A Lightweight Teacher-Student Deep Learning Pipeline for ADHD Classification Using EEG-Derived Features

Grace M Kalonji
Independent Research
graceklnj@gmail.com

January 21, 2026

Abstract

Attention Deficit Hyperactivity Disorder (ADHD) is one of the most prevalent neurodevelopmental disorders, yet its diagnosis remains mainly clinical and subjective. This study presents a lightweight, two-staged teacher-student deep learning pipeline for automated ADHD classification using electroencephalogram (EEG) derived feature vectors. I implement and evaluate three distinct teacher architectures—a 1D ResNet, a Time Series Transformer, and an EEGNet—each trained to identify discriminative EEG features through gradient-based saliency mapping. These saliency maps are used to filter low-importance features, and the resulting masked data is evaluated using EEGNet as the student architecture. My experiment on a publicly available preprocessed EEG dataset show that the Transformer-Teacher with EEGNet-Student pipeline achieves the best performance with an accuracy of 80.46%, F1 score of 80.57%, and inference time of 0.33 seconds which is better than the benchmark that trained EEgnet on unfiltered preprocessed data achieving an accuracy of 78%. These results suggest that feature masking guided by deep learning saliency can improve classification efficiency while maintaining competitive accuracy, offering a promising direction for objective, data-driven ADHD screening tools.
Keywords: ADHD, EEG, Deep Learning, Teacher-Student Learning, Feature Selection, Saliency Maps, EEGNet, Transformer

1 Introduction

Over the past two decades, the rates of Attention Deficit Hyperactivity Disorder (ADHD) diagnosis have risen significantly across all age groups [1]. ADHD is a neurodevelopmental disorder characterized by persistent patterns of inattention, hyperactivity, and impulsivity that interfere with daily functioning and development. Despite this rise in prevalence, ADHD screening methods remain almost entirely clinical, relying heavily on behavioral assessments, questionnaires, and clinical interviews [2].

This predominantly subjective approach to diagnosis presents several challenges:

- **Diagnostic burden:** Clinical assessments are time-consuming and require specialized expertise
- **Subjectivity:** Different clinicians may reach different conclusions based on the same symptoms
- **Delayed diagnosis:** The lack of objective biomarkers can lead to delayed or missed diagnoses
- **Comorbidity confusion:** ADHD symptoms often overlap with other conditions, complicating diagnosis

These limitations have created a pressing need for more objective, data-driven diagnostic tools. Electroencephalography (EEG) has emerged as a promising modality for ADHD detection because of its non-invasive nature and ability to capture neural activity patterns associated with attention and executive function [3].

While research has been conducted on utilizing physiological data, particularly EEG signals, to diagnose ADHD, there remains significant room for improvement in terms of both accuracy and computational efficiency. Many existing approaches either require extensive computational resources or fail to achieve clinically relevant accuracy levels.

1.1 Contributions

This paper makes the following contributions:

1. I propose a lightweight, two-staged teacher-student deep learning pipeline for ADHD classification using EEG-derived features

2. I implement and compare three teacher architectures: 1D ResNet, Time Series Transformer, and EEGNet

3. I demonstrate that gradient-based saliency mapping can effectively identify discriminative features for feature selection

4. I show that the proposed approach achieves competitive accuracy (80.46%) with efficient inference times suitable for real-time applications

2 Related Work

2.1 EEG-Based ADHD Detection

EEG has been extensively studied for ADHD detection due to its ability to capture the neurophysiological abnormalities associated with the disorder. Traditional approaches have focused on extracting hand-crafted features such as power spectral density, connectivity measures, and event-related potentials [4].

Recent work has shifted toward deep learning approaches that can automatically learn relevant features from raw or minimally preprocessed EEG data. Convolutional Neural Networks (CNNs) have shown promise in capturing spatial and temporal patterns in EEG signals [5], while recurrent architectures such as LSTMs have been used to model the sequential nature of EEG data [6].

2.2 EEGNet Architecture

EEGNet [7] is a compact convolutional neural network architecture specifically designed for EEG-based brain-computer interfaces. It employs depthwise and separable convolutions to efficiently learn spatial and temporal features from EEG data while maintaining a small parameter count. This makes it particularly suitable for deployment in resource-constrained environments.

2.3 Teacher-Student Learning

Teacher-student learning, also known as knowledge distillation [8], is a training paradigm where a smaller "student" network is trained to mimic the behavior of a larger "teacher" network. This approach has been successfully applied to compress large models while maintaining performance. In this work, I extend this paradigm by using teacher networks to identify important features through saliency mapping, rather than for traditional knowledge distillation.

2.4 Gradient-Based Saliency

Gradient-based saliency methods compute the importance of input features by examining the gradients of the model's output with respect to its inputs [9]. These methods have been widely used for model interpretability and can reveal which input features most strongly influence the model's predictions.

3 Methodology

3.1 Dataset

I utilized a publicly available dataset of preprocessed EEG-derived feature vectors from Kaggle [10]. The dataset contains feature vectors extracted from EEG recordings of subjects diagnosed with ADHD and healthy controls. The preprocessing pipeline applied to the original EEG data included:

- Band-pass filtering to remove artifacts
- Segmentation into fixed-length epochs
- Feature extraction including spectral and temporal features
- Normalization and standardization

3.2 Proposed Pipeline

Our approach consists of a two-staged pipeline:

3.2.1 Stage 1: Teacher Training and Saliency Extraction

In the first stage, we train three different "teacher" models on the full feature set:

1. **1D ResNet:** A one-dimensional residual network adapted for time-series classification. The architecture includes residual blocks with skip connections to facilitate gradient flow and enable training of deeper networks.

2. **Time Series Transformer:** A transformer-based architecture designed for sequential data. It employs self-attention mechanisms to capture long-range dependencies in the EEG feature sequences.

3. **EEGNet:** A compact CNN architecture specifically designed for EEG classification. It uses depthwise separable convolutions to efficiently learn spatial and temporal features.

After training, we compute gradient-based saliency maps for each teacher model. The saliency score S_i for each feature x_i is computed as:

$$S_i = \left| \frac{\partial L}{\partial x_i} \right| \tag{1}$$

where L is the loss function. These saliency scores indicate the importance of each feature for the classification task.

3.2.2 Stage 2: Feature Masking and Student Evaluation

In the second stage, we use the saliency maps to create a feature mask. Features with saliency scores below a threshold τ are masked (set to zero), effectively filtering out low-importance features:

$$x_i^{masked} = \begin{cases} x_i & \text{if } S_i \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The masked feature vectors are then used to train and evaluate an EEGNet "student" model. This approach allows us to:

- Reduce the effective dimensionality of the input
- Focus the student model on the most discriminative features
- Potentially improve generalization by removing noisy or irrelevant features

Algorithm 1 Teacher-Student Pipeline

Input: Features X, labels Y
Output: Trained student model
// Stage 1: Teacher Training
for $T \in \{ResNet1D, Transformer, EEGNet\}$ **do**
    Train teacher T on (X, Y)
    Compute saliency maps S_T
end for
// Stage 2: Feature Masking
Compute mask M from saliency
$X_{masked} = X \odot M$
Train student on (X_{masked}, Y)
return Student model

3.3 Model Architectures

3.3.1 1D ResNet

Our 1D ResNet architecture consists of:

- Initial convolutional layer with batch normalization
- Multiple residual blocks with skip connections
- Global average pooling
- Fully connected classification head

3.3.2 Time Series Transformer

The transformer architecture includes:

- Positional encoding layer
- Multi-head self-attention layers
- Feed-forward networks with GELU activation
- Classification token for final prediction

3.3.3 EEGNet (Student)

The EEGNet student architecture follows the original design [7]:

- Temporal convolution layer
- Depthwise spatial convolution
- Separable convolution
- Classification layer

3.4 Training Details

All models were trained using the following configuration:

- Optimizer: Adam with learning rate 10^{-3}
- Loss function: Cross-entropy loss
- Batch size: 32
- Early stopping with patience of 10 epochs
- 80-20 train-test split with stratification

4 Results

4.1 Performance Comparison

Table 1 presents the comprehensive performance comparison of our models across different configurations.

4.2 Key Findings

4.2.1 Best Performing Configuration

The Time Series Transformer teacher with EEGNet student pipeline achieved the best overall performance:

- **Accuracy:** 80.46%
- **F1 Score:** 80.57%
- **Precision:** 80.79%
- **Recall:** 80.46%
- **Inference Time:** 0.33 seconds

Table 1: Performance metrics for different teacher-student configurations

Model	Dataset	Accuracy	F1	Precision	Recall
EEGNet (Teacher)	Masked	0.6115	0.4641	0.3739	0.6115
ResNet1D (Teacher)	Masked	0.7904	0.7871	0.7884	0.7904
Transformer (Teacher)	**Masked**	**0.8046**	**0.8057**	**0.8079**	**0.8046**
EEGNet (Baseline)	Raw	0.7814	0.7695	0.7904	0.7814

Table 2: Inference time comparison (s/sample)

Model	Data	Time (s)
EEGNet	Masked	0.3206
ResNet1D	Masked	0.3258
Transformer	Masked	0.3301
EEGNet	Raw	0.3694

4.2.2 Effect of Feature Masking

Comparing the EEGNet baseline (trained on raw preprocessed data) with the masked configurations reveals interesting patterns:

- The EEGNet-Teacher with masked data performed worse (61.15%) than the baseline (78.14%), suggesting that EEGNet's saliency maps may not effectively capture discriminative features
- The ResNet1D and Transformer teachers both improved upon the baseline when used to guide feature masking
- The Transformer's attention mechanism appears to identify the most relevant features, leading to the best performance on masked data

4.2.3 Inference Efficiency

All masked configurations achieved faster inference times compared to the raw data baseline:

- Average improvement: 10-13% reduction in inference time
- The efficiency gain comes from processing a reduced effective feature space
- Inference times remain suitable for real-time applications

5 Discussion

5.1 Interpretation of Results

Our results demonstrate that the choice of teacher architecture significantly impacts the quality of the learned saliency maps and, consequently, the effectiveness of feature masking. The Time Series Transformer's self-attention mechanism appears particularly well-suited for identifying the most discriminative EEG features for ADHD classification.

The poor performance of EEGNet as a teacher for feature masking (61.15% accuracy) suggests that not all architectures produce equally useful saliency maps. This finding has important implications for the design of teacher-student pipelines: the teacher's architecture should be chosen not only for its classification performance but also for its ability to produce interpretable and useful feature attributions.

5.2 Clinical Implications

The achieved accuracy of 80.46% represents a promising step toward objective ADHD screening tools. While not yet sufficient for standalone diagnosis, such a system could serve as:

- A preliminary screening tool to identify individuals who should undergo comprehensive clinical evaluation
- A supplementary data point to support clinical decision-making or just a tool for monitoring treatment response over time

5.3 Limitations

This study has several limitations that should be addressed in future work:

1. **Dataset size:** The relatively small dataset may limit generalizability
2. **Preprocessed features:** Using preprocessed feature vectors rather than raw EEG may miss important information
3. **Single dataset:** Validation on multiple independent datasets is needed
4. **Binary classification:** ADHD presentations are heterogeneous and may benefit from multi-class or severity-based classification

5.4 Future Directions

the results still felt alittle underwhelming compared to larger deep networks. promising directions for future research include:

- Exploring other saliency methods such as integrated gradients or SHAP values
- Investigating the neurophysiological interpretation of the selected features
- Extending the approach to raw EEG data using end-to-end learning
- Developing ensemble methods combining multiple teacher architectures
- Validating on larger, multi-site datasets with diverse populations

6 Conclusion

This paper presented a lightweight, two-staged teacher-student deep learning pipeline for ADHD classification using EEG-derived features. By training teacher models to identify discriminative features through gradient-based saliency and using these insights to guide feature masking, we achieved competitive classification performance with improved computational efficiency.

Our experiments demonstrated that the Time Series Transformer serves as the most effective teacher architecture, achieving 80.46% accuracy with an inference time of 0.33 seconds. These results suggest that attention-based models may be particularly well-suited for identifying clinically relevant patterns in EEG data.

The proposed approach represents a step toward more objective, data-driven tools for ADHD screening. While further validation is needed before clinical deployment, the combination of reasonable accuracy and efficient inference makes this approach suitable for real-world screening applications. Future work will focus on validating these findings on larger datasets and exploring the neurophysiological interpretation of the selected features.

References

[1] Danielson, M. L., et al. (2018). Prevalence of parent-reported ADHD diagnosis and associated treatment among US children and adolescents, 2016. *Journal of Clinical Child & Adolescent Psychology*, 47(2), 199-212.

[2] American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

[3] Snyder, S. M., et al. (2015). Integration of an EEG biomarker with a clinician's ADHD evaluation. *Brain and Behavior*, 5(4), e00330.

[4] Lenartowicz, A., & Loo, S. K. (2014). Use of EEG to diagnose ADHD. *Current Psychiatry Reports*, 16(11), 498.

[5] Schirrmeister, R. T., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11), 5391-5420.

[6] Bashivan, P., et al. (2015). Learning representations from EEG with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*.

[7] Lawhern, V. J., et al. (2018). EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5), 056013.

[8] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

[9] Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

[10] EEG Dataset for ADHD. Kaggle. Available at: `https://www.kaggle.com/datasets/danizo/eeg-dataset-for-adhd`