 subjects, each with multiple EDF files. Each file follows a standardized structure and contains several key attributes:

File Structure: Each EDF file contains EEG data recorded over a one-hour period. The files are named sequentially and include metadata about the recording start and end times, as well as annotations for seizure events.

Example File Information:

File Name	Start Time	End Time	Number of Seizures	Seizure Start Time (s)	Seizure End Time (s)
chb01_01.edf	11:42:54	12:42:54	0	-	-
chb01_02.edf	12:42:57	13:42:57	0	-	-
chb01_03.edf	13:43:04	14:43:04	1	2996	3036
...	...	...	...	...	...
chb01_21.edf	07:33:46	08:33:46	1	327	420

Table 2: Example of EDF file information with seizure annotations.

Data Attributes:

- **Epochs:** Segments of EEG data for analysis, typically split into 16-second intervals.
- **Channels:** Electrode channels as per the 10-20 system (e.g., FP1-F7, F7-T7, etc.).
- **Seizure Annotations:** Time-stamped annotations indicating the start and end of seizure events within each file.

Data Example: Each EEG recording can be visualized and analyzed to extract meaningful features such as Power Spectral Density (PSD) and Shannon entropy, which are crucial for seizure prediction.

Attribute	Description	Data Type	Example
Epoch	16-second segment of EEG data	Array	[0.1, 0.2, 0.3, ...]
Channel	EEG electrode channel	String	FP1-F7
Seizure Start	Start time of seizure event	Timestamp	12:15:00
Seizure End	End time of seizure event	Timestamp	12:15:30

Table 3: Field Descriptions

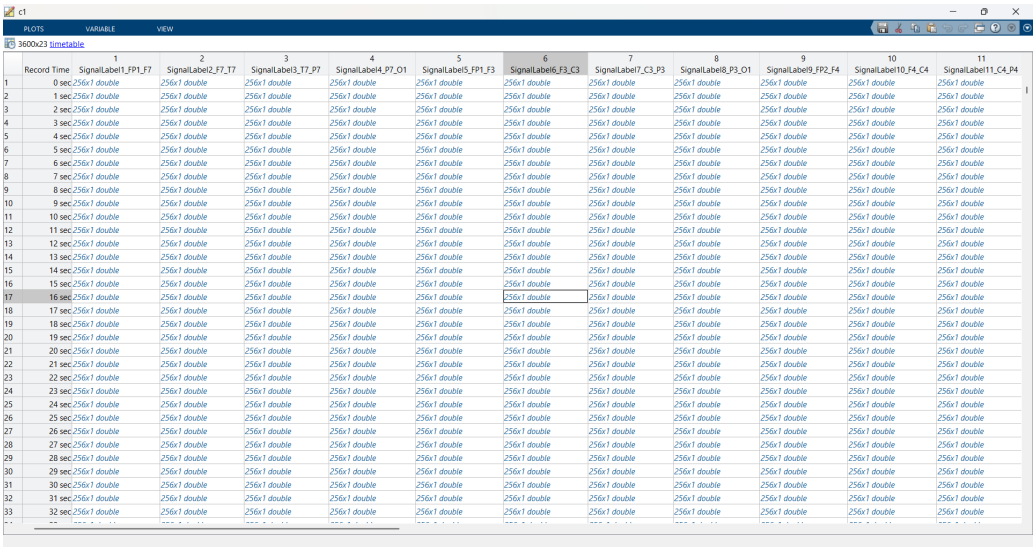

Figure 3: Example EEG Data Structure in MATLAB

3.3 Pre-Processing

Using a standard pipeline in EEG signal preprocessing is crucial for ensuring consistency, reproducibility, and objectivity in research. It reduces bias, enhances the reliability of results, and provides established best practices for addressing common challenges. A popular and widely used pipeline for EEG signal preprocessing is Makoto's pipeline (Makoto's preprocessing pipeline - SCCN).

The collected raw data from all participants were preprocessed following a simplified version of Makoto's pipeline using EEGLAB, as described in the following steps:

1. Apply 1 Hz high pass filter to remove baseline drifts.
2. Apply relevant notch filter to remove the 50 Hz line noise.
3. Reject bad channels as a critical step before average referencing using the clean_rawdata() EEGLAB plugin.
4. Interpolate the removed channels.
5. Re-reference the data to the average of all channels to obtain a good estimate of reference-independent potentials.
6. Apply clean_rawdata() for cleaning the data by running artifact subspace reconstruction (ASR).
7. Re-reference the data to the average again to compensate for any potential changes in the data caused by the previous step.
8. Run independent component analysis (ICA) to identify EEG sources as well as the sources associated with noise and artifacts.
9. Fit single and bilateral (if available) current dipoles.
10. Further clean the data by source (dipole) selection using IClabel() plugin in EEGLAB.

4 Phase 2 Task

In this phase, you will build upon the preprocessing work completed in Phase 1 and delve into the advanced analysis of EEG data for predicting epileptic seizures. The methods and steps you will follow are based on the article "Epileptic Seizure Prediction Using Spectral Entropy-Based Features of EEG" by Amirhossein Ahmadi and Hamid Soltanian-Zadeh. This phase involves feature extraction, feature selection, and applying machine learning algorithms to classify and predict seizure events.

4.1 Data Explanation

You will work with the EEG data related to epilepsy provided in Phase 2. Here, the focus will be on extracting meaningful features from the preprocessed EEG signals and using these features to predict seizures.

We will provide you with seven EDF files as data, all corresponding to one subject (person). One of these files (chb01_02.edf) is non-seizure data, while the other six files contain one seizure each. The details and times of seizures are provided in a text file (chb01-summary.txt) included alongside this documentation.

Note: Any further and more detailed information about the data has been explained earlier in the document. Please refer to the previous sections for comprehensive details.

4.2 Steps for Phase 2

4.2.1 Step 1: Load Database

- **Task**: Load the provided EEG data from the CHB-MIT Scalp EEG Database into MATLAB. The data files are included alongside this documentation.
- **Explanation**: The EEG data from the CHB-MIT Scalp EEG Database, recorded at the Boston Children's Hospital, will be used. These recordings are made after several days of drug withdrawal to capture natural seizure activity. The dataset includes recordings from 23 cases involving 22 subjects, aged 1.5 to 22 years. The data is recorded at a sampling rate of 256 Hz with 16-bit resolution, using the International 10-20 system for electrode placement. Ensure the data is correctly loaded into MATLAB to facilitate further processing and analysis.

4.2.2 Step 2: Calculating Power Spectral Density (PSD)

- **Task**: Calculate the Power Spectral Density (PSD) of each epoch.
- **Explanation**: PSD shows how the average power of the signal is distributed in different frequencies. This step involves using either direct calculation or autocorrelation methods to obtain the PSD. The formula for PSD calculation is:

$$S_x(f) = \lim_{T\to\infty} E[\frac{1}{2T}|\int_{-T}^{T} x(t)e^{-j2\pi ft}dt|^2]$$

Where $S_x(f)$ is the PSD of the signal $x(t)$.

4.2.3 Step 3: Calculating Shannon Entropy

- **Task**: Calculate Shannon Entropy for each epoch.
- **Explanation**: Entropy measures the randomness or disorder within a signal. Shannon entropy is calculated by treating the PSD values as probability density functions. The formula for Shannon Entropy is:

$$H(X) = -\Sigma_{i=1}^{n} p(x_i) \log p(x_i)$$

Where $p(x_i)$ is the probability of occurrence of x_i.

4.2.4 Step 4: Feature Extraction

- **Task**: Extract features from the EEG signals.
- **Explanation**: Each 10-minute interval prior to a seizure is broken into 16-second epochs using a moving window. For each epoch, calculate the FFT, then compute the PSD and normalize it. Calculate Shannon entropy and other statistical measures (average, standard deviation, minimum, and maximum) for each epoch.

4.2.5 Step 5:Feature Selection

- **Task**: Perform feature selection using statistical tests.
- **Explanation**: Apply a one-sample t-test to the extracted features and select those with a p-value less than 0.001. The test statistic is calculated as:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Where $\bar{X}$ is the mean of the features, μ_0 is the hypothesized mean (considered as zero), s is the standard deviation, n is the size of the feature vector.

4.2.6 Step 6: Support Vector Machine (SVM) Classifier

- **Task**: Train and test an SVM classifier.
- **Explanation**: An SVM classifier is used to distinguish between seizure and non-seizure epochs. Use a linear kernel function and employ a least squares method to find the separating hyperplane. The SVM classifier aims to maximize the margin between different classes.

4.2.7 Step 7: K-Nearest Neighbor (KNN) Classifier

- **Task**: Train and test a KNN classifier.
- **Explanation**: KNN is a non-parametric method that classifies each object based on the majority vote of its neighbors. Use the Minkowski distance metric for measuring the distance between data points.

4.2.8 Step 8: Performance Measures

- **Task**: Evaluate the performance of the classifiers.
- **Explanation**: Calculate sensitivity, specificity, and latency to assess the performance of your seizure prediction algorithm. Sensitivity measures the true positive rate, specificity measures the true negative rate, and latency measures the time required for the algorithm to generate an output.

4.3 Additional Instructions

You can follow these additional steps and guidelines:

1. Data Extraction and Preparation
   - Use functions such as ‘getDataBeforeTime‘ to extract 10-minute intervals before seizure events.
   - Convert the data into a matrix format using ‘getMatrix‘.
   - Use ‘getFeature‘ to extract features from the epochs, including entropy, mean, standard deviation, maximum, and minimum.
2. Train and Test Data Recommendation
   - Choose:
     - A: 6 epochs (10 minutes duration) from non-seizure data.
     - B: 2 epochs that have no seizure from seizure data.
     - C: 6 epochs containing seizures, with the seizure happening in the last minute.
   - Assign data for training and testing:
     - Train: 4 epochs from C, 5 epochs from A, 1 epoch from B.
     - Test: 2 epochs from C, 1 epoch from A, 1 epoch from B.
3. Feature Selection
   - Apply a one-sample t-test (‘TTestSelection‘) to select features with a p-value less than 0.001.
   - Ensure the selected features are consistent across different subjects and epochs.
4. Classifier Training and Testing
   - Train the SVM and KNN classifiers using the selected features.
   - Evaluate the performance using accuracy, false positive rate (FP rate), and true positive rate (TP rate).
5. Handling NaN Values
   - Ensure that any NaN values in the input channels are properly handled to avoid issues during classification.

6. K-Fold Cross-Validation

   - Implement K-fold cross-validation to evaluate the classifier performance more robustly. This involves splitting the data into k subsets, training the model on k-1 subsets, and testing it on the remaining subset. Repeat this process k times and average the results.
     Example:

```
k = 5; % Number of folds
indices = crossvalind('Kfold', data, k);
for i = 1:k
    test = (indices == i);
    train = ~test;
    % Train and evaluate your model here
end

```

7. Leave-One-Out Cross-Validation

   - For more detailed evaluation, use leave-one-out cross-validation, where each sample is used once as a test sample while the rest are used for training.

4.4 Deliverables

- **Power Spectral Density Plots:** Present figures showing the PSD of the EEG signals.
- **Shannon Entropy Calculations:** Provide a figure showing the Shannon entropy for different epochs.
- **Classifier Performance:** Report on the performance of the SVM and KNN classifiers, including sensitivity, specificity, and latency metrics.
- **Selected Features:** Present the selected features with their p-values and explain the significance of each feature.
- **Cross-Validation Results:** Include the results of K-fold and leave-one-out cross-validation.