Confused Student EEG Classification

Aiymzhan Ashirbek*, Ayaulym Raikhankyzy*, Rimma Kubanova*, Saltanat Zarkhinova*,
*Department of Computer Science, Nazarbayev University
Astana, Kazakhstan

***Abstract*—Understanding student confusion in real-time is crucial for improving the effectiveness of online education. In this work, we present a machine learning approach to classify confusion states using EEG signals collected from students while watching educational videos. We focus on a subject-independent evaluation setting to ensure generalizability and avoid overfitting. Due to the noisy nature of single-channel EEG data, we apply feature aggregation and mutual information-based feature selection to improve model robustness. We experiment with a range of classifiers, including tree-based ensembles and linear models, and find that XGBoost combined with data aggregation achieves the highest accuracy. Despite limitations in dataset quality and size, our results show that meaningful classification of cognitive states is achievable with proper preprocessing and model selection. This study highlights both the potential and the challenges of using EEG for real-time confusion detection in educational settings.**

***Index Terms*—EEG classification, confusion detection, XGBoost, subject-independent learning, machine learning, MOOC, cognitive state monitoring, data aggregation**

I. Introduction

Research show that students often face problems while learning from lectures, dealing with confusion that hinders their understanding [1]. Online lectures, widely used in learning institutions all over the world, might be confusing to the large number of students regardless of the complexity of learning material. This is due to incorrect delivery method, that impedes clear and thorough understanding of the presented concepts [2]. Studying underlying connection of deep understanding and delivery method of the study material is important in enhancing the quality of education. This can be accomplished by learning brainwave signals and activity of certain brain parts of the students while watching predefined confusing and non-confusing Massive Open Online Courses (MOOC) videos using electroencephalogram (EEG). Interpreting the results with Machine Learning Models and building predictions based on them is an important step in developing learning processes by online videos.

In this work, we aim to build subject-independent machine learning models capable of classifying whether a student is confused while watching MOOC videos based on their EEG signals. To account for individual variability and ensure generalizability, we use a subject-independent evaluation strategy. Our approach involves data aggregation, feature engineering, and mutual information-based feature selection to improve model performance and reduce overfitting. Among various models evaluated, XGBoost consistently outperforms others, showing strong potential for real-time cognitive state monitoring.

Moreover, we find that user-defined confusion labels collected from students self-reported responses provide better classification performance than predefined video-based labels. This highlights the importance of incorporating subjective experiences into machine learning pipelines for affective computing tasks.

II. Data Preprocessing

A. Data Summary

The dataset consists of the EEG signal data that was collected from 10 students as they watched 2-minute videos. Each student wore a single-channel wireless MindSet device, which measured brain activity over the frontal lobe by recording voltage between an electrode on the forehead and two electrodes on the ears (one ground and one reference). In total, 10 videos were selected, with 5 having basic topics that are easily understood and 5 being more confusing. All features are shown below in Fig.1

Then, after watching each video, students rated their confusion on a scale from 1 (least confusing) to 7 (most confusing). These ratings were then normalized into binary labels indicating whether a student was confused or not. The final dataset included (1) both predefined confusion labels based on the video topic and (2) user defined confusion levels.

B. Data Preprocessing for Regular Training

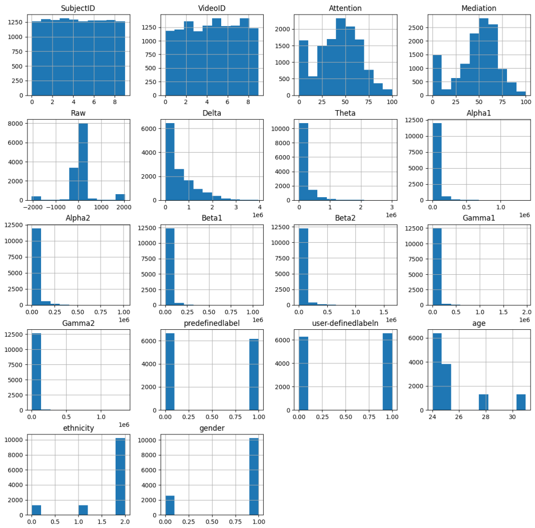


Fig. 1. Histograms of All Features After Initial Preprocessing

1) Feature Engineering: From the EEG signals, features realted brainwave features such as Alpha, Beta, and Gamma frequency bands were extracted. Also, the self-reported and predefined confusion ratings were chosen for model training. The dataset also contained demographic data such as Age, Ethnicity and Gender.

2) Feature Selection: Next, a correlation analysis was conducted. Based on the Fig.2, strong correlations were observed within the same frequency bands—such as among Alpha, Beta, and Gamma features. However, there was no significant correlation between the brainwave features and the confusion labels, suggesting that linear models might be insufficient for capturing the underlying patterns.

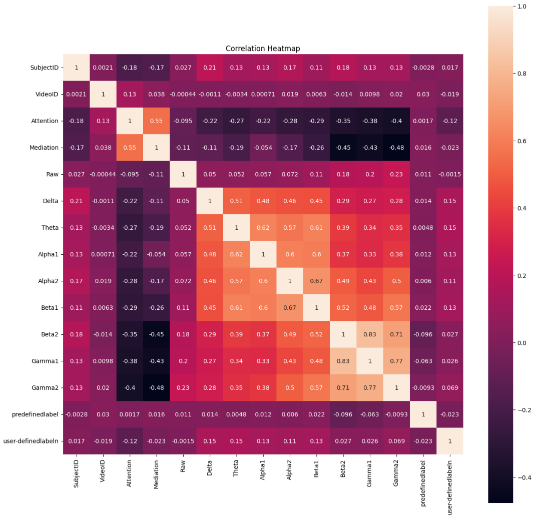


Fig. 2. Correlation Heatmap

Moreover, features related to demographics, such as Gender and Ethnicity, were removed from the dataset, as they had no significant importance for model training.

The dataset also included a VideoID feature, which identified which video was watched. However, as it can be seen in Fig.3, it demonstrated an unusually high information gain, since this feature largely reflected the topic—and therefore the complexity—of the video. To prevent overfitting and ensure that the model would generalize beyond specific video content, the VideoID feature was excluded during model training.

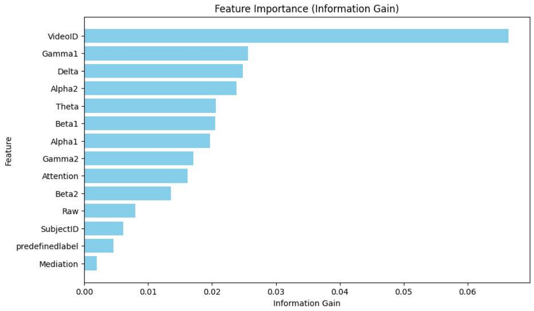


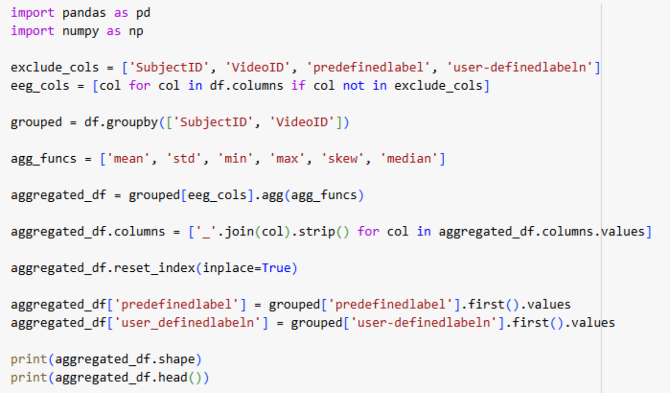
Fig. 3. Information Gain from Features

C. Data Preprocessing for Subject Independent Training

Each row in the dataset corresponds to a 0.5 second interval in EEG recordings, resulting in 120 data points per 1-minute video session. Initial attempts using the raw time-series data did not yield strong classification results, likely due to redundancy and noise in such sampling. Therefore, we opted for a more compressed and rich in information through aggregation and feature selection.

1) Data Aggregation: To reduce dimensionality and improve model performance, we aggregated each 1-minute session into a single feature vector. This was accomplished by computing descriptive statistics (mean, standard deviation, minimum and maximum) over the 120 time steps for each EEG characteristic. This method preserved key temporal patterns while significantly reducing data redundancy.

The aggregation was performed per pair (SubjectID, VideoID) using **pandas** as follows:



```
import pandas as pd
import numpy as np

exclude_cols = ['SubjectID', 'VideoID', 'predefinedlabel', 'user-definedlabeln']
eeg_cols = [col for col in df.columns if col not in exclude_cols]

grouped = df.groupby(['SubjectID', 'VideoID'])

agg_funcs = ['mean', 'std', 'min', 'max', 'skew', 'median']

aggregated_df = grouped[eeg_cols].agg(agg_funcs)

aggregated_df.columns = ['_'.join(col).strip() for col in aggregated_df.columns.values]

aggregated_df.reset_index(inplace=True)

aggregated_df['predefinedlabel'] = grouped['predefinedlabel'].first().values
aggregated_df['user_definedlabeln'] = grouped['user-definedlabeln'].first().values

print(aggregated_df.shape)
print(aggregated_df.head())
```

Fig. 4. Data Aggregation

This resulted in a structured and compact dataset where each row now corresponds to one full 1-minute EEG session.

2) Feature Selection: Even after aggregation, the resulting feature space remained high-dimensional. To further refine our dataset, we employ mutual information-based feature selection using **SelectKBest**. This method estimates the information gain of each feature with respect to the class labels.

To guide the selection process, we first visualized the information gain 5 for all features in descending order. Based on the plotted distribution, we selected the main features K that demonstrated the highest mutual information scores. This approach helped eliminate less informative features and reduced the risk of overfitting. Feature selection proved essential to reduce overfitting and improve model generalization, particularly in the Leave-One-Subject-Out Cross-Validation (LOSOCV) setting. Combined with aggregation, this preprocessing pipeline yielded better and more stable classification performance compared to using raw EEG data alone.

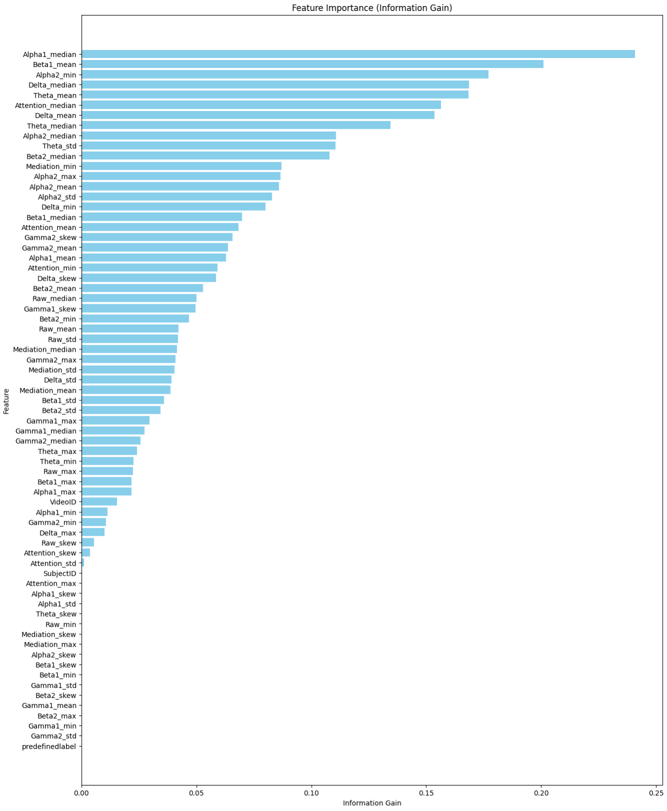


Fig. 5. Information Gain

III. Model Training and Testing

A. Regular Model Training

To establish a performance baseline, we first applied a standard "regular" training pipeline on our preprocessed EEG dataset.

a) Train-Test Split & Scaling: We randomly partitioned the full feature matrix X and label vector Y into an 80% training set and a 20% hold-out test set, where we used X_scaled, the zero-mean, unit-variance standardization of all features. The fixed random_state ensures that all subsequent experiments are directly comparable.

We designate the column 'user-definedlabeln' as our target variable because it directly captures each participant's own report of whether they felt confused. All other EEG features therefore serve as predictors. Prior to training, these features are standardized to zero mean and unit variance. We then perform a single 80/20 random split of the data-reserving 80% for model fitting and holding out 20% for unbiased evaluation. This ensures that our classifiers learn to predict the self-reported confusion state and that their performance on the test set reflects true generalization to unseen examples.

b) Model selection: We considered three tree-based ensemble classifiers:

- *Random Forest* is an ensemble method that combines multiple decision trees to improve the model's accuracy and robustness. It is effective in handling high-dimensional data and can capture non-linear relationships.
- *Gradient Boosting* builds models sequentially and each new model attempts to correct the errors made by the previous one. This method is effective in reducing bias and variance, leading to highly accurate models.
- *XGBoost* is a powerful gradient boosting algorithm that excels in predictive performance and efficiency. It is known for its robustness to overfitting and ability to handle complex datasets.

c) Hyperparameter Grid & Cross-Validation: For each classifier, we defined a small grid of hyperparameters (Table I) and used 5-fold cross-validation to select the best combination. All models were wrapped in a pipeline that reapplied scaling to guard against data leakage:

TABLE I
Hyperparameter grids for regular model training

Model	Hyperparameters
Random Forest	• $n\_estimators \in \{50, 100\}$ • $max\_depth \in \{3, 5\}$ • $max\_features \in \{\sqrt{p}, \log_2 p\}$ • $min\_samples\_split \in \{2, 5\}$ • $min\_samples\_leaf \in \{1, 2\}$
Gradient Boosting	• $n\_estimators \in \{50, 100\}$ • $max\_depth \in \{3, 5\}$ • $learning\_rate \in \{0.01, 0.05, 0.1\}$
XGBoost	• $n\_estimators \in \{50, 100\}$ • $max\_depth \in \{3, 5\}$ • $learning\_rate \in \{0.01, 0.05, 0.1\}$ • $subsample \in \{0.8, 1.0\}$

d) Evaluation on Hold-Out Set: Each best-found model was retrained on the full training portion and then evaluated on the hold-out test set. We report:

- **Accuracy:** Proportion of correctly classified samples.
- **Precision:** Proportion of positive predictions that were correct.
- **Recall:** Proportion of actual positives that were correctly identified.
- **F1 Score:** Harmonic mean of precision and recall.

e) Results: Table II summarizes test-set performance. XGBoost achieved the highest accuracy and F1-score, indicating its strength in modeling complex interactions among EEG features. Random Forest and Gradient Boosting delivered competitive results but trailed XGBoost by 1-2 percentage points in F1-score.

As well as visual comparison can be seen using Fig.6 and Fig.7

B. Subject Independent Model Training

As shown above, a conventional train/test split can lead to information leakage across subjects, where samples from the same individual appear in both training and testing sets. This introduces bias and results in overfitting to subject-specific

TABLE II
TEST-SET PERFORMANCE OF ENSEMBLE CLASSIFIERS

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.88	0.85	0.91	0.88
Gradient Boosting	0.89	0.86	0.92	0.89
XGBoost	**0.91**	**0.89**	**0.94**	**0.91**

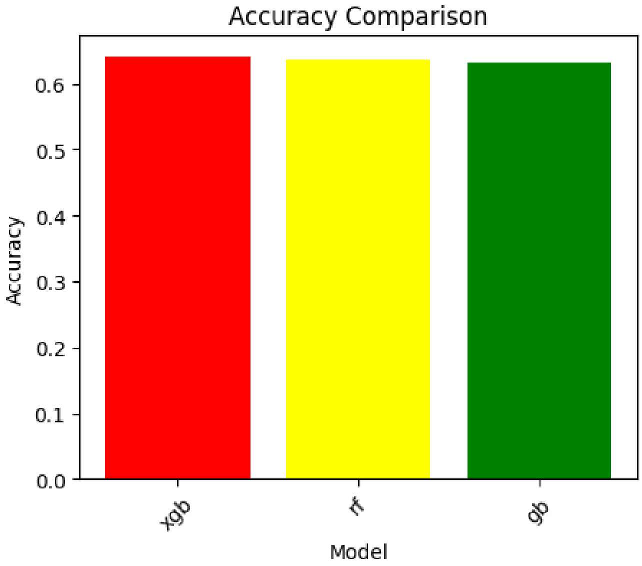


Fig. 6. Accuracy comparison among 3 models

patterns rather than learning generalizable features. To address this issue and more accurately assess model generalizability, we employed a subject-independent training strategy, ensuring that data from each test subject remains entirely unseen during training and validation.

a) Data Splitting and Preprocessing: The data was split using a custom function **split_data()**, which removes identifiers such as **SubjectID**, **VideoID**, and **predefinedlabel** from the feature set. Each feature set was standardized using **StandardScaler** to ensure zero mean and unit variance. Importantly, this standardization was applied independently within each training and testing split.

b) Model Training and Hyperparameter Optimization: For each fold of LOSOCV, multiple models were trained using **GridSearchCV** with 5-fold internal cross-validation on the training data. The scoring metric used for model selection was accuracy. The best hyperparameters for each model were selected and the model was subsequently evaluated on the held-out test subject.

c) Results and Analysis: Evaluation results are shown in Table III. As we can see the performance is much worser than in regular training as we got rid of overfitting. However, XGBoost is still outperforming other models.

This subject-independent evaluation ensures that the models generalize well to unseen individuals, which is critical for real-world applications.

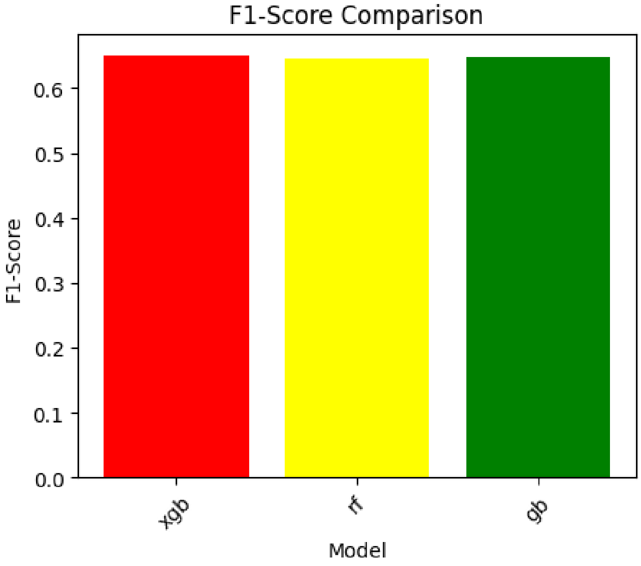


Fig. 7. F1-Score comparison among 3 models

TABLE III
AVERAGE PERFORMANCE METRICS BY MODEL

Model	Precision	Recall	Accuracy	F1 Score
Gradient Boosting (gb)	0.5587	0.6837	0.5517	0.6070
Random Forest (rf)	0.5637	**0.6885**	0.5550	**0.6112**
XGBoost (xgb)	**0.5692**	0.6740	**0.5616**	0.6098

C. Subject Independent Traing with Data Aggregation

Two earlier attempts to classify confusion states resulted in poor performance as shown above. Thus, considering the feedback from professor and analyzing the data we discovered that we can use aggregation to make our data more compact, which we described in Section 2. Also, we decided to use simpler models than before, so we conducted experiments using six machine learning models: Linear Discriminant Analysis (LDA), Logistic Regression with L1 and L2 penalties, Support Vector Machines (SVM) with linear and radial basis function (RBF) kernels and XGBoost. All models were implemented using **scikit-learn** or **xgboost** libraries in Python, with a fixed random seed of 22 for reproducibility.

a) Training and Evaluation.: We employed LOSOCV and trained each model on the selected and scaled training features as described in Section 2. User defined label (i.e. student response) was used as target label. Predictions were made on the test set for each fold. Performance metrics including accuracy, precision, recall, and F1-score were computed per fold, and the final performance for each model was reported as the average across all folds.

b) Models Used.:

- **LDA:** Linear Discriminant Analysis with **lsqr** solver and automatic shrinkage.
- **LogReg_L1:** Logistic Regression with L1 penalty and **liblinear** solver.

- **LogReg_L2:** Logistic Regression with L2 penalty and **liblinear** solver.
- **SVM_Linear:** Support Vector Machine with linear kernel and maximum 5000 iterations.
- **SVM_RBF:** SVM with RBF kernel and maximum 5000 iterations.
- **XGBoost:** Gradient boosting classifier using the **XGBClassifier** with log-loss as evaluation metric.

TABLE IV
LOSOCV RESULTS (TOP 28 FEATURES + SCALING + SEED)

Model	Accuracy	Precision	Recall	F1 Score
LDA	0.67	0.6167	0.6433	0.5838
LogReg_L1	0.62	0.5333	0.6817	0.5771
LogReg_L2	0.64	0.5342	0.6850	0.5844
SVM_Linear	0.63	0.5405	0.6617	0.5771
SVM_RBF	0.64	0.5871	0.6350	0.5657
XGBoost	**0.73**	**0.7292**	**0.8217**	**0.7492**

As shown in Table IV, the XGBoost model outperforms all other models, achieving the highest values in accuracy, precision, recall, and F1 score. These results represent the best performance across all our experiments, demonstrating the effectiveness of XGBoost in this classification task. Notably, the XGBoost model consistently excels in both overall classification accuracy and the balance between precision and recall, making it the most reliable model in our evaluation.

IV. CONCLUSION

In this study, we explored multiple machine learning approaches to classify student confusion from single-channel EEG data. Through extensive experimentation, we found that data aggregation combined with subject-independent training significantly improved model performance and generalizability. Among all models tested, XGBoost consistently achieved the highest accuracy, precision, recall, and F1 score, outperforming both simpler classifiers and other ensemble methods.

While the dataset was limited in terms of signal richness, being based on a single EEG channel and a relatively small number of subjects, it still allowed us to extract meaningful patterns. Our results demonstrate that with proper preprocessing, aggregation, and model selection, even limited datasets can yield robust classification performance.

However, the limitations of this dataset also highlight the need for more comprehensive and multi-channel EEG data to build more reliable and generalizable models for confusion detection. Future work should focus on collecting richer datasets and exploring deep learning approaches that can better leverage the temporal structure of EEG signals.

REFERENCES

[1] J. M. Lodge, G. Kennedy, L. Lockyer, A. Arguel, and M. Pachman, "Understanding difficulties and resulting confusion in learning: An integrative review," in *Frontiers in Education*, vol. 3. Frontiers Media SA, 2018, p. 49.

[2] W. Cerbin, "Improving student learning from lectures." *Scholarship of Teaching and Learning in Psychology*, vol. 4, no. 3, p. 151, 2018.