A Representative Subset of EEG Datasets for Generalizable Deep Learning Models

Nhi Nguyen

Hochschule Darmstadt - Fachbereiche Mathematik und Naturwissenschaften & Informatik - Studiengang DataScience

Motivation

Electroencephalography (EEG) is a widely used neuroimaging technique, thanks to its non-invasiveness, portability, and high temporal resolution. However, EEG analysis remains challenging due to its complexity, high dimensionality, and low signal-to-noise ratio. Deep Learning (DL) models address these challenges by automatically extracting features and achieving high classification accuracy.

Most DL models, however, are task-specific and struggle to generalize across diverse EEG datasets. Generalizable models could identify patterns across tasks and domains, enhancing the scalability of EEG analysis for broader applications. Given the diversity of EEG datasets, achieving generalization is resource-intensive. A promising solution is identifying a small, representative subset of datasets that balances diversity and redundancy. Such a subset would enable systematic testing and training of DL models, reducing computational costs while improving understanding of model performance across EEG data.

Within the scope of this thesis, generalizability refers to a model's ability to perform well across multiple datasets and domains without architectural changes, learning dataset-specific patterns while avoiding overfitting. It reflects the model's capacity to extract relevant features and classify new data accurately.

Research Goal

This thesis investigates the feasibility of constructing a small yet representative subset of EEG datasets that encapsulates the diversity of a broader dataset collection. Such a subset would support the development and evaluation of generalizable DL models that perform well across various EEG classification tasks without requiring domain-specific adjustments, while also significantly reducing computational costs. An initial exploration of a principled approach to selecting a representative subset of EEG datasets across multiple domains was undertaken, inspired by the "Atari-5" approach in reinforcement learning [1]. Additionally, the thesis explores the generalizability and compatibility of state-of-the-art end-to-end DL models across multiple EEG datasets spanning different domains.

The main research question of this study is:
"Which collection of EEG datasets can best represent the entire spectrum of EEG activity for evaluating and developing generalizable Deep Learning models?"

Methodology

The framework consists of three key components: (i) a collection of EEG data, (ii) DL methods for classification, and (iii) an approach for subset selection. The datasets and DL models were selected based on prominent literature review articles in the field, citation frequency, performance on tasks, and reproducibility.

- **EEG Dataset Selection:** The selected EEG datasets are publicly available and have been widely used as benchmarks for state-of-the-art classification algorithms, including STEW, EEGMAT, SEED, SEED IV, DEAP, DREAMER, BCIC-IV-2a, High-Gamma, PhysioNetMI, BCIC-III-2, CHB-MIT, Siena Scalp Dataset, The TUH-Abnormal EEG Corpus, Sleep-EDF. The collection of datasets represented a broad spectrum of common EEG-based classification paradigms.
- **Deep Learning Model Selection:** We selected only end-to-end DL models capable of decoding raw EEG signals without any preprocessing and exploiting hierarchical structure on the data. To evaluate their generalizability, we tested the models across data from various categories, assessing their performance beyond their intended domains. The 11 selected DL models consist of EEGNet, DeepConvNet, ShallowConvNet, CNN-FC, CNN-LSTM, MMCNN, ChronoNet, EEGTCNet, BLSTM-LSTM, Attention-1DCNN, DeepSleepNet
- **Performance Evaluation:** Each of the DL models will be trained and tested on all EEG datasets to evaluate their ability to generalize across different types of data. The performance of the models was assessed using $F1$-score, since the datasets included both balanced and unbalanced class distributions.
- **Representative Subset Selection Procedure:** The procedure evaluated dataset subsets based on their ability to predict a target summary metric, the median $F1$-score, using regression modeling with Ridge regularization. To address redundancy, a correlation analysis clustered and pruned highly correlated datasets, ensuring diversity. A brute force search tested all possible subsets, selecting the one minimizing prediction error.
- **Experiment Design:** We used a 2×2 factorial design to evaluate the subset selection procedure, examining two factors:
  1. Target Metric Computation
     - *Inclusion*: The median $F1$-score includes the candidate subset.
     - *Exclusion*: The median $F1$-score excludes the candidate subset.
  2. Dataset Preprocessing
     - *Full Dataset Corpus*: Direct application to the full dataset pool.
     - *Pruned Dataset Corpus*: Clustering and pruning to remove redundant datasets.

  This resulted in four experimental conditions, combining preprocessing and metric computation approaches.

Results

Key insights into the generalizability and compatibility of DL models with EEG datasets include:

- **Top Performers**: EEGNet and ShallowConvNet achieved the highest mean $F1$-score ($\overline{M} = 0.69$), while BLSTM-LSTM and DeepSleepNet showed less stable results.
- **Generalizability**: No model significantly outperformed others in variability, with standard deviations ranging from 0.14 to 0.18.
- **Compatibility**: While each model was originally tailored for specific EEG data categories, none of the models achieved their best performance on the datasets they were designed for. EEGNet and ShallowConvNet consistently outperformed other models across multiple EEG datasets and domains. In contrast, models like DeepSleepNet struggled to generalize effectively.

By combining Mean Squared Error (MSE) results with correlation analysis, we gained insights into the trade-offs between diversity and prediction accuracy. Figure 1 visualizes the correlation matrices for the selected subsets under the four experimental conditions.

- **Performance of Inclusion and Exclusion Approaches:** The *exclusion* approach consistently achieved lower MSE values compared to the *inclusion* approach, while the *inclusion* approach effectively preserved dataset diversity. The *exclusion* approach exhibited inconsistent behavior, producing fewer correlated datasets without pruning but higher correlations after pruning, even surpassing the *inclusion* approach.
- **Impact of Dataset Preprocessing:** Pruning enhanced diversity in the *inclusion* approach but disrupted the *exclusion* approach by making the dataset pool more sensitive to exclusions.
- **Trade-Offs Between Diversity and Prediction Accuracy:** The *Full Dataset Corpus + Exclusion* condition achieved the lowest MSE (0.0001) but showed redundancy due to high correlations among selected datasets. In contrast, the *Pruned Dataset Corpus + Inclusion* condition achieved higher diversity at the cost of slightly higher MSE (0.001), emphasizing a trade-off between diversity and predictive accuracy.

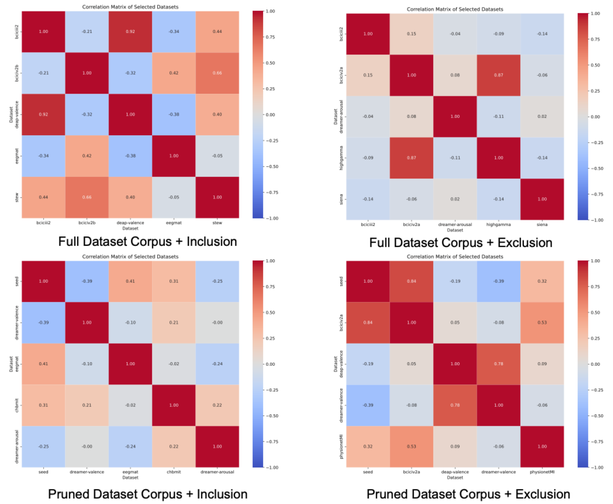

Figure 1. Correlation Matrix of Datasets in the Selected Subsets

Conclusion

The findings indicate that none of the selected DL models demonstrated outstanding generalizability across all EEG datasets, reflecting limitations in their ability to adapt to diverse domains. By minimizing redundancy while preserving diversity of EEG signals, this work demonstrates the feasibility of constructing a subset that captures the variability across multiple EEG classification tasks, paving the way for more efficient and scalable solutions. The representative subset, comprising SEED, DREAMER-Valence, DREAMER-Arousal, EEGMAT, and CHBMIT datasets, demonstrated a low prediction error and weak pairwise correlations, highlighting its diversity. Despite limitations, such as the absence of standardized pipelines for loading, preprocessing, and training DL models across EEG datasets, as well as the lack of suitable validation methods for the selected subset, this work provides a foundational framework for developing generalizable EEG classification models. Future research should focus on exploring alternative subset selection strategies and expanding the study with more datasets and models to further improve diversity and predictive performance.

References

[1] Matthew Aitchison, Penny Sweetser, and Marcus Hutter. "Atari-5: Distilling the arcade learning environment down to five games". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 421–438.

[2] Yannick Roy et al. "Deep learning-based electroencephalography analysis: a systematic review". In: *Journal of neural engineering* 16.5 (2019), p. 051001.