Benchmarking Deep Learning Models Performance for Raw EEG Classification: Separating Noise from Brain Activity

Fabrice Gagnon, Mathis Rezzouk, Marc-Antoine Tremblay

***Abstract*—This project aims to enhance the understanding of how deep learning models differentiate between genuine brain activity and noise, with a specific focus on binary classification of brain regions associated with aversive audio pain and thermal pain. By establishing a clear ground truth, this study simplifies interpretation and reduces complexities related to model performance. We compare the performance of various deep learning models in classifying raw EEG data collected from noisy environments, employing evaluation methods to analyze their decision-making mechanisms. The expected outcomes will provide insights into how these models discern authentic brain activity from noise artifacts within a binary classification framework. Furthermore, by contrasting the performances of models trained on raw EEG data with those trained on filtered data, we aim to gain critical insights into the decision-making processes that enable the discrimination of relevant neural signals from interference. Ultimately, this research proposes a refined understanding of the behavior of binary classification models in the presence of noise, paving the way for future studies focused on improving noise management strategies in EEG signal analysis. Additionally, we highlight the importance of advancing the comprehension of EEG signal classification models, suggesting that our findings could serve as a reference point for analyzing model interpretability in noisy contexts, applicable across various classification domains.**

I. Introduction

Electroencephalography (EEG) is a powerful tool for measuring electrical activity in the brain, widely used in neuroscience, clinical diagnostics, and brain-computer interface applications. By capturing voltage fluctuations generated by neural activity, EEG provides a non-invasive window into brain function. However, these signals are highly susceptible to noise, stemming from both external sources (e.g., electrical interference) and internal artifacts (e.g., eye movements or muscle activity). This inherent vulnerability to noise poses significant challenges for signal processing and analysis, particularly when leveraging EEG data for machine learning and deep learning applications.

In recent years, the application of deep learning to EEG analysis has gained significant traction due to its potential for automating complex tasks such as seizure detection, emotion recognition, and cognitive state classification. Unlike traditional methods that rely heavily on preprocessing to filter noise, deep learning models offer the promise of directly learning meaningful patterns from raw data. However, this paradigm shift comes with its own set of challenges. The reliance on raw signals increases the risk of models inadvertently learning spurious correlations or artifacts rather than genuine neural patterns. Moreover, the ”black box” nature of deep learning raises concerns about the interpretability and reliability of these models, particularly in critical applications such as healthcare.

To address this issue, several approaches have been developed. Some neuroscience experts choose to manually filter the data using commonly used filters to mitigate unwanted noise from external or internal sources. Other approaches involve training machine learning models without removing noise artifacts, hoping that the algorithms can learn to differentiate noise from relevant brain signals. This method is gaining popularity. Indeed, according to a recent study [1], 60 to 70% of scientific papers do not remove noise artifacts when processing signals with deep learning models.

Moreover, this approach without manual filtering introduces an additional challenge regarding the interpretability of classification algorithms. The inherent abstraction in deep learning models makes it difficult to understand how they manage to differentiate noise from a relevant brain signal. This difficulty raises important questions about the reliability and robustness of the results produced by these models.

Interpretability is a key discipline in artificial intelligence that aims to make the process by which an algorithm transforms input data into a meaningful output more transparent to the user. In the field of EEG signal analysis, this discipline has yet to reach its full potential. To date, most research focuses primarily on improving the performance of deep learning models for EEG signal classification, emphasizing indicators such as accuracy. However, few studies delve deeply into how these models process raw signals and distinguish real brain activity from noise artifacts. As a result, it remains challenging to determine whether the results obtained by these models are based on authentic neural signals or on artifacts produced by external or internal noise.

The interpretability of EEG signal classification models remains an underexplored field and represents a crucial puzzle for the scientific community [2]. It is this lack of documentation and understanding that motivates our team to focus on this subject. Our goal is to explain how EEG classification models manage to distinguish noise from real brain activity during classification tasks. By providing a better understanding of these mechanisms, we hope not only to improve the robustness of the models, but also to enhance their ability to accurately interpret neural signals while reducing reliance on artifacts.

II. Literature review

Recent studies in the field of EEG signal analysis have utilized both raw and preprocessed data for training machine learning models. Some works have focused on training models directly on raw EEG data, bypassing traditional preprocessing steps [3], while others have applied extensive preprocessing techniques to remove noise and artifacts [4]. However, few studies have systematically compared the performance of models trained on raw versus preprocessed data. This gap is particularly relevant in determining whether preprocessing truly enhances model performance, or if models can generalize well even without it.

Furthermore, in the studies that employ deep learning models trained on raw EEG data, there is often limited exploration of the reasons behind the model's classification decisions. As a result, it remains unclear whether these models are learning from meaningful brain signals or merely capitalizing on noise present in the data. This raises important concerns about the interpretability of such models.

To address these concerns, several post-hoc explainability methods have been proposed in the literature. Techniques like saliency maps[2], Layer-wise Relevance Propagation (LRP) [1], and Grad-CAM [1] have been used to better understand the features that deep learning models focus on during classification. These methods aim to provide insights into whether the learned features are indeed related to brain activity or are driven by irrelevant or noisy components[5]. However, the application of these explainability techniques to models trained on raw EEG data is still relatively limited and warrants further investigation.

III. Dataset

For the dataset, the input will be raw EEG signals recorded from 64 channels, captured at 500 Hz. Each sample represents brain activity over time from participants exposed to auditory and thermal stimuli. The data, for each 33 participants, is segmented into 2 sessions 8 minutes each.

The model will predict a classification label corresponding to the task performed by the subject during the recording. Specifically, the goal is to classify whether the EEG signals reflect brain activity related to a specific stimulus (e.g., auditory vs. thermal). The data will be divided into two parts:

- **Training Data**: A subset of participants' sessions will be used to train the model. We plan to use one session per participant for training.
- **Testing Data**: The second session for each participant will be used for testing to evaluate the model's ability to generalize to new data from the same participant.

The dataset contains annotations corresponding to different stimuli (auditory and thermal), which will serve as supervision for the model during training. We will first conduct a Train/Test experiment using raw data, and then compare the model's performance with that obtained from preprocessed data. EEG preprocessing involves downsampling the data to 128 Hz, applying a band-pass filter (1-100 Hz), and using a notch filter to remove 60 Hz noise. Artifacts like eye blinks are removed through ICA, and the data is re-referenced to the average of all channels. The signals are then segmented into relevant time windows for analysis and model training.

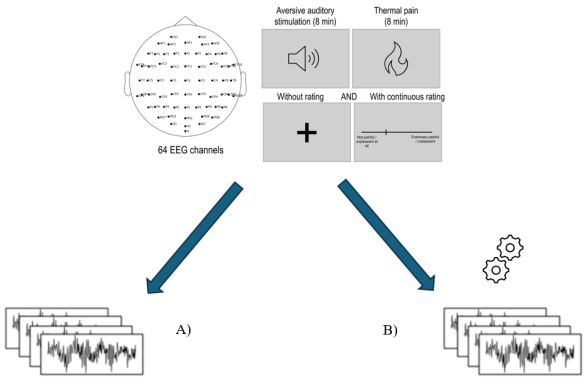


Fig. 1: Diagram of Inputs Acquisition Method **A)** Raw EEG data collected from participants for initial training phase. **B)** Preprocessed EEG data used for the second training phase, including downsampling, filtering, artifact removal, and re-referencing.

IV. Methodology

The training methodology implemented is based on a process applied to various machine learning models to compare their performance and behavior. For each subject, EEG recordings obtained during experimental sessions are collected and subjected to a preprocessing step. This step aims to clean the data by removing artifacts, environmental noise, and interferences while preserving relevant features for analysis. After preprocessing, the data is split into two distinct sets: a training set and a test set, following a standardized distribution.

A first machine learning (ML) model is then trained on this cleaned data. The predictions generated by this model are analyzed using performance metrics such. These indicators quantitatively evaluate the model's ability to correctly classify the EEG data. In parallel, post-hoc visualizations are produced to interpret the model's decisions. Among these visualizations are saliency maps and techniques like Grad-CAM (Gradient-weighted Class Activation Mapping), which are used to identify the temporal and spectral regions of the EEG data that contribute most to the predictions. Additionally, attention graphs are generated in the case of attention-based models to visualize the weights assigned to different parts of the EEG signals.

To assess the model's robustness to perturbations, Gaussian noise is artificially injected into the preprocessed data to create a new noisy dataset[6], [7]. The use of Gaussian noise is justified by its well-documented role in the state of the art as an effective method for simulating random perturbations and evaluating the robustness of machine learning models. Unlike specific artifacts or structured noise, Gaussian noise introduces subtle, normally distributed perturbations into the data [8].

This approach aims to replicate a more realistic environment where minor yet pervasive variations can influence model performance.

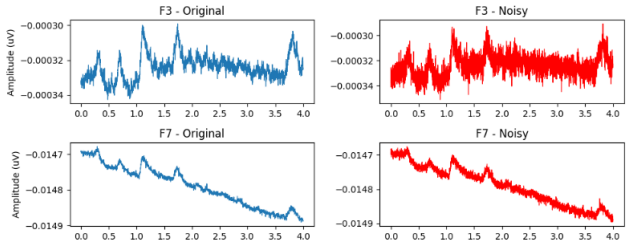


Fig. 2: Visualization of gaussian noise added to EEG signals for channels F3 and F7 of subject 3. The plots on the left show the original signals, while the plots on the right show the corresponding signals with added noise. The top row corresponds to channel F3, and the bottom row corresponds to channel F7.

The model is then reset to avoid any bias from prior training and retrained on the noisy data. The new predictions are subjected to the same analysis process to examine the impact of noise on performance and learning mechanisms. A systematic comparison is then conducted between the results obtained on the clean and noisy datasets, thereby characterizing the model's sensitivity to perturbations and potential performance degradation.

This process is repeated for multiple machine learning models applied to the same clean and noisy datasets. This systematic approach allows for a comparison, under identical conditions, of the models' abilities to extract meaningful information from EEG data and their resilience to noise. For each subject, the models' performances are analyzed individually, and the results are aggregated across all participants. Global metrics, such as average performance and standard deviations, are calculated for each model to provide an overall view of observed trends.

Additionally, the graphs generated by post-hoc methods, including saliency maps, Grad-CAM maps, and attention graphs, are examined to visually explore the effects of noise and the regions of interest exploited by each model. The models' performances are compared across raw, cleaned, and noisy data to identify the most performant and robust model. This integrative approach provides a detailed and comparative evaluation of the models' learning capabilities, their sensitivity to perturbations, and their robustness under varying conditions. It also sheds light on the underlying mechanisms of learning and decision-making in the studied machine learning models.

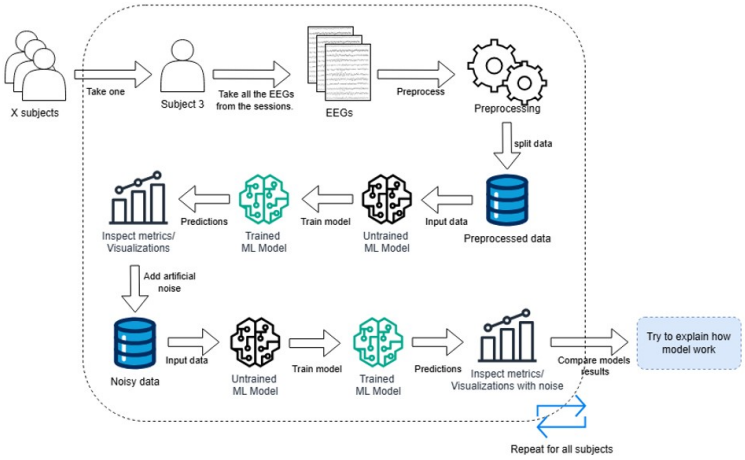


Fig. 3: Diagram of the methodology used to compare and explain models.

V. Results

A. Models performance on subjet 3

The results highlight the impact of data preprocessing and noise on model performance for Subject 3. On noisy data, the EEGDeformer outperformed other models with a significantly higher accuracy (0.9772), showcasing its robustness to noise and ability to capture meaningful global patterns in the signal. The SVM and EEGNet also performed well on noisy data, with accuracies of 0.9244 and 0.9042, respectively, but their reliance on localized or predefined features likely limited their ability to handle noise effectively. On processed data, the performance gap between the models narrowed, with EEGDeformer achieving an accuracy of 0.9607, slightly outperforming SVM (0.9223) and EEGNet (0.8962). This suggests that preprocessing effectively mitigates the challenges posed by noise for all models, reducing the comparative advantage of EEGDeformer. The GGN model achieved an accuracy of 86.2745 on processed data and 60.7843 on noisy data for subject 3. While GGN performed moderately well on processed data, its performance on noisy data was notably lower compared to EEGDeformer, SVM, and EEGNet. This indicates that GGN has limited robustness to noise and may struggle to maintain high accuracy when the data quality deteriorates. The disparity in performance between processed and noisy data suggests that GGN could benefit from enhanced noise mitigation strategies or more sophisticated feature extraction methods to improve its resilience in noisy environments.

Sujet 3	SVM	EEGNet	EEGDeformer	GGN
Processed	0.9223	0.8962	0.9607	86.2745
Noisy	0.9244	0.9042	0.9772	60.7843

TABLE I: Model performance on noisy and processed data for Subject 3.

Data Type	SVM	EEGNet	EEGDeformer	GGN
Preprocessed	0.7873	0.7542	0.8969	80.9826
Noisy	0.8547	0.7053	0.8674	64.8971

TABLE II: Validation accuracy of models on preprocessed and noisy data.

B. Models Average Performance Across All Subjects

VI. Discussion

A. Interpretation of Model Learning Mechanisms

The models benchmarked in this study utilize distinct approaches to learn from EEG signals. SVM relies on hyperplane separation in feature space, making it interpretable but limited in capturing nonlinear or complex patterns in raw EEG data. EEGNet, as a convolutional neural network (CNN), extracts temporal and spatial features from EEG signals through successive convolutional and pooling layers, highlighting important patterns across channels and time windows. EEGDeformer, leveraging transformer-based architectures, excels in capturing global dependencies within the signals, offering flexibility in analyzing long temporal sequences. The superior performance of EEGDeformer over SVM and EEGNet lies in its ability to capture long-range dependencies in EEG data through self-attention mechanisms. Unlike SVM, which relies on static feature separations, and EEGNet, which focuses on local patterns via convolutions, EEGDeformer dynamically weighs the importance of different features across time and channels. This allows it to model global dependencies, making it more robust to noise and better suited to handle the distributed and intricate nature of brain activity. Its adaptability explains its higher accuracy and generalization, even in noisy conditions, compared to the other models.The GGN model, on the other hand, utilizes both spatial and temporal features to handle the tasks it is designed for. A notable aspect of this model is its ability to account for the connectivity of electrodes during recording, enabling it to recognize connectivity patterns rather than solely relying on temporal feature analysis. Figure 4 presents a topographic representation of the electrodes used to classify the recordings across epochs for Subject 3. Although the representation does not change significantly in this case, it is noticeable that the FT9 and Fp1 electrodes are the most influential in the decision-making process and that the strongest connection is between Fp1 and Fz as seen in Figure 6. While the results were not as strong as those of the other models, the GGN model performed well in tasks that require a focus on electrode connectivity and involve significant changes in patterns between epochs and subjects.

B. Implications for Preprocessed Data vs Noisy Data

The comparison between preprocessed and noisy data underscores several key implications for the design and deployment of EEG classification models. Preprocessed data, with artifacts and noise removed, provides a cleaner input that reduces the burden on models to learn to distinguish relevant neural signals from interference. This typically leads to more consistent performance across models, as demonstrated by the relatively narrower performance gaps in preprocessed conditions (Table I).

However, the reliance on preprocessed data also introduces dependencies on the quality and type of preprocessing techniques. Manual or algorithmic removal of noise artifacts, such as ICA or band-pass filtering, may inadvertently eliminate subtle but relevant neural features, potentially limiting the model's ability to generalize to new datasets or experimental paradigms. This limitation highlights the trade-off between ensuring data quality and preserving neural signal diversity.

The results of the GGN model underscore its reliance on preprocessed data. While the model performed moderately well on clean data, its performance dropped significantly when tested on data with Gaussian noise. This can be attributed to the fact that, although the noise altered the temporal features, the spatial features remained largely similar to those of the noise-free data, as illustrated in Figure 5. When the model concatenated these unchanged spatial features with the altered temporal features, the attention in the frequency domain became diluted. This reduced focus on frequency-specific details caused the model to miss critical patterns, ultimately compromising its effectiveness. This diluted attention is further illustrated in Figure 7, where fewer regions of the noisy recording stand out during the learning process compared to the noise-free recording.

In contrast, training on noisy data forces models to learn more robust features, as evidenced by the superior results of EEGDeformer in noisy conditions. Models trained on noisy datasets tend to develop enhanced resilience to perturbations, which is critical for real-world applications where EEG signals are rarely pristine. The ability of EEGDeformer to maintain high accuracy despite noise suggests that transformer-based architectures are well-suited for scenarios with minimal preprocessing, leveraging their attention mechanisms to identify and focus on relevant signal components while disregarding noise.

These findings suggest that preprocessing may not always be necessary for achieving high classification performance, particularly when using advanced architectures capable of robust feature extraction. For applications prioritizing deployment speed and scalability, models trained on raw or minimally processed EEG data could offer practical advantages, reducing preprocessing overhead while maintaining competitive performance. Nonetheless, preprocessing remains valuable for traditional machine learning models, such as SVM, which benefit from clean input features for optimal performance.

While preprocessing can enhance performance consistency, it may also restrict the adaptability of models to varied and noisy environments, a key consideration for real-world EEG applications.

VII. Limitations and futur work

While the current study demonstrates promising results, several limitations and future directions warrant consideration. First, the focus on Within-Session evaluation restricts the generalizability of the models; a Cross-Session study would better

assess their robustness to inter-session variability, which is crucial for real-world applications. Second, exploring the performance of the models on additional EEG datasets with diverse experimental paradigms and population demographics would help validate their adaptability. Third, while this work utilized Gaussian noise to simulate real-world conditions, incorporating other noise types such as motion artifacts or electrode drifts could provide a more comprehensive understanding of model robustness. Lastly, integrating self-supervised learning approaches could further enhance performance by enabling the models to learn generalizable features from unlabeled EEG data, reducing reliance on extensive labeled datasets and improving their scalability to new tasks and domains.

REFERENCES

[1] A. Sujatha Ravindran and J. Contreras-Vidal, "An empirical comparison of deep learning explainability approaches for eeg using simulated ground truth," *Scientific Reports*, vol. 13, p. 17709, 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-43871-8

[2] J. Cui, "Towards best practice of interpreting deep learning models for eeg-based brain computer interfaces," *Front Computer Neuroscience*, 2023.

[3] A. Craik, "Deep learning for electroencephalogram (eeg) classification tasks: a review," *Journal of Neural Engineering*, 2019. [Online]. Available: https://doi.org/10.1088/1741-2552/ab0ab5

[4] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human Brain Mapping*, aug 2017. [Online]. Available: http://dx.doi.org/10.1002/hbm.23730

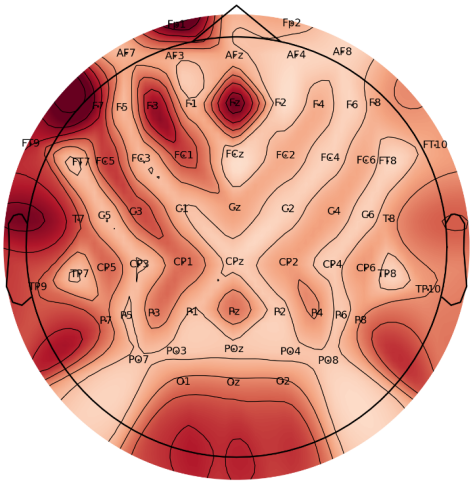
[5] Y. Ding, "Eeg-deformer: A dense convolutional transformer for brain-computer interfaces," *IEEE*, 2024.

[6] J. Park, H. Kim, and S. Yoo, "Orthogonal transform-driven data augmentation for limited gaussian data," in *Proceedings of the IEEE International Conference on Communications*, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10668869

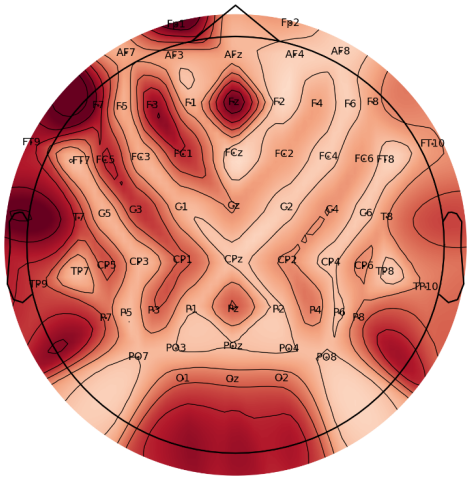
[7] H. Zhang, M. Zhao, C. Wei, D. Mantini, Z. Li, and Q. Liu, "Eegdenoisenet: a benchmark dataset for deep learning solutions of eeg denoising," *Journal of Neural Engineering*, vol. 18, no. 5, p. 056057, 2021. [Online]. Available: https://doi.org/10.1088/1741-2552/ac2bf8

[8] Y. Zhang, W. Ma, M. Li, Y. Zhu, and D. Tang, "Patchmask: A data augmentation strategy with gaussian noise in hyperspectral image denoising," *Remote Sensing*, vol. 14, no. 24, p. 6308, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/24/6308

ANNEX

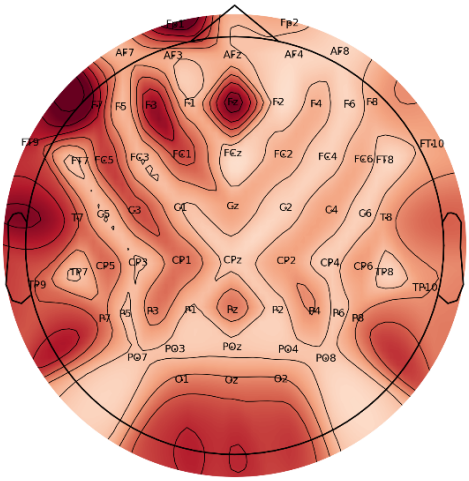


(a) Topographic map at epoch 1

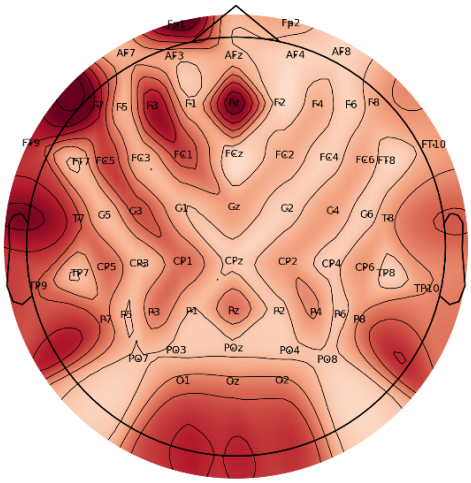


(b) Topographic map at epoch 50

Fig. 4: Topographic map of subject 3 without noise.



(a) Topographic map at epoch 1



(b) Topographic map at epoch 50

Fig. 5: Topographic map of subject 3 with noise.

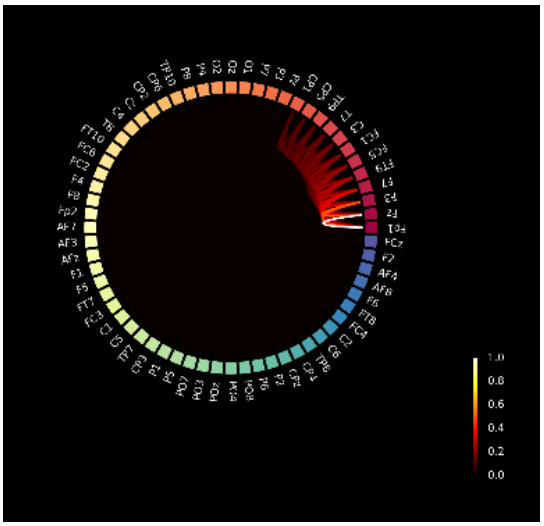
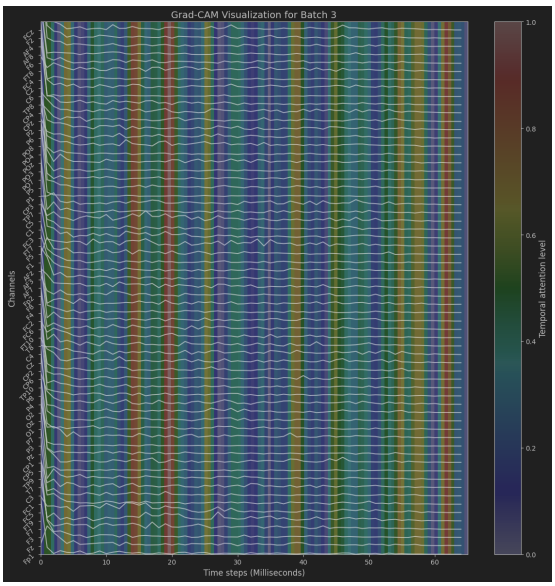
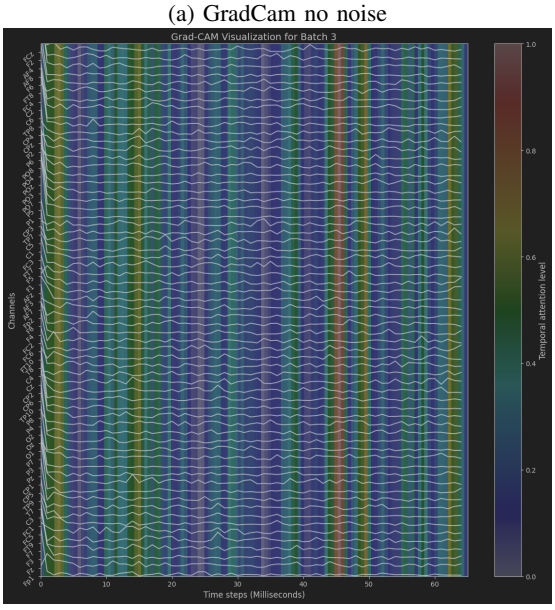


Fig. 6: Connectivity of electrodes for subject 3 without noise at epoch 50.



(a) GradCam no noise



(b) GradCam noise

Fig. 7: GradCam for subject 3.