Advancing EEG-Based Gaze Prediction: Depthwise Separable Convolution and Pre-Processing Enhancements in EEGViT

Matthew L Key*, Tural Mehtiyev*, and Xiaodong Qu

Department of Computer Science
The George Washington University
Washington DC, USA

Abstract. In the field of EEG-based gaze prediction, the application of deep learning to interpret complex neural data poses significant challenges. This study evaluates the effectiveness of pre-processing techniques and advancements in model architecture on EEG Vision Transformers (ViTs) pretrained model. We introduce a novel method, which combines depthwise separable convolutional CNNs with Vision Transformers, enriched by a pre-processing strategy involving data clustering. The new approach demonstrates superior performance, establishing a new benchmark with a Root Mean Square Error (RMSE) of 51.5 mm. This achievement underscores the impact of pre-processing and model refinement in enhancing EEG-based applications.

Keywords: EEG, Gaze Prediction, Machine Learning, Vision Transformer, EEGEyeNet, Depthwise Separable Convolution.

1 Introduction

Electroencephalogram (EEG) data, with its multidimensional architecture, captures an abundance of details regarding brain functions, providing various perspectives on numerous neurological events [42]. Despite the widespread use of machine learning regression models for EEG data, their complexity and expensive data collection process often hinder these models from effectively understanding the data's complex structures [8]. The EEGEyeNet dataset, with its extensive collection of EEG and eye tracking (ET) data, emerges as a significant asset in this field, enabling in-depth gaze behavior study and laying the groundwork for benchmarking gaze prediction approaches [19]. Leveraging the EEGEyeNet dataset, the Hybrid Vision Transformer (ViT) has showcased its potential in gaze prediction, challenging conventional convolution-based approaches [45]. As a contribution to the field, our study delves into how alterations in EEGViT design, combined with pre-processing techniques, can amplify the accuracy in predicting absolute eye position. Following our findings, we propose a new model which obtains better than state of the art performance on EEGEyeNet abosolute eye position.

* Both authors contributed equally to this work, Prof Qu is the research mentor for this project.
** Full source code is available at this Github Link

1.1 Research Questions

To further elucidate our direction within this evolving landscape, we formulate two pivotal research questions (RQs):

RQ 1: In what ways does incorporating depthwise separable convolution into EEG-based gaze prediction models influence their predictive accuracy?

RQ 2: What impact do advancements in pre-processing techniques have on the accuracy of EEG-based gaze prediction models?

2 Related Work

The analysis of EEG data has a long history and has been employed in various aspects of human-computer interaction (HCI), including research on sleep, emotion, gaming, and clinical applications[9,39,1,12,15,23,20,29,34,46,11,52,43,32,36,37,35,38,33]. Gaze prediction, with its extensive applications in human behavior analysis, advertising, and human-computer interactions, has garnered significant attention. The traditional reliance on Convolutional Neural Networks (CNNs) for this task has been reconsidered due to their limitations in capturing complex EEG patterns [4][28][30]. The introduction of the EEGViT model, incorporating Transformer blocks, represents a significant shift, offering a promising alternative to conventional convolutional approaches [45].

The integration of Vision Transformers (ViTs) with EEG-based gaze prediction marks a notable advancement, utilizing deep learning to navigate the intricacies of brain data interpretation. The effectiveness of both pure and hybrid transformer models in gaze estimation has been showcased, illustrating their capability in extracting detailed spatial features [7]. Such models, including the GazeTR, demonstrate the versatility of transformers, adapting well across different data modalities and enhancing the accuracy of human gaze prediction.

Transformers have also been applied beyond gaze prediction, notably in EEG signal analysis for tasks like epileptic seizure prediction. This broadens the scope of transformer applications from their origins in NLP to encompass the analysis of temporal and spatial EEG signal features, highlighting their adaptability and potential in handling complex EEG data [13].

Furthermore, exploring the synergy between CNNs and transformers has opened new avenues for EEG data processing. This combined approach leverages CNNs for local feature extraction and transformers for global dependency modeling, as demonstrated in Transformer-guided CNNs (TGCNN) for seizure prediction. Such innovations underline the potential of integrating CNN and transformer architectures to achieve higher accuracy and better generalization in EEG-based applications, including gaze prediction [13] [7].

This evolving landscape underscores the promise of combining CNNs and transformers in EEG data analysis, guiding our research towards optimizing such integrations. By harnessing the strengths of both architectures, we aim to set new standards in EEG-based gaze prediction and neural data interpretation, contributing to the field's advancement.

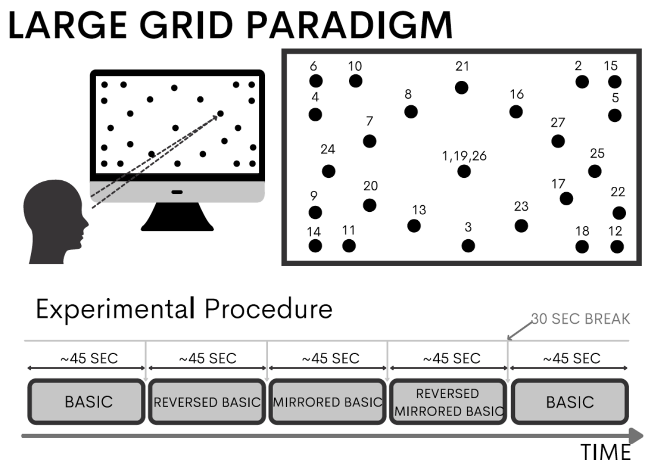

Fig. 1: **Large Grid Experimental Setup:** This image illustrates the schematic view of the experimental setup and the stimuli placement on the screen. It gives a visual representation of how participants interacted with the stimuli during the eye-tracking events [19].

3 Methods

Our research extends the work presented in [45], focusing on the utilization of pre-processing and depthwise-separable convolution techniques in EEG-based gaze prediction methodologies.

Data Pre-Processing: Pre-processing techniques have become crucial in enhancing the performance of pre-trained Vision Transformer models, as noted in studies by Chen et al. (2021) [5] and Li et al. (2021) [24]. We observed that the EEGEyeNet dataset contains noise. During the original data collection, the EEGEyeNet procedures required participants to focus on specific target positions. Kastrati et al. (2021) [19] reported that, with the computer monitor used in the experiment, 1 pixel equates to 0.5 mm. However, we identified x and y label positions in the dataset that are as much as 100 pixels (or 5 cm) away from any known target position (see Figure 2). This significant discrepancy led us to hypothesize that participants were indeed looking at the target positions, suggesting a potential issue with the eye tracking system. The inaccuracies could be due to the system's malfunction or improper calibration.

Another potential source of error might stem from the disparity in the granularity of the data collected. The EEG data were captured at a frequency of 500 Hz, equivalent to 500 times per second. In contrast, the eye tracking data were recorded at a much lower frequency, once per second [19]. Therefore, if a participant's gaze was in transit to a target point when captured, the recorded eye position might not accurately represent the entire second during which the brainwave data were collected. Unfortunately, with the available data, it is impossible to determine the exact position of the participant's eyes throughout each sample.

To address the discrepancy in eye tracking location, we employed K-means clustering to reconcile the differences between the label position and the actual target position. By updating the true label position with the centroids, as illustrated in Figure 3, we align it with the cluster center position, thereby enhancing the accuracy of our dataset.

Fig. 2: Clustering showing the difference between the label positions and the target position

Depthwise-separable convolutional neural networks (DS-CNNs): Early studies [21] highlighted that initial layers of CNNs are adept at detecting edges or specific colors in natural images. In recent years, research aiming to gain a deeper understanding of how Convolutional Neural Networks (CNNs) operate has largely shifted towards analyzing the features learned by convolutional layers rather than the weights themselves [48], [47]. While examining the learned features of convolutional layers is a logical approach, the interpretation of the filter weights in the deeper layers of CNNs remains a challenge. Meanwhile, Depthwise-Separable Convolutional Neural Networks (DS-CNNs) have been rising in prominence within the field of computer vision and demonstrated state-of-the-art accuracy while requiring significantly fewer parameters and computational operations than traditional CNNs, owing to the reduced computational demands of DS-CNNs [16].

The application of depthwise separable convolution in EEG data analysis, as demonstrated in various studies, shows its potential in enhancing model performance through efficient feature extraction from multichannel EEG signals. The high accuracy rates achieved in emotion recognition tasks using publicly available EEG datasets, as cited in the works by Li et al. [22] and further supported by studies [17], [44], underscore its effectiveness in reducing computational load while maintaining or improving performance score.

Building on these findings, the application of depthwise separable convolution could be extended to the EEGEyeNet dataset. EEGEyeNet, being a comprehensive dataset for gaze estimation and other EEG-based analyses, could benefit significantly from the effective feature extraction capabilities of depthwise separable convolution. This approach may enhance the accuracy, especially in tasks requiring the analysis of spatial EEG signal characteristics. The potential for improved performance in EEG-based predictive modeling with reduced computational demands makes depthwise separable convolution a promising technique for exploration in this dataset.

We apply depthwise separable convolution by expanding the the previous work [45] where the authors developed a hybrid Vision Transformer architecture named EEGViT, specifically tailored for EEG analysis. This model integrates a traditional two-step convolution operation during the patch embedding process. The first step involves a convolutional layer employing a 1×T kernel to capture temporal events across channels, acting as band-pass filters for EEG signals. This layer emphasizes the extraction of frequency bands of interest in the EEG data. Following this, the second step involves a depthwise convolutional layer with a C×1 kernel, designed to filter inputs across multiple channels at the same point in time. The model segments input images into C×T patches, which undergo a row-by-row linear projection, transforming each column vector into a scalar feature.

Building on previous study, we introduce an additional depthwise separable convolution layer in our approach. This layer incorporates both depthwise and pointwise convolutions. Following this enhancement, our systematic approach for EEG data classification begins with a 2D convolution layer employing 256 filters of size (1, 36), featuring a stride of (1, 36) and padding of (0, 2). This layer is tasked with extracting temporal features from EEG signals. Subsequently, the depthwise separable convolution layer, comprising 256 filters for the depthwise part and 512 filters for the pointwise part, processes spatial information across channels. The architecture further integrates a Vision Transformer (ViT), modified with a custom depthwise convolution layer using 512 filters of size (8, 1). The process concludes with a classifier that includes a linear layer, a dropout layer, and a final linear layer, responsible for outputting logits that indicate class probabilities in a binary classification task. The incremental addition of the depthwise separable convolution layers in on the previous approach has proven to be effective in generating enhanced spatial features. These improved features effectively contribute to the model's ability to refine its performance and improve its scoring accuracy.

Evaluation Metrics: To maintain consistency and ensure comparability with prior work, all methods, whether applied individually or in combination, will be gauged using the root mean squared error (RMSE).

Early Stopping: In addition to these techniques, we employ a type of early-stopping during training for improved performance. The SOTA EEG-ViT model was trained on a static number of 15 epochs [45]. However, the authors did not take advantage of the

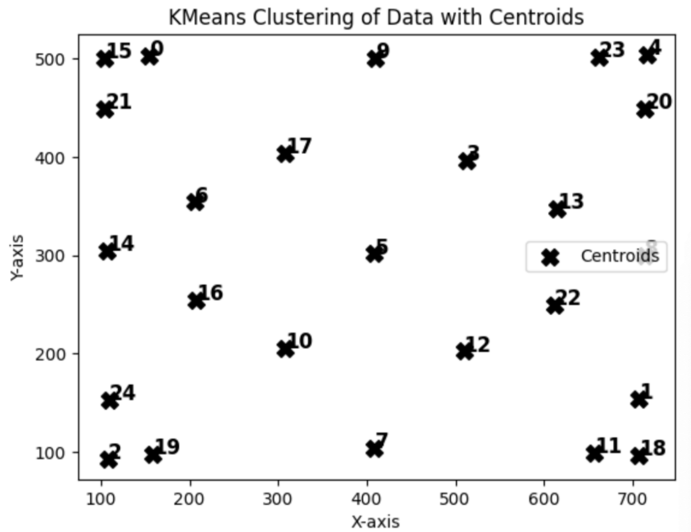

Fig. 3: The centroids used to correct training data labels.

Method	Description
Method 1	EEGVIT Pre-trained with DS-CNNs
Method 2	EEGVIT Pre-trained with Clustered Data
Method 3	EEGVIT Pre-trained with Clustered and DS-CNNs

Table 1: Descriptions of the various methods used in the study.

validation set to detect when the model was overfitting to the training data. During training, our algorithm will run for 15 epochs and then output the trained model based on the epoch that has the best validation score. This will protect against overfitting and encourage higher overall accuracy.

As outlined in Table 1, our study employs several methods to address the problem at hand. Each method has been tailored to optimize performance based on the specific characteristics of the dataset and the goals of the analysis.

Method 1: EEGVIT Pre-trained with DS-CNNs: This approach leverages a pre-trained EEGVIT model, further refined using depthwise-separable convolutional neural networks as an additional layer. Known for their superior spatial feature extraction capabilities (citation is needed), DS-CNNs enable the model to effectively identify and process complex patterns in EEG channels. This method addresses Research Question 1 by demonstrating the impact of depthwise separable convolutions techniques on the accuracy score in EEG data.

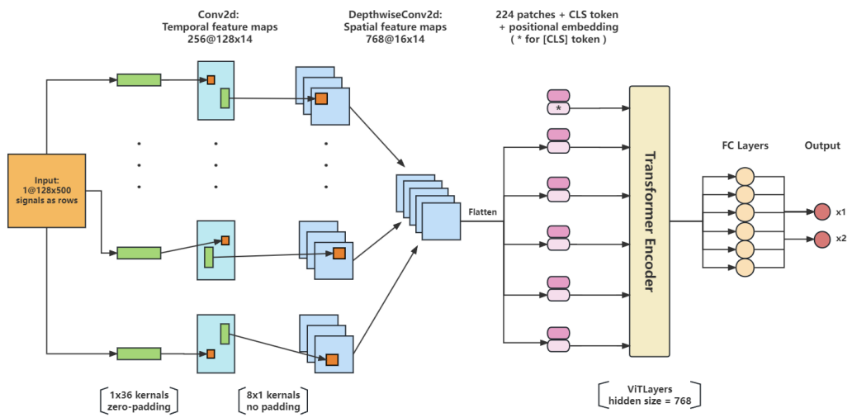

Fig. 4: **EEGViT Architecture Overview:** A specialized Vision Transformer (ViT) structure tailored for raw EEG signal input. This architecture utilizes a dual-step convolution process to produce patch embeddings. After this initial step, positional embeddings are integrated and the combined sequence is subsequently passed through the ViT layers [45]. The design of the positional embedding and ViT layer is adapted from [10].

Method 2: EEGVIT Pre-trained with Clustered Data: By clustering the data prior to training, we can ensure that the model is exposed to the most representative and diverse examples. This pre-processing step helps in improving the generalization capability of the EEGVIT model by focusing on the underlying distribution of the dataset. This method addresses Research Question 2 by demonstrating the impact of data processing techniques on the accuracy score in EEG data.

Method 3: EEGVIT Pre-trained with Clustered and DS-CNNs: This method integrates the techniques of data clustering with depthwise-separable convolutional neural networks (DS-CNNs) to harness the advantages of both approaches. By clustering the EEG data, the model can focus on learning from more homogeneous subsets, which improves its efficiency in recognizing underlying patterns. When combined with the DS-CNNs, known for their enhanced feature extraction with fewer parameters and computational efficiency, this strategy significantly boosts the model's capacity to identify intricate and subtle patterns within the EEG channels. This dual approach integrates the findings from both research questions to enhance the pre-training phase, laying a robust foundation for the model. This integration aims to boost the accuracy and improve the generalization capabilities of EEG data analysis.

4 Dataset

The EEGEyeNet dataset comprises data from 27 participants with a total of 21,464 samples [19]. The primary focus is on the "Absolute Position" task where the objective is to ascertain the exact gaze position in terms of XY-coordinates on the screen. Each sample corresponds to a one-second duration where a participant engages in a single fixation on the Large Grid paradigm (Figure 1). The performance is assessed by measuring the Euclidean distance between the actual and predicted gaze positions in the XY-plane.

5 Results

As shown in Table 2, the previous highest achievement on the EEGEyeNet dataset's absolute position task was an RMSE (Root Mean Square Error) of 55.4 ± 0.2 mm, as reported by [45]. The results from all three methods, as described in Table 2, demonstrated improved performance in terms of RMSE. In Method 1, where we implemented depthwise separable convolution, we achieved 53.5 mm. In Method 2, which applied 'EEGVIT Pre-trained with Clustered Data,' we achieved an RMSE of 52.9 mm. This result indicates the positive impact of data clustering on model accuracy. Finally, in Method 3, where we combined both methods by training the model with depthwise separable convolution on the clustered data, we achieved an even better RMSE of 51.5 ± 0.2 mm, reinforcing the effectiveness of these combined strategies.

Model	**Absolute Position RMSE (mm)**
Naive Guessing	123.3 ± 0.0
CNN	70.4 ± 1.1
PyramidalCNN	73.9 ± 1.9
EEGNet	81.3 ± 1.0
InceptionTime	70.7 ± 0.8
Xception	78.7 ± 1.6
VIT - Base	61.5 ± 0.6
VIT - Base Pre-trained	58.1 ± 0.6
EEGVIT	61.7 ± 0.6
EEGVIT Pre - trained	55.4 ± 0.2
Method 1	53.5 ± 0.6
Method 2	53.4 ± 0.8
Method 3	**51.5 ± 0.2**

Table 2: **RMSE Comparisons for Absolute Position Task:** Root Mean Squared Error (RMSE) was converted to millimeters at a ratio of 2 pixels/mm. Lower RMSE values signify better accuracy, aligning closer to true values. Displayed values represent the average and standard deviation from 5 trials. [45].

6 Discussion

These results collectively suggest that specialized pre-training involving data clustering and DS-CNNs can significantly improve the accuracy of deep learning models in estimating absolute positions from EEG data.

	precision	recall	f1-score	support
0	0.64	0.67	0.66	127
1	0.67	0.76	0.71	126
2	0.66	0.84	0.74	89
3	0.59	0.51	0.55	120
4	0.92	0.64	0.75	152
5	0.50	0.45	0.47	328
6	0.48	0.55	0.51	125
7	0.88	0.85	0.87	124
8	0.58	0.38	0.46	112
9	0.80	0.83	0.81	124
10	0.49	0.56	0.52	118
11	0.64	0.71	0.67	121
12	0.57	0.59	0.58	119
13	0.38	0.35	0.36	120
14	0.45	0.50	0.47	121
15	0.57	0.41	0.48	111
16	0.43	0.53	0.48	117
17	0.49	0.51	0.50	120
18	0.40	0.55	0.46	109
19	0.75	0.58	0.66	142
20	0.61	0.50	0.55	114
21	0.62	0.83	0.71	122
22	0.46	0.51	0.49	121
23	0.54	0.68	0.60	100
24	0.56	0.39	0.46	137
accuracy			0.58	3219
macro avg	0.59	0.59	0.58	3219
weighted avg	0.59	0.58	0.58	3219

Fig. 5: **Classification Performance Metrics by Cluster:** This figure presents a detailed breakdown of classification metrics including precision, recall, F1-score, and support for 25 clusters, highlighting the performance of each cluster in the model evaluation.

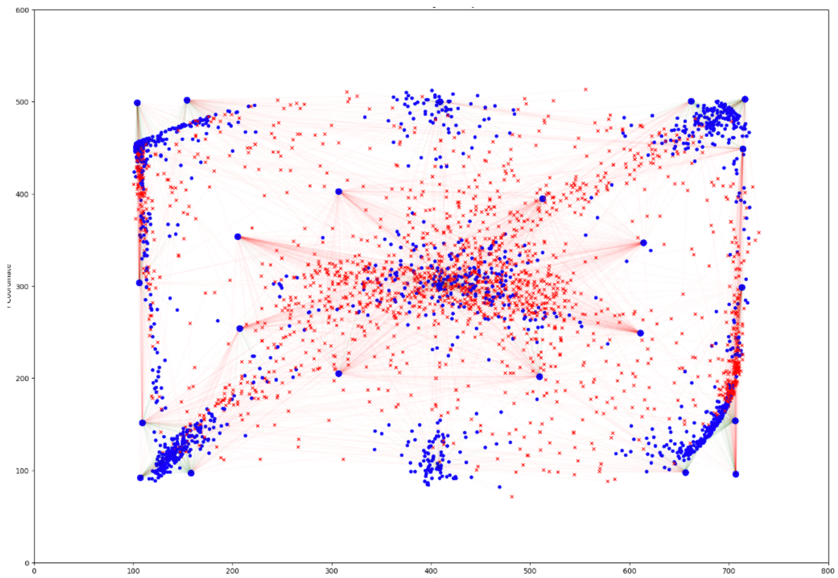

Fig. 6: Visual of Test Error for Absolute Eye Position Showing Positions within 55.4 RMSE (Blue) and Positions Above 55.4 RMSE (Red).

Understanding Test Error: One of the pivotal aspects of our study was the introduction of new visualization techniques that will help both computer scientists and neuroscientists understand the test error. During our training, we discovered a way to better understand the test error. Where is the test error coming from? Which eye positions have more error? Which eye positions had less error? We created a new visual in order to help us answer these questions (See Figure 6). For example, in Figure 6, we see that the eye positions on the top left and bottom right are more difficult for the model to perform well on compared to the bottom left and upper right-hand corners. Insights from neuroscientists and other subject matter experts will be critical in order to improve performance in these positions. In this same figure, faint lines between test locations and true labels show the distance between the target and predicted values. Notably, there are fewer red lines between the "inner" positions and the "outer" positions. This could mean that the model is good at determining the difference between someone looking at the center of the screen as opposed to the outside of the screen, though we did not quantify these results.

Understanding EEG-ViT Performance: In order to understand the original EEG-VIT model, our team expanded the use of clustered eye positions shown in Figure 3 by converting the model into a classifier. So, instead of predicting a location on a screen, the adjusted classification model would predict one of the 25 centroids shown in Figure 3.

In the given classification report in 5, the original EEGViT model's discriminative ability is quantified across multiple classes, with individual performance metrics

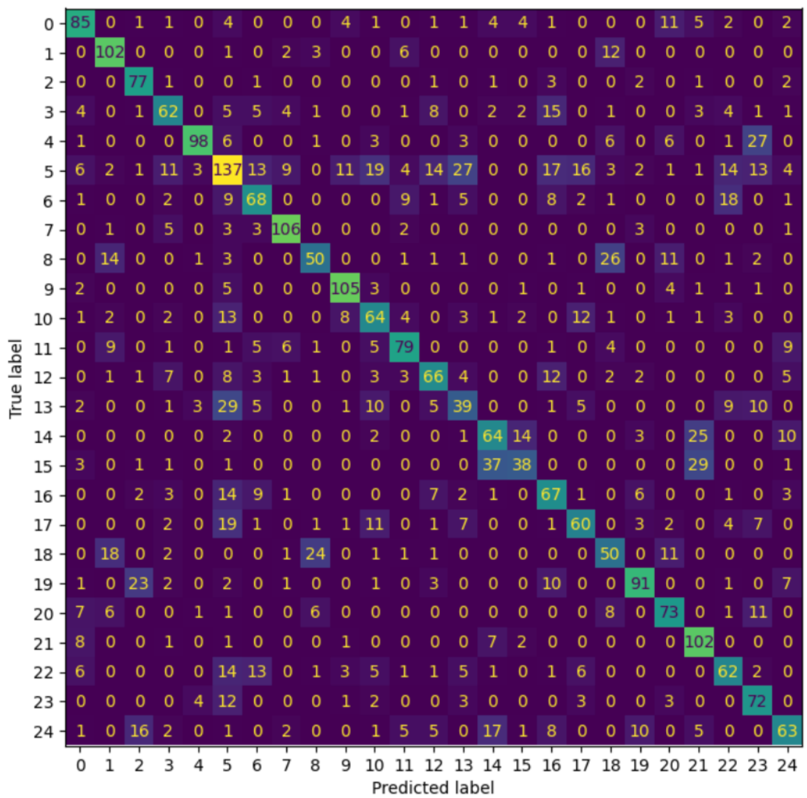

Fig. 7: **Confusion matrix across 25 clusters:** On the x-axis, we have the predicted values, which represent the outcomes as forecasted by our model (with method 3). The y-axis, on the other hand, displays the true labels for each data point.

presented for each class. Precision, recall, and F1-scores are provided, alongside the 'support' figure, which denotes the actual number of samples for each respective class. Classes 7 (participant looking straight down) and 9 (participant looking straight up) are noteworthy, with F1-scores of 0.87 and 0.81 respectively, indicating a robust predictive performance for these categories. However, there are classes with notably lower F1-scores, such as class 13, indicating potential areas for model improvement. Similarly, the confusion matrix in Figure 7 reveals that categories 7 and 9 closely match their predictions with the true labels, while class 13 has the least number of matched predictions. The high number of matched predictions in category 5 is attributed to its larger sample size in the dataset. Notably, the central category, represented three times more frequently than others, may skew the model's predictive distribution. Future iterations of the model could benefit from a more targeted approach in feature engineering and class-specific parameter tuning to uplift the predictive accuracy for underperforming classes.

Furthermore, our team evaluated samples that were predicted with high confidence scores by EEGViT. We hypothesised this would uncover patterns detected by EEGViT model that are also interpretable with the human eye. We discovered there are similarities in samples classified with high confidence using EEG-ViT. As shown in Figure 8 there are obvious similarities in the EEG samples. The cause of the similarities in these data are undetermined. It could be leakage from ocular artifacts or valuable data which requires insight for neuroscientists.

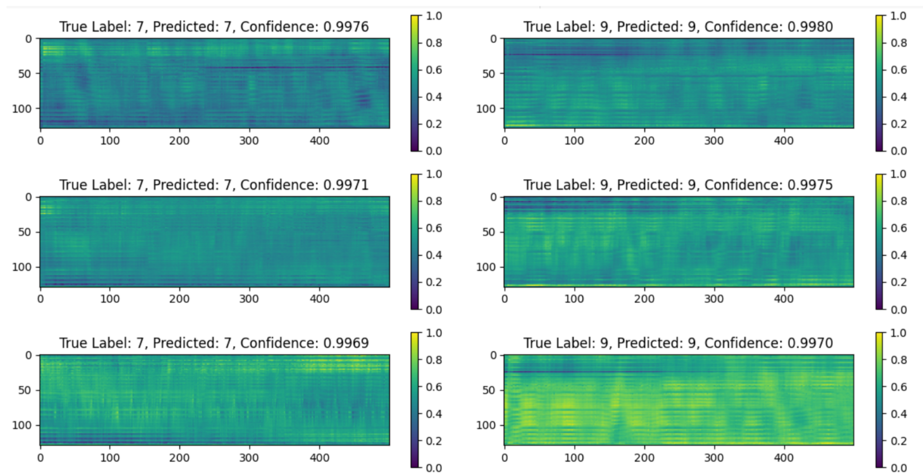

Fig. 8: **Heat Map of EEG Data Samples Predicted with High Confidence by EEG-ViT Model:** This figure shows samples with high confidence from class 7 (left) and class 9 (right). Class 7 displays similar bright spot in the upper left corner. Class 9 displays a similar, larger, dark spot in the upper left.

These visual tools not only facilitated a deeper understanding of the model's performance but also provided insights into the complex interplay of data features and model predictions. Visuals like this could be useful also as a communication tool between computer and neuroscientists. However, there is still a vast scope for innovation in this domain. Future research can focus on developing more advanced visualization techniques and tools that can provide even deeper insights into the workings of EEG data analysis models. This direction holds the promise of not only enhancing the interpretability of complex models but also fostering a more collaborative and intuitive approach to understanding neuroscience data.

Our investigation applied enhanced pre-processing strategies and architectural improvements to a pre-trained EEG-ViT model, resulting in notable performance enhancements. The integration of Vision Transformers with EEG data analysis in our EEG-ViT model has demonstrated a powerful synergy. Importantly, the potential of these pre-processing techniques, when applied to Convolutional Neural Networks (CNNs), should not be overlooked. Future studies could explore how these strategies might elevate the performance of CNNs in EEG data analysis without the addition of a Vision Transformer model. This approach could offer valuable comparative insights between Transformer-based and convolutional architectures. Additionally, the further exploration of depthwise separable convolution's potential to enhance performance warrants attention. Another promising avenue for research involves the use of Generative Adversarial Networks (GANs) in generating synthetic EEG datasets. This could potentially address the challenges of data scarcity and diversity in EEG analysis. Future studies can also investigate other potential deep learning approaches on various datasets for comparative analysis [3,2,18,14,25,6,41,26,27,40,31,49,50,51].

7 Conclusion

In conclusion, the deployment of pre-processing and using DS-CNNs has markedly improved the performance of EEG-based predictive models. Our proposed model, in particular, has established new state-of-the-art results, achieving a benchmark RMSE of 51.5 mm. We are optimistic that the significant performance leap made by our model will serve as a cornerstone for future developments in EEG-based brain-computer interfaces and machine learning, inspiring continued innovation and research in the field.

References

1. Altaheri, H., Muhammad, G., Alsulaiman, M., Amin, S.U., Altuwaijri, G.A., Abdul, W., Bencherif, M.A., Faisal, M.: Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. Neural Computing and Applications **35**(20), 14681–14722 (2023)
2. An, S., Bhat, G., Gumussoy, S., Ogras, U.: Transfer learning for human activity recognition using representational analysis of neural networks. ACM Transactions on Computing for Healthcare **4**(1), 1–21 (2023)
3. An, S., Tuncel, Y., Basaklar, T., Ogras, U.Y.: A survey of embedded machine learning for smart and sustainable healthcare applications. In: Embedded Machine Learning for Cyber-Physical, IoT, and Edge Computing: Use Cases and Emerging Challenges, pp. 127–150. Springer (2023)
4. Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F.: A review on vision techniques applied to human behaviour analysis for ambient-assisted living. Expert Systems with Applications **39**(12), 10873–10888 (sep 2012). `https://doi.org/10.1016/j.eswa.2012.03.005`
5. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. arXiv preprint arXiv:2012.00364 (dec 2021), `https://doi.org/10.48550/arXiv.2012.00364`
6. Chen, J., Hu, Y., Wang, Y., Lu, Y., Cao, X., Lin, M., Xu, H., Wu, J., Xiao, C., Sun, J., et al.: Trialbench: Multi-modal artificial intelligence-ready clinical trial datasets. arXiv preprint arXiv:2407.00631 (2024)
7. Cheng, Y., Lu, F.: Gaze estimation using transformer. arXiv preprint arXiv:2105.14424 (2021), `https://ar5iv.labs.arxiv.org/html/2105.14424`
8. Craik, A., He, Y., Contreras-Vidal, J.L.: Deep learning for electroencephalogram (eeg) classification tasks: a review. Journal of Neural Engineering **16**(3) (2019). `https://doi.org/10.1088/1741-2552/aba0b5`

9. Craik, A., He, Y., Contreras-Vidal, J.L.: Deep learning for electroencephalogram (eeg) classification tasks: a review. Journal of neural engineering **16**(3), 031001 (2019)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, G., Heigold, G., Gelly, S., Uszkoreit, J., Housby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (jan 2021)
11. Dou, G., Zhou, Z., Qu, X.: Time majority voting, a pc-based eeg classifier for non-expert users. In: International Conference on Human-Computer Interaction. pp. 415–428. Springer (2022)
12. Gao, Z., Dang, W., Wang, X., Hong, X., Hou, L., Ma, K., Perc, M.: Complex networks and deep learning for eeg signal analysis. Cognitive Neurodynamics **15**(3), 369–388 (2021)
13. Godoy, R.V., Reis, T.J.S., Polegato, P.H., Lahr, G.J.G., Saute, R.L., Nakano, F.N., Machado, H.R., Sakamoto, A.C., Becker, M., Caurin, G.A.P.: Eeg-based epileptic seizure prediction using temporal multi-channel transformers. arXiv preprint arXiv:2209.11172 (2022), https://ar5iv.labs.arxiv.org/html/2209.11172
14. Gui, S., Song, S., Qin, R., Tang, Y.: Remote sensing object detection in the deep learning era—a review. Remote Sensing **16**(2), 327 (2024)
15. Hossain, K.M., Islam, M.A., Hossain, S., Nijholt, A., Ahad, M.A.R.: Status of deep learning for eeg-based brain–computer interface applications. Frontiers in computational neuroscience **16**, 1006763 (2023)
16. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017), https://arxiv.org/pdf/1704.04861.pdf
17. Huang, Z., Ma, Y., Su, J., Shi, H., Jia, S., Yuan, B., Li, W., Yang, T.: Cdba: a novel multi-branch feature fusion model for eeg-based emotion recognition. Frontiers in Physiology **14** (2023). https://doi.org/10.3389/fphys.2023.1200656, https://www.frontiersin.org/articles/10.3389/fphys.2023.1200656/full
18. Jiang, C., Hui, B., Liu, B., Yan, D.: Successfully applying lottery ticket hypothesis to diffusion model. arXiv preprint arXiv:2310.18823 (2023)
19. Kastrati, A., Płomecka, M.B., Pascual, D., Wolf, L., Gillioz, V., Wattenhofer, R., Langer, N.: Eegeyenet: a simultaneous electroencephalography and eye-tracking dataset and benchmark for eye movement prediction. ETH Zurich (nov 2021)
20. Koome Murungi, N., Pham, M.V., Dai, X., Qu, X.: Trends in machine learning and electroencephalogram (eeg): A review for undergraduate researchers. arXiv e-prints pp. arXiv–2307 (2023)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. NeurIPS (2012), https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
22. Li, Q., Liu, Y., Liu, Q., Zhang, Q., Yan, F., Ma, Y., Zhang, X.: Multidimensional feature in emotion recognition based on multi-channel eeg signals. Entropy **24**(12), 1830 (2022). https://doi.org/10.3390/e24121830, https://www.mdpi.com/1099-4300/24/12/1830
23. Li, W., Zhou, N., Qu, X.: Enhancing eye-tracking performance through multi-task learning transformer. In: International Conference on Human-Computer Interaction. pp. 31–46. Springer (2024)
24. Li, W., Lu, X., Qian, S., Lu, J., Zhang, X., Jia, J.: On efficient transformer-based image pretraining for low-level vision. arXiv: Computer Vision and Pattern Recognition (Dec 2021), https://doi.org/10.48550/arXiv.2112.10175

25. Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Wei, W.: Machine learning for synthetic data generation: a review. arXiv preprint arXiv:2302.04062 (2023)
26. Ma, X.: Traffic performance evaluation using statistical and machine learning methods. Ph.D. thesis, The University of Arizona (2022)
27. Ma, X., Karimpour, A., Wu, Y.J.: Data-driven transfer learning framework for estimating on-ramp and off-ramp traffic flows. Journal of Intelligent Transportation Systems pp. 1–14 (2024)
28. Majaranta, P., Bulling, A.: Eye tracking eye-based human-computer interaction (mar 2014). https://doi.org/10.1007/978-1-4471-6392-3_3
29. Murungi, N.K., Pham, M.V., Dai, X.C., Qu, X.: Empowering computer science students in electroencephalography (eeg) analysis: A review of machine learning algorithms for eeg datasets. SIGKDD (2023)
30. Okada, G., Masui, K., Tsumura, N.: Advertisement effectiveness estimation based on crowd-sourced multimodal affective responses. CVPR Workshop (2023)
31. Qiu, Y., Zhao, Z., Yao, H., Chen, D., Wang, Z.: Modal-aware visual prompting for incomplete multi-modal brain tumor segmentation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 3228–3239 (2023)
32. Qu, X.: Time Continuity Voting for Electroencephalography (EEG) Classification. Ph.D. thesis, Brandeis University (2022)
33. Qu, X., Hall, M., Sun, Y., Sekuler, R., Hickey, T.J.: A personalized reading coach using wearable eeg sensors. CSEDU (2019)
34. Qu, X., Hickey, T.J.: Eeg4home: A human-in-the-loop machine learning model for eeg-based bci. In: Augmented Cognition: 16th International Conference, AC 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings. pp. 162–172. Springer (2022)
35. Qu, X., Liu, P., Li, Z., Hickey, T.: Multi-class time continuity voting for eeg classification. In: Brain Function Assessment in Learning: Second International Conference, BFAL 2020, Heraklion, Crete, Greece, October 9–11, 2020, Proceedings 2. pp. 24–33. Springer (2020)
36. Qu, X., Liukasemsarn, S., Tu, J., Higgins, A., Hickey, T.J., Hall, M.H.: Identifying clinically and functionally distinct groups among healthy controls and first episode psychosis patients by clustering on eeg patterns. Frontiers in psychiatry **11**, 541659 (2020)
37. Qu, X., Mei, Q., Liu, P., Hickey, T.: Using eeg to distinguish between writing and typing for the same cognitive task. In: Brain Function Assessment in Learning: Second International Conference, BFAL 2020, Heraklion, Crete, Greece, October 9–11, 2020, Proceedings 2. pp. 66–74. Springer (2020)
38. Qu, X., Sun, Y., Sekuler, R., Hickey, T.: Eeg markers of stem learning. In: 2018 IEEE Frontiers in Education Conference (FIE). pp. 1–9. IEEE (2018)
39. Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T.H., Faubert, J.: Deep learning-based electroencephalography analysis: a systematic review. Journal of neural engineering **16**(5), 051001 (2019)
40. Tan, J., Shen, X., Zhang, X., Wang, Y.: Multivariate encoding analysis of medial prefrontal cortex cortical activity during task learning. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 6699–6702. IEEE (2021)
41. Tan, J., Zhang, X., Wu, S., Song, Z., Chen, S., Huang, Y., Wang, Y.: Audio-induced medial prefrontal cortical dynamics enhances coadaptive learning in brain–machine interfaces. Journal of Neural Engineering **20**(5), 056035 (2023)
42. Teplan, M.: Fundamentals of eeg measurement. Measurement Science Review **2**(2) (2002)
43. Wang, R., Qu, X.: Eeg daydreaming, a machine learning approach to detect daydreaming activities. In: International Conference on Human-Computer Interaction. pp. 202–212. Springer (2022)

44. Wang, X., Shi, R., Wu, X., Zhang, J.: Decoding human interaction type from inter-brain synchronization by using eeg brain network. IEEE Journal of Biomedical and Health Informatics (2023). https://doi.org/10.1109/JBHI.2023.3239742, https://pubmed.ncbi.nlm.nih.gov/37917521/, epub ahead of print
45. Yang, R., Modesitt, E.: Vit2eeg: Leveraging hybrid pretrained vision transformers for eeg data (2023)
46. Yi, L., Qu, X.: Attention-based cnn capturing eeg recording's average voltage and local change. In: Artificial Intelligence in HCI: 3rd International Conference, AI-HCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings. pp. 448–459. Springer (2022)
47. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579 (2015), https://arxiv.org/pdf/1506.06579.pdf
48. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. arXiv preprint arXiv:1311.2901 (2013), https://arxiv.org/pdf/1311.2901.pdf
49. Zhang, Z., Tian, R., Ding, Z.: Trep: Transformer-based evidential prediction for pedestrian intention with uncertainty. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3534–3542 (2023)
50. Zhang, Z., Tian, R., Sherony, R., Domeyer, J., Ding, Z.: Attention-based interrelation modeling for explainable automated driving. IEEE Transactions on Intelligent Vehicles **8**(2), 1564–1573 (2022)
51. Zhao, S., Yang, X., Zeng, Z., Qian, P., Zhao, Z., Dai, L., Prabhu, N., Nordlund, P., Tam, W.L.: Deep learning based cetsa feature prediction cross multiple cell lines with latent space representation. Scientific Reports **14**(1), 1878 (2024)
52. Zhou, Z., Dou, G., Qu, X.: Brainactivity1: A framework of eeg data collection and machine learning analysis for college students. In: International Conference on Human-Computer Interaction. pp. 119–127. Springer (2022)