TO UNDERSTAND OR NOT TO UNDERSTAND

Analysis of EEG Data for Subject Understanding Prediction and Clustering

ABSTRACT

This report investigates the potential of electroencephalography (EEG) to predict subject comprehension during educational video viewing. Using a dataset collected during the COVID-19 lockdown with 14-channel EEG recordings and self-reported understanding labels, we implemented both supervised and unsupervised learning methods. Logistic Regression and Support Vector Machines (SVM) achieved high classification performance, with SVM reaching a balanced accuracy of 88.66%. Principal Component Analysis (PCA) and K-Means clustering revealed latent structure in the data, with optimization techniques enhancing cluster interpretability in relation to comprehension. The findings support the viability of EEG-based models for adaptive learning technologies and cognitive state detection in educational settings.

Casper van Laar - 2440678

casperdvanlaar@hotmail.com
University of Wolverhampton
MSc Artificial Intelligence
7CS074/UZ3: Data Mining & Informatics 23/04/2025
Word count: 3037

Contents

Introduction

Electroencephalography (EEG) is a non-invasive neurophysiological technique that records the brain's electrical activity through sensors placed on the scalp. This activity reflects underlying neural processes involved in cognitive functions such as attention, memory, and comprehension. Analysing EEG data enables researchers to gain valuable insights into how individuals process and understand information. Understanding neural signatures of comprehension holds particular importance for fields like cognitive neuroscience, where brain activity patterns are linked to internal mental states (Song, Shim, Rosenberg, 2023; Keller, *et al 2024*).

Predicting whether a subject has understood educational content based on EEG data can have positive implications for the development of adaptive learning technologies, human-computer interaction, and individualized education (Ramírez-Moreno et al., 2021). This report aims to explore this potential by developing and evaluating classification models that predict subject understanding based on EEG recordings. Specifically, the objectives include the implementation of Logistic Regression and Support Vector Machine (SVM) classifiers, analysis of EEG data through Principal Component Analysis (PCA) and K-means clustering. These tasks are intended to uncover latent patterns in brain activity associated with learning and to compare the efficacy of different machine learning methods in modelling such complex neural data.

Data collection & understanding

Dataset Description:

The dataset comprises EEG recordings collected from students as they watched educational videos. The data is organized into three CSV files: EEG_data.csv, Subject_details.csv, and Video_details.csv. The EEG_data.csv file contains 68,831 rows and 87 columns, including video and subject identifiers, raw EEG data from 14 electrodes (AF3, F7, etc.), and the power of five different brain wave frequencies (Theta, Alpha, BetaL, etc.) derived from those electrodes (Omar et al., 2022). It also includes a binary label (subject_understood) indicating whether the subject understood the video (1 = understood, 0 = did not understand). The Subject_details.csv file provides demographic information for the participants (n=8), including age (ranging from 11 to 24 years), gender (male and female), education level (spanning from middle school to university), field of interest, and ethnicity (predominantly Arab) (Appendix, A3). The Video_details.csv file includes 11 rows and 4 columns with information like Video_ID, Title, URL, and Instructor. Together, the subject and video metadata provide context for deeper analysis of the EEG recordings, enabling exploration of how different demographics or video content influence learning and comprehension. The dataset was collected using the Emotiv Epoc X 14-channel EEG headset during the COVID-19 lockdown, where students of various educational backgrounds were invited to watch lectures online. The researchers first assessed the students' knowledge and selected videos based on topics they were likely to understand or find challenging. After viewing, subjects were asked if they understood the lecture, and in some cases, comprehension was verified through testing (e.g., Khan Academy quizzes).

Data Source Justification:

This dataset was chosen for its significance in evaluating cognitive engagement within online learning environments. The presence of a clear understanding label facilitates classification, enabling the training of machine learning models to identify learning states. Moreover, the dataset provides a diverse set of EEG features, supporting detailed exploratory data analysis (EDA), signal processing, and clustering techniques. Compared to other public EEG datasets, this one stands out due to its context-specific nature (online learning), the availability of both subjective (self-reported understanding) and objective features (EEG), and the inclusion of contextual metadata. These factors make it ideal for developing educational technology tools and studying brain-based learning indicators.

Ethical Considerations:

Several ethical considerations were addressed. Data was collected with informed consent, and all personally identifiable information was anonymized. While the dataset is openly shared under the CC0 license, caution is needed regarding potential biases from subject demographics (e.g., cultural background, education, gender) and the limited range of videos. The findings may not generalize to other populations or contexts. To mitigate these biases, future research should include diverse participants and materials and employ stratified sampling. The current study is also limited by its reliance on self-reported understanding, which may introduce subjectivity
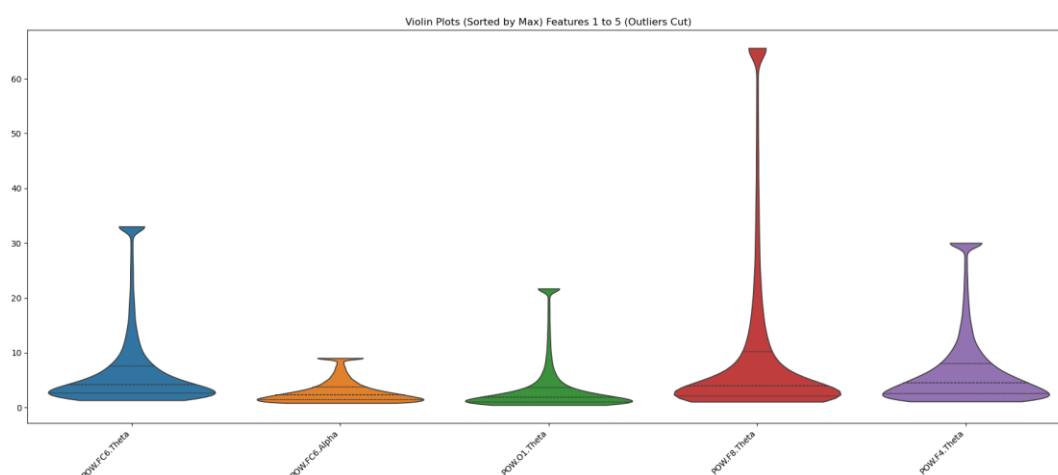
Initial Data Exploration:

Initial exploration of the EEG_data.csv file revealed a clean and well-structured dataset, free of missing values and duplicate rows—simplifying preprocessing. However, Z-score analysis (threshold ±3 SD) identified several outliers across EEG channels and power features. These may represent atypical neural activity or noise, warranting further examination based on the modelling approach.

Descriptive statistics (Appendix Table 1) for channels like AF3, F7, FC5, and T7, and wave bands (Theta, Alpha, BetaL, BetaH, Gamma) revealed substantial variability. For example, the AF3 channel had a mean of 4276.5 µV (SD = 115.6 µV), ranging from 1030.8 to 6238.1 µV. Likewise, power features like POW.F8.Theta ranged from 0.18 to 12,533, with an SD of 146.1, indicating potential skewness or bursts of cognitive activity.
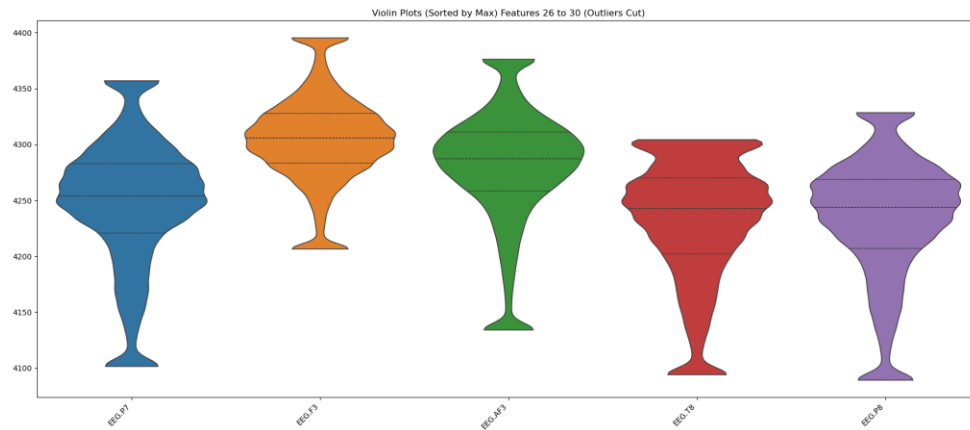
Outliers were visualized using violin plots (Figure 1, Figure 2, and Appendix A1), showing the median, quartiles, and data distribution. The top and bottom 5% of data points were capped to improve the visualization of the central tendency. Features were sorted by their maxima to highlight their ranges. Despite their extremity, outliers were retained for initial analysis, as they may represent meaningful cognitive phenomena.

The correlation matrix (Figure 3) revealed strong intra-group correlations within EEG channels and power features, potentially reflecting the spatial proximity of sensors or the shared physiological origin of certain EEG components. However, correlations between EEG channels and power features were generally weaker, suggesting that they represent different aspects of neural activity—perhaps localized versus distributed processes. Notably, POW.P7.BetaH, POW.O1.Theta, and POW.O1.BetaL showed stronger interrelations, forming a distinct subgroup, which could indicate functional coupling between these brain regions and frequency bands during the task. The significant correlations observed among alpha, beta, and gamma waves likely reflect coordinated communication between different neural populations operating at various frequencies, supporting information processing and integration.
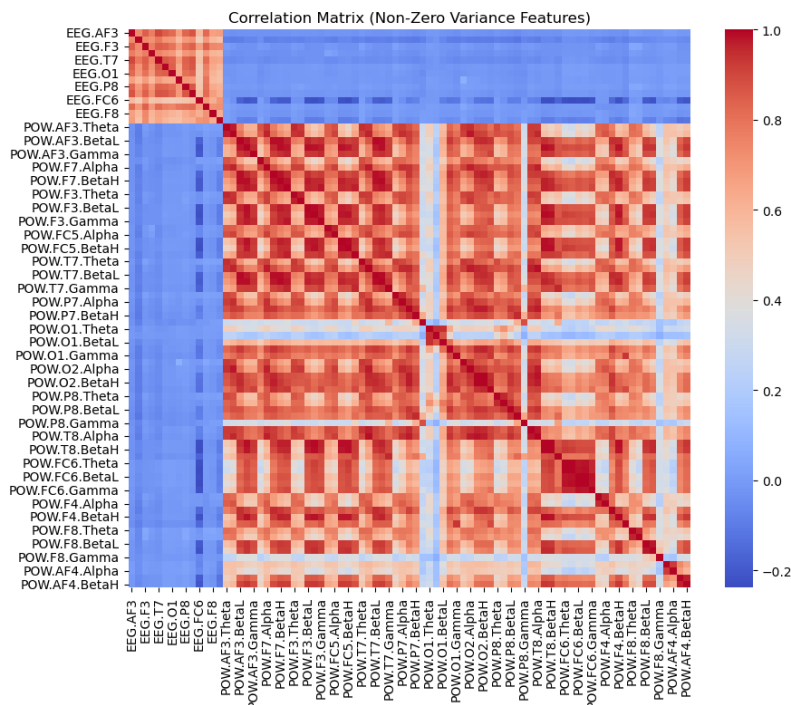
Finally, EDA identified a significant class imbalance in the subject_understood(Figure 4), necessitating resampling or cost-sensitive strategies in model training. Overall, the dataset's richness, variability, and structure support its use in machine learning to classify cognitive states such as understanding vs. non-understanding.
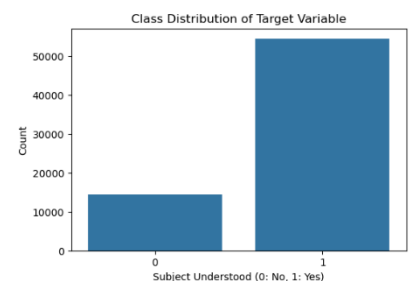

Figure 1, Violin plot of power spectral (POW) features. *The top and bottom 5% of the data capped, explaining the clustering at the high end. As usual, the middle line represents the median, the upper line indicates the third quartile, and the lower line indicates the first quartile. As there was a wide range among the waves they were sorted by their maxima.*

Figure 2, Violin plot of EEG. *with the top and bottom 5% of the data capped, explaining the clustering at the high end. As usual, the middle line represents the median, the upper line indicates the third quartile, and the lower line indicates the first quartile. As there was a wide range among the waves they were sorted by their maxima.*

Figure 3, Correlation matrix of EEG and power spectral (POW) features.
*The matrix highlights strong intra-group correlations within EEG and POW.O1.Theta, and POW.O1.BetaL forming a distinct subgroup.
Coordinated activity among alpha, beta, and gamma bands is also evident.*

Figure 4, class distribution bar.

Implementation & Results

Data Preprocessing

The EEG dataset was loaded from a CSV file, where each row represented a sample, and each column contained a feature or label. EEG features were extracted from columns 2 to 85, and the binary target label—"understand" (1) versus "not understand" (0)—was taken from column 86. A preliminary check confirmed that samples were stored as rows; if not, the DataFrame was transposed accordingly.

The dataset was then split into training and test sets using an 70/30 split. Standard scaling was applied to normalize the feature values, ensuring each feature had zero mean and unit variance—a critical preprocessing step for distance-based algorithms such as K-Means. No explicit feature selection was performed at this stage; all EEG and power features were retained for initial modelling. While this approach ensures that no potentially relevant information is discarded upfront, future work could benefit from feature selection techniques such as PCA or feature importance ranking to reduce dimensionality, mitigate multicollinearity, and improve model interpretability.

Classification techniques

In this study, Logistic Regression and Support Vector Machine (SVM) were selected for their proven effectiveness in binary classification tasks (Leevy et al., 2023). Logistic Regression offers probabilistic outputs, while SVM constructs an optimal hyperplane to separate classes. Both models were configured with class_weight='balanced' to account for class imbalance. Additionally, Synthetic Minority Over-sampling Technique (SMOTE), a method to address class imbalance, was applied to the training set. Which originally comprised 12,241 "not understand" and 34,109 "understand" samples. SMOTE generated 21,868 synthetic samples for the minority class, yielding a balanced set of 34,109 samples per class (Appendix B2).

Logistic Regression achieved an accuracy of 88.22% and a balanced accuracy of 87.54%, with a precision of 0.96 for the majority class and a recall of 0.86 for the minority class, resulting in F1-scores of 0.92 ("understand") and 0.75 ("not understand"). The confusion matrix (Table 2) showed 14,505 true positives and 3,712 true negatives, with 585 false positives and 1,848 false negatives.

SVM slightly outperformed Logistic Regression with an accuracy of 89.63% and balanced accuracy of 88.66%. It maintained a precision of 0.96 for the majority class, improved recall for the minority class (0.87), and F1-scores of 0.93 and 0.78 for classes 1 and 0, respectively. The confusion matrix (Table 3) reported 14,503 true positives, 3,749 true negatives, 548 false positives, and 1,850 false negatives. The SVM used an RBF kernel, with C=1.0 and gamma='scale'

Logistic regression (balanced accuracy: 87.54%)	Precision	Recall	F1-Score	Support
Not Understand	0.67	0.86	0.75	4297
Understand)	0.96	0.89	0.92	16353
SVM (balanced accuracy: 88.66%)				
Not Understand	0.78	0.87	0.82	4297
Understand	0.96	0.89	0.93	16353

Table 1, Performance metrics for Logistic Regression and SVM classifiers.
This table presents the classification performance of both models, including accuracy, balanced accuracy, precision, recall, and F1-scores for each class. SVM slightly outperformed Logistic Regression, particularly in balanced accuracy and F1-score for the minority class.

True Positives (TP): 14,505	**True Negatives (TN)**: 3,712
False Positives (FP): 585	**False Negatives (FN)**: 1,848

Table 2, Confusion matrix for the Logistic Regression model.
This confusion matrix shows the number of true positives, true negatives, false positives, and false negatives predicted by the Logistic Regression model, highlighting its classification strengths and areas of misclassification.

True Positives (TP): 14,503	**True Negatives (TN)**: 3,749
False Positives (FP): 548	**False Negatives (FN)**: 1,850

Table 3, Confusion matrix for the SVM model.
This confusion matrix presents the performance of the SVM model, with 14,503 true positives, 3,749 true negatives, 548 false positives, and 1,850 false negatives, reflecting slightly better overall classification than Logistic Regression.

Sampling Strategy

To manage computational efficiency and balance class distributions, 20,000 samples were drawn via random sampling for clustering and evaluation. A smaller subset of 10,000 samples was later used in optimization experiments involving target labels.

K-Means Clustering

K-Means clustering was chosen to explore potential groupings within the EEG data based on the similarity of feature vectors. While K-Means is known for its efficiency and simplicity, it has limitations, such as sensitivity to initial centroids and the assumption of spherical clusters. These limitations will be considered in the evaluation of the clustering results. The optimal number of clusters was estimated using the Silhouette Score, computed for cluster counts ranging from 2 to 5. The highest score before a significant drop off was 3 (figure 5), suggesting that the data naturally organizes into three distinct groupings.

To further refine clustering toward our application-specific goal—differentiating "understand" vs "not understand" states—we implemented an optimization-based strategy. Specifically, the objective function penalized cluster configurations that failed to produce pure "not understand" clusters (i.e., clusters without any "understand" samples). Using Scipy's `minimize` with SLSQP (appendix, B1), we optimized the number of clusters to maximize the number of such pure "not understand" clusters. The optimization converged at **k = 3**, aligning with the silhouette-based selection.

Principal Component Analysis & Visualization

To examine the clustering structure within the EEG data, Principal Component Analysis (PCA) was used to reduce the high-dimensional feature space to five components, capturing approximately 90% of the total variance (PC1: 60%, PC2: 15%, PC3: 5%, PC4/5: <5%; see Figure 6). Visualized in 2D and 3D scatter plots (Figures 7–8), samples were color-coded by cluster, with "understand" marked by circles and "not understand" by crosses.

K-means clustering was then applied to the reduced space. While no broad separations clearly aligned with cognitive state, a consistent split emerged along the PC2 axis: negative PC2 values were predominantly linked to "not understand", and positive values to "understand". This suggests PC2 captures a cognitively relevant signal, likely contributing to the strong performance of earlier supervised models.

A dense region near PC1 = 0 indicated overlap or ambiguity in the feature space. This aligns with the pattern in performance metrics—"understand" samples were more prevalent but diffuse, whereas "not understand" samples were fewer yet more distinct, aiding in their identification.

The k-means results supported this: Cluster 0 (n = 9700) was 80.91% "understand", while Cluster 2 (n = 296) was 92.23% "not understand". This implies that "not understand" states, though underrepresented, can form compact, well-separated clusters—consistent with extreme PC2 values marking cognitive confusion. In contrast, "understand" samples were more widely dispersed, reflecting a dominant but less sharply defined state.

To quantify these patterns, point-biserial correlations were calculated. PC3 (r = -0.1588), PC2 (r = 0.1107), and PC5 (r = -0.0697) showed the strongest relationships with cognitive state (Table

1), reinforcing PC2's relevance. Notably, although PC2 was positively correlated with "understand", visual inspection revealed a non-linear trend: both high and low extremes in PC2 corresponded to confusion, while values near zero aligned with understanding.

Since correlation captures only linear patterns, mutual information was computed to assess non-linear dependencies. PC3 (0.1455) and PC2 (0.1389) showed the highest scores (Table 2), confirming that PC2 reflects a non-linear cognitive axis where deviations from central values—whether high or low—indicate confusion.

This non-linear structure explains why "not understand" samples cluster tightly and why classifiers perform well despite feature space overlap. PCA, correlation, and mutual information together suggest that PC2 encodes a meaningful, curved trajectory of cognitive state, essential for distinguishing understanding.

Finally, hierarchical clustering using Ward's method on a 5,000-sample subset (Figure 9) produced only limited structure. Leaf-level intermixing of labels indicated poor alignment with cognitive state, likely due to high data dimensionality, which hampers the effectiveness of hierarchical methods.

PC	Correlation	p-value	abs_corr
PC3	-0.158837	1.67e-57	0.158837
PC2	0.110721	1.19e-28	0.110721
PC5	-0.069738	2.92e-12	0.069738
PC1	-0.028814	3.96e-03	0.028814
PC4	0.016469	9.96e-02	0.016469

Table 1, Point-biserial correlations between principal components and target variable.
PC3 shows the strongest correlation (r = -0.1588), followed by PC2 (r = 0.1107) and PC5 (r = -0.0697), suggesting that these components are most relevant for distinguishing understanding from non-understanding.
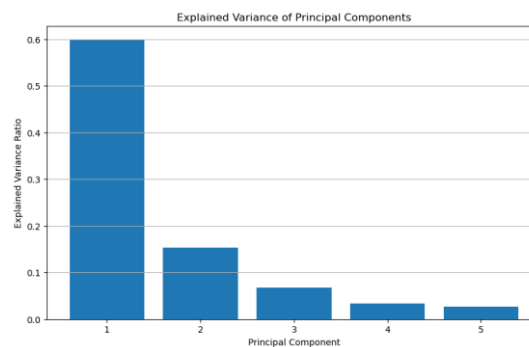
Principal Component	Mutual Information
PC3	0.145546
PC2	0.138908
PC5	0.124838
PC1	0.080381
PC4	0.058283

Table 2, Mutual information between principal components and the cognitive state target.
PC3 and PC2 contain the most information (0.1455 and 0.1389, respectively), supporting the interpretation of PC2 as a non-linear but informative cognitive axis.
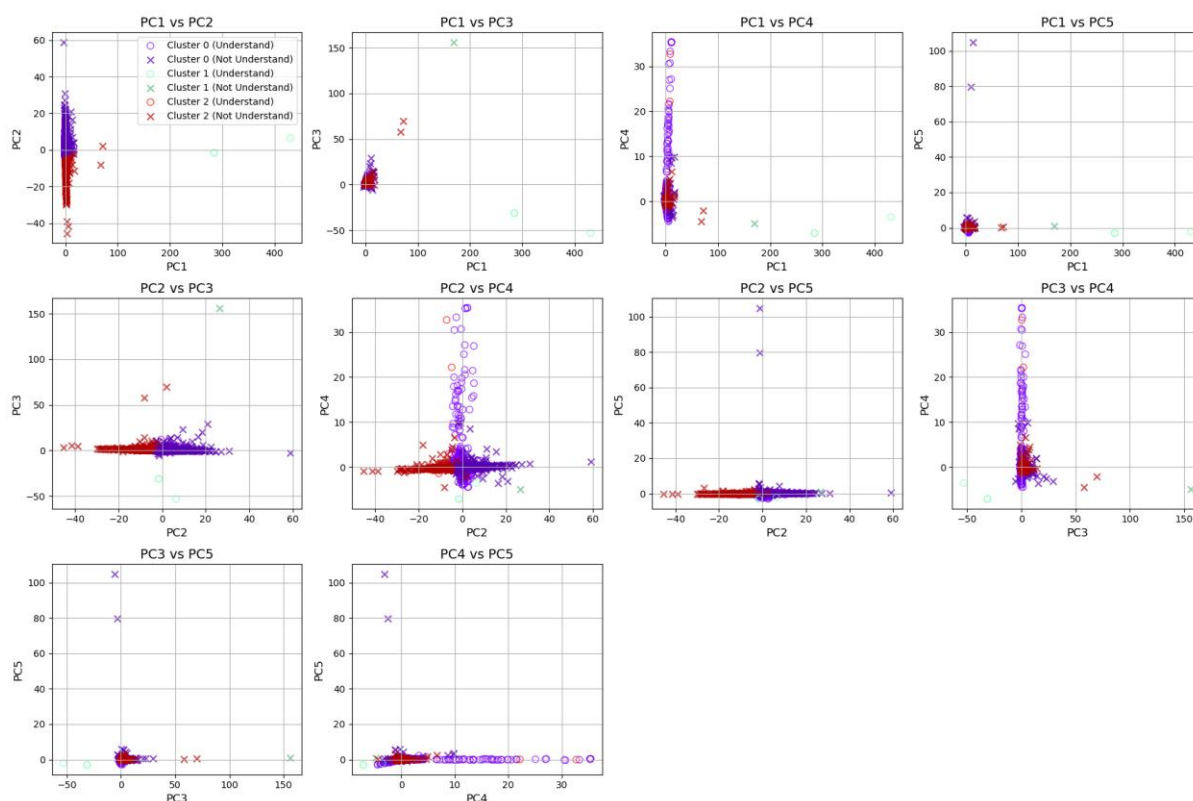
Figure 5, Silhouette Scores for K-Means clustering with 2 to 5 clusters.

This figure shows the average Silhouette Scores for different cluster counts. The peak score at 3 clusters suggests an optimal grouping, beyond which clustering quality noticeably declines
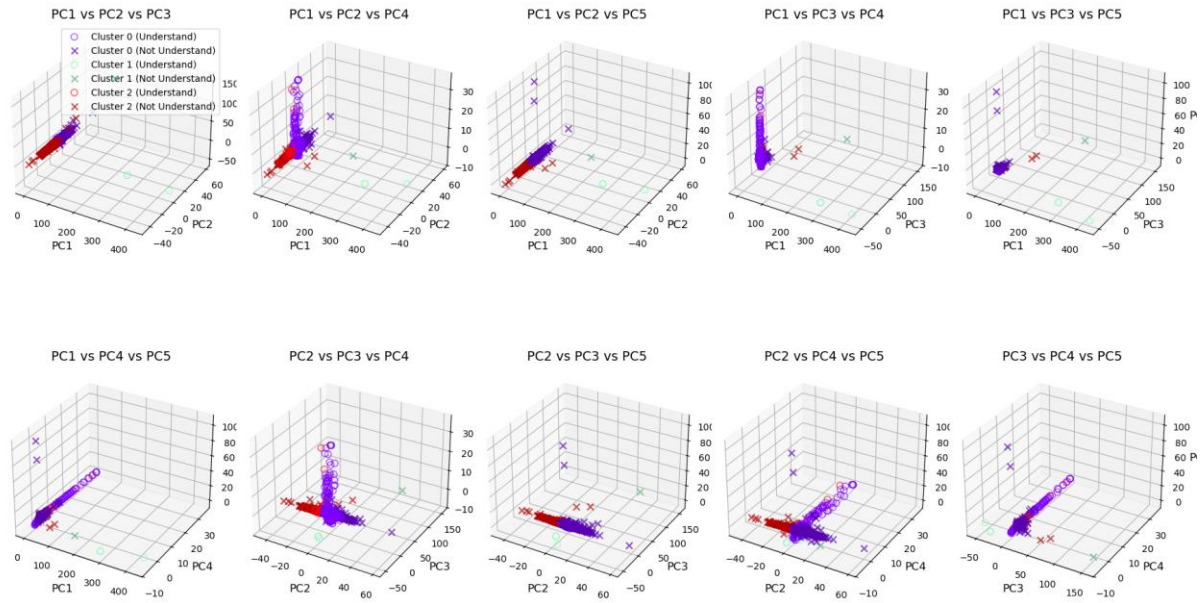

Figure 6 Explained variance by principal components (PCs) in PCA.

PC1 captures 60% of the total variance, followed by PC2 (15%), and PC3 (5%). PC4 and PC5 explain less than 5% each, with the first five components together accounting for ~90% of the total variance.
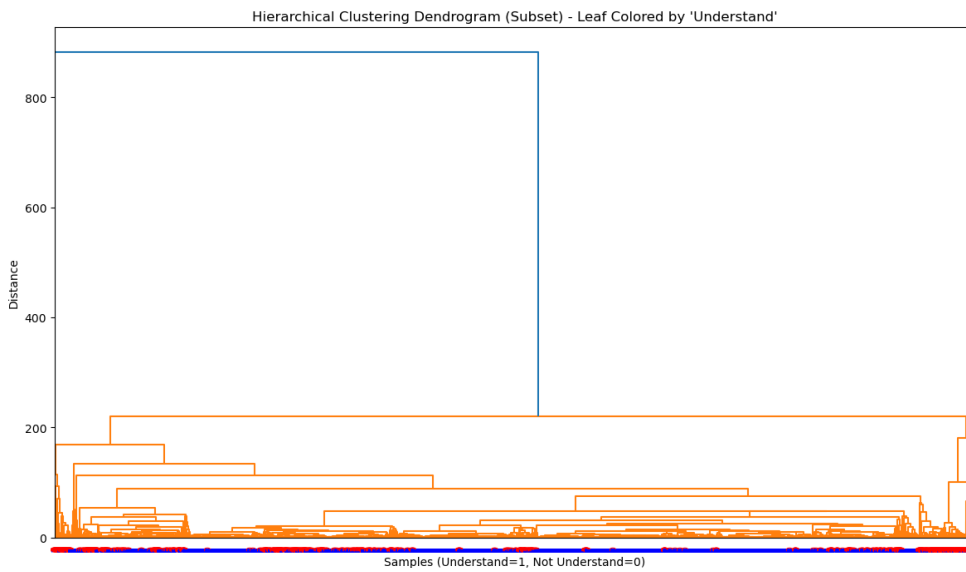

Figure 7, 2D PCA plot of EEG samples with k-means cluster labels.

Samples are projected onto the first two principal components and coloured by cluster. "Understand" states are marked by circles, and "not understand" by crosses.

Figure 8, 3D PCA plot of EEG samples across PC1–PC5.
Samples are color-coded by cluster and shaped by cognitive state. The structure reveals that extreme PC2 values correlate with "not understand" states, suggesting PC2 captures a meaningful cognitive signal.

Figure 9, Hierarchical clustering dendrogram using Ward's method on 5,000 samples.
Despite some local clustering, leaf-level class labels are heavily intermixed, indicating poor alignment with cognitive state and limited interpretability compared to k-means.

Implementation & Results

The classification task aimed to predict whether a subject was in the "understand" or "not understand" state based on EEG features. The dataset was split into training and testing sets using an 70-30% split, and the features were standardized using StandardScaler. To address the class imbalance between the two categories, SMOTE was applied to the training data to synthetically generate samples for the underrepresented "not understand" class. Two models were trained on the resampled data: Logistic Regression and Support Vector Machine (SVM). Evaluation metrics included classification reports, confusion matrices, accuracy, and balanced accuracy. The Logistic Regression model achieved an accuracy of 88.22% and a balanced accuracy of 87.54%, with a precision of 0.67, recall of 0.86, and F1-score of 0.75 for the "not understand" class, and 0.96, 0.89, and 0.92 respectively for the "understand" class (table 1). Its confusion matrix showed 3,712 true negatives, 585 false positives, 14,505 true positives, and 1,848 false negatives (table 2). In comparison, the SVM model performed slightly better, achieving an accuracy of 89.63% and a balanced accuracy of 88.66%. For the "not understand" class, SVM reached a precision of 0.78, recall of 0.87, and F1-score of 0.82, while for the "understand" class, it maintained a high precision of 0.96, recall of 0.89, and F1-score of 0.93 (table 1). Its confusion matrix included 3,749 true negatives, 548 false positives, 14,503 true positives, and 1,850 false negatives (table 2). These results indicate strong predictive performance from both models, with SVM showing a slight advantage in handling the class imbalance and overall classification accuracy.

Evaluation & Comparison of Approaches

The performance of the Logistic Regression and Support Vector Machine (SVM) models was assessed using precision, recall, F1-score, accuracy, balanced accuracy, and confusion matrices. While both models demonstrated strong classification capability in predicting subject understanding, SVM consistently outperformed Logistic Regression across most metrics. SVM achieved a slightly higher overall accuracy (89.63% vs. 88.22%) and better-balanced accuracy (88.66% vs. 87.54%), suggesting it was more effective in handling the imbalanced classes. Logistic Regression displayed higher recall for the minority class ("not understand"), but SVM achieved better precision and F1-score, particularly in correctly identifying the majority class ("understand") while still maintaining respectable performance on the minority class. This marginal improvement in balanced accuracy, however, was not substantial. These results aligned with expectations, as SVM often performs better in high-dimensional spaces like EEG data, while Logistic Regression tends to be more interpretable due to its linear nature, but may struggle with non-linear separability (Arora et al., 2018). Furthermore, Logistic Regression is computationally more efficient than SVM, particularly with higher-dimensional data, which can be a significant consideration in real-time applications or when dealing with limited computational resources. Overall, the choice between the two models depends on the specific priorities of the application. If interpretability and computational efficiency are paramount, Logistic Regression may be preferred. If maximizing predictive accuracy is the primary goal, and computational cost is less of a concern, SVM could be considered.

Two unsupervised clustering methods—K-Means and Hierarchical Clustering—were explored to assess whether EEG feature patterns reflected the cognitive states of understanding.

K-Means Clustering identified three clusters as optimal using the Silhouette Score, and PCA visualizations revealed some distinguishable grouping, although overlapping regions indicated a degree of noise or complexity in the data. K-Means served as a fast and interpretable baseline for clustering EEG data, but its limitations include sensitivity to initialization and assumptions about spherical clusters. Alternative clustering methods may better capture the underlying structure.

Hierarchical Clustering was implemented using Ward's linkage on a 5,000-sample subset of the scaled EEG data. The dendrogram revealed limited separation, and although some grouping was evident, the leaf labels showed considerable intermixing. This suggests the hierarchical structure of the data does not cleanly correspond to cognitive state labels, limiting the method's utility in this context. While useful for visual inspection of small datasets, Hierarchical Clustering did not yield meaningful cognitive groupings in this case and was therefore dismissed from further analysis.

Reflection on the Process

The data analysis process presented several challenges that influenced model performance and interpretation. One of the most significant was the class imbalance in the target variable, where the "understand" state was overrepresented. To address this, SMOTE (Synthetic Minority Oversampling Technique) was applied during training. While SMOTE effectively balanced the class distribution and improved classification outcomes, it also introduced synthetic samples that may not fully capture the underlying data distribution. Future work may benefit from exploring alternative strategies such as cost-sensitive learning or adaptive resampling methods that preserve the integrity of the original data.

Another major hurdle was the high dimensionality of EEG data, which increases the risk of overfitting and complicates both model training and interpretation. Although Principal Component Analysis (PCA) was utilized to reduce dimensionality for clustering and visualization, it was not incorporated into the classification pipeline. Future research could assess the impact of PCA, feature selection, or autoencoder-based dimensionality reduction on both model generalization and explainability.

In the context of unsupervised learning, the use of K-Means clustering revealed important limitations. Its assumption of spherical clusters and sensitivity to initialization hindered its ability to capture the complex structure of cognitive states in EEG. More sophisticated methods, such as DBSCAN, spectral clustering, or nonlinear embeddings like t-SNE or UMAP, could provide more faithful representations of the latent structure in EEG signals.

Despite these challenges, the study demonstrates the potential of machine learning techniques to decode cognitive states from EEG data. To enhance performance and interpretability, future research should consider integrating temporal dynamics using recurrent neural networks or attention-based models and leveraging domain-specific features such as frequency bands, event-related potentials, or source localization. Deep learning models like CNNs or autoencoders may further aid in extracting informative patterns from raw EEG.

Moreover, validating these findings on larger, more diverse datasets and incorporating multimodal information—such as video features or student interaction metrics—could improve prediction accuracy.

References

Arora, A., Lin, J.J., Gasperian, A., Maldjian, J., Stein, J., Kahana, M. and Lega, B., 2018. Comparison of logistic regression, support vector machines, and deep learning classifiers for predicting memory encoding success using human intracranial EEG recordings. *Journal of Neural Engineering*, 15(6), p.066028. https://doi.org/10.1088/1741-2552/aae131

Hayoung Song, Won Mok Shim, Monica D Rosenberg (2023) Large-scale neural dynamics in a shared low-dimensional state space reflect cognitive and attentional dynamics eLife 12:e85487

Keller, T.A., Mason, R.A., Legg, A.E. *et al.* The neural and cognitive basis of expository text comprehension. *npj Sci. Learn.* 9, 21 (2024). https://doi.org/10.1038/s41539-024-00232-y

Leevy, J.L., Hancock, J., Khoshgoftaar, T.M. *et al.* Investigating the effectiveness of one-class and binary classification for fraud detection. *J Big Data* **10**, 157 (2023). https://doi.org/10.1186/s40537-023-00825-1

Omar, M., 2020. *EEG data / Distance learning*. [dataset] Kaggle. Available at: https://www.kaggle.com/datasets/madyanomar/eeg-data-distance-learning [Accessed 15 Apr. 2025].

Ramírez-Moreno, M.A., Díaz-Padilla, M., Valenzuela-Gómez, K.D., Vargas-Martínez, A., Tudón-Martínez, J.C., Morales-Menendez, R., Ramírez-Mendoza, R.A., Pérez-Henríquez, B.L. and Lozoya-Santos, J.J., 2021. EEG-based tool for prediction of university students' cognitive performance in the classroom. *Brain Sciences*, 11(6), p.698. https://doi.org/10.3390/brainsci11060698.