School of Information Technology and Engineering at the ADA University

School of Engineering and Applied Science at the George Washington University

Advancing EEG-Based Gaze Prediction Using Pre-trained Vision Transformers

A Thesis
Presented to the Graduate Program of Computer Science and Data Analytics
of the School of Information Technology and Engineering
ADA University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Computer Science and Data Analytics
ADA University

By
Tural Mehtiyev

Supervisors: Dr. Xiadong QU and Dr. Jamaladdin Hasanov

April 2024

THESIS ACCEPTANCE

This Thesis by: Tural Mehtiyev
Entitled: *Advancing EEG-Based Gaze Prediction Using Pre-trained Vision Transformers*

has been approved as meeting the requirement for the Degree of Master of Science in Computer Science and Data Analytics of the School of Information Technology and Engineering, ADA University.

Approved:

(Adviser)	(Date)
(Program Director)	(Date)
(Dean)	(Date)

ACKNOWLEDGMENTS

This study was conducted at the "Applied Machine Learning Lab" at The George Washington University under the supervision of Dr. Xiaodong Qu. Together with another co-author, Matthew L. Key, the study—where Mr. Mehtiyev and Mr. Key contributed equally—was accepted to the HCI International 2024 conference, which will be held in Washington, D.C., USA. It has been peer-reviewed and assigned to be published in the LNAI 14695 volume of the proceedings by Springer, corresponding to the thematic area of the 18th International Conference on Augmented Cognition.

ABSTRACT

In the Brain-Computer Interface (BCI) field, applying deep learning to interpret neural data for Electroencephalogram (EEG) based gaze prediction is challenging due to complexity of EEG data. This study focuses on a hybrid deep learning model that combines Convolutional Neural Networks with Vision Transformers pre-trained on the ImageNet dataset. It focuses on the EEGeyenet dataset, targeting the absolute gaze prediction task. We evaluate the effectiveness of pre-processing techniques and depthwise separable convolution on EEG Vision Transformers (ViTs) within a pre-trained architecture. We introduce the EEG Deeper Clustered Vision Transformer (EEG-DCViT), an approach combining depthwise separable CNNs with Vision Transformers, enhanced by data clustering in pre-processing. This method sets a new benchmark by outperforming the state-of-the-art result (55.4 mm) with a Root Mean Square Error (RMSE) of 51.6 mm. This result validates the impact of preprocessing techniques and the potential of depthwise separable CNNs on EEG datasets. Details on experiment implementation are available at github.com/tmehtiyev2019/EEG-Gaze-Prediction.git repository.

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

Contents

List of Figures

List of Tables

Abbreviation	Explanation
BCI	Brain Computer Interface
EEG	Electroencephalogram
ET	Eye Tracking
ViT	Vision Transformer
EEGEyeNet	EEG and Eye Tracking Integrated for Gaze Estimation Dataset
CNN	Convolutional Neural Networks
DS-CNNs	Depthwise-Separable Convolutional Neural Networks
RMSE	Root Mean Squared Error
SOTA	State Of The Art
ICA	Independent Component Analysis
NLP	Natural Language Processing
EEGViT	EEG Vision Transformer
SOTA	State of the Art

1 Introduction

1.1 Background on Brain-Computer Interfaces

Brain-Computer Interfaces (BCIs) are at the forefront of neurotechnology, bridging the gap between the human brain and computers. They translate neurological signals into commands, enabling direct communication without any physical movement. An Electroencephalogram (EEG) is one of the primary modalities used in BCIs. It records the brain's electrical activity using a network of sensors placed on the scalp and maps the collective electrical activity of the brain. The data collected from this method are encoded into waveforms that provide a cortical snapshot of brain activity (see figure 1 and figure 2). The waveforms recorded in an EEG reflect the cortical electrical activity and are measured in microvolts, indicating the activity of large groups of neurons. This makes EEG an essential tool in the study of brain function and in the field of brain-computer interfaces.[2].

1.2 Applications of Machine Learning in EEG Data

The capture of detailed brain functions through EEG data enhances our understanding of brain dynamics across various states and conditions, and it opens up opportunities to apply machine learning algorithms to predict cognitive events using EEG data [30, 35, 37]. Various Machine Learning models such as Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs) are extensively utilized due to their effectiveness in handling the complex, high-dimensional nature of EEG data [3, 32]. EEG data finds applications in various predictive settings, including medical diagnostics, cognitive state assessments, and even in consumer research through neuromarketing. For instance, SVMs have been successfully used to classify EEG signals for epilepsy detection, differentiating seizure episodes from normal brain activity [3, 32]. ANNs have shown promising results in predicting emotional states from EEG, which can be used in enhancing user experience in interactive computing or in monitoring mental health [41]. Convolutional Neural Networks (CNNs) are particularly adept at spatial feature extraction, making them suitable for EEG data's spatial patterns [25]. Long Short-Term Memory networks (LSTMs) and Gated Recurrent Units (GRUs), on the other hand, are effective for temporal data analysis because of their capacity to remember and utilize past information, which is beneficial for the temporal dynamics inherent in EEG signals [40]. These models find utility across a range of applications, from diagnosing neurological disorders to brain-computer interfaces and beyond. For instance, one study presents a CNN-Bi-LSTM model with an attention mechanism specifically designed for EEG-based emotion recognition. The model effectively extracts

features from EEG signals and classifies different emotions by using the original EEG data as input. It employs both CNN and Bi-LSTM networks for feature extraction and fusion, which shows the method's ability to classify EEG emotions [18]. Furthermore, attention mechanisms, which enable the model to focus on the most relevant parts of the data sequence, have been integrated into deep learning models to enhance their interpretative ability and performance in complex tasks such as sentiment analysis and even EEG signal analysis [46]. Such advancements in machine learning techniques provide innovative pathways for analyzing the rich, high-dimensional data that EEGs offer, potentially transforming the landscape of neurological research and practical applications in the field of Brain Computer Interfaces.

1.3 Challenges in EEG Data

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 [10, 34]. The data itself is highly dimensional and contains patterns that are often subtle and varied across individuals. Furthermore, EEG signals are noisy, with potential interference from various sources, including electrical devices, muscle movements, and even the subject's own blink responses. Collecting high-quality, artifact-free EEG data necessitates advanced equipment and controlled settings, which can be prohibitively expensive and logistically complex. Another important issue with the dataset is the non-stationary and subject-dependent nature of EEG signals, meaning that patterns may vary over time even for the same individual, and differ from one person to another [38, 20]. Machine learning models often require retraining with updated data to maintain accuracy, and the need for participant-specific calibrations can make the process tedious and expensive, hindering their practical application outside of laboratory settings[38]. This complicates the task of developing models that are both effective and adaptable across diverse environments and populations.

1.4 Review of Significant Research

The EEGEyeNet dataset, with its extensive collection of EEG and eye tracking (ET) data, emerges as a significant asset in BCI field, enabling in-depth gaze behavior study and laying the groundwork for benchmarking gaze prediction approaches using EEG data. With recordings from 356 individuals, it covers a wide age range and includes over 47 hours of high-density, 128-channel EEG data synchronized with eye-tracking inputs [21]. In exploring the potential of EEG data for eye-tracking, benchmarks established by deep learning models, particularly CNNs, have marked a significant advancement. While tra-

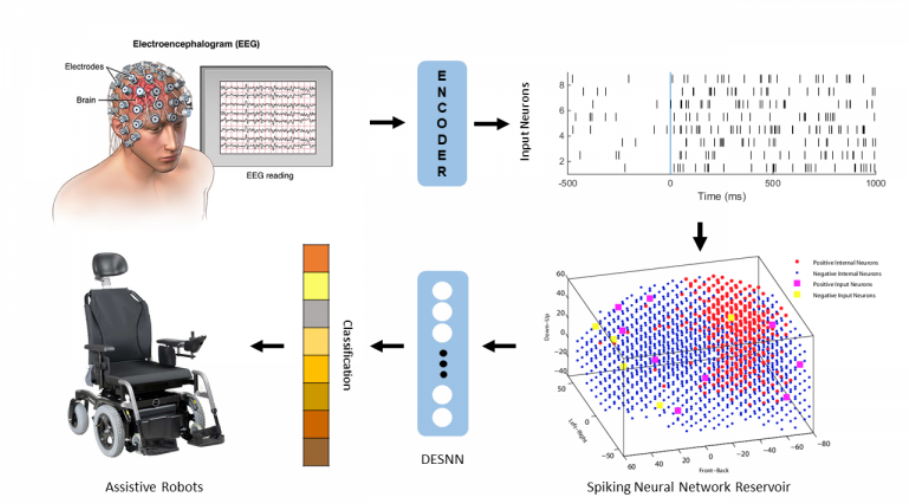


Figure 1: **EEG-Assisted Robotic Control:** The visual representation here outlines the flow from EEG signal capture, through advanced deep learning algorithms, to the direct control of assistive robots. We see a schematic EEG cap setup, highlighting electrode placements crucial for accurate data acquisition. The data flow diagram illustrates the transformation of raw EEG signals into a format suitable for machine learning models, culminating in the precise manipulation of an assistive robotic device [24].

ditional statistical methods (see Table 5) hovered near a baseline, the best performance in gaze prediction was achieved by CNN model (RMSE 70.4mm) which still left a room for advancement in the field using this dataset. Leveraging the EEGEyeNet dataset, the EEGViT, a hybrid vision transformer (ViT) that utilizes two convolutional layers and a Vision Transformer model pre-trained on the ImageNet dataset, has demonstrated its potential in gaze prediction, setting a new benchmark with an RMSE of 55.4 mm [47].

1.5 Contribution of This Study

As a contribution to the field, our study delves into how alterations in EEGViT design with additional depthwise separable convolution, combined with pre-processing techniques, can amplify the accuracy in predicting absolute eye position. We present a new approach employing clustering techniques and depthwise separable convolution on EEG Vision Transformers (ViTs) within an established framework. This method achieves a new standard by surpassing the best-known result with a Root Mean Square Error (RMSE) of 51.6 mm, compared to the previous 55.4 mm, on the EEGEyeNet Dataset for absolute gaze prediction tasks.

1.6 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

2.1 Introduction to Gaze Prediction Technologies

Gaze prediction has evolved over time with its applications spreading across human behavior analysis, advertising, and human-computer interactions. Recent advancements highlight an understanding of the human visual system and eye modeling essential for improving gaze estimation technologies. For instance, the modeling of gaze direction relative to the optical axis of the eye and the complexities of eye movements like saccades and smooth pursuits are crucial for interpreting gaze accurately [13].

Gaze estimation technologies also play role in developing intuitive human-computer interaction systems, particularly in virtual and augmented reality. By understanding the user's gaze behavior, systems can dynamically adapt interfaces or content to improve user engagement and interaction quality. This technology is instrumental in applications where real-time gaze tracking is essential, such as in driving simulators or during complex task executions where gaze direction can enhance performance and safety [13].

In human-robot interaction, gaze prediction plays a crucial role in understanding human intentions and enhancing interaction quality. Studies have shown that gaze-based controls and the ability to predict human intent can greatly improve a robot's responsiveness and effectiveness in collaborative tasks. For instance, gaze prediction can inform a robot about a human's next action or focus, allowing for smoother and more efficient cooperation [4].

Despite significant progress, gaze estimation still faces challenges, particularly in uncontrolled environments where variations in lighting, head poses, and individual differences can affect accuracy. Recent studies have introduced methods that address these challenges, using deep learning to adapt gaze estimation models to handle diverse conditions and user behaviors. These innovations continue to push the boundaries of what's possible with gaze estimation, paving the way for even more personalized and responsive technology [9].

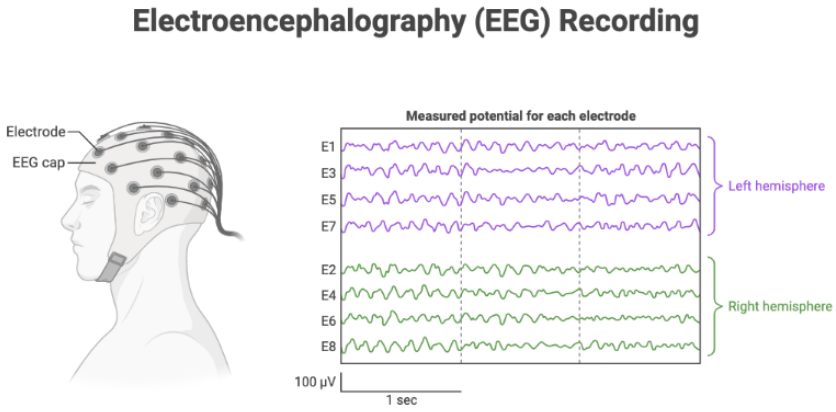


Figure 2: **Electroencephalography (EEG) Waveform Display:** This figure presents a schematic representation of EEG data acquisition, showing an individual wearing an EEG cap with electrode placement. Adjacent to this, a graphical display of EEG waveforms demonstrates the measured potential from each electrode, segmented into traces associated with the left and right hemispheres of the brain [24].

2.2 Advances in Gaze Prediction Using Deep Learning

Recent advancements have leveraged deep learning to enhance gaze estimation techniques, moving beyond traditional models to more sophisticated neural network architectures. These include convolutional neural networks (CNNs) and metric learning approaches, which have improved the accuracy of gaze prediction in both constrained and unconstrained environments. Specifically, models like EM-Gaze utilize eye context correlations and leverage metric learning to enhance gaze estimation accuracy, providing detailed insights into the correlations between eye features and facial cues [51]. This method aligns with trends in using appearance-based methods that directly learn from visual data to predict gaze direction, without needing dedicated devices to capture geometric features of the eye [9].

In addition to leveraging traditional CNN architectures, the field is seeing innovations with architectures designed to be more computationally efficient, making them suitable for mobile and embedded systems. For example, LiteGaze uses neural architecture search to find efficient network designs for gaze estimation, aiming to reduce computational demands while maintaining high performance[15].

Furthermore, recent methodologies have explored the enhancement of gaze estimation through the integration of deep feature extraction and neural network architecture optimizations. These approaches have tackled inherent challenges such as head motion and varying subject appearances that typically impact the performance of gaze estimation

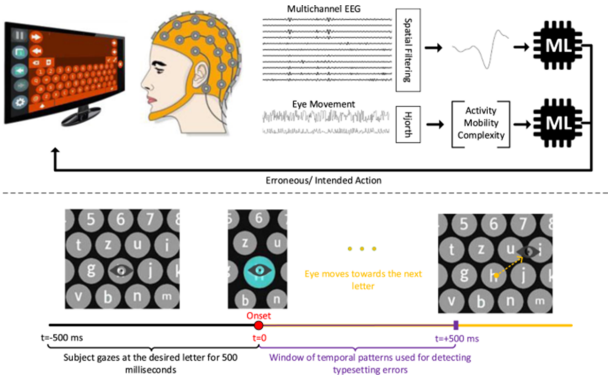


Figure 3: **Hybrid BCI System for Gaze-Based Typing Error Detection:** This figure illustrates the integration of gaze-based typing and EEG monitoring to identify and correct typing errors. It depicts an individual using a gaze-based keyboard displayed on a monitor, wearing an EEG cap for neural activity measurement [19].

systems in real-world settings [9].

2.3 Integration of EEG and Eye-Tracking Datasets

Transitioning to the utilization of EEG datasets for gaze prediction advancements are being driven by the comprehensive analysis of available datasets (see figure 3). For instance, by integrating EEG and eye-tracking technologies, the HCI-Tagging and EYE-EEG datasets has made several contributions in the field of gaze prediction and cognitive research. The HCI-Tagging dataset, specifically designed for studying human-computer interaction, captures how participants respond to visual tags associated with images or videos. This dataset helps researchers understand the cognitive processes involved in interpreting and reacting to visual stimuli, which are crucial for designing more intuitive user interfaces and advertising content [1].

Additionally, the EYE-EEG Integration approach allows for the detailed study of natural viewing behaviors, such as saccades and fixations, alongside brain activity. This toolbox facilitates precise synchronization and analysis of eye movements and EEG, providing tools for improving ocular artifact correction and analyzing eye-movement-related potentials, thereby enhancing the understanding of cognitive and perceptual dynamics during visual exploration [12]. Meanwhile, UnEye, a deep neural network tool for eye movement detection, offers capabilities for analyzing eye movements using machine learning. This tool supports the analysis of horizontal and vertical eye positions and can be integrated with EEG data, offering researchers a state-of-the-art technology to explore gaze and brain activity comprehensively [5].

Moreover, the EEGEyeNet has several contributions in advancing machine learning models that predict gaze direction based on EEG signals. By training deep learning

algorithms on the rich, multimodal data provided by EEGEyeNet, researchers have significantly improved the accuracy of gaze prediction models. This dataset includes synchronized EEG and eye-tracking data from 356 participants across various experimental paradigms. The diversity and volume of data available have enabled researchers to develop robust models that not only predict the direction of gaze accurately but also analyze the visual attention mechanisms in the brain [21].

2.4 Machine Learning Models Applied to the EEGEyeNet Dataset

In the landmark research conducted on the EEGEyeNet dataset, the models deployed covered a broad spectrum of machine learning disciplines. The traditional machine learning models included KNN, SVM with different kernels, Regression. Ensemble learning models brought into play were Random Forest, Gradient Boost, AdaBoost, and XGBoost. Regarding deep learning models, CNN, PyramidCNN, EEGNet, InceptionTime, and Xception were investigated. These models were applied to the tasks of Left-Right discrimination, Angle/Amplitude estimation, and Absolute Position prediction, offering a comparative landscape for performance evaluation. The performance of the machine learning models varied across the tasks. As we can see from the Table 1, for the Left-Right task, where models discerned whether a person was looking right or left, nearly all classification models, including both traditional and ensemble learning models, performed well, with most achieving over 90% accuracy. Remarkably, deep learning models such as CNN, PyramidCNN, EEGNet, InceptionTime, and Xception surpassed this, demonstrating around or over 98% accuracy [21].

However, when it came to regression tasks like Angle/Amplitude and Absolute Position prediction, traditional and ensemble models did not outperform the Naive Baseline by a notable margin. In the realm of absolute gaze prediction, CNNs showcased a performance nearly twice as accurate as the Naive Baseline, with a RMSE of 70.2. This leap in accuracy highlights the potential of deep learning models in tasks requiring fine-grained prediction of eye movements [21].

2.5 Vision Transformer (ViT) Models for Gaze Prediction on the EEGEyeNet Dataset

The performance of CNNs in absolute gaze prediction, while an improvement over traditional methods, has led researchers to reconsider their effectiveness in handling the complex patterns seen in EEG data associated with gaze tracking. Limitations in capturing the full spectrum of spatiotemporal EEG dynamics have been highlighted in studies, suggesting that CNNs alone might not be sufficient for this intricate task [6, 48, 29, 31].

Model	Left-Right Accuracy	Angle RMSE	Amp. RMSE	Abs. Position RMSE
KNN	90.7	1.26	59.3	119.7
GaussianNB	87.7	–	–	–
LinearSVC	92.0	–	–	–
RBF SVC/SVR	89.4	1.88	75.9	123
Linear Regression	–	1.39	64.6	118.3
Ridge Regression	–	1.39	64.2	118.2
Lasso Regression	–	1.38	63.9	118
Elastic Net	–	1.38	63.6	118.1
Random Forest	96.5	1.09	59.8	116.7
Gradient Boost	97.3	1.11	60	117
AdaBoost	96.3	1.43	65	119.4
XGBoost	97.9	1.11	61.3	118
CNN	98.3	0.33	32	70.2
PyramidalCNN	98.5	0.34	30.7	73.6
EEGNet	98.6	0.70	46	81.7
InceptionTime	97.9	0.44	43.6	70.8
Xception	98.8	0.47	32.2	78.7
Naive Baseline	52.3	1.90	74.7	123.3

Table 1: **Comparison of machine learning model performances on EEGEyeNet Dataset** The table compares various machine learning models on eye-tracking data, focusing on left-right classification accuracy and error metrics like Angle RMSE, Amplitude RMSE, and Absolute Position RMSE. This comparative analysis helps to understand each model's precision in predicting eye movement and its reliability in different aspects of gaze estimation [21].

Consequently, there's a growing interest in models that can better handle these complexities. The integration of Transformer architectures, particularly the Vision Transformer (ViT), into EEG data analysis for gaze prediction has its roots in their initial success in handling complex patterns in image data. The Vision Transformer (ViT) model is a pioneering work that applies the mechanisms of transformers, which have been very successful in natural language processing (NLP), to the domain of computer vision. The original paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" by Alexey Dosovitskiy et al. marked a paradigm shift by demonstrating that the same architecture used for sequence-to-sequence predictions in NLP could effectively process images, divided into patches, as sequences. This work sparked significant interest and subsequent research into the application of transformer models to a variety of vision

tasks [11]. The core capability of these architectures to process sequential information and recognize patterns over time is similarly applicable to EEG signals, which are fundamentally temporal sequences with rich spatial information. The self-attention mechanism of Transformers effectively captures temporal dependencies within EEG signals, making it exceptionally suitable for tasks like gaze prediction [47].

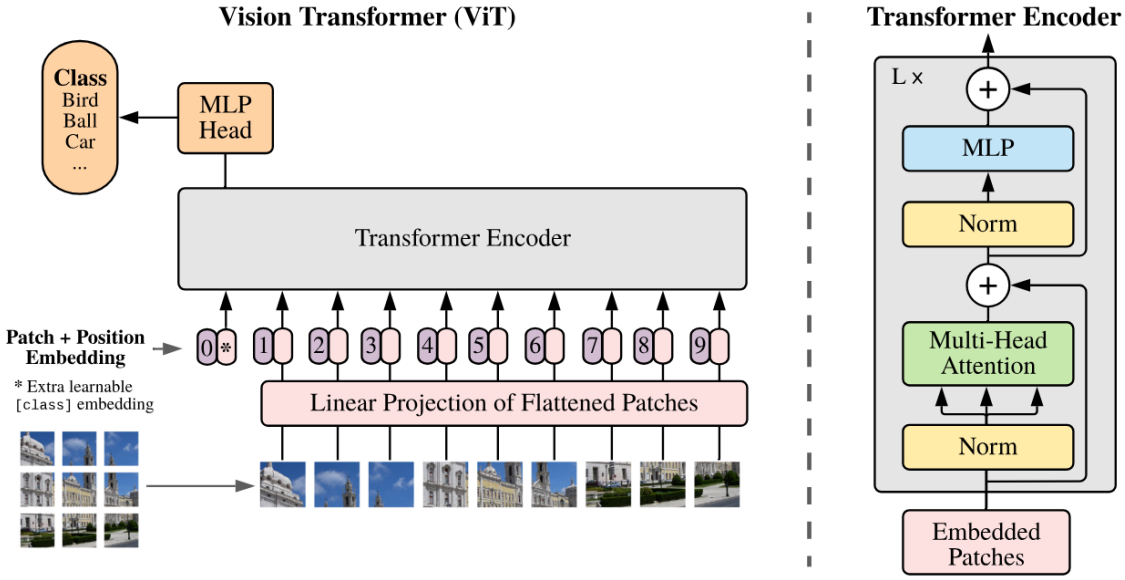


Figure 4: **Vision Transformer (ViT) Architecture:** This illustration details the architecture of a Vision Transformer, a model for image recognition tasks. At the base, flattened image patches are combined with position embeddings and then linearly projected. These projections are processed by a series of Transformer encoders, which apply self-attention mechanisms to capture the relationships between different parts of the image. On the right, the detailed structure of a single Transformer Encoder block is depicted, featuring layers of multi-head self-attention and multilayer perceptron (MLP), with normalization steps in between. The overall architecture demonstrates a shift from conventional convolutional networks to attention-based models in computer vision [11, 39].

In a recent study [47] utilizing the EEGEyeNet dataset for gaze prediction, researchers explored the effectiveness of transformer models, specifically Vision Transformers (ViT). They implemented both a standard ViT-Base model and a pre-trained version of the ViT-Base model, which was initially trained on the ImageNet dataset. The study revealed that the pre-trained model achieved a lower root mean square error (RMSE) of 58.1 mm, compared to the non-pretrained model's 61.5 mm RMSE. This outcome highlights the benefits of transfer learning, where pre-training on image datasets can significantly enhance the model's performance in analyzing EEG data, a testament to the adaptability of transformers across different domains of data analysis. This approach leverages the

intricate patterns learned from image data to improve the precision in predicting gaze directions from EEG signals.

2.6 EEG-ViT Models for Enhanced Gaze Prediction

In the same study [47] that highlighted the benefits of transfer learning for analyzing EEG data with transformers, researchers further innovated by incorporating two convolutional neural network (CNN) layers into the Vision Transformer (ViT) Base model, resulting in a custom model named EEGViT. They tested the EEGViT model in both its standalone form and when pre-trained with the ImageNet dataset. While the standalone EEGViT did not show improvements over the standalone ViT-Base, the pre-trained EEGViT model performed significantly better, achieving a 55.4 mm RMSE, compared to the pre-trained ViT-Base model. This result underscores the effectiveness of integrating CNN layers for feature extraction prior to transformer processing when combined with transfer learning from image datasets. This hybrid approach set a new benchmark for gaze prediction accuracy on the EEGEyeNet dataset, demonstrating the best performance to date with an RMSE of 55.4 mm.

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 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 [14, 8].

2.7 Optimizing Integration of CNNs and Vision Transformers in EEG Analysis

This evolving landscape underscores the promise of combining CNNs and Vision 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.

3 Methods

Our research extends the work presented in [47], focusing on the utilization of preprocessing and depthwise-separable convolution techniques in EEG-based gaze prediction methodologies.

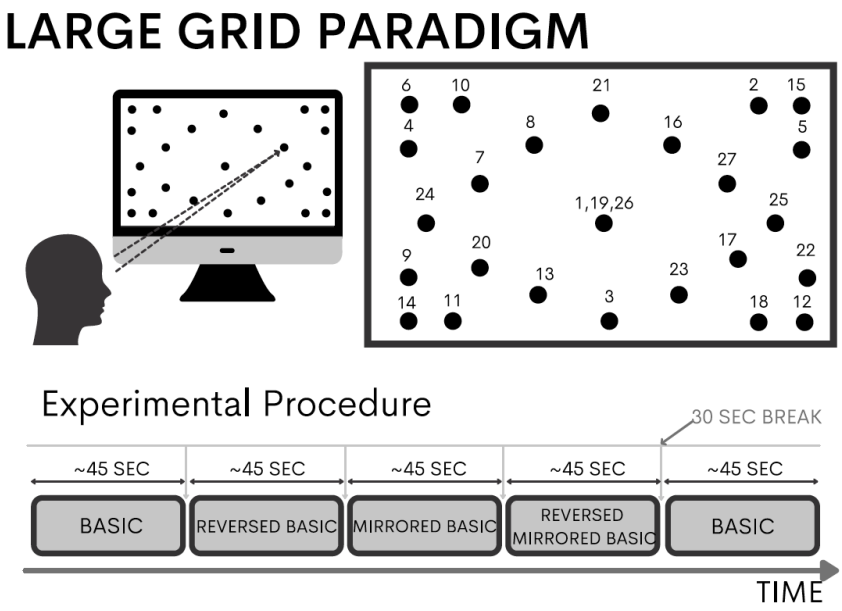


Figure 5: **Large Grid Experimental Setup:** This image illustrates the schematic view of the experimental setup and the stimuli placement on the screen. It gives a visual representation of how participants interacted with the stimuli during the eye-tracking events [21].

3.1 Data Clustering

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) [7] and Li et al. (2021) [28]. In our analysis of the EEGEyeNet dataset, we noted the presence of significant noise. During the original data collection, the EEGEyeNet procedures required participants to focus on specific target positions. Kastrati et al. (2021) [21] 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 50 mm) away from any known target position (Figure 6). 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. This inaccuracy leads to inherent biases in the label positions which cannot be learned during model training. These errors could be the result of 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 [21]. Therefore, if a participant's gaze was in transit towards a target point when captured, the recorded eye position might not accurately represent the

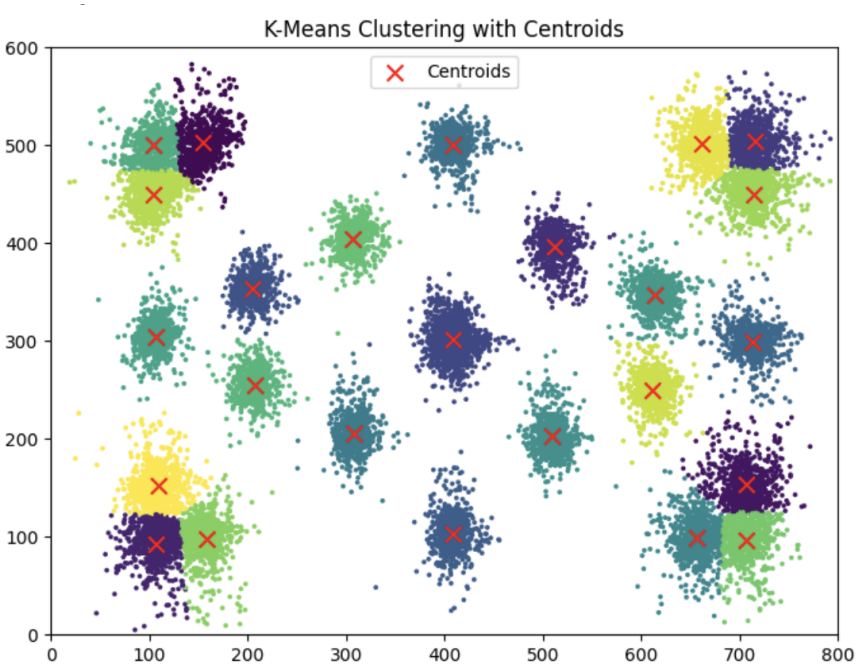


Figure 6: **Positional Discrepancy in Clustering:** Clustering illustrates the discrepancy between labeled positions and actual target positions.

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 labeled position and the actual target position. By updating the true label position with the centroids, as illustrated in Figure 7, we aligned it with the cluster center position, thereby enhancing the accuracy of our dataset.

3.2 Incorporation of Depthwise-Separable Convolutions

Early studies [23] 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 [50] [49]. 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 oper-

ations than traditional CNNs, owing to the reduced computational demands of DS-CNNs [16].

The application of depthwise separable convolution in EEG data analysis 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. [26] and further supported by studies [17], [42], underscore its effectiveness in reducing computational load while maintaining or improving performance score.

Building on these findings, we extend the application of depthwise separable convolution 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 [47] 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. 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, as shown in figure 8, 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 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

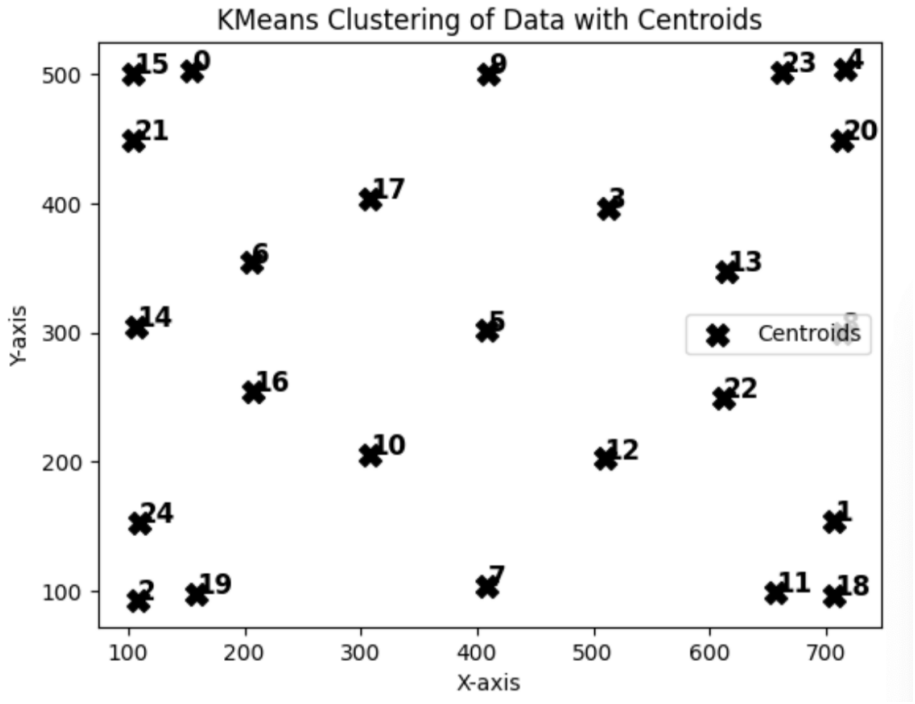


Figure 7: **Centroid Correction for Training Data:** The figure illustrates the use of centroids to refine and correct the labels of training data.

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 accuracy.

3.3 Model Training and Evaluation

We employ a type of early-stopping during training to improve model performance. The SOTA EEG-ViT model was trained on a static number of 15 epochs [47]. However, the authors did not take advantage of the validation set to detect when the model was overfitting to the training data. During training, our algorithm 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.

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). The initial error measurement given in pixels is converted to millimeters using a scale of 2 pixels per millimeter, which aids in clearer understanding.

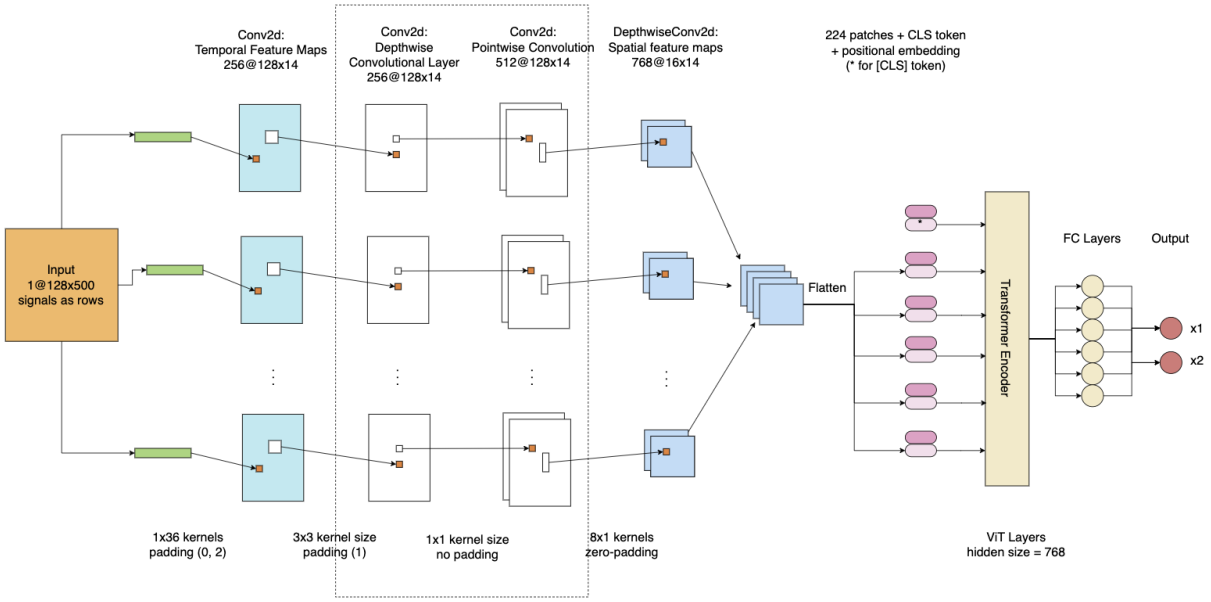


Figure 8: **EEG Vision Transformer with Depthwise Separable Convolution:** A specialized ViT structure tailored for raw EEG signal input. This architecture utilizes a quad-step convolution process to produce patch embeddings. The dotted outline highlights the depthwise separable convolution. After this initial step, positional embeddings are integrated and the combined sequence is subsequently passed through the ViT layers [11, 47].

$$RMSE = \left(\sqrt{\frac{1}{N} \sum_{i=1}^{N} ((x_i^{true} - x_i^{pred})^2 + (y_i^{true} - y_i^{pred})^2)} \right) \times 0.5 \tag{1}$$

The RMSE (in mm) equation quantifies the average magnitude of error between pairs of observed and predicted values across a dataset, specifically for gaze position predictions in the XY plane. It is calculated using the Euclidean distance formula, considering the true (x_i^{true}, y_i^{true}) and predicted (x_i^{pred}, y_i^{pred}) coordinates. The inclusion of a conversion factor of 0.5 adjusts the unit from pixels to millimeters since the original values were in pixels. Smaller values of the Root Mean Square Error (RMSE) are indicative of improved accuracy, as they reflect predictions that are closer to the true measurements.

3.4 Methods Employed

As outlined in Table 2, 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.

3.4.1 Method 1: EEGViT 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, 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.

3.4.2 Method 2: EEGViT 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 exploring the impact of data processing step on the model performance in EEG data.

3.4.3 Method 3 (EEG-DCViT): EEGViT Trained with Clustered and DS-CNNs

This method, EEG Deeper Clustered Vision Transformer (EEG-DCViT), 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 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.

Method	**Description**
Method 1	EEGViT Trained with DS-CNNs
Method 2	EEGViT Trained with Clustered Data
Method 3 (EEG-DCViT)	EEGViT Trained with Clustered and DS-CNNs

Table 2: **Methodology Overview:** Descriptions of the methods used in the study.

3.5 Dataset

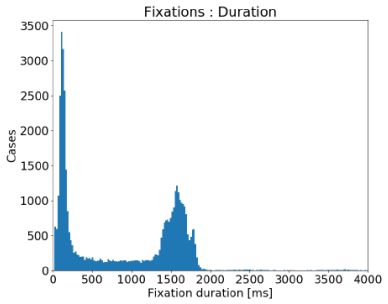
The EEGEyeNet dataset comprises data from 27 participants with a total of 21,464 samples [21]. The primary focus is to ascertain the exact gaze position in terms of XY-coordinates on the screen. Each sample corresponds to a one-second duration where a participant engages in a single fixation on the Large Grid paradigm (Figure 5). The performance is assessed by measuring the Euclidean distance in millimeters between the actual and predicted gaze positions in the XY-plane. In the Large Grid Paradigm as described in figure 5, participants fixated on a series of dots sequentially presented at 25 different screen positions, with the central dot appearing thrice, resulting in 27 trials per block. The dot positions covered all screen corners and the center, with each dot displayed for 1.5 to 1.8 seconds. To record a larger number of trials and reduce predictability, different pseudo-randomized orderings of dot presentations were used across five experimental blocks, repeated six times during the measurement, resulting in 810 stimuli for each participant. The recording setup involved high-density EEG data collected at a sampling rate of 500 Hz using a 128-channel EEG Geodesic Hydrocel system. Participants were seated 68 cm from a 24-inch (609.6 mm) monitor with a resolution of 800 $\times$ 600 pixels, with their head position stabilized using a chin rest [21].

3.5.1 Data Analysis and Visualization

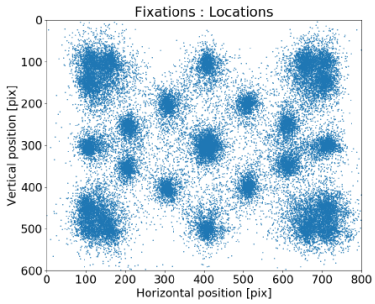
In figure 9, detailed analysis of eye tracking metrics in the Large Grid Paradigm reveals specific patterns of visual engagement and ocular activity.

In figure 9 graph (a), the fixation duration distribution exhibits a substantial portion of fixations falls within the 0 to 500 ms range. However, these instances are excluded from the dataset analysis, adhering to the criterion that each data sample must consist of at least a one-second fixation duration. This ensures that within the provided one-second timeframe, the participant is engaged in a singular fixation event. In addition, the data reveals a significant cluster of fixations around 1500 and 1800 milliseconds, which correlates with the experimental parameters that stipulate each dot's presentation time within the grid, ranging from 1.5 to 1.8 seconds. Consequently, the fixation durations closely mirror the exposure time of each visual stimulus, with the number of fixations tapering sharply beyond the 1.8-second mark. This is consistent with the experimental design, where participants are prompted to redirect their gaze upon the disappearance of one dot and the appearance of another, leading to a negligible number of fixations exceeding the maximum stimulus exposure time.

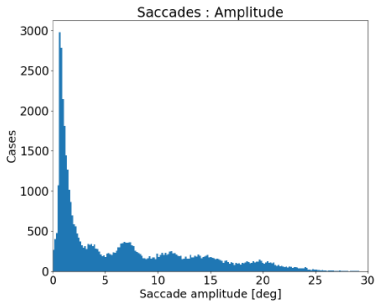
The distribution of fixation positions in graph (b) of Figure 9 reveals a pattern of concentrated clusters that likely correspond to the 25 distinct dots depicted in Figure 6



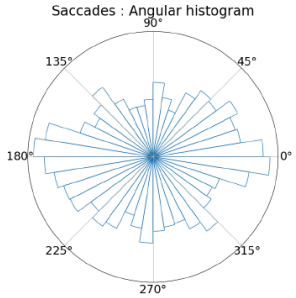
(a) Distribution of the fixation duration



(b) Distribution of the fixation positions



(c) Distribution of the saccade amplitude



(d) Distribution of the saccade angle

Figure 9: **Eye Movement Event Characteristics:** The figure illustrates the distribution patterns for eye movement events in the Large Grid Paradigm. It includes a histogram for fixation duration (a), saccade amplitude (b), a scatter plot for fixation positions (b), a histogram for fixation duration (c) and a polar histogram for saccade angles (d), providing a comprehensive visual representation of the eye tracking data characteristics [21].

on a display grid measuring 800 by 600 pixels. These dense accumulations suggest areas within the grid that repeatedly captured participants' visual attention or represented significant visual cues. Observation of the graph uncovers not only the concentrated clusters around the predefined dot locations but also instances of fixations scattered further away from these central points. These outlying fixations imply that participants occasionally directed their gaze to areas outside the immediate vicinity of the displayed locations, with a variable range of fixation durations. To address the dispersion of fixation positions in our analysis, we have implemented a methodological adjustment by aligning fixations that deviate from the central points back to their nearest centroid. This corrective step ensures that our analysis predominantly reflects the participant's engagement with the intended stimuli. A comprehensive explanation of this methodological approach, including the rationale and implications for our findings, is elaborated upon in the methods section of our report.

In graph (c) of Figure 9, there's a clear predominance of shorter saccade amplitude over longer ones, suggesting that participants made restrained eye movements. The specific placements of the consecutive dots likely influenced the saccadic amplitudes, as the

eye movements required to shift gaze from one dot to the next were relatively small. Expanding upon this observation, we notice a substantial volume of saccadic amplitudes concentrated within the 1-3 degree range. There are also discernible peaks at 4, 7-8, 11-12, 15, and 20 degrees. As we move from less to larger amplitude across the graph, the number of cases progressively decreases. This distribution indicates that during the experimental tasks, the consecutively shown dots were often located within a visual angle of 30 degrees of each other.

The polar histogram in figure 9 graph (d) shows more frequent horizontal saccades at 0° and 180°, relative to vertical saccades, which can be attributed to the horizontal layout of the grid or high occurences of the horizontal switches during during pseudo-randomized switches between 25 different positions in the screen. These observations together paint a picture of how subjects navigate and process visual information in a structured task environment.

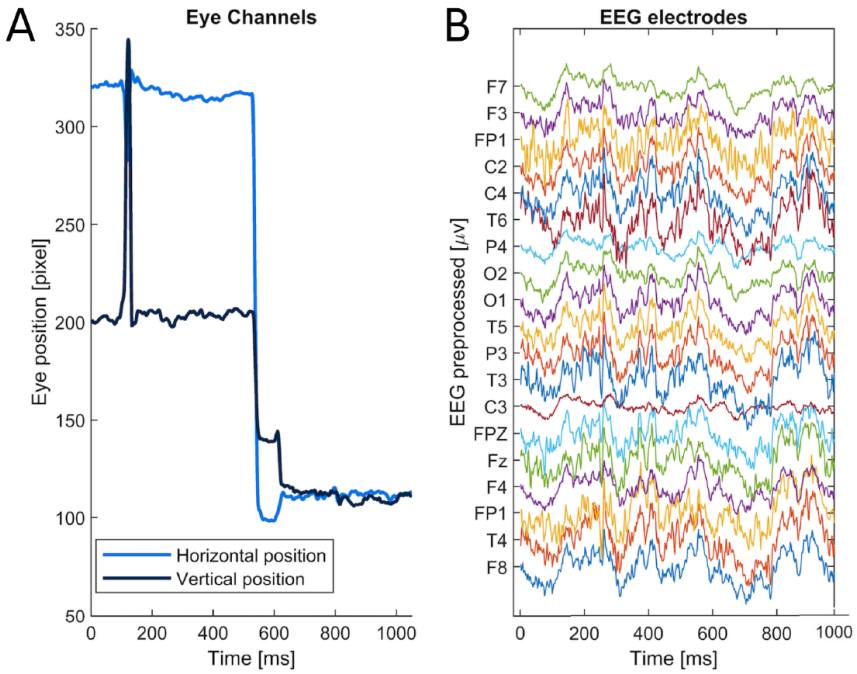


Figure 10: **EEGEyeNet Sample Data Visualization:** This figure displays a sample of the EEGEyeNet dataset. Panel A shows the gaze data trajectory on the XY-plane, representing the eye movement across time. Panel B exhibits a segment of the EEG data with preprocessed waveforms from selected electrodes. These visualizations are from a one-second interval, outlining the structure and patterns within the dataset, with Panel A focusing on eye tracking and Panel B on EEG channel activity [21].

Figure 10 demonstrates the dual aspects of the EEGEyeNet dataset, integrating eye movement metrics with brainwave signals. Panel A, on the left, plots the gaze trajectory over time, where the horizontal axis represents time in milliseconds and the vertical axis

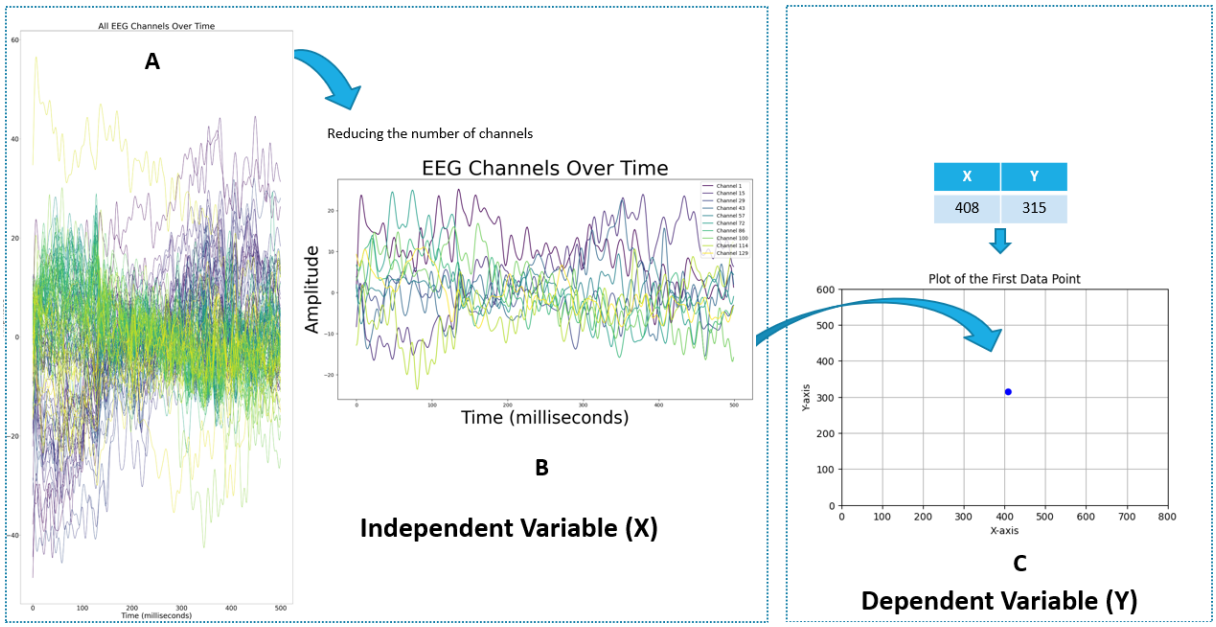


Figure 11: **Dataset Exploration:** This visualization show the relationship between independent variables (EEG data) and dependent variables (eye movement), with each panel designed to provide distinct yet complementary perspectives on the data. The entire training dataset comprises 21,464 records, exemplifying the synchronization and analysis of EEG and eye-tracking data within the EEGEyeNet framework.

depicts eye position in pixels. The separate lines trace the horizontal (x-axis) and vertical (y-axis) positions of the gaze, signifying the dataset's target variable, which reflects the outcome of eye movements. These are the variables that the EEG data may help in predicting. In the context of the "Large Grid" paradigm, the one-second samples are designed to capture a single fixation without any saccadic movements. The shape of the target data, in this case, would be 21464x2, representing the horizontal and vertical positions for each of the 21464 cases. Observing Panel A closely, we notice a saccade occurring within a one-second window, which, according to the dataset's protocol for the "Absolute Position" task, would be considered an artifact. Any such cases where a saccade is present after 500 ms would be treated as outliers and not representative of the intended fixation data, thereby being removed or otherwise accounted for in the preprocessing stage to ensure data integrity for the task.

Panel B in figure 10, on the right, captures the independent variables: the electrical brain activity from multiple scalp electrodes. The data is arranged in a matrix of 500 time points by 128 channels, encapsulating one second of EEG data at a temporal resolution of 2 milliseconds. This equates to 500 distinct time points within the span of 1000 milliseconds, or one second, for each of the 128 channels, illustrating the EEG's high temporal granularity. The dataset records such high-resolution data for each case,

resulting in a comprehensive EEG matrix with the dimensions of 21464x500x128. Such detailed temporal data, when coupled with spatial precision, provides a comprehensive view of the interplay between ocular and cerebral dynamics.

Similarly figure 11 illustrates a sample from the EEGEyeNet dataset. Panel A depicts all EEG channels over time from the first record in the training dataset, showcasing the extensive brain activity captured. Panel B zooms in on a subset of EEG channels, presenting preprocessed waveforms from selected electrodes to emphasize specific signal patterns and facilitate detailed analysis. Panel C displays the corresponding eye gaze position on an 800x600 board, illustrating the eye movement data as the dependent variable in relation to EEG activity.

Data for each experimental condition is processed using specialized modules that segment data into one-second slices at 2ms intervals, yielding a dataset matrix of 500x128; thus, each of the 128 EEG channels provides 500 temporal data points. This structured approach affords a detailed analysis of eye-brain interactions, capturing the high-resolution temporal dynamics and spatial precision of the neural and ocular data within the EEGEyeNet dataset.

Preprocessing Methods	**# Fixations**	**# Saccades**	**# Blinks**	**Total time**
min	68075	68245	11108	7 h 52 min
max	69013	69185	11237	7 h 58 min

Table 3: **Pre-Processing Data Summary:** This table presents the minimum and maximum values of fixations, saccades, blinks, and total observation time across different eye-tracking data preprocessing methods. [21].

3.5.2 Preprocessing Techniques

The EEGEyeNet dataset had already undergone preprocessing in a prior study [21]. EEG data preprocessing is frequently subject to interference from external factors such as temperature and air humidity, as well as electromagnetic disturbances like line noise. These artifacts often interact with participant-specific physiological signals, such as eye movements, blinks, muscle activity, heart rhythms, and sweating, which differ from one participant to another. Typically, these artifacts are more prominent than the brain activity signals of interest, necessitating preprocessing to clean or correct these distortions. In the pre-processing of the dataset the publicly available toolkit [33]. to further preprocess the EEGEyeNet dataset in two different ways: minimal and maximal preprocessing. The minimal preprocessing included identifying and repairing faulty electrodes and applying high-pass and low-pass filters with cutoff frequencies of 40 Hz and 0.5 Hz, respectively.

The maximal preprocessing approach removed a broader range of artifacts, such as muscle, heart, eye movements, line noise, and channel noise using independent component analysis (ICA) coupled with IClabel, a classifier that evaluates the probability of a component being an artifact. Components with a probability higher than 0.8 of being an artifact were excluded. Minimally processed data retains ocular artifacts to assist in assessing gaze positions, whereas maximal preprocessing, typically employed in neuroscientific studies, focuses on retaining only neurophysiological information.

After these preprocessing steps, both EEG and eye-tracking data were synchronized using the "EYE EEG" method. This synchronization allowed for the analysis of EEG data at specific time points corresponding to significant events in the experimental paradigm. The synchronization quality was confirmed by comparing trigger latencies recorded in the EEG and eye-tracker data, with synchronization errors kept below 2 ms. Table 3 displays the frequencies of the three extracted events: fixations, saccades, and blinks, following both minimal (min) and maximal (max) preprocessing stages. In our study, we utilized both minimally and maximally preprocessed EEG data and achieved higher performance on the minimally preprocessed version. This result aligns with the previous study, which reported superior performance of minimally preprocessed data compared to maximally preprocessed data [21].

3.5.3 Experimental Setup and Data Split

To ensure the experiments align with established benchmarks, we've adopted the same participant data split used in the EEGEyeNet dataset. This involves allocating 70% of the participant data to the training set, with the remaining 30% equally divided between validation and testing. This division guarantees each participant's data is exclusive to a single set, eliminating overlap across different phases of model evaluation. A detailed account of the data distribution for the dataset is provided in Table 4.

# Participants				**# Samples**			
Total	**Train**	**Validation**	**Test**	**Total**	**Train**	**Validation**	**Test**
27	19	4	4	21464	14706	3277	3481

Table 4: **Benchmark Data Analysis:** The table enumerates participants and samples distributed across training, validation, and test phases for the benchmark dataset. [21].

4 Results and Analysis

As shown in Table 5, 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 [47]. The results from all three methods, as described in Table 5, 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 Trained with Clustered Data,' we achieved an RMSE of 53.4 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 51.6 ± 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	53.6 ± 0.6
Method 2	53.4 ± 0.8
Method 3 (EEG-DCViT)	**51.6 ± 0.2**

Table 5: **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. [47].

The graph in Figure 12 illustrates the learning trajectory of a model as it is being trained to minimize loss over a series of epochs, where each epoch represents a complete pass through the training dataset. The X-axis denotes the number of epochs the model has undergone, while the Y-axis measures the mean squared error in terms of pixels, with the established conversion ratio being that 1 mm equates to 2 pixels.

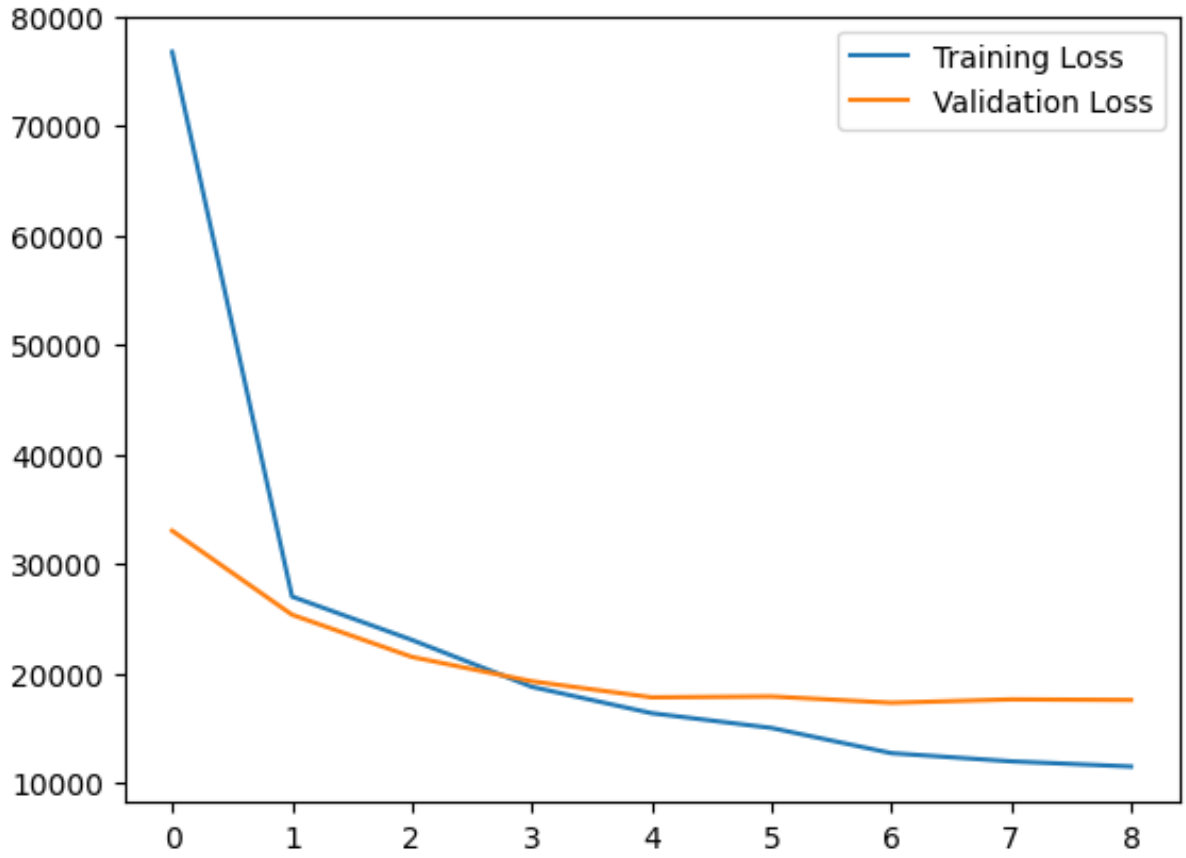


Figure 12: **Training and Validation Loss Over Epochs:** The graph depicts the learning curve of the model with the x-axis representing the number of epochs. Each epoch corresponds to a complete pass through the entire training dataset. The y-axis quantifies the mean squared error in terms of pixels, where the conversion ratio is set such that 1 mm is equivalent to 2 pixels. The blue line illustrates the training loss, indicating the model's performance on the dataset it learns from. In contrast, the orange line reflects the validation loss, which provides insight into the model's generalization capabilities on unseen data.

From the blue line, representing training loss, we observe a rapid decline in the initial epochs, indicating that the model is quickly learning from the training data. As epochs progress, the rate of decrease in training loss slows down, which is expected as the model starts to converge to a solution.

The orange line depicts the validation loss, offering a view of how well the model generalizes to unseen data. Initially, it too drops steeply alongside the training loss, but as training progresses, the validation loss stabilizes. This divergence between training and validation loss is a classic indicator that the model might be beginning to overfit to the training data; it's improving on the data it has seen, but its performance on new, unseen data isn't keeping pace.

We implemented early stopping to prevent overfitting. With a 'patience' parameter set at three iterations, training halts if there's no validation loss improvement after three consecutive epochs. On the chart, early stopping would trigger after epoch 4, where the validation loss stops decreasing. This indicates that the model has achieved its best generalization ability, and further training could cause overfitting. Hence, the weights at the early stopping point are optimal for predictions, striking a balance between learning from the training data and generalizing to new data.

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

4.1 Computational Complexity

Traditionally, adding depth to vision transformers by increasing the number of convolutional layers adds computational complexity. Although our work does not include a comprehensive analysis, EEGViT has 86.0M trainable parameters, while EEG-DCViT has 86.2M trainable parameters. This results in insignificant differences in training time and memory usage. Depthwise separable convolutions are highly regarded in the literature for their efficiency in reducing computational costs while maintaining or even enhancing model performance. This efficiency is primarily due to the separation of the convolution operation into two parts: depthwise convolutions that apply a single filter per input channel, and pointwise convolutions that combine the outputs of the depthwise convolutions across channels. This structure helps the computation by reducing the number of multiplicative operations compared to standard convolutions. For instance, in vision transformers, such as the SepViT and Convolutional Vision Transformer (CvT), depthwise separable convolutions have been adapted to further reduce computational overhead while capturing both local and global dependencies effectively. SepViT integrates depthwise self-attention mechanisms that allow local processing within each attention window, which is analogous to depthwise convolutions in CNNs like MobileNets. This method enhances local feature extraction while reducing complexity [27]. The CvT model further explores this by introducing convolutional projections that replace the standard linear projections in the multi-head self-attention (MHSA) modules of transformers. This adaptation not only preserves the local spatial relationships but also reduces the parameter count and computational expense, making these models more suitable for tasks requiring high efficiency[45].

The clustering technique runs in $\mathcal{O}(ndki)$ where n is the number of points, k is the number of clusters, d is the dimensionality of x, and i is the number of iterations that the algorithm takes to converge. In this case, the number of clusters is 25 and the number

of dimensions is 2. Since these are constant, our algorithm runs in $\mathcal{O}(ni)$. Given only 21,000 data points and the hardware requirements to train EEG-DCViT, the algorithm converges within seconds.

The clustering algorithm used in our study can be represented by the following formula:

$$\text{Clustering} = \sum_{j=1}^{k} \sum_{x \in S_j} \|x - \mu_j\|^2$$

where S_j is the set of data points in cluster j, μ_j is the mean of points in S_j, k is the number of clusters (fixed at 25), and x are the data points with 2-dimensional features. This formula aims to minimize the variance within each cluster, effectively grouping similar data points together.

The computational complexity of our clustering algorithm is represented by:

$$\text{Complexity} = \mathcal{O}(ndki)$$

where n is the number of points, d is the dimensionality of the points, k is the number of clusters, and i is the number of iterations for convergence. Given $k = 25$ and $d = 2$, both constants, the complexity simplifies to:

$$\text{Simplified Complexity} = \mathcal{O}(ni)$$

4.2 Understanding Test Error

One of the 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? We created a new visual in order to help us answer these questions (See Figure 13 and 14). In mentioned figures, the representation of errors using blue and red lines serves a specific purpose. The blue lines denote instances where the model's predictions were within a 55.4 mm distance from the true eye position, indicating a smaller error margin. Conversely, red lines signify predictions that exceeded this 55.4 mm threshold, representing a larger discrepancy between the predicted and actual eye positions. For example, in Figure 14, 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. There are fewer red lines between the "inner" positions and the "outer" positions. This could mean that

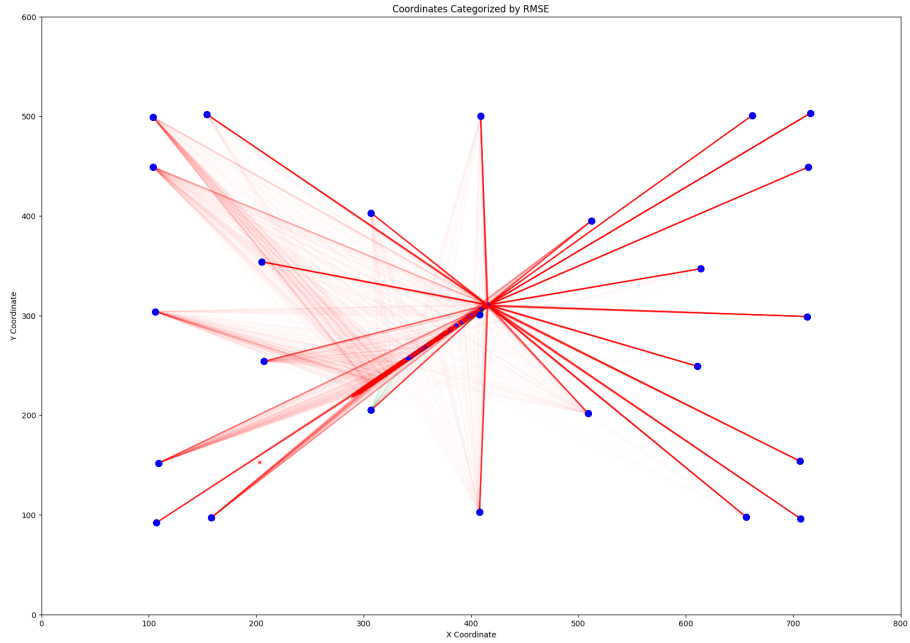


Figure 13: **Visual Test Error Distribution in the first epoch:** The visualisation shows positions within 55.4 mm RMSE (Blue) and positions above 55.4 mm RMSE (Red).

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.

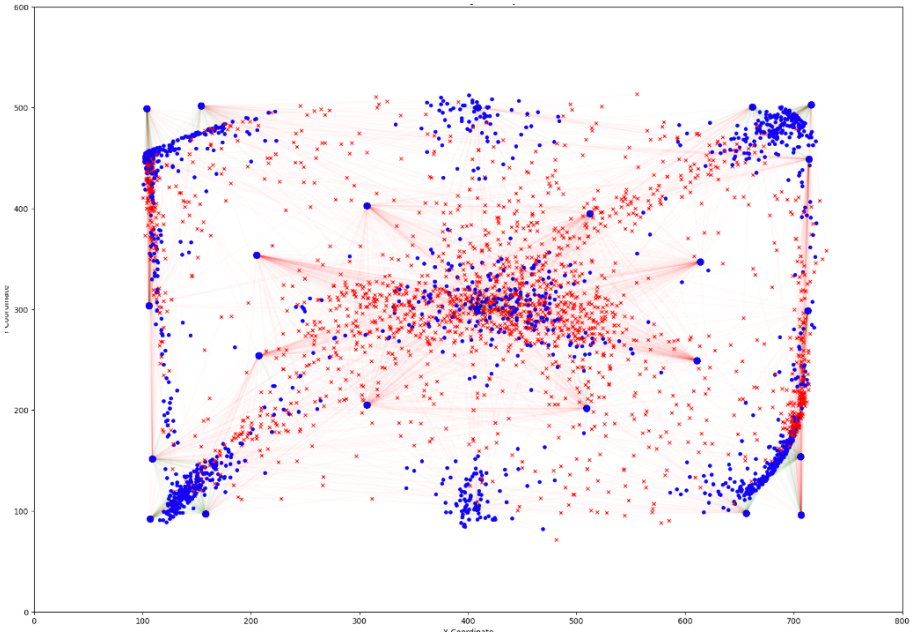


Figure 14: **Visual Test Error Distribution closer to the convergence:** The visualisation shows positions within 55.4 mm RMSE (Blue) and positions above 55.4 mm RMSE (Red).

As we can see from Figure 13, many predictions are clustered in the center in the

first epoch, whereas the true locations are at the corners. This indicates that the initial model has difficulty distinguishing and detecting corner points. Additionally, there is an imbalance in the dataset, as the central location appeared three times more frequently than the other 24 points, which were shown only once. In the subsequent chart, nearing convergence, an increase in blue colors is evident, signifying a reduction in the error distance between the true and predicted locations, with a concentration of blue lines nearer to the corners.

4.3 Understanding EEG-ViT Performance

In order to understand the original EEG-ViT model, we expanded the use of clustered eye positions shown in Figure 7 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 7.

	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

Figure 15: **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.

In the given classification report in figure 15, the original EEGViT model's discriminative ability is quantified across multiple classes, with individual performance metrics presented for each class. Precision, recall, and F1-scores are provided, alongside the 'support' column, 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 confu-

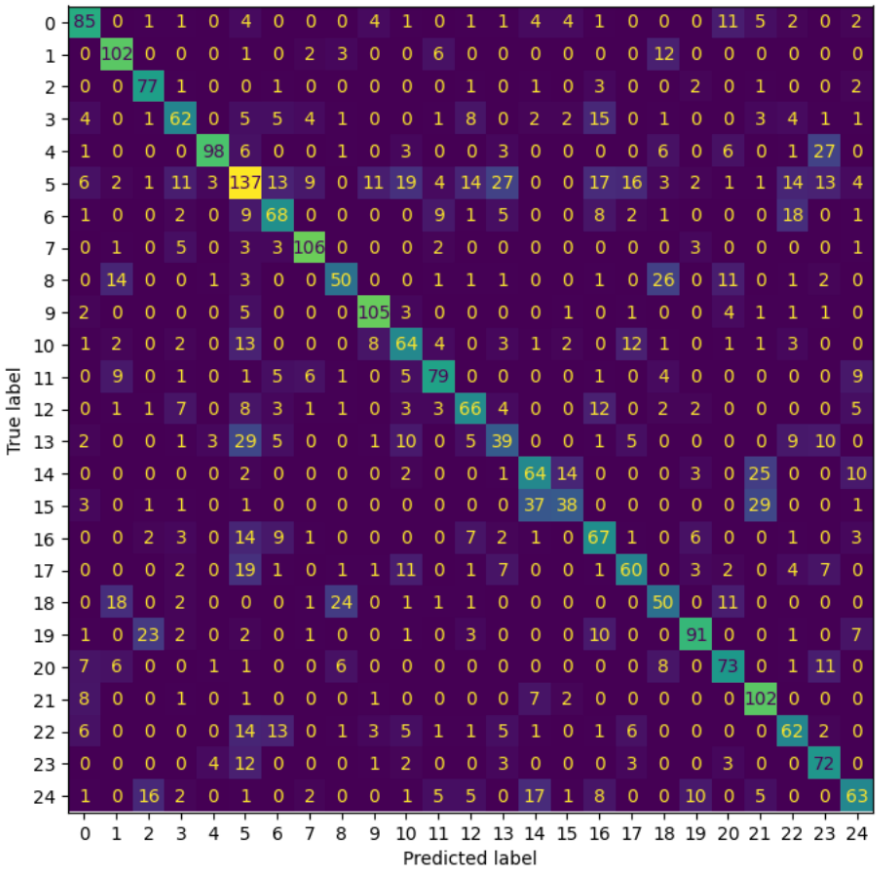


Figure 16: **Confusion matrix across 25 clusters:** On the x-axis, we have the predicted values, which represent the outcomes as forecasted by our model. The y-axis, on the other hand, displays the true labels for each data point.

sion matrix in Figure 16 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.

4.4 Additional Methodological Explorations

In addition to the methodologies mentioned in this study, we have experimented with several other techniques to assess their effectiveness. We share detailed results from these

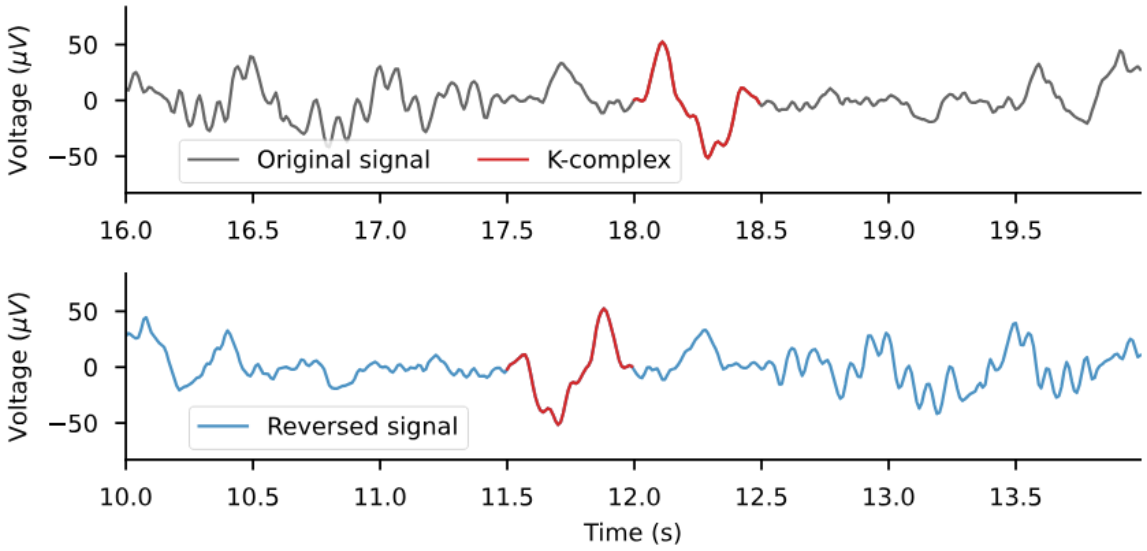


Figure 17: **Time Reversal Augmentation Visualized in EEG Signals:** This illustration compares an original EEG signal with its time-reversed counterpart. The top graph shows the original EEG trace with a prominent K-complex highlighted in red. The bottom graph presents the same EEG signal reversed in time, illustrating that the overall waveform is preserved but mirrored along the time axis. This visualization emphasizes how time reversal can affect asymmetric EEG patterns like the K-complex, which appears inverted after the transformation [36].

supplemental experiments, as we believe such insights could be useful for future researchers to avoid repetition.

4.4.1 Data Augmentation

According to a study by Rommel et al. (2022), a comprehensive analysis of data augmentation techniques including time reversal for EEG signals, revealing their positive impact on classifier training for various tasks like sleep staging and brain-computer interface (BCI) operations [36]. Moreover, Data augmentation has been shown to enhance model robustness by introducing variability in the training data, which is especially beneficial in scenarios where the available dataset is limited or imbalanced which is very relevant for EEG data [36].

This research emphasizes that choosing the correct type of augmentation and its magnitude can help enhance EEG gaze prediction performance. We have implemented time reversal data augmentation (see figure 18) and compared it using RMSE. This approach leveraged a pre-trained EEGVIT model and further expanded the EEGEyeNet dataset by using time reversal data augmentation techniques.

Equation below represents the time reverse function, describing the reversal of an

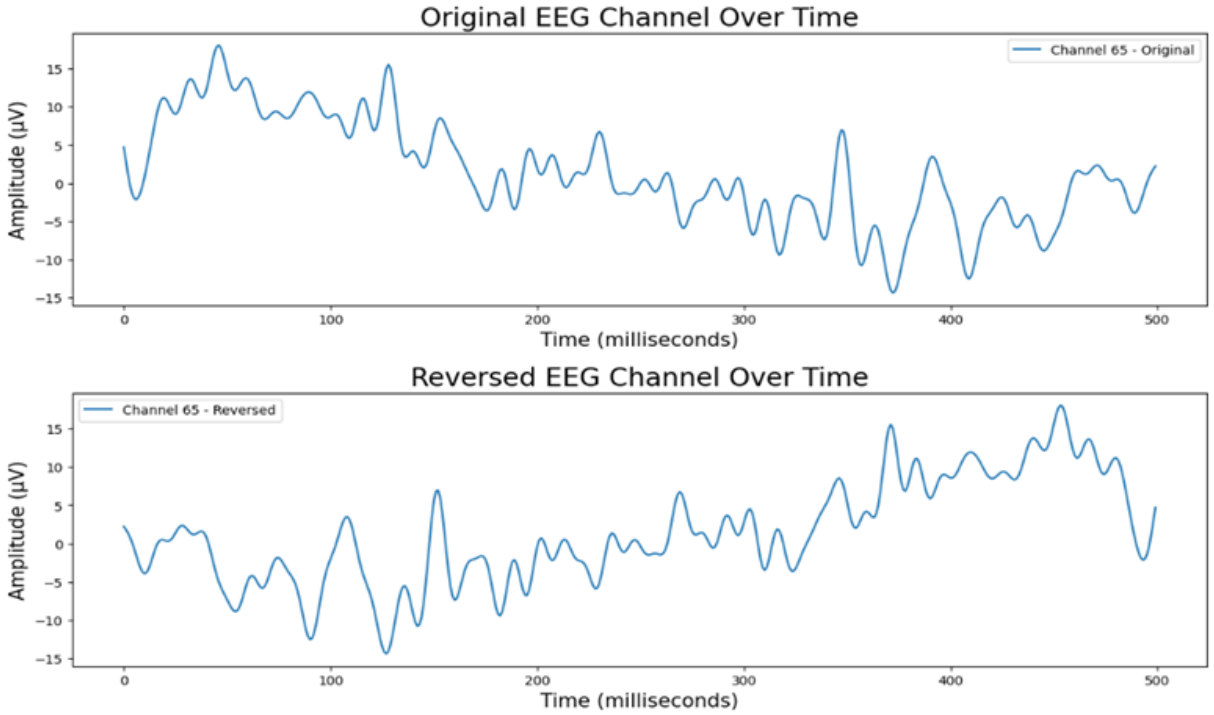


Figure 18: **Time Reversal Augmentation Visual on Sample Filter of EEGEyeNet Dataset:** This chart illustrates the time-reversed augmentation applied to a sample filter of EEGEyeNet Dataset.

interval with probability p_{aug}:

$$T_{Rev}[X](t) = \begin{cases} X(t_{max} - t) & \text{with } p_{aug}, \\ X(t) & \text{with } 1 - p_{aug}, \end{cases} \tag{2}$$

The function $T_{Rev}[X](t)$ defines a time-reversal transformation applied to a signal $X(t)$. It's a conditional function that operates with a certain probability p_{aug}. When the condition is met (with probability p_{aug}), the function outputs the value of the original signal at a mirrored time point $t_{max} - t$, effectively reversing the signal in time. Conversely, with probability $1 - p_{aug}$, the original signal is left unchanged. The parameter t_{max} represents the maximum time point in the signal, which is used to calculate the mirrored time point. This type of augmentation can be useful in signal processing to increase the diversity of data during model training, potentially leading to more robust pattern recognition.

From our analysis, we observed a slight downward trend after implementing data augmentation, through which we were able to double the size of the dataset. However, in terms of performance, we observed an RMSE of 54.95 mm, which was not consistent across the five runs (in some runs it even exceeded the 55.4 mm benchmark). Thus, we concluded that in our case, time reversal augmentation did not significantly aid the process.

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 6: **Vision Transformer Model Specifications:** The table outlines distinct configurations of the Vision Transformer (ViT) models varying by complexity. The "Layers" column indicates the depth of each model, while "Hidden size D" and "MLP size" reflect the dimensionality of the embeddings and the size of the MLP layers, respectively. "Heads" refers to the number of attention heads, and "Params" indicates the total number of parameters. This breakdown facilitates a clear comparison of the models' scalability and potential computational requirements [11].

4.4.2 Model Scaling

A model's performance is intrinsically linked to its scale, where scale encompasses data volume, model depth, and computational capacity. In the realm of neural networks, particularly those designed for complex tasks such as interpreting EEG data, increased depth allows for the extraction of nuanced features that are often pivotal for robust performance. As Wu et al. (2020) elucidate, a deeper model has the propensity to generate more intricate representations of data, leading to potentially more insightful inferences [44].

Within this framework, our investigation into model scaling concentrated on the impact of increasing the size of a pre-trained model. The baseline, represented by Google's "vit-base-patch16-224", consists of a configuration with 12 attention heads and 12 layers, trained on image inputs of 224x224 dimensions. This stands in contrast to Google's "vit-large-patch16-224", which, while trained on identically sized images, boasts an expanded architecture with 16 heads and 24 layers, as reflected in the table 6 . Despite the intuitive notion that larger models correlate with enhanced performance, our findings were contrary when scaling to "ViT-Large". A marginal increase in RMSE by 0.04 mm suggested that for our specific application in EEG data interpretation, the benefits of model scaling are not effective, aligning with recent insights into the limitations of model scaling [22, 43].

4.5 Limitations of This Study

In light of the complexities inherent in EEG data analysis discussed previously, our research encounters a primary limitation stemming from the scarcity and costliness of large, accurate datasets. The complex and controlled conditions required for EEG data acquisition, alongside its high dimensionality and susceptibility to noise, impede the assembly of

vast, diverse datasets essential for robust model training and validation. Particularly, the EEGEyeNet dataset, crafted under stringent task-specific protocols, underscores the limitation in model generalization. Training on such a dataset restricts the model's inferential applicability to external datasets, which may not adhere to identical collection protocols, thus confining the model's efficacy to a narrow experimental context and diminishing its extrapolative value in broader, real-world applications.

5 Summary and Future Work

Our study introduced enhanced pre-processing strategies and depthwise-separable convolutions to the pre-trained EEG-ViT model, which includes two convolutional layers and a ViT base model pre-trained on the ImageNet dataset. This resulted in performance improvements, specifically, the implementation of depthwise-separable convolutions demonstrated effectiveness on EEG datasets. Our model (EEG-DCViT) established a new benchmark with an RMSE of 51.6 mm on the EEGEyeNet dataset's absolute gaze prediction task, an improvement from the previous 55.4 mm. Our findings are a potential milestone for future work in EEG-based interfaces and machine learning. Additionally, we explored various visualization techniques, gaining further insights into EEG datasets. By transforming the gaze prediction task from regression to classification, we uncovered nuances in gaze detection with EEG data. Although using the ViT-Large model and time reversal augmentation did not surpass the state-of-the-art results, these methodologies contribute valuable insights for future research. Our contributions to the field are twofold: advancing the state-of-the-art in EEGEyeNet gaze prediction and providing insights into the EEG dataset and the EEG-ViT model through our comprehensive experiments.

For future studies, the potential of these pre-processing techniques and depthwise-seperable convolutional layer additions, when applied to directly Convolutional Neural Networks (CNNs) without ViT model, 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.

Furthermore, future research could focus on developing more advanced visualization techniques and tools that 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. Another promising avenue for research involves the use of Generative Adversarial Networks (GANs) to generate synthetic EEG datasets, which

could potentially address the challenges of data scarcity and diversity in EEG analysis. Finally, experiments with Recurrent Neural Networks or LSTMs, which are designed for processing sequential data, could be implemented on the EEGEyeNet dataset for the task of absolute gaze prediction.

References

[1] Hci tagging database. HCI Tagging Database (2023), https://mahnob-db.eu/hci-tagging/, accessed: 2023-04-27

[2] Al-Nafjan, A., Hosny, M., Al-Ohali, Y., Al-Wabil, A.: Review and classification of emotion recognition based on eeg brain-computer interface system research: A systematic review. Applied Sciences **7**(12), 1239 (2017). https://doi.org/10.3390/app7121239

[3] Althaeri, 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**, 14681–14722 (aug 2021)

[4] Belardinelli, A.: Gaze-based intention estimation: principles, methodologies, and applications in hri. Submitted to ACM Transactions on Human-Robot Interaction (2023). https://doi.org/10.48550/arXiv.2302.04530, https://arxiv.org/abs/2302.04530, preprint submitted on 9 Feb 2023, available at arXiv:2302.04530

[5] Berenslab: uneye: Deep neural network for eye movement detection. GitHub Repository (2024), https://github.com/berenslab/uneye, accessed: 2024-04-27

[6] 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

[7] 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

[8] Cheng, Y., Lu, F.: Gaze estimation using transformer. arXiv preprint arXiv:2105.14424 (2021), https://ar5iv.labs.arxiv.org/html/2105.14424

[9] Cheng, Y., Wang, H., Bao, Y., Lu, F.: Appearance-based gaze estimation with deep learning: A review and benchmark. Accepted by TPAMI (2021). https://doi.org/10.48550/arXiv.2104.12668, https://arxiv.org/abs/2104.12668

[10] 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

[11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020). https://doi.org/10.48550/arXiv.2010.11929, https://arxiv.org/abs/2010.11929, last revised on 3 Jun 2021 [version v2]

[12] EYE-EEG Project Team: Eye-eeg: Open source matlab toolbox for simultaneous eye tracking & eeg. EYE-EEG Official Website (2024), https://www.eyetracking-eeg.org, accessed: 2024-04-27

[13] Ghosh, S., Dhall, A., Hayat, M., Knibbe, J., Ji, Q.: Automatic gaze analysis: A survey of deep learning based approaches. Preprint submitted to arXiv (2021). https://doi.org/10.48550/arXiv.2108.05479, https://arxiv.org/abs/2108.05479, version 3, last revised 21 Jul 2022

[14] 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

[15] Gu, X., Wu, Y., Miao, J., Chen, Y.: Litegaze: Neural architecture search for efficient gaze estimation (2023). https://doi.org/10.1371/journal.pone.0284814, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0284814, open Access, Published: May 1, 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] Huang, Z., Ma, Y., Wang, R., Li, W., Dai, Y.: A model for eeg-based emotion recognition: Cnn-bi-lstm with attention mechanism. Electronics **12**(14), 3188 (2023). https://doi.org/10.3390/electronics12143188

[19] Kalaganis, F.P., Chatzilari, E., Nikolopoulos, S., Kompatiarsis, I., Laskaris, N.A.: An error-aware gaze-based keyboard by means of a hybrid bci system. Scientific

Reports **8**(1), 13176 (2018). https://doi.org/10.1038/s41598-018-31425-2, https://www.nature.com/articles/s41598-018-31425-2, pMC6123473

[20] Kamrud, A., Borghetti, B., Schubert Kabban, C.: The effects of individual differences, non-stationarity, and the importance of data partitioning decisions for training and testing of eeg cross-participant models. Sensors **21**(9), 3225 (2021). https://doi.org/10.3390/s21093225

[21] 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)

[22] Katsaris, V.R., Tsanousa, A., Georgakopoulou, N., Diplaris, S., Vrochidis, S., Kompatsiaris, I.: Graph theoretical analysis of eeg functional connectivity patterns and fusion with physiological signals for emotion recognition. Sensors **22**(21), 8198 (oct 2022). https://doi.org/10.3390/s22218198

[23] 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

[24] Laboratory, A.R.T.: Spiking neural network based brain-computer-interface. Purdue Engineering, ARTLab (2024), available online: https://engineering.purdue.edu/artlab/?page_id=574

[25] Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: Eegnet: a compact convolutional neural network for eeg-based brain-computer interfaces. Journal of Neural Engineering **15**(5), 056013 (oct 2018). https://doi.org/10.1088/1741-2552/aace8c

[26] 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

[27] Li, W., Wang, X., Xia, X., Wu, J., Li, J., Xiao, X., Zheng, M., Wen, S.: Sepvit: Separable vision transformer (2023)

[28] Li, W., Lu, X., Qian, S., Lu, J., Zhang, X., Jia, J.: On efficient transformer-based image pre-training for low-level vision. arXiv: Computer Vision and Pattern Recognition (Dec 2021), https://doi.org/10.48550/arXiv.2112.10175

[29] Majaranta, P., Bulling, A.: Eye tracking and eye-based human–computer interaction. In: Fairclough, S., Gilleade, K. (eds.) Advances in Physiological Computing, pp. 39–65. Human–Computer Interaction Series, Springer, London (2014). https://doi.org/10.1007/978-1-4471-6392-3_3

[30] Murungi, N.K., Pham, M.V., Dai, X., Qu, X.: Trends in machine learning and electroencephalogram (eeg): A review for undergraduate researchers. arXiv preprint arXiv:2307.02819 (2023)

[31] Okada, G., Masui, K., Tsumura, N.: Advertisement effectiveness estimation based on crowdsourced multimodal affective responses. CVPR Workshop (2023)

[32] Paul, Y.: Various epileptic seizure detection techniques using biomedical signals: a review. Brain Informatics **5**(6) (jul 2018). https://doi.org/10.1007/s40708-018-0085-2, https://braininformatics.springeropen.com/articles/10.1007/s40708-018-0085-2

[33] Pedroni, A., Bahreini, A., Langer, N.: Automagic: Standardized preprocessing of big eeg data. NeuroImage **201**, 460–473 (2019). https://doi.org/10.1016/j.neuroimage.2019.06.046, https://pubmed.ncbi.nlm.nih.gov/31233907/

[34] 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)

[35] 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)

[36] Rommel, C., Paillard, J., Moreau, T., Gramfort, A.: Data augmentation for learning predictive models on eeg: a systematic comparison. Journal of Neural Engineering **19**(6), 066020 (Nov 2022). https://doi.org/10.1088/1741-2552/aca220, http://dx.doi.org/10.1088/1741-2552/aca220

[37] Teplan, M.: Fundamentals of eeg measurement. Measurement Science Review **2**(2) (2002)

[38] The Bitbrain team: How deep learning is changing machine learning ai in eeg data processing. Bitbrain Blog (2020), https://www.bitbrain.com/blog/deep-learning-eeg-data-processing, available online: https://www.bitbrain.com/blog/deep-learning-eeg-data-processing (Accessed on April 24, 2020)

[39] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017), https://arxiv.org/abs/1706.03762

[40] Velarde, G., Brañez, P., Bueno, A., Heredia, R., Lopez-Ledezma, M.: An open source and reproducible implementation of lstm and gru networks for time series forecasting. Engineering Proceedings **18**(1), 30 (2022). https://doi.org/10.3390/engproc2022018030

[41] Wang, T., Huang, X., Xiao, Z., Cai, W., Tai, Y.: Eeg emotion recognition based on differential entropy feature matrix through 2d-cnn-lstm network. EURASIP Journal on Advances in Signal Processing **2024**(49) (apr 2024), https://link.springer.com/article/10.1186/s13634-024-0088-5

[42] 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

[43] Wang, X., Ren, Y., Luo, Z., He, W., Hong, J., Huang, Y.: Deep learning-based eeg emotion recognition: Current trends and future perspectives. Frontiers in Psychology **14** (feb 2023). https://doi.org/10.3389/fpsyg.2023.1126994

[44] Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., Vajda, P.: Visual transformers: Token-based image representation and processing for computer vision (2020)

[45] Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers (2021)

[46] Xu, R., Miao, F., Cong, P., Zhang, F., Xin, Y., Feng, X.: Multiview feature fusion attention convolutional recurrent neural networks for eeg-based

emotion recognition. Journal of Sensors **2023**, Article ID 9281230 (2023). https://doi.org/10.1155/2023/9281230

[47] Yang, R., Modesitt, E.: Vit2eeg: Leveraging hybrid pretrained vision transformers for eeg data (2023)

[48] 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)

[49] 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

[50] Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. arXiv preprint arXiv:1311.2901 (2013), https://arxiv.org/pdf/1311.2901.pdf

[51] Zhou, J., Li, G., Shi, F., Guo, X., Wan, P., Wang, M.: Em-gaze: eye context correlation and metric learning for gaze estimation. Visual Computing for Industry, Biomedicine, and Art **6** (2023). https://doi.org/10.1186/s42492-023-00135-6, https://vciba.springeropen.com/articles/10.1186/s42492-023-00135-6, open Access, Published: 05 May 2023