Seizure Detection Using SVM and EEG Signals

BE 175, Fall 2024, June 11, 2025
Lab Section 1C

Adhitya Ram, Yuta Kiami

Introduction

Epilepsy is a common disorder affecting 50 million people globally. Patients are faced with a constant risk of seizures, which can have many complications that can sometimes be fatal. Early detection and diagnosis of epilepsy and related seizures can greatly improve patient outcomes. Electroencephalography (EEG) has become a common way for physicians to diagnose epilepsy and detect seizures, due to the increased electrical activity in the brain that seizures cause. However, the current method for reading EEG signals involves doctors manually interpreting the signals, which requires a specialized technique, is time-consuming, and yields inconsistent results. Implementing machine learning to classify the signal could lead to faster and more accurate results, allowing for earlier diagnosis and improved patient outcomes. We propose the implementation of support vector machines to classify EEG signals

Previous research by Himalyan et. al. showed the viability of using discrete wavelet transform (DWT) and support vector machine (SVM) techniques to train a classifier that can detect epileptic seizures with extremely high accuracy. Using the Bonn EEG dataset, they were able to achieve a 99% accuracy across all non-seizure datasets compared to the seizure set. We aim to reproduce this result using the same data, then implement our own method of training the SVM using principal component analysis (PCA).

Problem Definition

The goal of this project is to reimplement the paper by using DWT to decompose the signal and extract features, then train an SVM model off of those features to classify between seizure and non-seizure. We then plan to use PCA to identify the components that contribute most to the variance in the model, and train the SVM model off of those components, and compare the performance of the two models. Overall, we hope to evaluate whether SVM is a good method for classifying EEG signals as epileptic or non-epileptic.

The ability to classify data this way would show promise for the use of machine learning in classifying many biological signals. Many biological signals are similar to EEG signals in that they can be noisy and unclear, and strong performance in this task would indicate that machine learning could allow for increased ability to read biological signals across the board.

Methods

1. Dataset

We used the publicly available Bonn EEG dataset, which contains single-channel EEG recordings classified into five sets (A through E). Each set consists of 100 samples, each lasting 23.6 seconds and sampled at 173.61 Hz, totaling to 4096 data points per sample.

- **Set A**: EEG recordings from healthy volunteers with eyes open.
- **Set B**: EEG from healthy volunteers with eyes closed.
- **Set C**: Interictal recordings from the hippocampal formation of epilepsy patients.

- **Set D**: Interictal recordings from the epileptogenic zone of epilepsy patients.
- **Set E**: EEG data recorded during seizure (ictal) episodes.

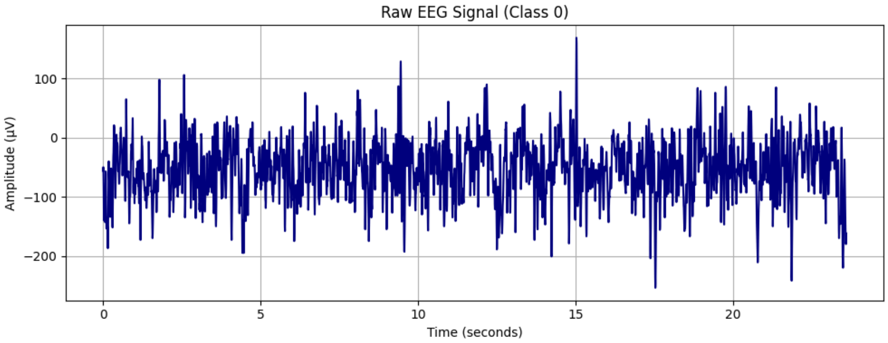


Figure 1a. Raw EEG signal of healthy brain (Set A)

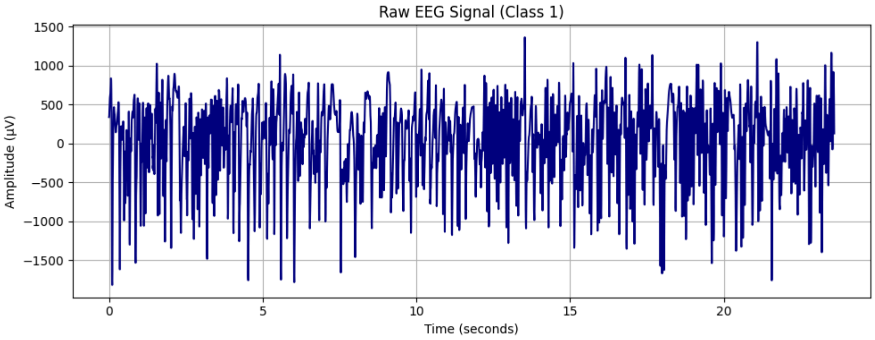


Figure 1b. Raw EEG signal during seizure episode (Set E)

The goal of our research was to develop binary classifiers that could effectively distinguish between seizure and non-seizure brain activity. To evaluate when seizure activity is distinguishable, we progressively added non-seizure data to the negative class and developed the following classification tasks:

- **A vs E**
- **AB vs E**
- **ABC vs E**
- **ABCD vs E**

2. Reference Method: DWT + SVM

We reimplemented the seizure detection pipeline proposed in the paper “Machine Learning-Based EEG Epileptic Seizure Detection: Review, Challenges, and Future Directions” (MDPI, 2023). The authors used the Discrete Wavelet Transform (DWT) to extract relevant features from raw EEG signals and classified them using a Support Vector Machine (SVM) with an RBF kernel. We replicated their preprocessing and modeling steps as a baseline for comparison, then proposed an extension using Principal Component Analysis (PCA) as an alternative feature extraction technique.

3. Preprocessing

(a) Discrete Wavelet Transform (DWT)

The paper used DWT to extract informative features from EEG signals because it can reveal details in both the time and frequency domains of non-stationary signals, such as EEG signals, in a way that other methods can't. This makes DWT an extremely powerful tool for biomedical engineering applications, such as detecting epileptic seizures.

Wavelet transforms are particularly effective for analyzing time-series data where important features may be localized in time, such as seizure onsets. We implemented a 3-level DWT using the Daubechies 8 (db8) wavelet on each 4096-point EEG signal and extracted 8 statistical features from the third-level detail coefficients:

- Mean Absolute Value (MAV)
- Standard Deviation (SD)
- Variance (VAR)
- Skewness (SKEW)
- Kurtosis (KURT)
- Peak Amplitude (PEAK)
- Signal Power (PWR)
- Shannon Entropy (ENT)

These features were then used as inputs to the SVM.

(b) Principal Component Analysis (PCA)

In our creative extension, we replaced the feature extraction through DWT performed by the paper with a Principal Component Analysis, to be applied directly to normalized raw EEG signals.

Pre-processing of EEG data for dimensionality reduction involved each 4096-point signal getting standardized through z-score normalization using scikit-learn's StandardScaler() function. Second, we applied PCA on the whole dataset and generated the cumulative explained variance (R^2X) to identify the amount of variance that was explained by consecutive components. To ensure that we captured as much information as possible without losing dimensionality, we selected the minimum number of principal components that would capture at least 95% of the total variance. Optimal number of components varied for each task of classification; e.g., A vs E required approximately 65 components. These reduced-dimensional representations were used as input features in SVM classifier training.

4. Classification Model: SVM Pipeline

After data preprocessing, we trained a Support Vector Machine (SVM) with a radial basis function (RBF) kernel to classify seizure versus non-seizure EEG activity. The modeling pipeline for every classification problem (A vs E, AB vs E, etc.) was identical.

The data was first split into an 80% training set and a 20% holdout test set, stratified for class balance. Preprocessing differed per model: the DWT pipeline employed wavelet-based feature extraction, while the PCA pipeline utilized dimensionality reduction with optimal component selection using R^2X threshold.

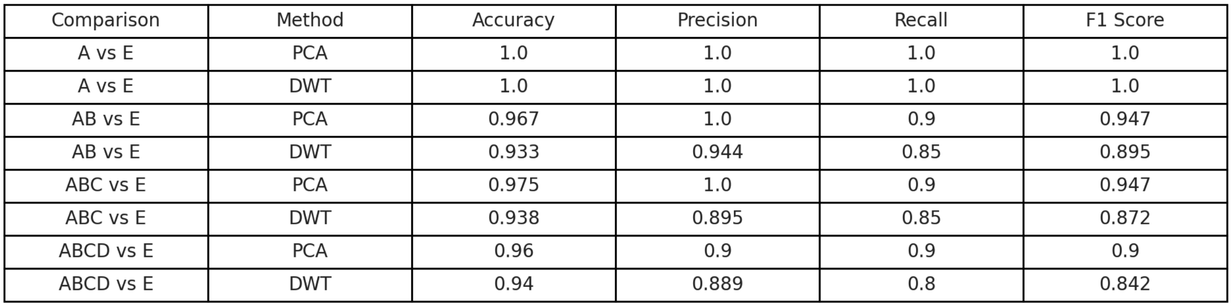
In an effort to optimize the performance of the classifier, we performed 5-fold cross-validation of the training data, which evaluated generalization and reduced the risk of overfitting. We then used GridSearchCV to perform a comprehensive search over combinations of hyperparameters, specifically testing C values of [0.1, 1, 10, 100] and gamma values of ['scale', 0.01, 0.1, 1]. After finding the best-performing hyperparameters, we re-trained the final model on the whole training set and evaluated its performance on the held-out test set. Performance measures employed are accuracy, precision, recall, and F1 score and are discussed further in the results section.

5. Software Packages and Libraries Used

We used Python for this project and included various standard computing and machine learning libraries. NumPy and Pandas allowed us to organize and handle data, PyWavelets allowed us to perform the DWT to extract the features, and Matplotlib allowed visualization. For machine learning libraries, Scikit-learn allows us to z-score, perform PCA, cross-validate, train our SVM model, and tune hyperparameters using gridsearch. It also allowed us to produce evaluation metrics used to evaluate our model, such as accuracy, precision, recall, and F1-score.

Results

1. Classification Performance



Comparison	Method	Accuracy	Precision	Recall	F1 Score
A vs E	PCA	1.0	1.0	1.0	1.0
A vs E	DWT	1.0	1.0	1.0	1.0
AB vs E	PCA	0.967	1.0	0.9	0.947
AB vs E	DWT	0.933	0.944	0.85	0.895
ABC vs E	PCA	0.975	1.0	0.9	0.947
ABC vs E	DWT	0.938	0.895	0.85	0.872
ABCD vs E	PCA	0.96	0.9	0.9	0.9
ABCD vs E	DWT	0.94	0.889	0.8	0.842

Table 1. Performance metrics

As shown in Table 1, the PCA-based SVM pipeline achieved strong performance across all classification tasks. It reached perfect scores (100% accuracy, precision, recall, and F1) in the simplest case and maintained strong performance even as the classification task became more challenging. For the most complex comparison (ABCD vs E), PCA achieved 96% accuracy and an F1 score of 0.90.

Compared to the original DWT-based pipeline, PCA consistently matched or exceeded performance across all tasks. In particular, the PCA pipeline demonstrated higher recall and F1 scores in the more difficult scenarios involving larger and more varied non-seizure data. These improvements suggest that PCA is not only a valid alternative to DWT preprocessing but may offer improved generalizability and robustness for EEG-based seizure classification.

However, we also must note that the paper's version of the DWT-based pipeline performed better than both of the pipelines that we developed. So, even though PCA had a very strong performance in our pipeline, DWT is still the preferred method, especially when analyzing high-dimensional time series data.

Overall, we can confidently say that regardless of using DWT or PCA, SVM shows very high promise in application to EEG signal classification in regards to epilepsy.

2. Dimensionality Reduction Results (PCA)

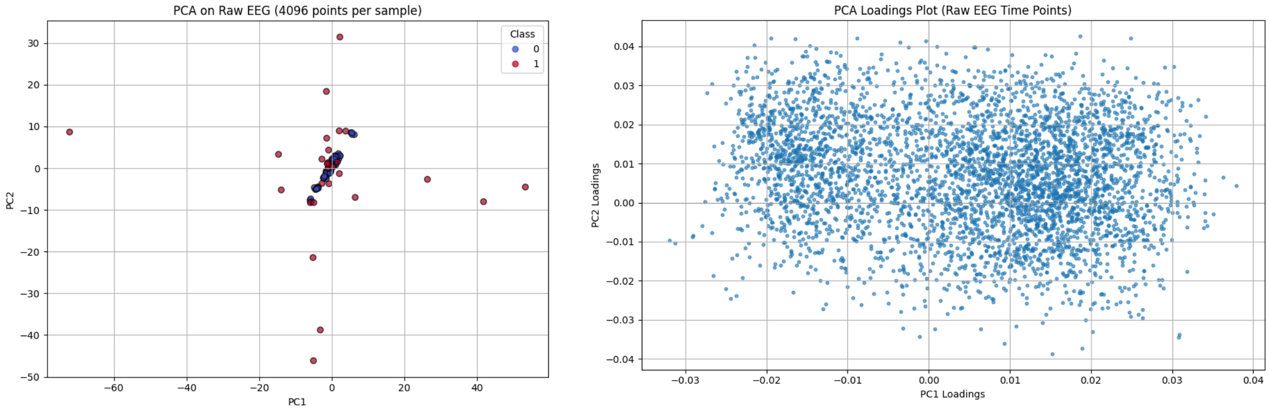


Figure 2. PCA scores and loadings plot

The scores plot shows the projection of EEG recording samples onto the first two principal components. These plots were also produced after the removal of outliers in order to promote better generalizability in our models and avoid skewing of our results. Each point represents a single EEG sample, colored by class (blue = non-seizure, red = seizure). The PCA-transformed data shows a concentrated cluster of non-seizure samples, while seizure samples exhibit greater spread and variability. This is an expected result as seizure activity and dynamics are highly variable across different people. Additionally, from more analysis we found that PC1 explained 16.5% of the variance while PC2 explained 10%.

This loadings plot visualizes the contribution of each of the 4096 time-domain EEG features to the first two principal components. Each point corresponds to a time index in the raw EEG signal. The symmetric and dispersed structure of the loadings suggests that variance is not dominated by any one region of the signal but is instead distributed across time.

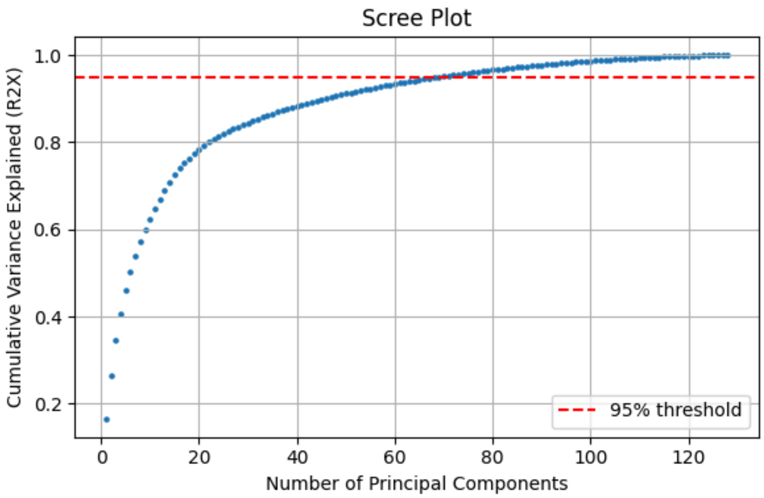


Figure 3. Scree plot

This scree plot for the A vs E classifier shows the cumulative variance explained (R^2X) as a function of the number of principal components. We selected the number of components needed to explain at least 95% of the total variance (denoted by the red dashed line). For the A vs E comparison, about 65 principal components were required to meet this threshold. This dimensionality reduction retained nearly all relevant information while reducing computational cost and overfitting in the SVM classifier.

3. Classifier Boundary Visualization

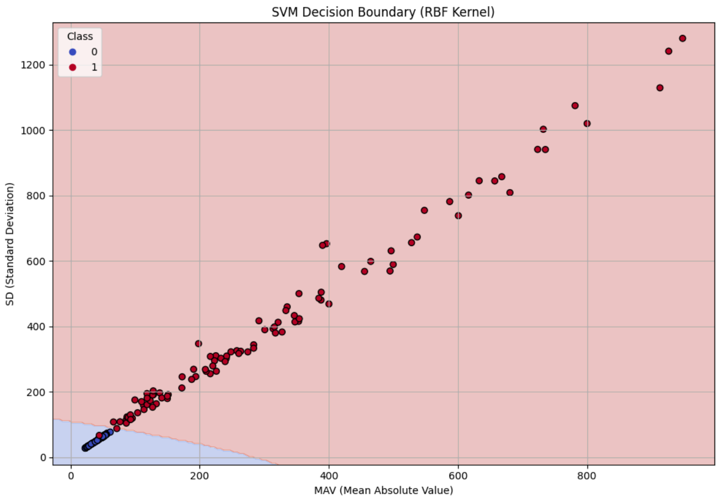


Figure 4. SVM decision boundary

This plot visualizes our decision boundary for our SVM in two of the dimensions. We can see from this plot that our decision boundary does extremely well in creating a boundary that accurately classifies points, misclassifying an extremely small number of points. The plot looks lopsided due to the exclusion of the other dimensions related to the decision boundary.

Bibliography

- Himalyan, S., & Gupta, V. (2024, November 26). *Support Vector Machine-based epileptic seizure detection using EEG signals*. MDPI. https://www.mdpi.com/2673-4591/18/1/73
- Andrzejak, R., Lehnertz, K., Mormann, F., Rieke, C., David, P., & Elger, C. (2001, November 20). *Phys. rev. E 64, 061907 (2001) - indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and Brain State*. Physical Review E. https://journals.aps.org/pre/abstract/10.1103/PhysRevE.64.061907