Sandia Data Challenge

MULTILINGUAL BRAIN RESPONSE CLASSIFICATION USING MACHINE LEARNING AND DEEP LEARNING

Ching-Yu Lin
Chen-Chin Lin
Zhi Li
Hou-Chun Liu
3 November, 2024

AGENDA

- Exploratory Analysis
- Model
- Result
- Conclusion

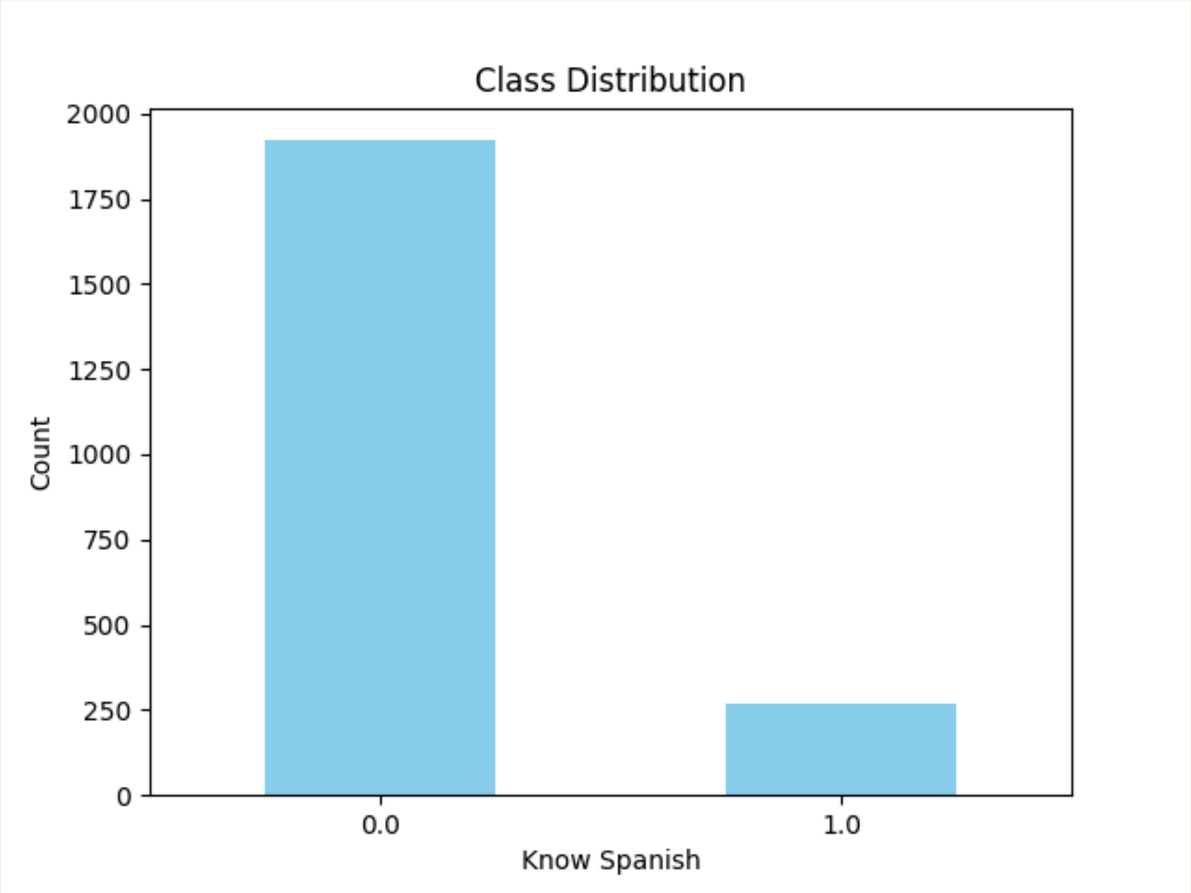
INTRODUCTION

Objective: Detect Spanish language proficiency through EEG data analysis

- Only consider Spanish related data.
- Using Spanish-specific neural responses, we employ machine learning and deep learning algorithms to analyze signals from multiple EEG electrodes.

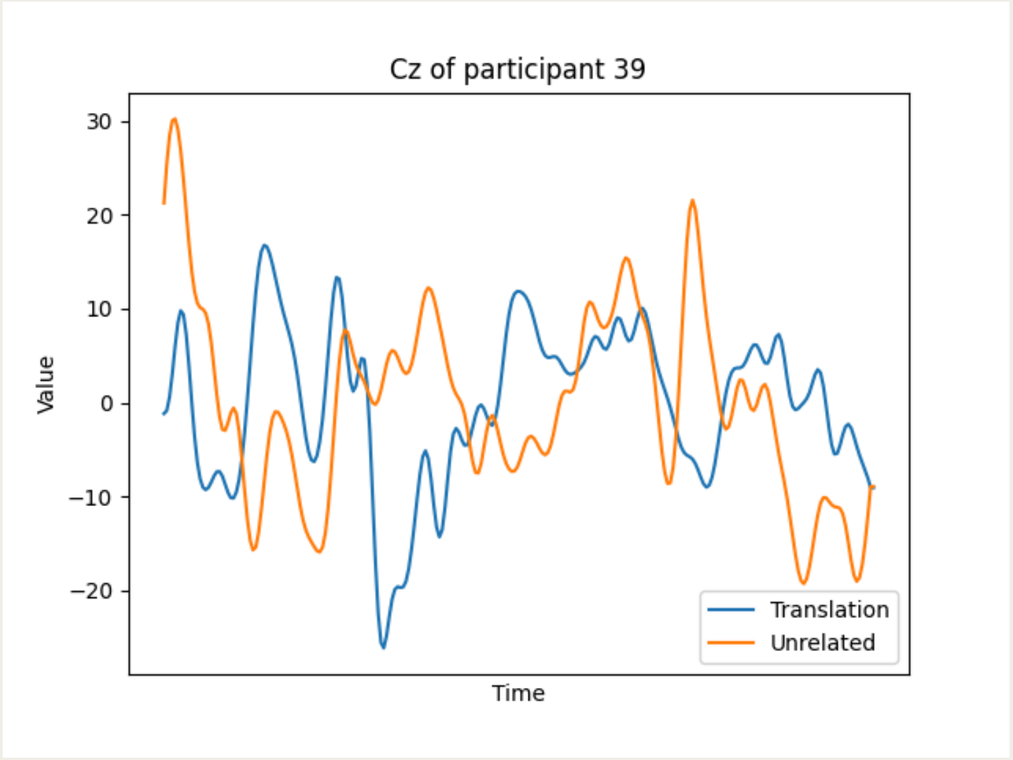
DATA

- The distribution of the data



EXPLORATORY ANALYSIS

- Calculate the difference between the translation and unrelated reaction
- XGBoost Classifier



EXPLORATORY ANALYSIS

- FC6 and CP6 are the two best electrodes to use as single input

Electrodes	Accuracy	Roc_auc	pr_auc
FC6	0.772	0.610	0.399
CP6	0.743	0.587	0.317
CP5	0.738	0.553	0.282
F8	0.733	0.489	0.278

EXPLORATORY ANALYSIS

- Baseline Models using only CP6
- Best parameters found:
  - colsample_bytree: 0.8
  - learning_rate': 0.01
  - max_depth: 3
  - n_estimators: 100
  - subsample: 1.0
- Best ROC AUC score: 0.64
- Cross-validated Accuracy: 0.79
- Cross-validated AUC: 0.64

MODEL SELECTION

	Random Forest	Gradient Boosting	XGBoost	LightGBM	SVM	Logistic Regression
Accuracy	.73	.77	.77	.76	.74	.76
Recall	.06	.32	.32	.31	.20	.35
Precision	.54	.64	.66	.62	.60	.60
F1-Score	.11	.43	.43	.42	.30	.44

MODEL SELECTION

	Neural Network	Naive Bayes	Decision Tree	AdaBoost	LDA	QDA	ExtraRF
Accuracy	.69	.63	.65	.77	.77	.67	.73
Recall	.37	.54	.39	.28	.44	.16	.03
Precision	.42	.37	.37	.66	.59	.29	.63
F1-Score	.39	.44	.38	.40	.51	.20	.06

MODEL SELECTION

- Accuracy: Gradient Boosting, XGBoost, LDA, and AdaBoost (0.77)
- Recall: Naive Bayes (0.54)
- Precision: XGBoost (0.66), AdaBoost (0.66)
- F1-Score: LDA(0.51)

- Overall: LDA

LONG SHORT-TERM MEMORY (LSTM)

- Why LSTM?
  - EEG data is sequential
  - LSTM can maintain memory of earlier responses while processing later ones
  - LSTM can capture both immediate and delayed neural responses
  - Support by previous works: “An accuracy of 98% and 93.67% were obtained with the complete feature set and the reduced feature set respectively.” *(Supakar, 2022)

*Supakar R. (2022). A deep learning based model using RNN-LSTM for the Detection of Schizophrenia from EEG data.

BIDIRECTIONAL LONG SHORT-TERM MEMORY

- Why Bi-LSTM?
  - Processes sequences in both **forward** and **backward** directions
    - Forward LSTM: learns from past information
    - Backward LSTM: learns from future information
  - Brain signals often have bidirectional temporal dependencies

DATA PREPROCESSING

- **Goal:** Predicting Spanish Bilingualism

- **Model:**
    a. LSTM
    b. Bi-LSTM

- **Approach:**
    a. Apply only EEG Data to the model (30 features)
    b. Apply EEG Data and Word-related Data to the model (30 + 4 features)

RESULT - PREDICTING SPANISH BILINGUALISM

	LSTM	Bi-LSTM
Accuracy (Apply only EEG Data)	**0.8161**	**0.8619**
Accuracy / Precision / Recall (Apply EEG + Word-related)	**0.9935 / 0.9887/ 0.9876**	**0.9267 / 0.7886/ 0.9736**

Initial Data Shape: 1,530,880 rows (Spanish, all participants)
Train/Test split shapes:

- X_train shape: 960 sequences, 1536 time points long/per sequence
- X_test shape: 240 sequences, 1536 time points long/per sequence

CONCLUSION

- FC6 and CP6 emerge as the most informative electrode
- Best traditional ML models (Gradient Boosting, XGBoost) achieve 0.77 accuracy
- Neural networks outperform traditional ML models
  - LSTM with EEG data: 0.816 accuracy
  - LSTM with EEG data + word features: 0.994 accuracy
- Combining EEG and word-related features significantly enhances prediction of Spanish bilingualism

THANKS FOR LISTENING

APPENDICES

EXPLORATORY ANALYSIS

- Baseline Models using only CP6
- Best parameters found:
  - colsample_bytree: 0.8
  - learning_rate': 0.01
  - max_depth: 3
  - n_estimators: 100
  - subsample: 1.0
- Best ROC AUC score: 0.64
- Cross-validated Accuracy: 0.79
- Cross-validated AUC: 0.64