Abstract: In this project, the electroencephalogram (EEG) channel(s) is used to better characterize post-traumatic stress disorder (PTSD). For this aim, we applied boosting methods along with a combination of k-means and Support Vector Machine (SVM) models to find the diagnostic channels of PTSD cases and healthy subjects. We grouped 32 channels and 12 subjects (6 PTSD and 6 healthy controls) using k-means. Channels of the brain are grouped by the k-means clustering method to find the most similar part of the brain. This approach uses SVM by performing classification based on cluster classes are been mapped to EEG channels. This mapping uses information across all samples without the bias of using the outcome variable. The linear SVM found weights that distinguished channels within each subject for each cluster to compare the PTSD cases and healthy controls' channel weights. It was found that the significant SVM weights of F4, F8, and Pz were smaller in subjects with PTSD than in healthy subjects. This new method can be used as a tool to better understand the relationship between EEG signals and diagnosis.

Keywords: linear support vector machine; k-means clustering; PTSD; EEG

1. Introduction

Electroencephalography (EEG) records electrical activity produced in the cerebral cortex of the brain [1] and contains information about the brain’s neuronal electrical activities [2]. EEG has been used for clinical purposes to predict outcomes such as depression [3]. The extraction of time-series features from the EEG signal is an important approach to identify signals that can differentiate healthy subjects from patients. However, additional improvements are needed to improve sensitivity and specificity [1].

Diagnostic interviews are commonly used to assess PTSD and determine if an individual meets the criterion for a PTSD diagnosis. They can provide detailed information of the individual’s symptoms, including their overall functioning, medical and psychiatric history, and the presence of other conditions. Advantages of diagnostic interviews include their comprehensive nature, ability to evaluate the existence and seriousness of symptoms, and the fact that they are well validated and widely used in clinical settings. Disadvantages of diagnostic interviews include their time-consuming nature and the potential for bias in the interviewer’s assessment. Interviews provide an opportunity for the direct assessment of symptoms and diagnosis of PTSD based on specific criteria. Many discussions, like the Clinician-Administered PTSD Scale (CAPS), have established clinical validity and reliability. Interviews can be adapted to different populations and can be modified based on the specific needs of the individual being assessed. But interviews may be subject to bias, such as interviewer bias, recall bias, and social desirability bias. They can be time consuming and may require a trained interviewer. Interviews can be costly as a trained interviewer is required [4,5].
Psychological assessments refer to a range of measures that are used to evaluate an individual’s mental health and functioning. This can include self-reporting questionnaires, behavioral assessments, and clinical interviews. Advantages of psychological assessments include their ease of administration, cost effectiveness, and the fact that they can be administered in a variety of settings. Disadvantages of psychological assessments include the potential for response bias and limitations in the interpretation of results. The advantage of self-reported measures is that many psychological assessments for PTSD are self-reported measures, which can be completed in a relatively short amount of time; hence, they are less expensive than diagnostic interviews. Psychological assessments for PTSD often use standardized measures, which help ensure consistent and reliable assessments. Many psychological assessments for PTSD have established validity and reliability. But they may only provide a limited assessment of other factors, such as physical health, that may also impact PTSD symptoms [6,7].

Neuroimaging studies refer to the use of various imaging techniques. Disadvantages of neuroimaging studies include the need for specialized equipment and the potential for exposure to ionizing radiation. Neuroimaging studies can provide a direct assessment of brain activity and may help to identify specific brain regions associated with PTSD. Neuroimaging studies provide objective measures, which can help to reduce bias in assessments. They may have the potential to identify biomarkers for PTSD, which can help to diagnose the condition and guide treatment [8].

An important biomedical application of EEG signals is investigating PTSD cases, which often exhibit sleep disturbances and several symptoms related to a traumatic event [9]. Previously, we employed a time-series Hurst analysis of EEG signals to recognize healthy controls and PTSD subjects. We found the Hurst exponent in channel F3 to be smaller in PTSD subjects compared to healthy controls [1]. In another study, Kim et al. [10] applied non-linear interdependence (NI) determined from EEG data with 16 channels of 18 healthy controls and 18 PTSD cases. This study showed the growth of non-linear functional connectivity in cortical networks in PTSD in channels of F7, F3, T3, C3, T5, and P3 and decreases in F4, C4, P4, and O2 [10].

In addition to standard statistical methods and feature extraction, machine learning methods are increasingly popular for prediction and modeling in biomedical research. Many machine learning algorithms extract feature importance that indicates the role of variables in prediction models. In addition, some methods can compute indirect feature importance from machine learning models. In Amin Zandvakili’s study [11], Support Vector Machine (SVM) weights show that after treatment, depressive disorder symptoms reduced by half [11].

The study is organized with three main parts. First, we describe the EEG data that are available for the study. Next, the new analytical approach of k-mean clustering and linear SVM is explained. Finally, we apply this new approach to an EEG study of PTSD and discuss challenges and potential applications.

2. Data Description

The data were assembled at the Laureate Institute of Brain Research [12]. The ethical code of the World Medical Association (Declaration of Helsinki) for human tests was applied to the implements on human subjects.

The data included 12 male subjects, 6 PTSD (P) cases and 6 healthy (H) controls, with an average age of 27 ± 5 years. For each subject, there is a dataset with 31 EEG channels and 50,000 time points. The 31 channel names, 1 to 31, respectively, are Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T7, T8, P7, P8, Fz, Cz, Pz, Oz, FC1, FC2, CP1, CP2, FC5, FC6, CP5, CP6, TP9, TP10, POz. The signals of the channels were collected with electrodes placed at AFz and FCz and tagged at a testing space of 5000 samples/second and a resolution of 0.1 µV. The study used the software BrainVision Analyzer2 (https://brainvision.com/products/analyzer-2/) accessed on 1 February 2016. [13] to process EEG signals. For
an advanced assessment, we employed EEGLAB software (http://sccn.ucsd.edu/eeglab, accessed on 1 January 2024) [14].

There were millions of data points per channel. The first 50,000 data were cut off and the second 50,000 data points taken as a sample.

The study applied the template deduction method to remove the MRI gradient artifact and cardio ballistic (BCG) artifact for EEG processing [15]. A sample of 250 samples/second (4 ms temporal resolution) to 40 Hz was selected. We removed the artifacts and subject head movement for an independent component analysis (ICA). The scanning spanned 526 s. For signal consistency, the initial 6 s were removed. Each channel has 130,000 timepoints. We selected 50,000 reliable data points after correcting for subject movement.

3. Methods

The overall approach is first to identify clusters of EEG channels from the time series for each subject. In other words, each subject has a separate set of channel clusters. We then apply SVM to each subject, where the inputs are the cluster IDs of the 31 channels and the class label is the binary indicator of PTSD. We find the model and the weight of each channel. This dimensionality reduction maps each channel’s time series to an SVM weight and results in a derived data matrix with SVM weights as the predictor features. Finally, we use the clinical outcome variable to test the ability of the SVM-based features to discriminate between PTSD cases and healthy people.

3.1. K-Means Clustering

We use k-means to cluster channel time series into six groups within each subject. We computed k-means clusters for a different number of clusters (k), between 1 and 31, and we found that k = 6 provided the best cluster size balance. With more clusters, many groups had only 2 or 3 channels and some groups had none. Fewer clusters led to inferior diagnostic effectiveness in the following analysis. We used the Euclidean distance as the clustering metric. We calculated the distance between channels and used k-means clustering to classify the channels. This clustering is repeated to determine groups of channels for each subject.

3.2. Linear SVM Classification

SVM identifies a hyperplane or decision boundary that separates instances (in this case, channels) with a maximum margin of the support vectors to separate instances into classes (in this case, clusters) [15,16]. The hyperplane for 12 subjects is specified by weights, which are used to construct a new dataset.

The vector $\vec{x}$ represents a data point in the n-dimensional space. Here, x is the vector of 31 channel cluster ids. The values of x can be integers 0 to 5, corresponding to six clusters. A 12 * 31 matrix for 12 subjects and 31 channels are obtained. The main task of the linear SVM is finding the linear model [11]

$$f(\vec{x}) = \begin{cases} 
1 & \text{if } \vec{w} \cdot \vec{x} + b \geq 1 \\
-1 & \text{if } \vec{w} \cdot \vec{x} + b \leq -1 
\end{cases}$$

that is defined by the best hyperplane:

$$\vec{w} \cdot \vec{x} + b = 0.$$ 

$\vec{w}$ and b need to be determined. They are equivalent to the objective of maximizing the margin:

$$\text{Margin} = \frac{2}{||\vec{w}||}$$
So, the main factor that affects the classification is vector \( \vec{w} \) which is the weight in the linear model. An assessment of this vector \( \vec{w} \) can help to show how the data points are classified and how the classification is determined. The weights, \( \vec{w} \), indicate the importance of each channel for separating the channels into 6 groups for a given subject [3].

The k-mean SVM boosting method was defined by Yao et al. in Ref. [17]. Then, it was applied to predict cancer by Kim et al. in Ref. [18].

The SVM model is applied to each subject based on the k-means cluster labels of the EEG time series. Then, the weights (\( \vec{w} \)) can be aggregated across subjects to construct a subject by the weight dataset to test the most important channels for distinguishing between subjects with PTSD and the healthy control subjects (Figure 1).

![Methodology diagram](image)

**Figure 1.** Methodology diagram. The frequency of 31 channels is given weights by k-means clustering \((k = 6)\). The data-driven weights are used to train the support vector machine (SVM) model.

### 3.3. K-Mean SVM Method

For each subject, we cluster the EEG time series into six sets of channels by k-means. Figure 2 shows the subjects and sets of channels in different colors.

![Graph](image)

**Figure 2.** Six groups of channels for health controls H1 to H6 and PTSD cases P1 to P6 using 6-mean clustering method. Background colors indicated different clusters and red font values highlight the diagnostic channels.

We then apply SVM to each subject to classify the six sets of channels and identify hyperplane weights for each one. This results in a dataset \((12 \text{ subjects} \times 31 \text{ channels})\) where each subject has an SVM weight for each channel. Finally, we use this derived dataset to perform hypothesis testing \((t\text{-tests})\) to test channel weights that may distinguish between PTSD and controls (Figure 3).
Signals 2024, 5

4. Results and Discussion

There are six health controls and six PTSD cases. In each subject, the data of 31 channels is presented as a time-series dataset with a size of $31 \times 50,000$. After applying k-means ($k = 6$) we applied individual SVM models to obtain 31 value weights showing, respectively, the importance of 31 channels in classifying six clusters of 12 subjects. With the same combination of k-means and SVM, we generate the fixed matrix of weights (size $12 \times 31$ for 31 channels in 12 data subjects). To compute the importance of each channel for differentiating healthy subjects and PTSD subjects, we compare the different weights of each channel over 12 data subjects by 31-line graphs, matching, respectively, 31 channels.

In the current paper, we applied machine learning methods to construct the importance of each channel for differentiating healthy subjects and PTSD subjects, which needs to be checked by hypothesis testing. The weights of channel F4 (channel 4), channel F8 (channel 8), and channel P3 (channel 19) are significantly higher in the healthy control group (H) versus the PTSD group (P) (Figure 4). All of these channel weights are higher in healthy subjects than in subjects with PTSD except for three values: subject P1 in channel 4 (F4), subject P5 in channel 8 (P4), and subject P3 in channel 19 (Pz).

In channel F4, all weights in healthy subjects are equal to or greater than 0.01297, while all of the weights in subjects with PTSD are lower than this value (Figure 3). The pattern is

**Table 1**

| Channel | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 |
|---------|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| P1      | 0.0189579 | 0.00089 | 0.01374 | 0.0071 | 0.01355 | 0.0062 | 0.01797 | 0.0062 | 0.01297 | 0.0081 | 0.01972 | 0.0057 | 0.02305 | 0.0035 | 0.02878 | 0.0042 | 0.02705 | 0.0048 | 0.02634 | 0.0057 | 0.01984 | 0.0072 | 0.02411 | 0.0057 | 0.02219 |
| P2      | 0.00612 | 0.00012 | 0.00615 | 0.0081 | 0.0069 | 0.0086 | 0.0025 | 0.0017 | 0.00891 | 0.0016 | 0.00589 | 0.00932 | 0.00568 | 0.00996 | 0.0124 | 0.00524 | 0.0023 | 0.0015 | 0.0084 | 0.0022 | 0.00487 | 0.0016 | 0.00547 | 0.0016 | 0.00279 |
| P3      | 0.01265 | 0.00101 | 0.00245 | 0.0024 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 |
| P4      | 0.00719 | 0.00159 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 |
| P5      | 0.00651 | 0.00159 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 |
| P6      | 0.01845 | 0.00101 | 0.00245 | 0.0024 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 |
| P7      | 0.00719 | 0.00159 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 |
| P8      | 0.00651 | 0.00159 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 | 0.0015 | 0.0004 |

**Figure 3.** The weights of channels per subject for the SVM hyperplane model that divides channels of subjects into 6 groups.

**Figure 4.** Red numbers are minimum weight of each channel. (a) Box plot of weights of channel F4 in each group of subjects; (b) box plot of weights of channel F8 in each group of subjects; (c) box plot of weights of channel Pz in each group of subjects.
similar in channel F8 and Pz with the different bordering values of 0.019724 in channel F8 and 0.012953 in channel Pz.

We test the hypothesis of differences in channel weights between the healthy and PTSD groups using the Mann–Whitney U test (Table 1).

Table 1. Individual-level SVM channel weights from clustering and p-values for three significant channels by Mann–Whitney U test.

<table>
<thead>
<tr>
<th>Channel</th>
<th>Subject</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
<th>H4</th>
<th>H5</th>
<th>H6</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
</tr>
</thead>
<tbody>
<tr>
<td>F4</td>
<td>Weights</td>
<td>0.129</td>
<td>0.126</td>
<td>0.027</td>
<td>0.053</td>
<td>0.012</td>
<td>0.026</td>
<td>0.004</td>
<td>0.012</td>
<td>0.005</td>
<td>0.003</td>
<td>0.011</td>
<td>0.099</td>
</tr>
<tr>
<td></td>
<td>Group Average</td>
<td>0.06267 (S.D. = 0.04019)</td>
<td>0.02248 (S.D. = 0.01554)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F8</td>
<td>Weights</td>
<td>0.086</td>
<td>0.131</td>
<td>0.023</td>
<td>0.104</td>
<td>0.036</td>
<td>0.020</td>
<td>0.008</td>
<td>0.002</td>
<td>0.022</td>
<td>0.005</td>
<td>0.005</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>Group Average</td>
<td>0.06678 (S.D. = 0.01916)</td>
<td>0.00949 (S.D. = 0.00303)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pz</td>
<td>Weights</td>
<td>0.080</td>
<td>0.041</td>
<td>0.061</td>
<td>0.014</td>
<td>0.030</td>
<td>0.013</td>
<td>0.021</td>
<td>0.013</td>
<td>0.003</td>
<td>0.005</td>
<td>0.001</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>Group Average</td>
<td>0.03999 (S.D. = 0.0109)</td>
<td>0.00776 (S.D. = 0.00305)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The p-values of the Mann–Whitney U test for channels F4, F8, and Pz are statistically significant for average weights in healthy subjects when compared to average weights in PTSD cases (Table 1). These findings suggest that channels F4, F8, and Pz may differentiate PTSD and healthy controls based on our combination of machine learning-based weights.

These results are consistent with previous reports for channels F4 and Pz as important channels to differentiate healthy subjects and subjects with PTSD [12,18]. The study by Rutter [19] found F4 and Pz to be channels associated with the disorder [18].

The main function of channel F4 is like channel F3; namely, motor planning for the right arm movement while channel F4 mainly controls motor planning for the left arm [15]. This symmetry explains the similarity in channel F4 and channel F3 in the discriminating between healthy subjects and subjects with PTSD. Previous studies have reported the association of these channels with PTSD in Refs. [12,15]. Our previous Hurst exponent analysis found an association of channel F3 with PTSD. In that paper, we applied the Hurst exponent as a useful statistical method to find the distinctive channel of the brain between subjects with PTSD and healthy people based on long-term memory. Based on that paper, the F3 channel of all six healthy controls was higher than the PTSD cases [1]. In the current paper, we applied machine learning methods to confirm the previous work and find more diagnostic channels. The current study suggests that channels F4, F8, and Pz are important factors in discriminating between healthy subjects and subjects with PTSD. F3 and F4 are parallel channels, and both are related to long-term memory. These results may find practical application in the diagnosis of PTSD, and the channel weight construction approach may be useful in combination with other machine learning methods.

A study of veterans with PTSD syndrome confirmed our results about diagnostic F3-F4 channels [19].

The insights from this analysis have the potential to dramatically change the diagnosis of PTSD in patients. There are several ways that EEG evaluation could complement existing tools. One possibility is to use EEG screening as an initial tool to evaluate the possibility of PTSD. This could be conducted as part of general medical screening among those who may have undergone trauma. Another opportunity is to use EEG measures at an initial stage of PTSD diagnosis. Especially for patients who self-report symptoms, or those who have completed a questionnaire on their own, EEG evaluation is a cost-effective and unbiased step in the process. This allows reserving in-depth interviews for those patients where the severity of PTSD is present. Even for patients who have been diagnosed with PTSD,
EEG evaluation, using the method proposed here, provides an unbiased evaluation on the occurrence of PTSD.

**Author Contributions:** Conceptualization, P.N., E.S., S.T., B.M., B.R.; methodology, V.N., M.P., P.N., B.R.; software, M.P., B.M.; validation, P.N., E.S., B.M., B.R.; formal analysis, M.P., T.W., P.N., B.R.; investigation, P.N., E.S., B.M., B.R.; resources, P.N., B.M., B.R.; data curation, M.P., B.R.; writing—original draft preparation, M.P., T.W.; writing—review and editing, P.N., E.S., S.T., B.R.; visualization, M.P., T.W., S.T.; supervision, P.N., E.S., B.M.; project administration, B.R.; funding acquisition, E.S.; All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Data are contained within the article.

**Acknowledgments:** This paper and the research behind it would not have been possible without the exceptional support of Jerzy Bodurka who passed away on Friday, 13 August 2021. He provided the EEG data for our projects accessed on 1 February 2016. [https://www.laureateinstitute.org/in-memoriam.html](https://www.laureateinstitute.org/in-memoriam.html).

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

14. EEGLAB, Swartz Center for Computational Neuroscience, UC San Diego, Regents of the University of California. Available online: [http://sccn.ucsd.edu/eeglab](http://sccn.ucsd.edu/eeglab) (accessed on 1 January 2024).

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.