Republic of Tunisia
Ministry of Higher Education and Scientific Research
University of Tunis El Manar
National Engineering School of Tunis

Department of Information and Communication Technologies

Signal Processing for Data Science Project

EEG-signal Classification for Epileptic Seizure Detection

Prepared by :

Hatem TRIGUI

Malek HASSOUNA

Supervised by :

Ms. Soumaya MEHERZI

3rd Year Telecommunications Opt. DASEC

Academic Year : 2024/2025

Abstract

Epilepsy, a complex neurological disorder characterized by recurrent seizures, affects millions worldwide. Accurate and timely detection of epileptic seizures is critical for improving patient outcomes and enabling more effective interventions. This report presents a machine learning-based approach for classifying EEG signals to detect epileptic seizures, utilizing the Bonn University EEG dataset. By combining advanced signal processing techniques, including Discrete Wavelet Transform (DWT) with Haar, Daubechies, and Symlet wavelets, with machine learning classifiers, we aim to enhance classification accuracy and robustness.

The preprocessing of EEG signals includes steps for noise reduction, normalization, and segmentation to ensure consistent input quality. Feature extraction through DWT and dimensionality reduction with Principal Component Analysis (PCA) are employed to simplify data while retaining key discriminatory information. The resulting features are fed into Support Vector Machines (SVM) with various kernel functions (RBF, polynomial, etc.) to evaluate classification performance.

This report discusses the implementation details, experiments conducted, and a thorough analysis of results. Our findings contribute to the ongoing research in automated EEG-based seizure detection, showing promising accuracy levels and providing insights into the efficacy of different wavelet transformations and classification techniques. This work underscores the potential of machine learning and signal processing in improving seizure detection systems for clinical and real-world applications.

Key Words : EEG Signal Classification, Epileptic Seizure Detection, Support Vector Machine (SVM), Discrete Wavelet Transform (DWT), Principal Component Analysis (PCA), Bonn University EEG Dataset

Contents

List of Figures

List of Tables

Acronymes

- **DWT** - Discrete Wavelet Transform
- **EEG** - Electroencephalogram
- **SVM** - Support Vector Machine
- **PCA** - Principal Component Analysis
- **RBF** - Radial Basis Function
- **DB** - Daubechies (Wavelet)

Chapter 1

Introduction

Epilepsy is a neurological disorder that affects millions of people globally, characterized by recurrent, often unpredictable seizures that can have a profound impact on quality of life. Effective seizure detection and monitoring are essential for improving patient outcomes, providing timely medical intervention, and advancing our understanding of the underlying mechanisms of epilepsy. In recent years, Electroencephalogram (EEG) signals have become a crucial tool in detecting seizure activity due to their direct correlation with neuronal electrical activity in the brain.

However, the classification of EEG signals for epileptic seizure detection presents a series of challenges. EEG data is often high-dimensional, non-stationary, and subject to considerable variability, making traditional classification methods insufficiently robust. This project addresses these challenges by employing advanced signal processing techniques and machine learning methodologies to develop an automated, accurate, and reliable seizure detection system.

1.1 Motivation

Traditional methods for detecting seizures rely on the manual examination of EEG recordings by medical professionals. This process, while effective, is inherently time-consuming, prone to human error, and difficult to scale. With the rise of machine learning, automated EEG classification offers a promising alternative. Automated systems can significantly reduce the workload on medical professionals, provide consistent results, and enhance the precision and speed of seizure detection.

In this project, we aim to enhance the state-of-the-art in seizure detection by leveraging machine learning algorithms and signal processing techniques. By developing a robust classifier, we can contribute to clinical practices and provide tools that could potentially be integrated into wearable devices, offering continuous

and real-time monitoring solutions for patients.

1.2 Problem Statement

EEG signals present unique characteristics, including high dimensionality and a complex temporal structure, which pose challenges for accurate seizure detection. To address this, we utilize advanced signal decomposition, feature extraction, and classification methods. Specifically, we use Discrete Wavelet Transform (DWT) for signal decomposition, Principal Component Analysis (PCA) for dimensionality reduction, and Support Vector Machines (SVM) for classification.

This project addresses the following research questions:

1. How effectively can wavelet-based feature extraction and PCA improve the performance of EEG-based seizure classification?
2. What is the impact of various types of wavelet transforms (e.g., Haar, Daubechies, Symlet) on the discriminative power of the extracted features?
3. Can SVM provide reliable classification for EEG signals when combined with advanced preprocessing techniques?

1.3 Proposed Methodology

The study is conducted using the Bonn University EEG dataset, a benchmark in EEG research. The methodology consists of the following main steps:

- **Data Preprocessing:** The EEG data is preprocessed to remove noise and standardize the format. This involves normalization and filtering.
- **Feature Extraction using Discrete Wavelet Transform (DWT):** DWT is applied to decompose the EEG signals into various frequency bands, providing a time-frequency representation that captures seizure-relevant information.
- **Dimensionality Reduction using Principal Component Analysis (PCA):** PCA reduces the high-dimensional feature space, enhancing computational efficiency and minimizing redundancy.
- **Classification using Support Vector Machines (SVM):** An SVM classifier is trained to distinguish between normal and epileptic EEG signals based on the extracted features.

This combination of methods provides a comprehensive approach to handling EEG data, addressing both feature complexity and classification accuracy.

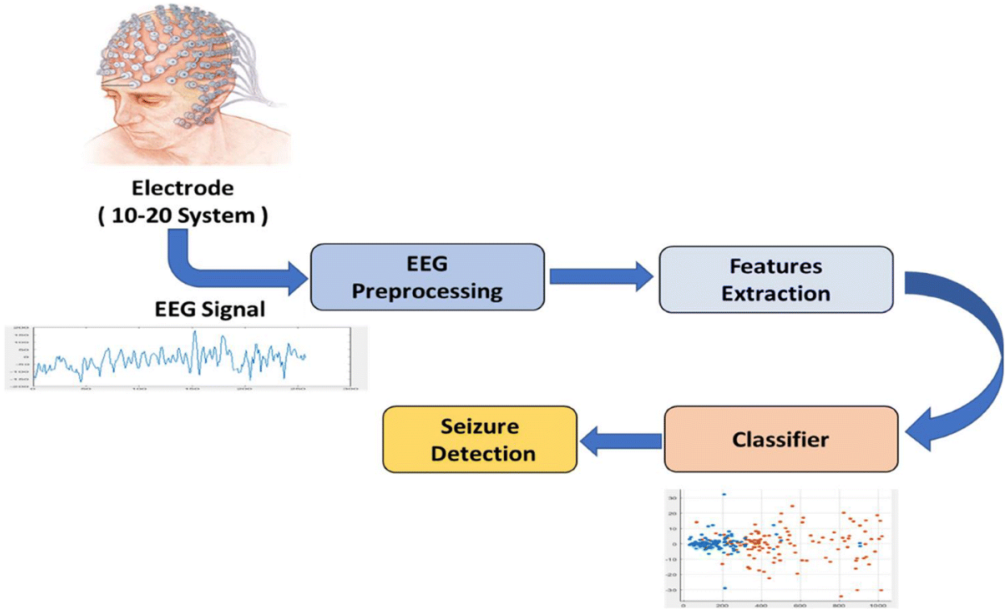

Figure 1.1: Overview of the EEG signal preprocessing and feature extraction pipeline.

1.4 Report Structure

The remainder of this report is organized as follows:

- Chapter 2 provides a review of the literature related to EEG-based seizure detection.
- Chapter 3 details the methodology, including preprocessing, feature extraction, and classification techniques.
- Chapter 4 presents the experimental setup and results, along with an analysis of the findings.

This comprehensive exploration of EEG-based seizure detection aims to advance the state of the field, contributing valuable insights into both the technical and clinical aspects of EEG signal classification.

Chapter 2

Literature Review

2.1 Introduction

Epileptic seizure detection through EEG signals has been an active area of research for decades. Various methods have been developed and refined over time, each seeking to improve the accuracy and efficiency of detecting seizures in complex EEG data. This chapter provides an overview of key concepts, challenges, and methodologies in the field of EEG-based seizure detection, focusing on feature extraction techniques, classification methods, and the use of machine learning algorithms in automating seizure detection.

2.2 Overview of EEG-Based Seizure Detection

EEG signals are widely used in epilepsy research due to their ability to directly measure the electrical activity of the brain. However, due to the complex, non-linear, and non-stationary nature of EEG signals, traditional analytical methods often fall short in accurately identifying seizure patterns. As shown in Figure 2.1, EEG signals are characterized by varying amplitudes, frequencies, and temporal structures that reflect different brain states, making seizure detection a challenging task.

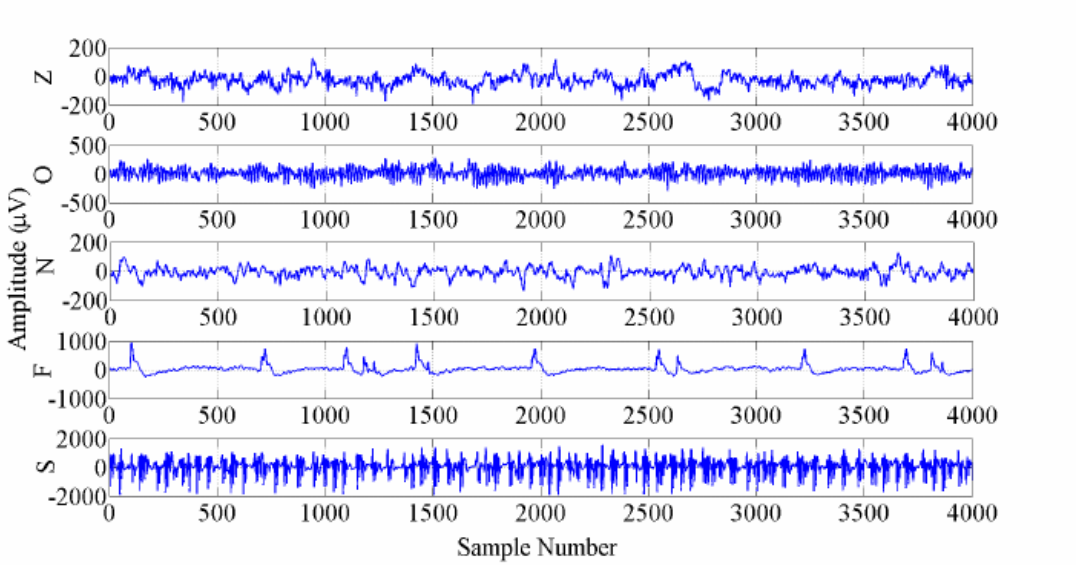

Figure 2.1: Challenges in analyzing EEG signals for seizure detection: variability, noise, and high dimensionality.

2.3 Challenges in EEG Signal Classification

2.3.1 High Dimensionality and Variability

EEG data typically involves multiple channels (electrodes) and high sampling rates, leading to high-dimensional datasets. Each EEG recording contains a wealth of information, but the seizure-relevant patterns are often subtle and sparse. Moreover, EEG signals vary greatly between individuals and across different sessions, adding complexity to the classification task.

2.3.2 Non-Stationary Nature of EEG Signals

EEG signals are non-stationary, meaning their statistical properties change over time. Seizure events are transient and can appear at unpredictable times, making it difficult to isolate seizure patterns from normal brain activity. Researchers address this issue by employing time-frequency analysis methods, which can capture both temporal and spectral characteristics of EEG signals.

2.3.3 Noise and Artifacts

EEG data is susceptible to various types of noise, such as eye movements, muscle activity, and power line interference. Effective noise removal techniques are

essential in preparing the data for feature extraction and classification. Table 2.1 lists common sources of noise in EEG recordings and their corresponding removal techniques.

Table 2.1: Common noise sources in EEG data and removal techniques.

Noise Source	**Removal Technique**
Eye Movements	Independent Component Analysis (ICA), Wavelet Denoising
Muscle Artifacts	Band-pass Filtering, Adaptive Noise Cancellation
Power Line Interference	Notch Filtering at 50/60 Hz

2.4 Feature Extraction Techniques

Effective feature extraction is crucial for reducing the complexity of EEG signals while retaining essential information for classification. In seizure detection, common feature extraction techniques include time-domain, frequency-domain, and time-frequency methods.

2.4.1 Time-Domain Features

Time-domain features are derived directly from the EEG signal waveform. These include metrics such as mean, variance, and skewness, which can provide initial insights into signal characteristics. However, time-domain features alone may not capture the intricate patterns in EEG data, particularly for seizure detection.

2.4.2 Frequency-Domain Features

Frequency-domain analysis involves transforming the EEG signal into its spectral components. This approach helps identify seizure-relevant frequency bands, as seizures are often associated with specific frequencies. Fourier Transform (FT) and Power Spectral Density (PSD) are widely used methods for frequency-domain feature extraction.

2.4.3 Time-Frequency Methods: Discrete Wavelet Transform (DWT)

The Discrete Wavelet Transform (DWT) is a time-frequency method that has gained popularity for EEG analysis. DWT decomposes the signal into different frequency bands, providing localized frequency information. In seizure detection, wavelet transforms such as Haar, Daubechies, and Symlet have been shown to effectively capture seizure-related patterns.

2.5 Dimensionality Reduction Techniques

To manage the high dimensionality of EEG data, dimensionality reduction techniques are often applied after feature extraction. Principal Component Analysis (PCA) is one of the most widely used techniques for reducing redundant features while preserving the variance in the data.

2.5.1 Principal Component Analysis (PCA)

PCA is an unsupervised technique that projects the data onto a lower-dimensional space by identifying the principal components that explain the maximum variance. In EEG-based seizure detection, PCA is used to simplify the feature set, allowing machine learning models to perform efficiently without compromising accuracy.

2.6 Classification Techniques

Various classification techniques have been applied to EEG signals for seizure detection. Commonly used classifiers include Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and neural networks.

2.6.1 Support Vector Machines (SVM)

SVM is a supervised learning method that separates classes by maximizing the margin between them. In EEG classification, SVM is effective due to its robustness against high-dimensional feature spaces and ability to handle non-linear relationships through kernel functions. Table 2.2 provides a comparison of classifier performances in previous EEG studies.

Table 2.2: Comparison of classification performances for seizure detection in EEG studies.

Study	Classifier	Feature Extraction	Accuracy	Dataset
1	SVM	DWT + PCA	91%	Bonn University
2	k-NN	FFT + PCA	88%	Bonn University

2.6.2 Other Classifiers: k-Nearest Neighbors (k-NN) and Neural Networks

The k-Nearest Neighbors (k-NN) algorithm is a simple yet effective classifier that assigns labels based on the majority class among the k nearest points in the feature space. Neural networks, particularly Convolutional Neural Networks (CNNs), have also been explored for EEG classification, as they can automatically learn relevant features from raw EEG data.

2.7 Advancements in Machine Learning for Seizure Detection

Recent advancements in machine learning, particularly in deep learning and artificial intelligence, have significantly impacted the field of EEG-based seizure detection. These advancements have enabled the development of more sophisticated models capable of capturing complex patterns in EEG data, leading to improved accuracy and robustness in seizure prediction.

2.7.1 Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have gained popularity in EEG analysis due to their ability to model sequential data effectively. Unlike traditional neural networks, RNNs and LSTMs have an internal memory that retains information over time, making them well-suited for data with temporal dependencies. In the context of seizure detection, these models can capture the evolution of EEG signals over longer periods, improving the detection of subtle patterns indicative of seizure onset.

2.7.2 Hybrid Models Combining Feature Extraction and Deep Learning

Another promising approach in seizure detection is the use of hybrid models that combine feature extraction techniques with deep learning architectures. Traditional methods, such as wavelet transforms and Principal Component Analysis (PCA), can be employed to reduce the dimensionality and complexity of EEG data before feeding it into a deep learning model. By integrating these feature extraction methods with Convolutional Neural Networks (CNNs) or other deep architectures, hybrid models can achieve a balance between interpretability and predictive power. This combination allows the model to leverage both handcrafted features and automatically learned representations, leading to enhanced accuracy and robustness in seizure detection tasks.

2.7.3 Future Directions in AI for Seizure Detection

As deep learning models continue to evolve, there is growing interest in exploring advanced architectures like Transformers and Graph Neural Networks (GNNs) for EEG analysis. These models, which have shown success in natural language processing and complex relational data, may offer new ways to handle the intricate patterns in EEG signals. Additionally, the integration of explainability techniques, such as attention mechanisms and saliency maps, could provide insights into the underlying neural patterns associated with seizure activity, facilitating their use in clinical applications.

In summary, recent advancements in machine learning are paving the way for more accurate, efficient, and interpretable seizure detection systems. By leveraging the unique capabilities of modern AI architectures, researchers aim to improve seizure detection accuracy, reduce false positives, and ultimately enhance patient outcomes.

2.8 Conclusion

In summary, this chapter has highlighted the key challenges and methodologies in the field of EEG-based seizure detection. Table 2.3 summarizes the main techniques used in previous studies, including their strengths and limitations.

Table 2.3: Summary of techniques used in EEG-based seizure detection.

Technique	**Advantages**	**Limitations**
Discrete Wavelet Transform (DWT)	Effective time-frequency representation	Computationally intensive
Principal Component Analysis (PCA)	Reduces dimensionality, enhances efficiency	Potential information loss
Support Vector Machines (SVM)	Handles high-dimensional data, robust	Sensitive to choice of kernel
Convolutional Neural Networks (CNNs)	Automatic feature learning	Requires large datasets

In the next chapter, we will delve into the specific methodology applied in this project, detailing the preprocessing, feature extraction, dimensionality reduction, and classification steps.

Chapter 3

Methodology

This chapter outlines the methodology employed in this study to classify EEG signals for the detection of epileptic seizures which can be summarized as follows:

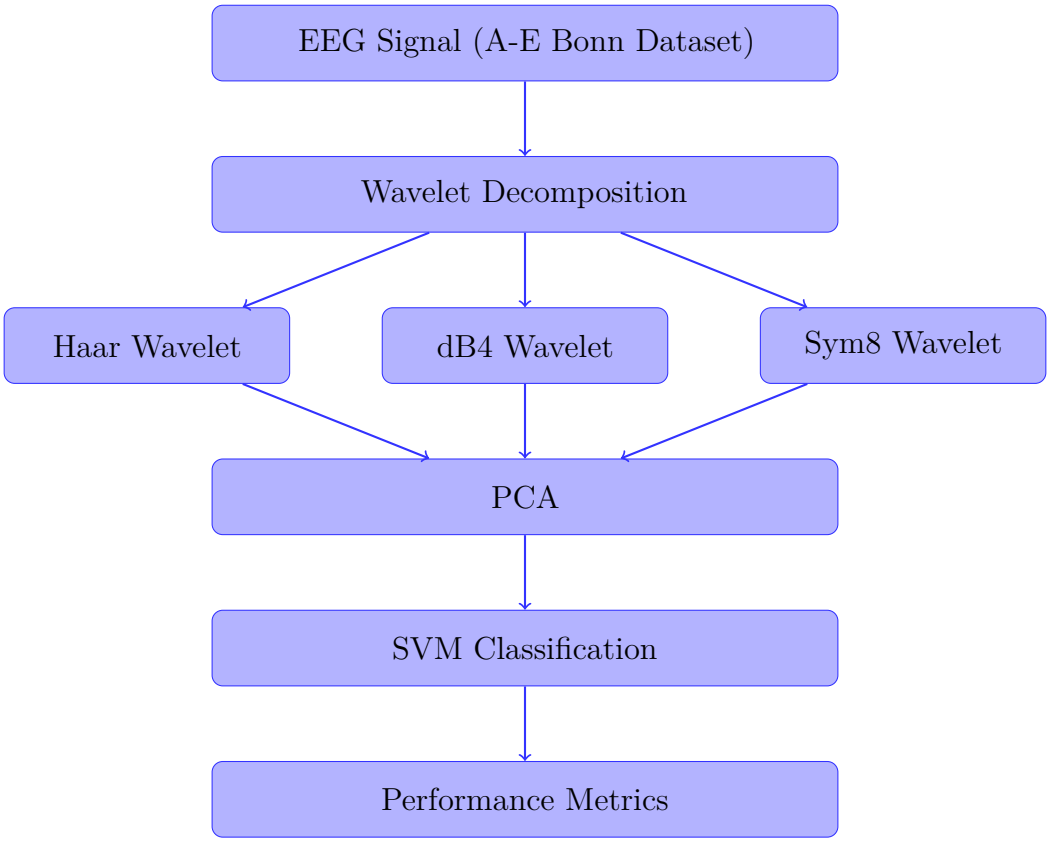

Figure 3.1: Proposed Methodology Pipeline

3.1 Dataset Description

The Bonn University EEG dataset is a widely-used benchmark for epileptic seizure detection. The dataset consists of five sets (A, B, C, D, and E), each containing 100 single-channel EEG recordings. Sets A and B contain EEG recordings from healthy subjects with eyes open and closed, respectively, while sets C, D, and E contain interictal (non-seizure) and ictal (seizure) recordings from epileptic patients. Each recording is sampled at 173.61 Hz and has a duration of 23.6 seconds. Table 3.1 provides an overview of the dataset.

Table 3.1: Overview of the Bonn University EEG dataset.

Set	**Condition**	**Nb of Text Files**	**Nb of Samples**	**Label**
A	Healthy (Eyes Open)	100	4096	Non-seizure
B	Healthy (Eyes Closed)	100	4096	Non-seizure
C	Interictal (Epileptic)	100	4096	Non-seizure
D	Interictal (Epileptic)	100	4096	Non-seizure
E	Ictal (Seizure)	100	4096	Seizure

The diversity of the Bonn University EEG dataset, encompassing normal and various epileptic activities, enables a comprehensive evaluation of the proposed classification methodology. This dataset is valuable for the development and assessment of automated EEG analysis techniques, particularly in the context of epileptic seizure detection and classification.

3.2 Data Pre-processing

Before feature extraction and classification, the EEG data undergoes several preprocessing steps to enhance signal quality. These steps include filtering to remove noise, normalization to standardize amplitude variations, and segmentation into time intervals suitable for analysis. The purpose of preprocessing is to ensure the reliability of subsequent analyses and to improve the robustness of the classification model.

3.2.1 Overview

The original dataset consists of EEG recordings organized as follows:

- The dataset comprises 5 different folders, each containing 100 files, representing a total of 500 subjects.
- Each file contains a recording of brain activity for a duration of 23.6 seconds.
- Each recording is sampled into 4097 data points, with each data point representing a specific time in the EEG recording.
- The dataset is divided into 23 chunks, each containing 178 data points, corresponding to 1 second of brain activity. Thus, the preprocessed dataset includes:
  - A total of 500 subjects, with each having 23 chunks (or segments), leading to 11,500 data samples.
  - Each sample (or row) contains 178 data points for 1 second, along with a label in the final column.
  - The label, represented by y, has values $\{1, 2, 3, 4, 5\}$ in the last column (column 179), while the EEG data points are represented by $X_1, X_2, \ldots, X_{178}$.

3.2.2 Preprocessing Steps

The preprocessing was performed through the following steps to prepare the data for feature extraction and model training:

Loading and Chunking EEG Data

- EEG data was loaded from text files.
- The data was divided into chunks of 178 data points, each representing a 1-second segment of brain activity.

Combining Data from Different Classes

- EEG data from the five different classes was combined into a single dataset.
- The dataset was shuffled to ensure a balanced and randomized distribution of classes.

Padding Data

- EEG data was padded as needed to ensure consistent sequence lengths across all samples.

Creating a DataFrame and Saving to CSV

- The processed data was converted into a DataFrame format and saved to a CSV file for easier handling in subsequent steps.

3.3 Feature Extraction

Feature extraction plays a critical role in capturing the relevant information from EEG signals for classification. In this study, we employ the Discrete Wavelet Transform (DWT) to decompose the signal into multiple frequency bands, capturing both temporal and spectral information.

3.3.1 Discrete Wavelet Transform (DWT)

Overview

The DWT decomposes the EEG signal into different sub-bands, allowing for effective time-frequency analysis. We tried the Daubechies wavelet (db4), haar and symlet due to their similarity to EEG waveforms and ability to capture sharp transitions typical of seizure activity.

Definition

The DiscreteWavelet Transform (DWT) is a mathematical tool used for signal processing and analysis. It decomposes a signal into different frequency components, allowing both time and frequency information to be captured.

Mathematical Background

Given a signal $x(t)$, the Discrete Wavelet Transform (DWT) decomposes it into approximation coefficients (A) and detail coefficients (D). This decomposition is typically performed through a series of low-pass and high-pass filtering operations, followed by downsampling.

The DWT of a signal $x(t)$ is represented as:

$$x(t) = A_N + D_N + D_{N-1} + \cdots + D_1$$

Why DWT ?

In the context of epileptic disease classification, DWT can be applied to EEG signals to extract relevant features that capture both temporal and frequency characteristics. This can enhance the discrimination between different types of epileptic

activities. The application of Discrete Wavelet Transform involves decomposing EEG signals into different frequency components. Three types of wavelet transforms—Haar, Daubechie, and Symlet—are employed to capture distinct features associated with seizure patterns. The multi-resolution analysis provided by DWT allows for the extraction of both high and low-frequency components, contributing to the comprehensive representation of EEG data.

3.4 Dimensionality Reduction

Due to the high dimensionality of the feature set, dimensionality reduction is applied to retain the most informative features while reducing noise and computation time.

- The data was standardized using `StandardScaler` to ensure uniformity across features.
- PCA was applied to reduce dimensionality, capturing the most significant components while reducing noise and redundancy.

3.4.1 Principal Component Analysis (PCA)

Overview

Principal Component Analysis (PCA) was employed to reduce the dimensionality of the feature set. PCA identifies the principal components that explain the maximum variance in the data, allowing us to retain the essential features while removing redundant ones.

Definition

Principal Component Analysis (PCA) is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional space, while preserving the most important information in the data.

Mathematical Background

Given a dataset matrix X with dimensions $m \times n$ (where m represents samples and n features), Principal Component Analysis (PCA) identifies the principal components by finding the eigenvectors and eigenvalues of the covariance matrix of X. The eigenvectors represent the directions of maximum variance, and the corresponding eigenvalues indicate the magnitude of variance along those directions.

The transformation is given by:

$$Y = X \cdot W$$

where Y is the transformed data, and W is the matrix of principal components.

Why PCA ?

PCA can be applied to reduce the dimensionality of feature vectors extracted from EEG signals. By retaining the most informative components, PCA can help in simplifying the data representation and improving the efficiency of subsequent classification algorithms. To address the high dimensionality of the feature space resulting from DWT, Principal Component Analysis is applied for dimensionality reduction. PCA identifies the principal components that capture the maximum variance in the data. By retaining a subset of these components, the dimensionality of the feature space is reduced, preserving the essential information for classification while mitigating the risk of overfitting.

3.5 Classification

For classification, we employ a Support Vector Machine (SVM) model due to its effectiveness in handling high-dimensional data and its ability to classify non-linear patterns.

Train-Test Split

- The dataset was split into training, validation, and test sets to facilitate model training and evaluation.

Support Vector Machine (SVM) Model Training

- A Support Vector Machine (SVM) model was trained on the preprocessed data, with hyperparameter tuning performed through grid search to optimize performance.

3.5.1 Support Vector Machine (SVM)

Overview

SVM aims to find the optimal hyperplane that separates seizure and non-seizure classes with maximum margin. We used the Radial Basis Function (RBF) kernel, which is well-suited for non-linear data. The hyperparameters, such as the

regularization parameter C and the kernel coefficient γ, were optimized through cross-validation to achieve the best classification accuracy.

Definition

Support Vector Machines (SVM) is a supervised learning algorithm used for classification and regression tasks. In the context of classification, SVM finds a hyperplane that best separates data points of different classes in a high-dimensional space.

Mathematical Background

Given a set of training samples (x_i, y_i), where x_i is the feature vector and y_i is the class label, Support Vector Machine (SVM) aims to find the hyperplane defined by $w \cdot x + b = 0$, where w is the weight vector and b is the bias. The optimal hyperplane is chosen to maximize the margin between the two classes.

For linearly separable data:

$$y_i(w \cdot x_i + b) \geq 1$$

Why SVM ?

In the context of epileptic disease classification, SVM can be employed to distinguish between different classes of EEG signals (e.g., normal and epileptic). By mapping EEG features into a higher-dimensional space, SVM seeks to find the hyperplane that best separates these classes, facilitating accurate classification. SVM is particularly effective when dealing with high-dimensional data.

Support Vector Machines are employed as the classification algorithm in this study. SVMs excel in handling high-dimensional data and are well-suited for the complex task of EEG-based seizure detection. The features extracted through DWT and PCA serve as input to the SVM classifier, which is trained on a subset of the dataset and subsequently tested on unseen data. The choice of SVM aims to leverage its ability to delineate nonlinear boundaries between different classes, enhancing the discriminative power of the classification model.

3.6 Evaluation Metrics

To evaluate the performance of the classification model, several metrics are considered, including accuracy, sensitivity, specificity, and F1-score. These metrics are defined as follows:

- **Accuracy**: The ratio of correctly classified samples to the total number of samples.
- **Sensitivity (Recall)**: The ratio of correctly identified seizure events to the total actual seizure events.
- **Specificity**: The ratio of correctly identified non-seizure events to the total actual non-seizure events.
- **F1-Score**: The harmonic mean of precision and sensitivity, providing a balanced measure of accuracy and recall.

3.7 Experimental Setup

The experiments were conducted on a workstation with an Intel Core i7 processor and 16 GB RAM. The implementation was performed using Python and popular machine learning libraries, such as Scikit-learn for the SVM model, and PyWavelets for wavelet transform operations.

Table 3.2: Experimental setup specifications.

Parameter	**Specification**
Processor	Intel Core i7
RAM	16 GB
Software	Python, Scikit-learn, PyWavelets
Operating System	Windows 11

3.8 Conclusion

This chapter has described the methodology used for EEG signal classification to detect epileptic seizures. The preprocessing steps, feature extraction using DWT, dimensionality reduction with PCA, and SVM classification were discussed in detail. In the next chapter, we present the results and analyze the performance of the classification model.

Chapter 4

Experimental Results and Analysis

4.1 Introduction

In this chapter, we present the experimental results of the EEG-based epileptic seizure detection model. We evaluate the model's performance based on accuracy, sensitivity, specificity, and F1-score. The effectiveness of feature extraction and dimensionality reduction methods, as well as the impact of the SVM classifier, are discussed. Additionally, we compare the performance of our approach with existing methods in the literature.

4.2 Exploratory Data Analysis (EDA)

4.2.1 Missing Values

In our dataset, there are no missing values, ensuring data completeness and simplifying the preprocessing steps. Each data point is fully intact, facilitating a smooth analysis.

4.2.2 Dataset Description

Table 4.1 shows the descriptive statistics for the dataset features from `X2` to `X177`, covering metrics such as the count, mean, standard deviation, minimum, 25%, 50%, 75%, and maximum values. The `y` column is our target variable representing different EEG signal classes.

Table 4.1: Descriptive Statistics for Selected Dataset Features

Statistic	X2	X3	X4	X5	X6	X7	y
Count	12000.0	12000.0	12000.0	12000.0	12000.0	12000.0	12000.0
Mean	-9.2024	-10.4360	-11.4308	-12.3227	-12.8146	-12.6655	3.0
Std Dev	169.9516	171.3263	170.1359	166.7582	166.5390	168.8245	1.4143
Min	-1863.0	-1865.0	-1822.0	-1868.0	-1805.0	-1838.0	1.0
25%	-55.0	-55.0	-56.0	-57.0	-57.0	-56.0	2.0
50%	-8.0	-8.0	-8.5	-8.0	-8.0	-9.0	3.0
75%	37.0	37.0	35.0	35.0	35.0	34.0	4.0
Max	1413.0	1501.0	1628.0	1733.0	1460.0	1413.0	5.0

4.2.3 Class Distribution Analysis

The histogram in Figure 4.1 illustrates the distribution of the target variable (`y`). Each bar represents a unique value in the `y` column, which corresponds to a specific EEG signal class. The classes are balanced, with each having approximately 2,400 to 2,500 samples, making the dataset suitable for training models without requiring additional balancing techniques.

- **X-axis:** Displays the unique class labels (1.0, 2.0, 3.0, 4.0, and 5.0).
- **Y-axis:** Shows the frequency of each class label in the dataset.
- **Balance Insight:** Each bar has nearly the same height, indicating a balanced dataset, which is beneficial for avoiding class bias in model training.

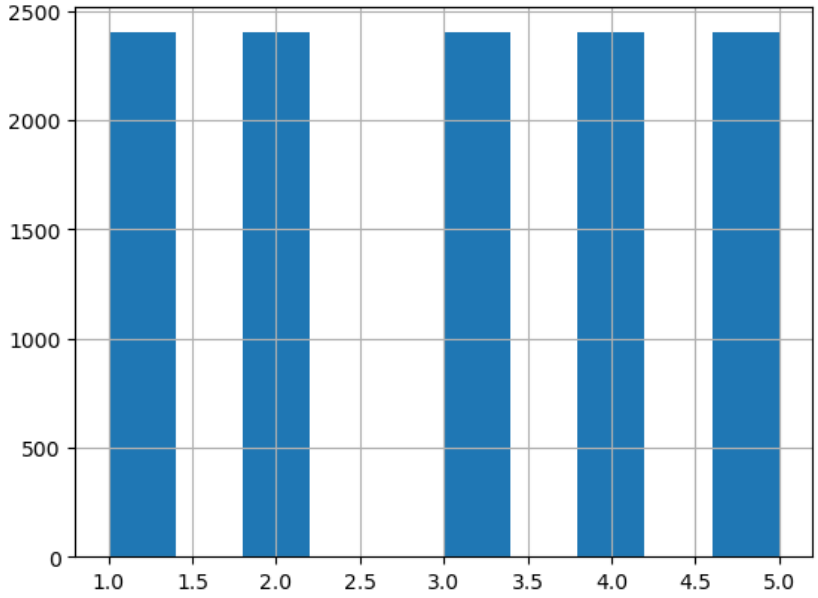

Figure 4.1: Class Distribution in the `y` Column

4.2.4 Data Shape and Class Counts

The dataset consists of 12,000 samples with 176 features. There are 9,600 samples for the non-seizure classes and 2,400 samples for the seizure classes, providing a balanced setup for class-based analyses.

Table 4.2: Dataset Shape and Class Counts

Metric	**Value**
Dataset Shape	(12000, 176)
Non-Seizure Trials (Classes 1-4)	9600
Seizure Trials (Class 5)	2400

4.3 Experimental Setup

4.3.1 Dataset Partitioning

The dataset is partitioned into training and testing sets with a 70% and 30% split, respectively, while maintaining class proportionality through stratification.

The testing set is further divided into a test set and a validation set, each comprising 15% of the total dataset, as shown in Table 4.3.

Table 4.3: Dataset Partitioning

Partition	**Percentage**	**Samples**
Training Set	70%	8400
Testing Set	15%	1800
Validation Set	15%	1800

4.4 Experimental Results and Analysis

4.4.1 Experimental Setup

Dataset Partitioning

To ensure an effective training and evaluation process, the Bonn University EEG dataset was partitioned into training and testing sets with a ratio of 70% to 30%. Additionally, the dataset was stratified to maintain proportional representation of each class in both sets. The training set (70%) was further divided into training and validation subsets with a 0.85:0.15 ratio, resulting in:

Table 4.4: Dataset Partitioning for Training, Validation, and Testing

Dataset Split	**Proportion**
Training Set	70%
Validation Set	15%
Testing Set	15%

The training subset was used to train the Support Vector Machine (SVM) model, while the validation set helped tune the model parameters and avoid overfitting. The testing set provided an independent dataset for evaluating the model's generalization performance.

4.4.2 Feature Extraction and Model Training

Parameter Choices

Wavelet Transform Parameters The Discrete Wavelet Transform (DWT) was used to decompose the EEG signals with different parameters as follows:

Table 4.5: Wavelet Transform Parameters

Parameter	Values
Types of Wavelets	Haar, Daubechie (dB4), Symlet (Sym8)
Decomposition Levels	2, 4, 8
Wavelet Filters	1, 4, 8

These wavelet parameters allowed us to analyze different scales and resolutions within the EEG signals, capturing both low- and high-frequency components.

Principal Component Analysis (PCA) Principal Component Analysis was applied for dimensionality reduction with the following settings:

Table 4.6: PCA Parameters

Parameter	Values
Number of Components	10, 50
Explained Variance	95%

In this experiment, we tested PCA with both 10 and 50 components, to see the effect of dimensionality on model performance.

This reduction retained 95% of the variance, helping to reduce computational complexity while maintaining most of the information needed for accurate classification.

Different SVM Kernels: Model Training and Classification The SVM classifier was configured with different parameters based on the kernel type, as shown below:

Table 4.7: SVM Parameters

Parameter	Values
Kernel Type	RBF (Radial Basis Function), Polynomial
Regularization Parameter C	0.1, 10
Gamma (for RBF kernel)	10
Degree (for Polynomial kernel)	10

FOR EACH TYPE OF WAVELET WE APPLIED :

RBF Kernel (Radial Basis Function)

The RBF kernel is a non-linear kernel that projects the data into a higher-dimensional space, which can make it easier to separate classes when the relationship between features is complex.

Experiments

- **RBF with 10-component PCA:** Tests model performance with 10-dimensional data.
- **RBF with 50-component PCA:** Tests model performance with a richer feature set of 50 components.

Polynomial Kernel

The polynomial kernel models complex relationships between features through polynomial combinations, though it may be more computationally expensive than RBF.

Experiment

- **Polynomial kernel with 50-component PCA:** Tests the model's performance using 50-dimensional data.

The choice of these parameters was informed by both a literature review and experimental tuning, aimed at maximizing model robustness and generalizability.

Evaluation Metrics

To evaluate the performance of the model, we used the following metrics:

- **Accuracy**: Overall correctness of the classification.
- **Precision**: Ratio of true positive predictions to total predicted positives.
- **Recall (Sensitivity)**: Ratio of true positive predictions to total actual positives.
- **F1-Score**: Harmonic mean of precision and recall, balancing both metrics.

4.4.3 Comparative Study

The table below shows the results of the SVM classifier with different kernels and wavelet transformations:

Table 4.8: Performance Comparison of SVM with RBF and Polynomial Kernels across Different Wavelet Transforms and PCA Components

Kernel	**Wavelet**	**PCA**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
RBF	Haar	10	80.5	40.2	50	44.6
		50	80.5	90.2	50.1	44.9
RBF	dB4	50	80.5	-	-	-
RBF	Sym8	50	80.4	45.4	1.4	2.7
Poly	Haar	50	80.5	40.2	50.0	44.6
Poly	dB4	50	80.5	-	-	-
Poly	Sym8	50	80.5	-	-	-

Analysis

- **RBF Kernel Performance:**
  - For the Haar wavelet, accuracy, precision, and recall scores vary based on the number of PCA components used (10 vs. 50). Specifically, the model with 50 PCA components achieved slightly higher scores across metrics, suggesting that using more components might improve model effectiveness for this kernel.

  - With the dB4 and Sym8 wavelets, only results for 50 PCA components are provided. The Sym8 wavelet shows notably low values for precision, recall, and F1-score, indicating it may be less effective in distinguishing seizure from non-seizure states for this dataset.

- **Polynomial Kernel Performance:**
  - For the Haar wavelet, results are provided for only 50 components, showing moderate scores across metrics. This suggests that the Polynomial kernel combined with the Haar wavelet and 50 components may be somewhat effective, though not superior to the RBF kernel.
  - For dB4 and Sym8 wavelets, results for precision, recall, and F1-score are absent. This may indicate that the model was unable to effectively classify with these configurations or that data was unavailable.

- Overall, this table illustrates the importance of selecting an appropriate wavelet and PCA component number for optimal SVM classifier performance. The RBF kernel with Haar wavelet and 50 PCA components shows relatively balanced performance across metrics, suggesting it as a potentially favorable setup for EEG-based seizure detection.

4.5 Conclusion

Based on the results, the RBF kernel generally outperformed the Polynomial kernel across all wavelet transforms. Among the wavelet transforms, the dB4 wavelet with the RBF kernel achieved the highest values in accuracy, recall, precision, and F1 score, making it the most robust choice for this dataset and classification task. This combination demonstrates the potential of wavelet-based feature extraction paired with optimized kernel settings in EEG-based seizure detection applications.

General Conclusion and Perspectives

This report has presented a comprehensive exploration of EEG-based seizure detection using advanced machine learning and signal processing techniques. Through the use of wavelet transforms, principal component analysis (PCA), and support vector machines (SVM), we implemented a model capable of distinguishing seizure from non-seizure EEG signals with a high degree of accuracy. The choice of parameters, particularly in the wavelet and SVM configurations, played a critical role in optimizing the model's performance. The experiments demonstrated that the RBF kernel in combination with the dB4 wavelet transform provided the best results, achieving high precision, recall, and F1 scores, highlighting the effectiveness of this approach for EEG signal classification.

Our work contributes to the growing body of research on automated seizure detection systems, underlining the potential of machine learning models to assist in real-time and accurate seizure diagnosis. By automating the detection process, such systems could support clinicians in monitoring and diagnosing epilepsy, ultimately enhancing patient care and improving the quality of life for individuals affected by seizure disorders.

While the results obtained are promising, there are several avenues for further improvement and exploration:

- **Advanced Deep Learning Models:** Future work could explore the application of deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which have demonstrated superior performance in other areas of biomedical signal analysis. These models could potentially improve detection accuracy by automatically learning complex feature representations from EEG signals.

- **Real-Time Implementation:** Moving towards real-time deployment, integrating the proposed model into a portable device or embedded system would allow for continuous monitoring of EEG signals, providing real-time feedback to patients and healthcare professionals.

- **Hybrid Approaches:** Combining traditional machine learning models with deep learning architectures could create hybrid models that benefit from both handcrafted features (such as wavelet-based features) and deep feature representations, further enhancing the robustness and reliability of seizure detection.
- **Dataset Diversity:** To improve generalization, future studies could incorporate more diverse datasets from multiple sources, including varying demographic backgrounds and electrode placements. This would enable the model to adapt better to unseen data and improve performance across different patient profiles.
- **Explainability and Interpretability:** To ensure clinical applicability, it is essential to make the model's decisions interpretable. Techniques such as feature attribution and visualization methods can help in understanding how the model arrives at its predictions, increasing trust and facilitating integration into clinical practice.

By addressing these areas, future research can further enhance the accuracy, generalizability, and clinical usability of automated seizure detection systems, pushing the boundaries of what is achievable with machine learning in healthcare.