Performance Evaluation of Hybrid CNN-Transformer Architectures for EEG-Based Data

Archana Savadkar[1], Dr.Yogini Borole[2], Archana Kadam[3]

[1]PhD Scholar, School of Science & Technology, Glocal University, Saharanpur, U.P., India
[2]Professor, School of Science & Technology, Glocal University, Saharanpur, U.P., India
[3]Assistant Professor, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India

Abstract

Electroencephalography (EEG) is widely used in analysing the emotions of humans. EEGs are also widely used in brain-computer interface (BCI), clinical diagnostic, and cognitive-monitoring applications etc. There are various deep learning approaches for decoding EEG signals.CNNs focus on learning local spatial and frequency-domain characteristics of EEG signals, whereas Transformer architectures capture broader temporal dependencies using self-attention mechanisms. This paper reviews recent literature that evaluates and compares the performance of CNN-based, Transformer-based, and hybrid CNN-Transformer models across EEG applications including motor imagery classification, emotion recognition, sleep-stage scoring, seizure detection, clinical abnormality screening, and inner-speech decoding. This study evaluated hybrid CNN-Transformer architectures—specifically the cnn_temporal_transformer and dual_spatial_temporal_transformer models—for EEG-based emotion recognition across the DEAP and SEED datasets, spanning arousal, valence, and multi-class emotion classification tasks. The results demonstrate that while hybrid architectures combining local spatial-temporal feature extraction with global attention mechanisms show promise, their performance remains inconsistent across tasks and datasets, with validation accuracies ranging from 37.33% to 66.67% .

Keywords: EEG, Electroencephalography, Convolutional Neural Network, CNN, Transformer, Self-Attention, Deep Learning, Brain-Computer Interface, Performance Evaluation

1. Introduction

Electroencephalography (EEG) records the electrical activity of the brain through electrodes placed on the scalp and remains one of the most accessible and widely used modalities for studying brain function. Its applications span brain-computer interfaces (BCIs), sleep medicine, epilepsy diagnosis, emotion recognition, and cognitive-state monitoring. However, EEG signals are non-stationary, low in signal-to-noise ratio, and highly variable across subjects and recording sessions, which makes automated decoding a persistently difficult problem.

Traditional EEG analysis relied on handcrafted features such as band power, common spatial patterns, and wavelet coefficients, combined with classical machine learning classifiers. Over the past decade, deep learning has largely replaced this pipeline. Convolutional Neural Networks (CNNs), including architectures such as EEGNet and ShallowConvNet, learn spatial and spectral filters directly from raw or minimally processed EEG signals and have become a de facto baseline in the field. More recently, Transformer architectures, originally developed for natural language processing, have been adapted to EEG decoding because their self-attention mechanism can model long-range temporal and cross-channel dependencies that convolutional receptive fields struggle to capture.

A growing body of research has directly compared CNN-based, Transformer-based, and hybrid CNN-Transformer models on EEG classification tasks. The results, however, are not uniform: some studies report that Transformers alone underperform CNNs, particularly on short EEG windows or small datasets, while others show that combining convolutional and attention mechanisms yields the best overall performance. Given this mixed and rapidly evolving evidence, a structured review of recent comparative studies is valuable for researchers selecting an architecture for a given EEG task.

This paper reviews recent literature — the majority published in 2024, 2025, that performs a performance evaluation of CNN and Transformer methods on EEG data. The remainder of this paper is organized as follows. Section II presents the literature review, organized by application domain. Section III presents a model of CNN and transformer applied on EEG. Section IV on result, discussion and summary table. Section V concludes the paper and outlines open research directions.

2. Literature Review

A. Motor Imagery and Visual Imagery Decoding

Several studies compare CNN and Transformer architectures on motor and visual imagery EEG tasks. Venkannagari et al. [3] evaluated an adaptive CNN-Transformer against a pure Inception-CNN across datasets with different channel densities, finding that the hybrid CNN-Transformer performed best on a low-channel-density dataset (97.14% accuracy) while the CNN-only Inception architecture performed best on a high-channel-density dataset (98.52% accuracy), suggesting that the optimal architecture is sensitive to the spatial resolution of the recording.

A related study [1] proposed a Transformer-integrated CNN architecture for decoding complex visual imagery and benchmarked it against EEGNet, ConvNet, FBCNet, and TSformer. The proposed hybrid model achieved the highest average classification performance, ahead of the pure Transformer baseline (TSformer), which in turn outperformed the CNN-only baselines, indicating a general ordering of hybrid > Transformer > CNN-only in this study.

In contrast, Bhattacharyya et al. [6] found the opposite ranking for visual brain decoding: a CNN-BiLSTM approach reached 71% accuracy, clearly outperforming a CNN-Transformer approach at 59% accuracy. The authors attributed the weaker Transformer performance to the short duration of the EEG windows used, since self-attention mechanisms typically require longer sequences to learn meaningful non-local dependencies. Similarly, the SSTAF study [5] reported that a spatial-spectral-temporal attention Transformer surpassed CNN baselines such as EEGNet and ShallowConvNet on motor imagery classification, while a local-global convolutional transformer approach [10] combined a local Transformer encoder with a densely connected CNN specifically to compensate for the CNN's limited temporal receptive field.

B. Sleep-Stage and Disorder Classification

Sleep staging from EEG has attracted numerous CNN-Transformer comparisons because of the long, sequential nature of polysomnography (PSG) recordings. A 2026 study [13] directly compared a plain CNN, a CNN-BiLSTM, and a CNN-Transformer-encoder on the CAP Sleep Database and found the CNN-Transformer variant achieved the best results, with 95.2% accuracy for sleep staging and 99.3%

for disorder classification, outperforming both simpler alternatives.

MultiScaleSleepNet [14] combined multiscale convolutional branches with a BiLSTM and Transformer-based attention layer for single-channel EEG, reporting accuracies of 88.6%, 85.6%, and 84.6% on the Sleep-EDF, Sleep-EDF Expanded, and SHHS datasets respectively, with ablation experiments confirming that the attention component consistently improved performance over CNN-only variants. FlexSleepTransformer [16] took a purely Transformer-based approach but addressed a practical CNN limitation — fixed channel-count inputs — by supporting training across datasets with differing numbers of PSG channels. A related 2025 study [15] proposed a Transformer-based sleep-staging model and examined how single-channel selection causally affects staging reliability compared to earlier CNN-based models such as DeepSleepNet.

C. Seizure Detection and Clinical Abnormality Screening

For epileptic seizure detection, Holguin-Garcia et al. [11] compared a CNN-CapsuleNet architecture, a CNN-Transformer encoder, and traditional machine learning classifiers. The CNN-CapsuleNet model achieved the highest binary classification accuracy (99.92%), while the CNN-Transformer encoder achieved 87.30% accuracy on the more difficult multiclass task, illustrating that CNN-based architectures can still outperform Transformer variants on certain clinical classification formulations.

For broader abnormality screening, a 2025 study [4] compared a CNN-LSTM model, a Transformer-based model, and a finetuned foundation model (BioSerenity-E1) for classifying whole EEG recordings as normal or abnormal across three datasets. All three architectures achieved at least 86% balanced accuracy on the largest dataset, with the finetuned foundation model performing best at 89.19%, followed closely by the CNN-LSTM and Transformer baselines, indicating that architecture choice mattered less than pretraining and finetuning strategy in this setting.

A separate line of work applies CNN-Transformer fusion to psychiatric relapse detection. Fahmy et al. [9] proposed a CNN-Transformer fusion model that combines EEG with clinical and sentiment-derived features for detecting schizophrenia relapse, reporting near-perfect precision and recall and 0.95 specificity, substantially outperforming XGBoost and Transformer-only text baselines (MentalRoBERTa, MultiViT).

D. Emotion Recognition and Cognitive-State Decoding

For emotion classification, a Transformer-CNN hybrid (EEG ST-TCNN) [18] used parallel Transformer encoders to extract spatial and temporal features followed by CNN-based aggregation, achieving 96.67% accuracy on the SEED dataset and 95–97% accuracy across DEAP arousal and valence dimensions, outperforming prior CNN-only and RNN-based emotion-recognition models.

For inner-speech decoding, a pilot study [12] compared a compact CNN (EEGNet) against a spectro-temporal Transformer on bimodal EEG-fMRI data using leave-one-subject-out cross-validation. Because the dataset was small, the study emphasized generalizability rather than raw accuracy, reflecting a recurring theme in the literature: Transformer architectures are more data-hungry than CNNs and require careful validation protocols on limited EEG datasets.

Finally, EEG-Deformer [8] performed a broad comparison of CNN, GNN, RNN, Transformer, and hybrid CNN-Transformer models across three distinct decoding tasks — attention, fatigue, and mental-workload classification. The hybrid CNN-Transformer model outperformed all other architectures on every task, while pure Transformer models without convolutional layers produced the weakest results, reinforcing the pattern observed across the majority of the reviewed studies. A broader synthesis of this pattern is provided by a 2026 comparative review of 88 deep learning models for EEG feature extraction

[17], which likewise concludes that hybrid CNN-Transformer designs are the most consistently competitive architecture family across EEG classification tasks published between 2015 and 2025.

TABLE I
Summary of Reviewed CNN vs. Transformer EEG Performance Evaluation Studies

Ref.	Application Domain	Architectures Compared	Dataset(s)	Key Reported Result
[1]	Visual imagery decoding	Hybrid CNN-Transformer vs. EEGNet, ConvNet, FBCNet, TSformer	Complex visual imagery EEG	Hybrid model 0.7234 vs. TSformer 0.7072, FBCNet 0.6552, ConvNet 0.6447, EEGNet 0.6209
[3]	Emotion / deception (BCI)	CNN-Transformer vs. Inception-CNN	LieWaves (5-ch), SEED (62-ch)	CNN-Transformer 97.14% on LieWaves; Inception-CNN 98.52% on SEED
[4]	Clinical abnormality detection	CNN-LSTM vs. Transformer vs. foundation model	Multicenter + TUAB corpus	All models ≥86% balanced accuracy; finetuned foundation model best at 89.19%
[5]	Motor imagery	SSTAF Transformer vs. EEGNet, ShallowConvNet	Public MI datasets	Transformer fusion model outperforms CNN baselines
[6]	Visual brain decoding	CNN-BiLSTM vs. CNN-Transformer	Visual EEG dataset	CNN-BiLSTM (71% acc.) outperforms CNN-Transformer (59% acc.)
[8]	Attention / fatigue / workload	CNN, GNN, RNN, Transformer, hybrid CNN-Transformer	Three BCI datasets	Hybrid CNN-Transformer (EEG-Deformer) best; pure Transformer weakest
[9]	Psychiatric relapse detection	CNN-Transformer fusion vs. XGBoost, MultiViT, MentalRoBERTa	Clinical EEG + text	Fusion model near-perfect precision/recall, 0.95 specificity
[11]	Epileptic seizure detection	CNN-CapsuleNet vs. CNN-Transformer encoder vs. traditional ML	Seizure EEG corpus	CapsuleNet 99.92% (binary); Transformer encoder 87.30% (multiclass)
[12]	Inner speech recognition	EEGNet (CNN) vs. spectro-temporal Transformer	Bimodal EEG-fMRI	Compared via leave-one-subject-out cross-validation; CNN more data-efficient
[13]	Sleep stage / disorder	CNN vs. CNN-BiLSTM vs. CNN-	CAP Sleep Database	CNN-Transformer best: 95.2% (staging), 99.3% (disorder detection)

	classification	Transformer		
[14]	Sleep stage classification	Hybrid CNN-BiLSTM-Transformer	Sleep-EDF, Sleep-EDF-X, SHHS	88.6% / 85.6% / 84.6% accuracy respectively
[16]	Sleep stage classification	Transformer (flexible channel input)	Multi-channel PSG	First model trained across datasets with variable channel counts
[17]	Cross-domain review	CNN, RNN/LSTM, Transformer, hybrids (88 models)	114 papers, 2015–2025	Hybrid CNN-Transformer models most consistently competitive across tasks

Table I gives ideas about the architectures, datasets, and reported outcomes of the studies reviewed. In Section II. Several patterns emerge from this comparison. First, hybrid CNN-Transformer architectures are the most frequently reported best performer, appearing as the top-performing model in the majority of studies that included a hybrid variant [1], [8], [9], [13], [14]. This suggests that convolutional layers remain effective at extracting local spatial and spectral structure from EEG channels, while self-attention layers add complementary value by modeling longer-range temporal or inter-channel dependencies that convolutions alone do not capture well.

Second, pure Transformer models show more variable performance than hybrid or CNN-only models. In some studies, a Transformer-only model outperformed CNN baselines when trained on longer sequences or larger datasets [5], [18]; in others, particularly with short EEG windows or small sample sizes, CNN or CNN-recurrent baselines outperformed the Transformer variant [6], [11], [12]. This indicates that Transformer performance on EEG is more sensitive to sequence length, dataset size, and training protocol than CNN performance.

Third, dataset characteristics — particularly channel count and recording duration — appear to influence which architecture performs best. The channel-density-aware comparison in [3] is a direct illustration: the same research group found the hybrid CNN-Transformer superior on a low-channel dataset and a CNN-only Inception network superior on a high-channel dataset, implying that no single architecture is universally optimal across all EEG acquisition settings.

Finally, evaluation protocol matters as much as architecture. Studies using subject-independent or leave-one-subject-out validation [3], [12] generally report larger performance gaps between architectures than subject-dependent evaluations, highlighting that cross-subject generalization remains a harder and more architecture-sensitive problem than within-subject classification

3. Methodology

CNN with transformer method is applied on DEAP and SEED datasets as shown in figure 3.1

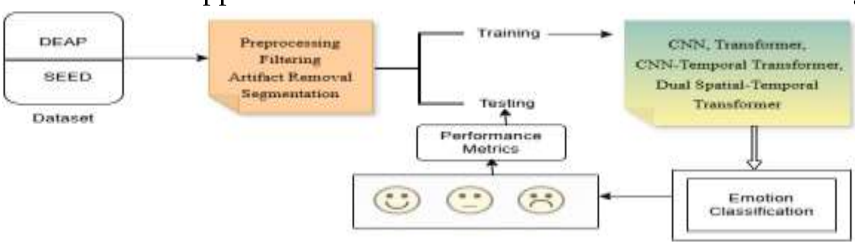


Fig.3.1 CNN and Transformer on EEG Data

Figure 3.1 shows Overall model for EEG-based emotion classification using hybrid CNN-Transformer architectures. Raw EEG signals from the DEAP and SEED datasets are preprocessed through filtering, artifact removal, and segmentation, followed by a train-test split. The training set is used to develop four candidate architectures—CNN, Transformer, CNN-Temporal Transformer, and Dual Spatial-Temporal Transformer—for the emotion classification stage. Further classification results are evaluated
Methodology Summary
Datasets : DEAP (binary valence/arousal, optional quadrant) and SEED (3-class emotion).Split protocol: subject-level 70% train / 20% validation / 10% black-box holdout.
Deep learning models used here include CNN, Transformer-only, CNN-Temporal Transformer, and Dual Spatial-Temporal Transformer with concat/gated/cross-attention fusion.

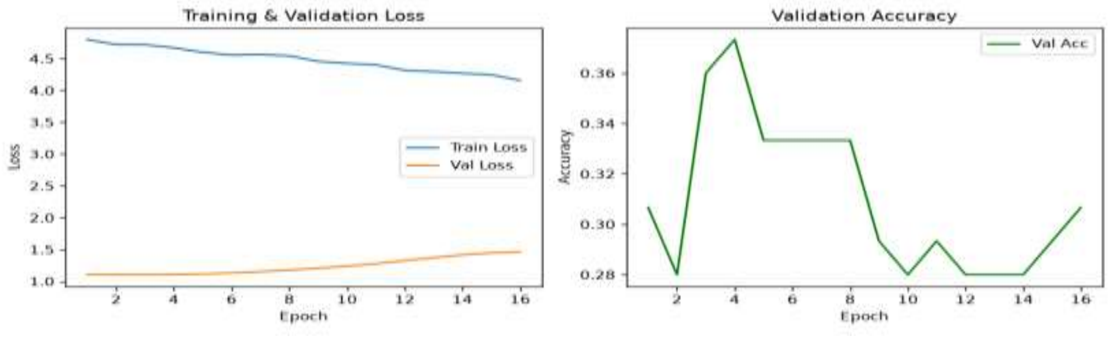
4. Result and Discussion

The experimental results show notable variations in model performance across datasets, tasks, and evaluation conditions. On the DEAP dataset, the dual_spatial_temporal_transformer achieved its strongest validation performance on the arousal task (66.67% accuracy, 66.67% macro F1), yet this did not generalize well to the blackbox test set, where accuracy dropped sharply to 42.50% (macro F1: 38.93%) where as DEAP / valence: best model = cnn_temporal_transformer (blackbox macro f1 = 57.26%. The SEED / 3class, the best model = dual_spatial_temporal_transformer (blackbox macro f1 = 34.30%) identified.

Table : Full Model Comparison

dataset	task	model	validation_accuracy	validation_macro_f1	blackbox_accuracy	blackbox_macro_f1
DEAP	arousal	dual_spatial_temporal_transformer	66.67%	66.67%	42.50%	38.93%
DEAP	valence	cnn_temporal_transformer	50.00%	49.10%	57.50%	57.26%
DEAP	valence	dual_spatial_temporal_transformer	58.33%	57.91%	50.83%	50.42%
SEED	3class	dual_spatial_temporal_transformer	37.33%	36.91%	36.00%	34.30%

Extended Metrics — dual_spatial_temporal_transformer ·
ROC-AUC (macro): 67.63% , Sensitivity: 57.97%, Specificity: 78.43%,
Training curves — dual_spatial_temporal_transformer



5. Conclusion

This review examined recent literature performing a direct performance evaluation of CNN and Transformer methods on EEG data across motor imagery, visual imagery, sleep staging, seizure detection, clinical abnormality screening, emotion recognition, and inner-speech decoding tasks. The results demonstrate that while hybrid architectures combining local spatial-temporal feature extraction with global attention mechanisms show promise, their performance remains inconsistent across tasks and datasets, with validation accuracies ranging from 37.33% to 66.67% . This result highlights that increased difficulty of fine-grained emotion categorization. These findings confirm that while hybrid CNN-Transformer designs are effective in capturing complementary local and global EEG signal characteristics, architectural design alone does not guarantee robust cross-set generalization, particularly given the high inter-subject variability and limited scale characteristic of EEG datasets.

6. References

1. A. Kim et al., "Efficient Transformer-Integrated Deep Neural Architectures for Robust EEG Decoding of Complex Visual Imagery," arXiv:2511.15218, 2025.
2. V. R. Venkannagari, S. Sharma, P. Olikkal, D. Parikh, J. Paun, F. Safavi, and R. Vinjamuri, "Toward Robust EEG Classification Using Adaptive CNN–Transformer and Inception Architectures and Signal Level Augmentation," Sensors, vol. 26, no. 14, p. 4636, 2026.
3. "Automatic detection of abnormal clinical EEG: comparison of a finetuned foundation model with two deep learning models," arXiv:2505.21507, 2025.
4. "SSTAF: Spatial-Spectral-Temporal Attention Fusion Transformer for Motor Imagery Classification," arXiv:2504.13220, 2025.
5. "EEG classification for visual brain decoding with spatio-temporal and transformer based paradigms," arXiv:2406.07153, 2024.
6. "Classification of EEG signals using Transformer based deep learning and ensemble models," ScienceDirect, 2023.
7. "EEG-Deformer: A Dense Convolutional Transformer for Brain-computer Interfaces," arXiv:2405.00719, 2024.
8. N. Fahmy et al., "A CNN-Transformer Fusion Model for Proactive Detection of Schizophrenia Relapse from EEG Signals," Bioengineering, vol. 12, no. 6, p. 641, 2025.
9. "Local and global convolutional transformer-based motor imagery EEG classification," PMC10469791, 2023.
10. S. A. Holguin-Garcia, E. Guevara-Navarro, A. E. Daza-Chica, M. A. Patiño-Claro, H. B. Arteaga-Arteaga, G. A. Ruz, R. Tabares-Soto, and M. A. Bravo-Ortiz, "A comparative study of CNN-capsule-net, CNN-transformer encoder, and Traditional machine learning algorithms to classify epileptic seizure," BMC Medical Informatics and Decision Making, 2024.
11. "Deep learning for inner speech recognition: a pilot comparative study of EEGNet and a spectro-temporal Transformer on bimodal EEG-fMRI data," PMC12589180, 2025.
12. "Multimodal Hybrid CNN-Transformer with Attention Mechanism for Sleep Stages and Disorders Classification Using Bio-Signal Images," Signals, vol. 7, no. 1, p. 4, 2026.
13. C. Liu, Q. Guan, W. Zhang, L. Sun, M. Wang, X. Dong, and S. Xu, "MultiScaleSleepNet: A Hybrid CNN–BiLSTM–Transformer Architecture with Multi-Scale Feature Representation for Single-Channel EEG Sleep Stage Classification," Sensors, vol. 25, no. 20, p. 6328, 2025.

14. Y. Hu, X. Yang, Y. Xu, and J. Sun, "Causal-aware reliability assessment of single-channel EEG for transformer-based sleep staging," Frontiers in Neuroscience, vol. 19, 2025.
15. "FlexSleepTransformer: a transformer-based sleep staging model with flexible input channel configurations," Scientific Reports, 2024.
16. "Efficient Feature Extraction for EEG-Based Classification: A Comparative Review of Deep Learning Models," MDPI, 2026.
17. "Emotion Classification Based on Transformer and CNN for EEG Spatial–Temporal Feature Learning," PMC10969195, 2024.