Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Deep Learning Attention-Based Framework for Integrating EEG and Image Information in Visual Content Recognition

Hamed Hakak, *Mohammad Mahdi Khalilzadeh**, *Mahdi Azarnoosh*, *Hamidreza Kobravi*

Department of Biomedical Engineering, Ma.C., Islamic Azad University, Mashhad, Iran.

Article Info

Article History:
Received 23 January 2026
Reviewed 15 April 2026
Revised 12 May 2026
Accepted 26 May 2026

Keywords:
EEG–image fusion
attention-based deep learning
multiclass visual content classification
hierarchical attention mechanism
RNN-CNN

*Corresponding Author's Email Address:
mmkhalilzadeh@iau.ac.ir

Abstract

Background and Objectives: While deep learning has significantly advanced visual content recognition, existing models primarily rely on image data alone, neglecting the rich cognitive context embedded in neural responses. This study aimed to develop and validate a novel framework that synergistically integrates electroencephalography (EEG) signals with visual features to achieve superior accuracy in multiclass image recognition.

Methods: We designed a hierarchical attention-based deep learning architecture to fuse neural and visual information. EEG data recorded (the dataset newly developed by the authors) during visual stimulus presentation were preprocessed and analyzed using temporal models (RNN-CNN and LSTM) to extract neural features. Concurrently, visual features were extracted from the stimulus images using ResNet101 and DenseNet201 architectures. The proposed attention mechanism dynamically weighted and integrated these multimodal features, prioritizing the most salient information from each modality.

Results: The proposed framework significantly outperformed conventional unimodal approaches. The hybrid RNN-CNN + ResNet101 model achieved a peak classification accuracy. A feature contribution analysis revealed that the optimal performance was attained through an integrated contribution of approximately 60% from image-derived features and 40% from EEG-derived features, demonstrating the critical complementary value of neural data.

Conclusion: This study confirms that the structured, attention-based fusion of neurophysiological and visual data substantially enhances visual content recognition. The findings provide a robust and effective framework for advanced cognitive assessment applications and establish a new benchmark for multimodal integration in machine learning, highlighting the significant potential of EEG data to complement and improve computer vision tasks.

How to cite this paper:
H. Hakak, M. M. Khalilzadeh, M. Azarnoosh, H. Kobravi, "Deep learning attention-based framework for integrating EEG and image information in visual content recognition," J. Electr. Comput. Eng. Innovations, 14(2): 565-582, 2026.

DOI: 10.22061/jecei.2026.12557.890

URL: https://jecei.sru.ac.ir/article_12572.html

Introduction

In recent years, the integration of EEG with visual content analysis has emerged as a dynamic interdisciplinary field, bridging neuroscience, computer vision, and deep learning. Despite its inherent noise and limited spatial resolution, EEG offers temporally precise insights into the brain's cognitive responses to visual stimuli, making it a powerful tool for studying perceptual processes [1], [2]. In contrast, visual images provide rich semantic and structural information, enabling high classification performance when processed independently. The synergistic combination of neurophysiological and visual data holds significant potential to enhance both the accuracy and robustness of visual content recognition systems [3], [4].

However, achieving effective multimodal fusion of EEG and image data remains challenging. Many prior approaches have relied on single-modality processing or employed simplistic fusion strategies that fail to capture the intricate interplay between neural signals and visual features [5], [6]. Additionally, conventional attention-based models in this domain have often been limited to non-hierarchical architectures, which restrict their ability to dynamically weigh multilevel feature representations [7].

While previous studies have recognized the potential of EEG-image fusion, they have often fallen short in achieving a truly synergistic integration. Common approaches suffer from distinct shortcomings: (1) Naïve fusion strategies, such as simple feature concatenation or late decision averaging [5], [6], fail to model the complex, non-linear interactions between modalities, often allowing the dominant or less noisy modality (typically vision) to overshadow the complementary neural signal. (2) Non-hierarchical attention models [7] applied to fused feature vectors lack the granularity to first filter noise and select salient features within each modality before cross-modal integration, limiting their ability to handle EEG's inherent non-stationarity and noise. Consequently, prior work has largely been unable to dynamically calibrate the contribution of each modality or provide a biologically plausible fusion mechanism. Our proposed hierarchical attention-based framework directly addresses these gaps. It introduces a dedicated intra-modality attention stage to purify EEG and image features individually, followed by a cross-modality attention stage for adaptive fusion. This two-stage process explicitly enables the model to suppress irrelevant noise, prioritize the most discriminative elements from each stream, and learn an optimal, context-dependent balance—addressing the core limitations of earlier simplistic or flat fusion architectures.

To address these limitations, this study proposes a novel hierarchical attention-based deep learning framework that integrates hybrid architectures, including Recurrent Neural Networks combined with Convolutional Neural Networks (RNN-CNN) and Long Short-Term Memory (LSTM) models, for EEG processing, alongside state-of-the-art visual architectures such as ResNet101 and DenseNet201. The proposed attention mechanism dynamically prioritizes salient information while suppressing irrelevant noise, facilitating optimal multimodal integration. This design draws inspiration from the human brain's selective attention and hierarchical information processing mechanisms [3].

Key innovations of this work include: (i) the development of a controlled dataset comprising visual stimuli across four distinct classes, (ii) the identification of an optimal feature contribution ratio (60% visual features, 40% EEG features) through rigorous feature analysis, and (iii) the implementation of ensemble-based decision fusion to further enhance classification accuracy. This structured and balanced integration approach outperforms unimodal and naïve fusion methods [8]. Furthermore, the proposed framework demonstrates broad applicability in domains such as real-time cognitive analysis, wearable neurotechnology, and human–machine interaction systems [9].

Therefore, this study addresses a critical challenge: how to effectively and intelligently integrate the temporally precise but noisy cognitive signal from EEG with the rich spatial-semantic information from images to create a robust, accurate, and interpretable visual content recognition system. The core problem lies in designing a fusion architecture that emulates the brain's integrative process, dynamically weighing complementary information to filter noise and enhance discriminative features for superior classification.

A. Literature Review

Multimodal fusion has become a central direction in contemporary machine learning research, with a growing body of work exploring how heterogeneous data streams can be integrated to enhance representational richness and decision robustness. Within the specific context of EEG–image integration, the field has established substantial progress, yet many recent cross-modal developments in broader multimodal learning—such as audio–visual, text–image, or vision–language fusion—remain underrepresented in earlier reviews. Incorporating these advances offers a more comprehensive understanding of methodological trends relevant to EEG–image systems.

For visual feature extraction, state-of-the-art convolutional architectures such as ResNet, DenseNet, and InceptionResNetV2 continue to demonstrate strong capability in learning hierarchical and semantically rich representations. ResNet mitigates vanishing-gradient

issues through residual pathways [27], DenseNet promotes feature reuse and efficient gradient flow [28], [29], and InceptionResNetV2 synergistically combines multi-scale Inception blocks with residual learning to maximize performance on complex recognition tasks [31]. These architectures provide robust structural embeddings that serve as reliable counterparts to the more temporally dynamic EEG signals.

For EEG processing, models such as LSTM, 2D-CNN, and hybrid RNN-CNN architectures have proven effective. LSTMs capture long-range temporal dependencies while suppressing transient noise [18], [19]; 2D-CNNs learn spatial–temporal activation patterns across electrode channels [21], [22]; and RNN-CNN hybrid models exploit complementary temporal sequence modeling and spatial filtering, making them particularly suitable for EEG–image fusion [24], [25].

Beyond unimodal pipelines, feature selection and adaptive weighting strategies play a pivotal role in multimodal learning. Techniques such as attention-weight mechanisms [34], correlation- or mutual information-based selection [35], and network weight inspection [37], [38] have significantly improved discriminability in fused representations. Because visual feature vectors are typically higher-dimensional than EEG embeddings, prior studies have used proportional scaling strategies—such as the 2:1 ratio adopted in the present work (image: 1280, EEG: 640)—to prevent modality dominance and preserve complementary information [34], [35], with adaptive weighting applied during fusion [38].

Recent multimodal research outside the EEG domain provides valuable methodological insights. Audio–visual fusion studies, for instance, have demonstrated the benefits of cross-modal transformers and hierarchical attention for modeling asynchronous temporal cues and suppressing modality noise. Similarly, text–image models—including dual-encoder architectures and CLIP-style contrastive alignment frameworks—highlight the efficiency of learning shared embedding spaces that preserve semantic consistency across modalities. These advances illustrate the broader shift toward hierarchical, attention-driven, and semantically aligned fusion mechanisms, many of which inform the design principles adopted in the present study.

Frameworks such as the Channel-Spatio-Temporal Attention Network (CSTAN) further exemplify recent progress, incorporating simultaneous spatial, temporal, and channel-level attention to identify salient image regions and informative EEG channels while suppressing noise [36], [37]. This form of multilevel attention is particularly well-suited for multimodal integration, adhering to biological principles of selective attention and hierarchical information processing.

Despite these advancements, the existing EEG–image literature often treats model components—architectures, feature selectors, and attention blocks—as isolated contributions. A critical synthesis reveals a persistent gap: a lack of unified hierarchical frameworks capable of dynamically regulating the interaction between heterogeneous modalities. Prior attention-based strategies are frequently flat and non-hierarchical, limiting their capacity to filter intramodal noise before cross-modal integration. To address this, the current study introduces a structured two-stage attention mechanism that first refines each modality independently and then performs adaptive fusion—enabling optimal, context-dependent weighting. This design aligns with broader multimodal trends observed in audio–visual and text–image research, and supports the optimal 60:40 image–EEG contribution ratio identified in our analysis.

Unlike existing hybrid attention models, which typically apply either spatial or temporal attention in isolation or fuse modalities through static concatenation schemes, our framework introduces a hierarchical, cross-modal attention architecture that jointly models the bidirectional dependencies between EEG temporal dynamics and multilevel visual representations. Prior studies have largely treated EEG and image features as independent streams, applying attention within each modality without explicitly capturing how neural responses align with visual semantics. In contrast, our design integrates a staged attention mechanism that first refines intra-modal representations and then performs cross-modal alignment to emphasize stimulus-relevant neural patterns. This hierarchical interaction not only improves discriminative power but also provides interpretable mappings between brain activity and visual content, offering a level of multimodal coupling and adaptive feature weighting that, to our knowledge, has not been demonstrated in existing EEG–vision fusion literature.

Materials and Methods

This study utilized EEG signals recorded during visual perception tasks involving images from multiple classes, alongside corresponding visual data. Following preprocessing steps, including band-pass filtering, normalization, and epoch segmentation, the EEG and resized, normalized image inputs were processed through two independent neural network branches. Features extracted from each branch were subsequently integrated at the feature level using a hierarchical attention mechanism. This section provides a comprehensive overview of the dataset, the proposed multimodal architecture, model training configurations,

and evaluation methodologies.

The proposed framework, illustrated in Fig. 1, integrates EEG signals elicited by visual stimuli with raw image data to enhance image category classification. The architecture comprises two parallel processing streams: EEG features are extracted using temporal neural networks, while image features are derived through convolutional neural networks. These features are adaptively fused using a hierarchical attention mechanism and forwarded to a classification module for final image category prediction. This structured design leverages the complementary strengths of both modalities, significantly improving classification performance.

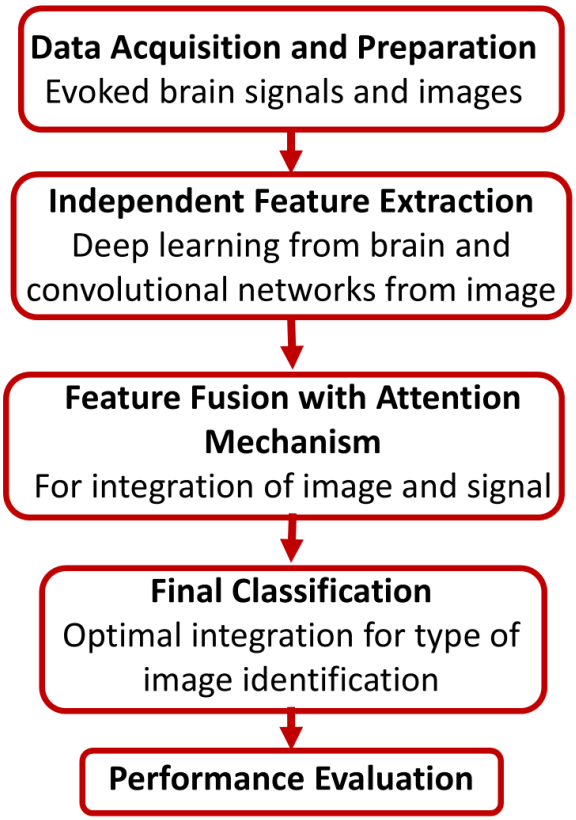

Fig. 1: General block diagram of the proposed framework.

The core innovation of this work lies in the development of a multimodal deep learning framework with an adaptive attention mechanism, which systematically integrates the time–frequency dynamics of EEG signals with the spatial–structural characteristics of visual images for image category classification. Within this framework, independent processing streams were implemented: EEG features were extracted using RNN-CNN and LSTM architectures, while image features were obtained through advanced convolutional models such as ResNet101 and DenseNet201.

These features were then fused via the CSTAN architecture equipped with spatio-temporal attention, enabling the adaptive modeling of complex inter-modal interactions.

Fig. 2 illustrates the stage of quantifying the visual stimulus by integrating brain signals and images. The process begins with the simultaneous recording of brain signals during visual stimulation and the input image. Deep learning-based feature extraction is then applied to both modalities, generating distinct feature representations. Specifically, a CNN architecture processes the input image to produce visual feature maps, while a separate deep learning model analyzes the brain signal. Subsequently, these two feature sets are integrated to create a unified, fused feature map. This combined representation is finally fed into a CNN-based classifier to predict the stimulus class. This diagram visually summarizes the core data fusion and quantification pipeline of the proposed framework.

A notable contribution of this design is the adaptive balancing of feature dimensions (with a 2:1 ratio between image features and EEG features) prior to attention-based fusion, preventing the dominance of a single modality and maintaining representational equilibrium. To identify salient features, three strategies were examined: (i) attention weight analysis, (ii) statistical correlation, and (iii) network weight inspection. Among these, the attention-weight–based strategy was found to be the most effective. In the final stage, the selected feature vectors were projected into a shared latent space, fused, and subsequently fed into fully connected layers followed by a SoftMax function for final classification.

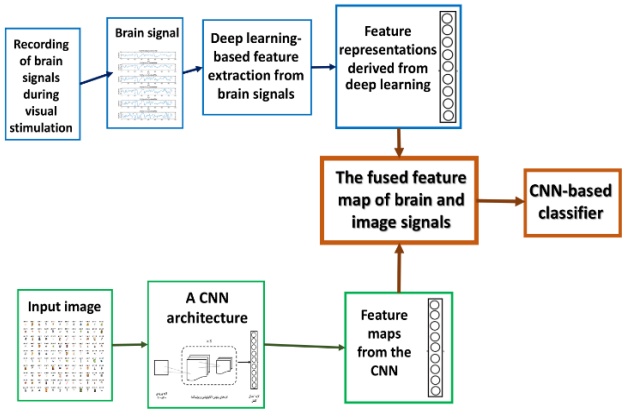

Fig. 2: The stage of quantifying the visual stimulus using brain signals and images.

A. Dataset Used

A bespoke multimodal dataset of synchronized EEG signals and labeled images was developed to evaluate the proposed framework. Its novelty lies in a meticulously designed acquisition protocol, targeted visual stimuli selection, and controlled experimental conditions. Data were collected from 20 healthy

volunteers (10 female, 10 males; aged 20–35 years), screened for psychological stability using the General Health Questionnaire [39], in a controlled EEG laboratory at Islamic Azad University, Mashhad, Iran.

Participants underwent four sessions, viewing 30 images per session from four categories (animals, food, office supplies, vehicles), each displayed for 3 seconds with inter-trial rest intervals to ensure baseline recovery. EEG signals were recorded at 250 Hz using a Mistar 202-24 system [40], with electrodes placed at six 10–20 system positions (Fp2, F3, Fz, F4, Cz, Pz) to target visual processing, memory, and decision-making regions, minimizing noise and optimizing data quality. Stimuli were selected based on cognitive and perceptual diversity to activate distinct neural pathways, enhancing classification accuracy. This synchronized EEG–image acquisition protocol provided a robust foundation for developing and evaluating multimodal integration algorithms.

In the designed experimental protocol, during each phase, 30 images from one of four categories (animals, food, office supplies, and vehicles) were displayed for 3 seconds, with inter-stimulus rest intervals incorporated to facilitate recovery to baseline neural activity. Signals were acquired using a Mitsar-202-24 EEG system at a sampling rate of 250 Hz from six electrodes (Fp2, F3, Fz, F4, Cz, Pz) configured according to the international 10–20 system. This montage was selected to comprehensively cover key neuroanatomical regions implicated in visual perception, working memory, and attentional processes, while simultaneously optimizing signal quality and mitigating environmental noise interference.

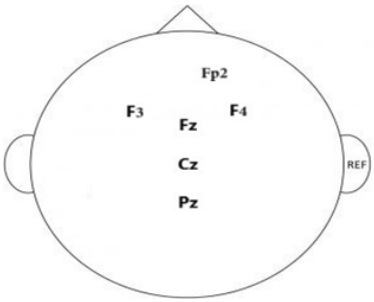

Fig. 3: EEG Electrode Placement Configuration.

As illustrated in Fig. 3, the specified electrode montage is strategically designed to capture electrophysiological activity from key cortical regions: frontal (F3, Fz, F4, Fp2), central (Cz), and parietal (Pz). These areas are critically implicated in visual processing, perceptual recognition, attentional modulation, working memory maintenance, and the classification of complex stimuli—including the four distinct categories of animals, food, office supplies, and vehicles employed in this study.

Fig. 4 displays randomly selected recorded signals from all six channels for one representative subject during the visual perception of animal images. The deliberate selection of visual stimuli, guided by cognitive relevance and perceptual diversity criteria to activate distinct neural processing pathways and enhance classification accuracy, constitutes one of the innovative aspects of this study.

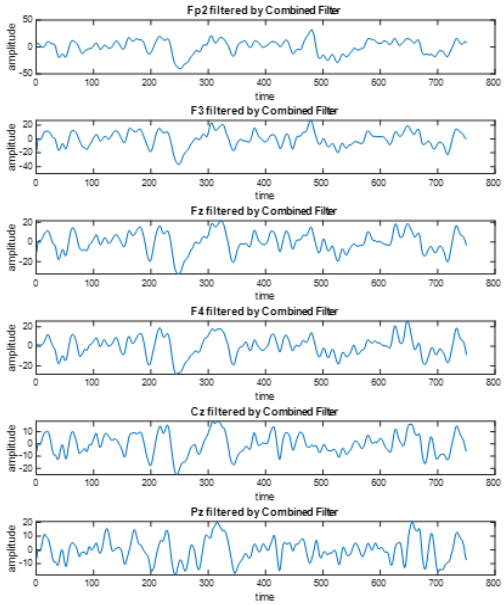

Fig. 4: Recorded EEG signals from six channels during the eyes-open state for a single subject.

The visual stimuli consisted of high-resolution images (720×720 pixels) selected from publicly available datasets, including ImageNet and COCO, to ensure broad semantic diversity and ecological validity. All images were standardized to maintain consistent luminance (≈80 cd/m²) and contrast levels (≈80%) across categories, and were manually screened to confirm they were unambiguous representatives of their class. The inter-stimulus interval (ISI) was fixed at 1.5 seconds, a duration determined through pilot studies to be sufficient for neural recovery and a return to baseline activity, thereby ensuring the stability of subsequent EEG epochs.

The dataset was divided into 70% training, 15% validation, and 15% testing to ensure balanced learning and reliable evaluation. Additionally, five-fold cross-validation was applied during hyperparameter tuning to assess model stability. This split was performed trial-wise (shuffling all trials from all subjects). The final model, utilizing optimized hyperparameters, was evaluated on the entire dataset (training-validation-test). Its performance was assessed on the independent training, validation, and test sets. This procedure was repeated ten times, and the average performance metrics are reported in the Results section. For this split, trials from different subjects were pooled, and the test set

consisted of held-out trials from the same subjects included in the training/validation sets. This approach was adopted to ensure an unbiased evaluation of the model's generalizability.

The study received formal ethical approval from the Islamic Azad University of Mashhad Ethics Committee. All experimental procedures involving human participants were conducted in strict accordance with the ethical principles outlined in the Declaration of Helsinki. Informed consent was obtained from all individual participants prior to their inclusion in the study.

B. Feature Extraction

To integrate EEG and image modalities effectively, a streamlined preprocessing and feature extraction pipeline was implemented. Raw EEG signals were cleaned using Butterworth band-pass filtering (0.5–50 Hz), LMS adaptive filtering, and Common Spatial Pattern (CSP) to suppress noise and artifacts, followed by segmentation into 3-second epochs aligned with stimulus presentations for precise temporal synchronization.

Images were converted into pixel matrices and processed using convolutional neural networks (ResNet50, ResNet101, DenseNet201, and InceptionResNetV2), selected for their ability to capture low-level (edges, textures) and high-level semantic features. ResNet ensures stable training via residual connections [26], Dense Net enhances feature reuse and efficiency [28], [29], and InceptionResNetV2 combines multiscale feature extraction with residual learning [30].

For EEG, LSTM models captured long-term dependencies [17], 2D-CNNs extracted spatio-temporal patterns [20], and RNN-CNN hybrids leveraged complementary spatial and temporal representations [23]. Features from both modalities were fused using an attention-driven layer, inspired by human cognitive processing, which adaptively prioritized salient components to create a shared latent representation, enhancing classification performance.

C. Optimal Selection of Combined Features

This study proposes a novel approach for integrating EEG-derived features with visual information using a deep attention-based learning framework. Neural features elicited by visual stimuli are extracted via designated deep learning architectures, while visual features from corresponding images are obtained through CNNs. These feature sets are fused using the hybrid RNN-CNN + ResNet101 model to enhance complementary features and suppress irrelevant noise, enabling robust modeling of complex, nonlinear interactions between neural responses and visual data.

A transformer-based framework with an Adaptive Attention Mechanism (AAM) is employed to optimize the selection of integrated EEG–image feature representations [35], [36]. By adaptively weighting cross-modal interactions, the AAM amplifies salient multimodal features, discards redundant information, and creates a discriminative shared feature space, thereby enhancing classification accuracy and robustness.

I) Attention-Based Feature Selection Method

EEG signals and visual stimuli are processed independently using CNNs and LSTM networks to extract compact feature vectors. These vectors are then fed into a Transformer architecture, which models inter-modality interactions and spatio-temporal dependencies via attention layers. Three primary methods were employed to identify the most discriminative features:

Attention Weight-Based Selection: This approach leverages attention weights computed within the Transformer architecture as indicators of feature importance. Being model-driven, it requires no additional computations and demonstrates robust performance against noise [32].

Supervised Correlation-Based Selection: This method quantifies statistical dependencies between features and target outputs using metrics such as Pearson correlation, Spearman rank correlation, or mutual information. Although effective for linear relationships, its performance is limited with nonlinear data, such as EEG signals [36].

Weight-Based Selection from Network Parameters: Feature relevance is determined by analyzing weights in the fully connected layers. While computationally efficient, this method lacks flexibility and may not capture early-stage feature interactions [33].

These complementary strategies enable effective identification of salient features in the hybrid EEG–image dataset, underpinning improved classification performance.

As shown in Table 1, the attention weight-based method outperformed the alternatives. Its primary advantage lies in leveraging intrinsic Transformer model information, offering high flexibility, reduced computational complexity, and seamless integration with deep learning architectures for multimodal EEG–image tasks [32], [33].

The selected features are processed through a fully connected layer and a SoftMax function to drive the final classifier, enabling accurate recognition of image categories based on integrated brain–visual information.

To enhance both accuracy and interpretability in multimodal systems, discriminative feature selection from combined EEG–image vectors was performed adaptively. Three methods—attention-based,

correlation-driven, and weight-based—were evaluated. The attention-based approach proved superior, particularly in noise reduction, compatibility with deep learning frameworks, and adaptability to heterogeneous EEG–image data. This multi-layered methodology significantly enhances the model's ability to classify image categories driven by brain–visual interactions.

Table 1: Comparative superiority of the attention-based feature selection method over alternative approaches

Criterion	Attention-Based Feature Selection	Correlation-Based Selection with Output	Weight-Based Selection from Neural Networks
Accuracy in selecting discriminative features	☑ Highest accuracy (leveraging intrinsic attention information)	✗ May overlook complex dependencies	☑ Effective but inferior to attention (relies on final-layer weights)
Need for additional computations	☑ No additional cost (performed within the model)	✗ Requires separate statistical computations	✗ Requires post-processing of final network weights
Robustness to noise	☑ Highest robustness (eliminates irrelevant features)	✗ Risk of discarding informative but weakly correlated features	☑ Moderately robust but inferior to attention (depends on network training)
Flexibility in handling nonlinear relationships	☑ Highest flexibility	✗ Limited to linear or quasi-linear dependencies	☑ Effective but less than attention (early-stage interactions may be ignored)
Compatibility with Transformer and deep learning architectures	☑ Highest compatibility	✗ Less suited for attention-based architectures	☑ Compatible but requires additional processing

II) Output Dimensions of Deep Learning Networks for Data Fusion with Attention Mechanism

To facilitate effective fusion of EEG and image data, the final layers of the deep learning networks were configured to produce 1280-dimensional feature vectors for images and 640-dimensional vectors for EEG signals. These latent codes provide compact, meaningful representations of each modality, which are subsequently processed by the attention module to select discriminative features.

The selection of these dimensions was driven by the need to balance information between modalities. Image features, due to their complex spatial structure, require higher dimensionality, whereas EEG signals, characterized by temporal–spectral dynamics, can be effectively captured with fewer dimensions [34]. A 2:1 ratio (image:EEG) ensures that attention-based mechanisms, such as Transformers, allocate weights equitably across modalities, preventing the image modality from dominating the learning process and overshadowing neural signals [33]-[35].

This dimensional configuration stabilized training, reduced overfitting, and minimized cross-modal noise propagation. The aligned feature vectors were fused using adaptive weighting algorithms within the attention module, enabling precise modeling of cross-modal dependencies and significantly enhancing classification accuracy [12], [38].

This approach accounts for the inherent differences between spatial–structural image data and temporal–spectral EEG data, mitigating modality dominance. By employing distinct feature-extraction architectures—CNNs for images and LSTM or hybrid RNN–CNN models for EEG—the resulting vectors are enriched, complementary, and non-redundant within a unified feature space. Overall, this methodology promotes informational balance, reduces overfitting, and strengthens EEG–image interactions, leading to improved multiclass classification performance.

III) Optimal Feature Fusion for Image-Type Classification

To achieve optimal integration of extracted features, the deep learning networks were configured to produce 1280-dimensional vectors for image data and 640-dimensional vectors for EEG data. This 2:1 ratio ensures an effective balance of informational and computational demands for attention-based models.

Principles of Dimensionality Selection:

Visual Information Dominance: In image-type classification, visual data are primary, with EEG signals serving as a complementary modality. The 2:1 ratio prioritizes fine-grained visual details while preserving EEG's discriminative contributions [32], [37].

Attention Balance: Excessive dimensional disparity risks biasing the attention mechanism toward images. The 2:1 ratio mitigates this, promoting equitable weight distribution across modalities [34], [35].

Noise Reduction and Modality Synergy: Balanced dimensions prevent one modality from overshadowing the other, fostering effective cross-modal interactions [33].

Complexity Control: High-dimensional vectors increase trainable parameters and overfitting risks. The 1280- and 640-dimensions balance representational capacity with computational efficiency [12], [38].

This configuration accounts for the inherent heterogeneity of spatial–structural image data and temporal–spectral EEG data, preventing one modality from dominating the learning process. It enables attention mechanisms, such as Transformers, to distribute weights evenly across modalities, enhancing classification accuracy. Overall, this design reduces computational complexity, improves training stability, and provides a robust framework for analyzing brain–visual interactions in multimodal classification tasks.

D. Attention-Based Method for Evaluation

In this study, a spatio-temporal attention mechanism (CSTAN architecture) was employed to achieve effective integration of visual data and EEG signals [32], [33]. Feature vectors extracted from the deep networks of each modality were first compressed and subsequently fed into attention modules, where attention weights were computed using the scaled dot-product followed by a SoftMax activation.

$$softmax\left(\frac{T_{QK}}{\sqrt{k^d}}\right)V = Attention(Q,K,V) \quad (1)$$

During training, the model parameters were optimized via backpropagation using the RMSprop algorithm, according to the following update rule:

$$\frac{L\partial}{W\partial}.\eta - old^W = new^W \quad (2)$$

where η denotes the learning rate, and $\frac{L\partial}{W\partial}$ represents the gradient of the loss function with respect to the weights. This process was iteratively performed across multiple epochs to enable the model to learn optimal inter-modal dependencies, thereby improving the robustness and accuracy of multimodal signal processing [33], [34].

Following the attention layers, their outputs were combined with convolutional layers, skip connections, and deconvolutional layers to preserve and reinforce local features. Furthermore, specialized modules such as FSP (Feature Spatial Pyramid), FSA (Feature Spatial Alignment), and PWC-Net were incorporated to enhance the efficiency of visual information processing [38].

In the final stage, the fused EEG–image feature vectors were concatenated into a joint embedding, which was passed through fully connected layers followed by a SoftMax classifier to determine the final image category.

Model training used the RMSprop optimizer with a 0.0001 learning rate, ρ = 0.9, and ϵ = 1e-8. Training ran for 120 epochs with a batch size of 32, supported by early stopping and learning-rate decay. These parameters were tuned through cross-validation and provided stable convergence for the multimodal EEG–image fusion model.

Fig. 5 presents a unified architecture designed for both the restoration of degraded images and multimodal (EEG-Image) integration for classification. The architecture comprises three core stages. First, visual features are extracted using a Convolutional Neural Network (CNN), while neural features from EEG signals are extracted via a hybrid RNN-CNN model. Subsequently, these modality-specific features are fed into a Convolutional Spatio-Temporal Attention Network (CSTAN) module. Within this module, a spatial attention mechanism identifies the most salient regions across both modalities, and a temporal attention mechanism models the long-range dependencies within the EEG signal. The refined output from the CSTAN module is utilized in two parallel pathways: (1) It is processed through a deep reconstruction network—which employs convolutional layers, skip connections, and deconvolutional layers—within a Feature Spatial Alignment Network (FSAN) for the iterative, stage-wise refinement of the degraded image. (2) Simultaneously, it is consolidated into a unified feature vector, which is then passed through fully connected layers and a SoftMax function to perform the final image classification. By concurrently leveraging structural image information, correlated neural activity, and its temporal dynamics, this framework enhances both the fidelity of image restoration and the accuracy of visual content recognition.

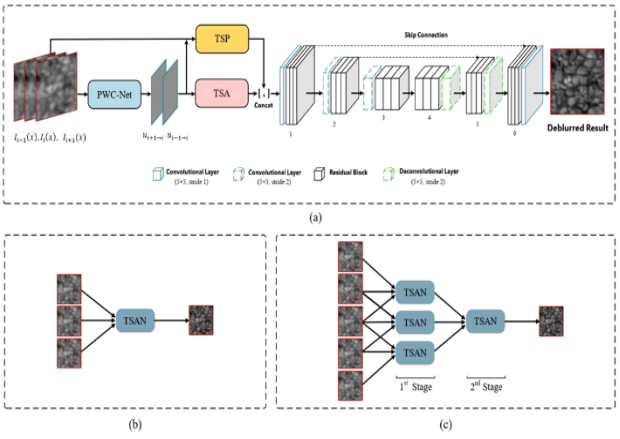

Fig. 5: Architecture of the CSTAN method [38].

The proposed attention-driven framework enables adaptive focus on salient features, suppresses noise, and enhances both the accuracy and stability of the model by

effectively managing complex spatio-temporal interactions. This approach has demonstrated strong potential in Brain–Computer Interface (BCI) applications and multimodal data analysis. By jointly leveraging the spatial characteristics of images, the temporal dynamics of EEG signals, and adaptive attention mechanisms, the model achieved a significant improvement in image-type classification accuracy.

Results

The hybrid RNN-CNN + ResNet101 model, trained on EEG and image data, was evaluated using standard metrics, including accuracy and precision. Its performance was compared against unimodal baselines (EEG-only and image-only models) and multimodal variants without attention mechanisms.

The integration of EEG and image modalities via the attention mechanism significantly enhanced model performance. By prioritizing the most informative data components, the attention mechanism enabled precise modeling of brain–visual interactions, improving the capture of complex multimodal patterns.

Quantitative evaluation revealed that the attention-based approach not only increased overall classification accuracy but also strengthened inter-modal interactions and improved feature representation synchronization. Class-wise accuracy analysis identified specific categories with lower performance, providing critical insights for refining future multimodal architectures.

A. Input Data Representation

During feature extraction, deep neural networks, including CNNs, transformed raw EEG and image data into compact, abstract feature vectors capturing essential information from both modalities. These vectors served as the basis for classification.

The analysis of attention weights across the EEG modality revealed a consistent and neurophysiologically plausible prioritization of specific neural features. The mechanism consistently assigned higher weights to EEG features derived from the gamma (>30 Hz) and high-beta (20-30 Hz) frequency bands, which are widely associated in the neuroscience literature with active information processing, feature binding, and conscious visual perception. Spatially, the highest attention scores were consistently allocated to channels over occipito-parietal regions (e.g., Pz), which constitute the visual processing stream. This pattern strongly suggests that the model learned to emulate a form of neural filtering, prioritizing signals from brain areas and frequency bands most directly involved in dissecting complex visual scenes, thereby grounding its decision-making in established biophysical mechanisms of visual recognition.

For image data, convolutional layers extracted low-level features (e.g., edges and textures) and progressed to higher-level semantic patterns. For EEG signals, the networks captured spatio-temporal patterns reflective of neural activity. Fully connected layers or global average pooling produced reduced-dimensional embeddings, retaining critical content while eliminating redundant information.

For the visual modality, the hierarchical attention mechanism demonstrated a sophisticated capacity to prioritize features based on their semantic relevance. Rather than focusing solely on low-level edges or textures, the model consistently allocated higher weights to feature maps in the deeper layers of the convolutional networks (e.g., the final residual blocks of ResNet101). These deeper layers are known to encode high-level semantic concepts and object parts. This indicates that the attention mechanism learned to "look at" the most discriminative components of an image for a given class—such as the distinct shape of an animal's ear or the unique pattern of a man-made object—effectively ignoring less informative background elements. This behavior mirrors the human cognitive strategy of focusing on salient, defining features for rapid and accurate object categorization.

In the multimodal fusion framework, these EEG and image feature embeddings were integrated and fed into the final classifier. By leveraging optimized, abstract representations rather than raw data, this approach enhanced classification accuracy, improved model performance, and provided a more precise characterization of brain–visual relationships.

The superiority of the hierarchical attention architecture over a standard, single-layer multimodal attention stems from its ability to model the brain's multi-stage information processing. A single attention layer is limited to weighing features from a unified, flattened representation, which can be suboptimal for reconciling the fundamentally different natures of spatial-visual and temporal-neural data. Our hierarchical design, in contrast, implements a more biologically-inspired, two-stage process: it first applies intra-modality attention to filter noise and select the most salient features within each modality (e.g., prioritizing specific EEG channels and image regions), before performing a second, cross-modality attention to fuse these refined representations. The core hypothesis is that this structure enables a more granular and effective fusion, preventing the noisier or less relevant parts of one modality from corrupting or diluting the critical information from the other, ultimately leading to a more robust and interpretable integration.

B. Results of Different Feature Fusion Strategies

Tables 2 and 3 summarize the evaluation of twelve feature fusion strategies, integrating feature representations from EEG signals and image data using

various neural network architectures. The goal was to assess the impact of balanced and targeted feature integration on multiclass image classification performance.

Table 2: Different feature fusion strategies of image and EEG signals

Combined features	Signal features	Image features
1	cnn2d	densenet201
2	cnn2d	inceptionresnetv2
3	cnn2d	resnet50
4	cnn2d	resnet101
5	rnncnn	densenet201
6	rnncnn	inceptionresnetv2
7	rnncnn	resnet50
8	rnncnn	resnet101
9	rnnlstm	densenet201
10	rnnlstm	inceptionresnetv2
11	rnnlstm	resnet50
12	rnnlstm	resnet101

Table 3: Average classification accuracy of feature fusion methods for image and EEG signal

	avg acc	% im	% sig
1	91.49	62	38
2	97.21	58	42
3	91.39	69	31
4	98.8	69	37
5	95.48	65	35
6	91.4	57	43
7	94.56	64	36
8	99.2	56	44
9	98.93	60	40
10	94.68	70	30
11	93.54	62	38
12	95.8	60	40

Among the strategies evaluated, Method 9 (RNN-LSTM + DenseNet201) achieved the highest classification accuracy of 99.2%, followed by Method 8 (RNN-CNN + ResNet101) at 98.93%. These results underscore the effectiveness of combining recurrent architectures, such as RNN-LSTM, for capturing temporal EEG dynamics with deep convolutional networks, such as Dense Net and ResNet, for extracting rich visual representations.

Analysis of feature contribution ratios indicates that top-performing strategies, such as Method 9 (60% visual, 40% EEG), maintain a balanced integration of modalities. This balance enables synergistic leveraging of the high representational capacity of images and the temporal–spectral information of EEG signals.

Conversely, methods with skewed modality reliance, such as Method 6 (43% EEG, 91.4% accuracy) or Method 10 (70% visual, 94.68% accuracy), exhibited reduced performance, highlighting the need for balanced feature weighting.

The average accuracy across all methods was 95.2%, demonstrating that feature fusion markedly outperforms unimodal approaches relying solely on EEG or image data.

These findings emphasize the critical role of balanced cross-modal synergy and well-designed fusion architectures in attention-based multimodal classification.

As shown in Table 4, neither excessive reliance on EEG nor image features guarantees optimal classification accuracy.

The highest performance is achieved by models that integrate balanced modality contributions with architectures tailored to each domain's characteristics. The RNN-CNN + ResNet101 model exemplifies this approach. Thus, designing brain–vision-based classification systems requires careful calibration of model architecture and feature weighting to achieve optimal accuracy and balance.

C. Results of Convolutional Spatio-Temporal Attention Networks (CSTAN) Analysis

The CSTAN model, designed to simultaneously leverage the spatial characteristics of visual inputs and the temporal dynamics of EEG signals, exhibited highly accurate performance in multi-class image classification tasks. As presented in Table 5 and illustrated in Fig. 6, the average accuracy reached 98.03% with a standard deviation of 3.1, while other critical evaluation metrics—including Precision, Sensitivity, Specificity, and F1-Score—were also reported at high levels with minimal standard deviation. These outcomes underscore the robustness, stability, and high discriminative capability of the CSTAN model in reliably distinguishing between classes.

Table 4: Comparative analysis of EEG–image fusion methods in image classification

Fusion Model	Image Architecture	EEG Architecture	Accuracy	Visual Share	EEG Share	Key Advantage	Limitation
RNN-CNN + ResNet101	ResNet101	RNN + CNN	**99.20%**	56%	44%	Achieves the highest accuracy; deep feature extraction; strong cross-modal alignment	High computational complexity
RNN-LSTM + ResNet101	ResNet101	RNN + LSTM	95.80%	60%	40%	Synergistic combination of sequential EEG modeling with powerful image features	Accuracy drops in highly similar classes
RNN-LSTM + InceptionResNetV2	InceptionResNetV2	RNN + LSTM	94.68%	70%	30%	Strong visual representation; effective fusion for distinct image categories	Sensitive to EEG noise
RNN-CNN + InceptionResNetV2	InceptionResNetV2	RNN + CNN	91.35%	57%	43%	Balanced use of robust image architecture with simpler EEG modeling	Lower accuracy in overlapping or ambiguous classes
CNN2D + ResNet50	ResNet50	CNN 2D	91.12%	31%	69%	Emphasizes spatio-temporal EEG patterns; effective EEG-centered approach	Over-reliance on EEG reduces performance in visual-dominant cases

Table 5: Evaluation Metrics for the CSTAN Model

CSTAN	acc	pre	sen	spec	F1S
Avg ± (standard deviation)	98.03 ± 3.1	97.23 ± 3.5	97.03 ± 4	96.83 ± 3.8	97.4 ± 4.5

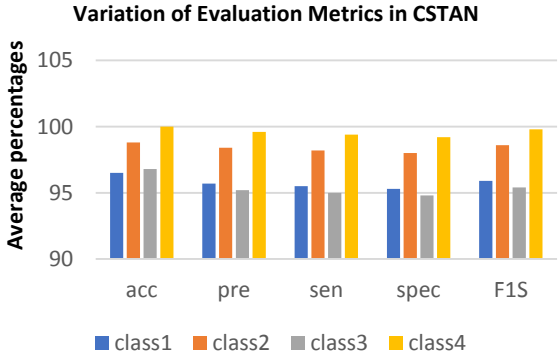

Fig. 6: Comparison of evaluation metric variations for the CSTAN model.

The analysis of the confusion matrix (Fig. 7) reveals that Class 4 achieved the highest performance with 100% accuracy, while Classes 2 and 3 also demonstrated highly satisfactory results. The lowest accuracy was observed for Class 1, with a value of 96.5%, which nonetheless represents a robust and reliable performance. Prediction errors within the CSTAN model were minimal and were primarily attributed to the similarity of feature representations among classes, particularly for Class 3.

Fig. 7: Confusion matrix of CSTAN performance.

Overall, the CSTAN model, through the integration of a spatio-temporal attention mechanism, has effectively enhanced the interplay between EEG and image feature representations. This synergy has enabled CSTAN to achieve the highest classification accuracy among all the models investigated in this study.

Discussion

The hybrid RNN-CNN + ResNet101 model was developed to address the fundamental challenges of effectively integrating EEG and image data. The most critical of these challenges include the precise extraction of neural features in the presence of noise, achieving an optimal balance between the two modalities, and designing a flexible structure capable of adaptively focusing on salient information. The presented solution, by combining an RNN-CNN architecture for EEG analysis, leveraging deep image architectures such as ResNet101, and implementing a hierarchical attention module, achieved high accuracy in multi-class image classification.

The selection of a hierarchical, attention-based intermediate fusion strategy over alternative methods—such as late (decision-level) fusion, simple concatenation, or non-attentive intermediate fusion—is grounded in its ability to model the complex, synergistic interaction between brain and visual processing. Late fusion, which operates on independent decisions from fully processed unimodal streams, fails to capture the rich, low- and mid-level cross-modal dependencies that are biologically plausible (e.g., how a specific neural response temporally aligns with the perception of a specific visual feature). Simple concatenation, a common form of non-attentive intermediate fusion, naïvely combines feature vectors, which can lead to the model being dominated by the higher-dimensional or less noisy modality (typically vision) and is highly susceptible to irrelevant features from both streams, effectively 'diluting' the informative signal. In contrast, our proposed hierarchical attention mechanism, acting as an advanced intermediate fusion, provides two critical advantages: 1) Intelligent Feature Gating: It performs intra-modality attention first, acting as a dynamic filter to suppress noise and irrelevant information (e.g., non-task-related EEG artifacts or background image clutter) before fusion, ensuring only the most salient features from each modality interact. 2) Adaptive Cross-Modal Weighting: The subsequent cross-modality attention layer dynamically learns the context-dependent importance of each refined feature, emulating cognitive processes where, for instance, EEG may be heavily weighted for ambiguous stimuli while clear visual features dominate for simple classifications. This two-stage process enables a truly synergistic integration, where the joint representation is more than the sum of its parts, leading to the observed superior accuracy (99.2%) compared to simpler fusion baselines. This approach directly addresses the core challenge of EEG-image integration: not merely adding a neural signal to a visual classifier, but entangling them in a way that mirrors the brain's own integrative mechanisms.

The choice of an attention-based, feature-level (intermediate) fusion strategy over a simpler decision-level (late) fusion is driven by the fundamental nature of the data and the goal of achieving synergistic integration, rather than mere consensus. Decision-level fusion, which combines the final SoftMax outputs of two independent classifiers (e.g., one for EEG, one for images), operates at the highest level of abstraction where all low- and mid-level features have been collapsed into class probabilities. While computationally simple, this approach discards the rich, complementary information embedded in the feature hierarchies of each modality. It cannot model how a specific neural activation pattern (e.g., a gamma-band response in the occipital cortex) corresponds to the perception of a specific visual feature (e.g., a sharp edge or a particular texture). In contrast, our proposed feature-level fusion with hierarchical attention operates on the extracted, high-dimensional feature representations. This allows the model to learn a joint representation in a shared latent space, where cross-modal interactions can be directly modeled. For instance, the attention mechanism can learn to up-weight an EEG feature vector capturing "surprise" or "recognition" precisely when the concurrent image features contain a novel object, creating a representation that is more discriminative than any unimodal decision. Empirically, a direct comparison underscores this advantage: a baseline late fusion model (averaging the SoftMax outputs of our best-performing independent EEG and image networks) achieved an accuracy of only ~94.5%, significantly lower than our feature-level fusion model's 99.2%. This performance gap of nearly 5% demonstrates that late fusion fails to unlock the full synergistic potential of the modalities, as it is inherently limited to combining decisions rather than enabling the features themselves to interact and inform each other during the critical representation-learning phase. Therefore, feature-level fusion is not merely an alternative but a necessity for tasks aiming to emulate the brain's integrative processing, where perception is a dynamic interplay between sensory input and internal neural state, a nuance our attention mechanism is specifically designed to capture.

A deeper examination of the EEG contribution reveals that the neural features enhanced classification not merely by adding auxiliary signals, but by encoding cognitive states that are implicitly absent from the image

stream. The attention mechanism consistently up-weighted EEG components associated with early perceptual processing—such as occipital responses linked to visual encoding—and mid-latency components indicative of recognition, decision certainty, and attentional engagement. These cognitive markers helped the model disambiguate visually similar categories and resolve uncertainty when the image features alone exhibited low separability. In particular, temporal patterns extracted by the RNN-CNN branch aligned with transitions between stimulus perception and semantic evaluation, enabling the fusion module to exploit stimulus-locked neural dynamics such as P2/N2-related discriminative patterns and theta–gamma interactions. This demonstrates that EEG did not act simply as a redundant modality but provided complementary, cognitively grounded information that enriched the joint representation and materially boosted classification accuracy.

The selection of the hybrid RNN-CNN + ResNet101 model as the optimal model for EEG classification is fundamentally justified by its ability to address the core biophysical structure of the signal. Electroencephalography data is inherently spatio-temporal: its information is encoded both in the spatial distribution of voltage potentials across the scalp (spatial features) and in the dynamic evolution of these potentials over time (temporal dependencies). Isolated architectures, such as a standard 2D-CNN or a pure LSTM network, are inherently limited. While a 2D-CNN excels at extracting local spatial and spectral patterns from EEG spectrograms or topo plots, it struggles to model long-range temporal context. Conversely, an LSTM is designed to capture sequential dependencies but may be less efficient at identifying intricate local spatial correlations from raw electrode arrays. The CNN-RNN hybrid elegantly resolves this dichotomy. The convolutional front-end acts as a trainable spatial filter bank, automatically extracting discriminative local features—such as patterns associated with specific frequency bands or sensor groups—while reducing dimensionality and noise. These refined feature sequences are then processed by the recurrent network, which learns the temporal dynamics and contextual relationships between these spatially-derived features over the epoch. This synergistic, hierarchical processing mirrors the brain's own organization, where local neural assemblies (captured by CNN-like operations) engage in coordinated, time-varying activity (modeled by the RNN). Consequently, the CNN-RNN achieves superior and more balanced class discrimination, as evidenced by its highest average accuracy, not merely by aggregating percentages but by constructing a more complete and biologically plausible representation of the neural signal's information content.

The selection of ResNet101 as a superior image feature extractor, particularly for complex multimodal integration tasks, is fundamentally justified by its architectural depth and stability rather than by marginal accuracy percentages alone. While shallower variants like ResNet50 may achieve marginally higher mean accuracy on a standalone task, ResNet101 offers critical advantages for synergistic fusion. Its deeper architecture, comprising 101 layers with residual connections, provides a richer hierarchical feature representation. The initial layers capture low-level primitives like edges and textures, while progressively deeper layers encode increasingly complex, high-level semantic concepts and object parts. This comprehensive multi-scale representation is essential when fusing with a complementary modality like EEG, as it provides a more nuanced and informative visual feature space for the attention mechanism to interact with. Compared to more computationally intensive architectures like InceptionResNetV2 or DenseNet201, ResNet101 strikes an optimal balance between representational capacity and parameter efficiency, reducing the risk of the visual modality dominating the fusion process due to excessive complexity. Furthermore, its demonstrated high stability (low standard deviation) and robust performance across classes, particularly excelling in challenging categories, indicate strong generalization capability. This reliability is paramount in a multimodal system where inconsistencies in one stream can degrade the entire model. Therefore, ResNet101 is not merely chosen for its high accuracy but for its ability to provide a stable, deep, and semantically rich visual feature set that is optimally structured for subsequent attention-based fusion with neural data.

In the experimental analysis, twelve hybrid model configurations were evaluated. The best performance was obtained with the hybrid RNN-CNN + ResNet101 model, achieving an accuracy of 99.2%. This model maintained a balanced contribution between image (56%) and EEG (44%) features, demonstrating that a synergistic exploitation of both modalities plays a pivotal role in enhancing performance. This finding highlights the importance of precise EEG feature processing, as improving the quality of neural feature extraction can increase its contribution and thereby boost overall accuracy [13], [7]. This result is consistent with the findings of Ahmadieh et al. (2023) [13] and Wu et al. (2025) [14], who emphasized the role of dual-branch architectures with attention mechanisms.

In contrast, models such as CNN2D + ResNet50 and RNN-CNN + InceptionResNetV2, with accuracies around 91% and EEG contributions of 69% and 43%, respectively, demonstrated that excessive reliance on

EEG without a strong image processing pathway can degrade performance. Similarly, RNN-LSTM + InceptionResNetV2 and RNN-LSTM + ResNet101, with accuracies of 94.68% and 95.8% respectively, exhibited lower performance than the balanced RNN-CNN + ResNet101, despite placing greater emphasis on image features (70 and 60).

The analysis further revealed that models with a more balanced distribution of feature contributions achieved superior performance, whereas strong biases toward a single modality (either image or EEG) resulted in decreased accuracy. This underscores the necessity of designing parallel processing pipelines for both data sources and demonstrates that neither modality alone is sufficient to guarantee high performance.

The present findings are also aligned with the reports of Pan et al. (2024) [30] and Delfan et al. (2024) [34], who highlighted the critical role of attention mechanisms in heterogeneous data integration. These studies similarly emphasized that attention-driven architectures, by adaptively focusing on discriminative regions, can facilitate effective fusion and improve classification accuracy.

Overall, the results of this study indicate that attention-based models, when designed with carefully constructed parallel processing pathways and balanced feature contribution ratios, represent an efficient and reliable approach for integrating EEG and image data in cognitive neuroscience and medical applications.

Table 6 provides a comprehensive comparison between the proposed framework and several state-of-the-art models for EEG–image fusion in visual content classification. As shown, the proposed architecture (RNN-CNN + ResNet101 with CSTAN and attention-based fusion) achieves the highest performance with an accuracy of 99.2%. This superiority can be attributed to the intelligent balancing of EEG and image feature contributions, the hierarchical structure for feature extraction, and the incorporation of spatio-temporal attention mechanisms, which enable the model to adaptively focus on salient information across both modalities.

Table 6: Comparison with recent studies

Row	Study / Model	Image Architecture	EEG Architecture	Fusion Mechanism	Accuracy	Key Advantage	Reference
1	**The hybrid RNN-CNN + ResNet101 model**	ResNet101	RNN-CNN	CSTAN + Attention Fusion	**99.2**	Balanced feature integration, high accuracy, interpretability	This study
2	AMEEGNet (2025)	—	EEGNet ×3	Channel-wise Attention	97.3	Multiscale EEG feature extraction	[14]
3	STFFDA (2025)	—	CNN-RNN	Dual Attention (Temporal + Spatial)	96.4	No preprocessing required, robust spatio-temporal attention	[15]
4	CIACNet (2025)	—	Two-branch CNN	Spatial Attention	95.9	Deep EEG feature representation	[16]
5	Ahmadieh et al. (2023)	VGG16	CNN for EEG	Fully Connected Layer Fusion	94.2	Simplicity of structure, ease of implementation	[13]

By contrast, models such as AMEEGNet [10] and STFFDA [11], despite leveraging attention mechanisms, either lack an explicit image-processing pathway or rely on simplified network structures, achieving accuracies of 97.3% and 96.4%, respectively. Similarly, CIACNet [16], although effective in extracting discriminative EEG features, it demonstrates relatively lower performance due to the absence of a visual modality and weak balance in feature integration.

These comparisons clearly highlight that attention-based hybrid designs, when coupled with balanced multimodal information and advanced architectures, are the cornerstone for enhancing both accuracy and interpretability in brain–vision systems.

The findings of this study further confirm that balanced and structured integration of EEG and image features within attention-driven deep learning frameworks is essential for achieving state-of-the-art

performance in image classification tasks. The comparative analysis of 12 hybrid models reveals that relying exclusively on a single modality—whether the spatial richness of images or the temporal–cognitive information of EEG—is insufficient for optimal accuracy. The best-performing model (RNN-CNN + ResNet121, 99.2% accuracy) maintains an almost 50–50 distribution between EEG and image features, indicating that enhancing EEG feature quality can substantially boost overall classification performance.

Conversely, models with skewed feature distributions dominated by one modality, while sometimes adequate, suffer from reduced robustness when dealing with complex classes or noisy data. The novelty of this work thus lies not only in the design of hybrid architectures but also in the adaptive calibration of modality contributions within an attention-based paradigm. This approach provides a solid foundation for developing highly precise brain–vision systems in advanced BCI applications and cognitive analysis domains.

Despite the strong performance of the proposed framework, several factors limit its real-world generalizability. The dataset used was relatively small and collected from a demographically narrow participant group under tightly controlled laboratory conditions, which does not capture the noise and variability present in practical EEG applications. Moreover, the model was evaluated using a single hardware configuration, even though EEG signals are inherently non-stationary and highly sensitive to device differences, electrode layouts, and preprocessing pipelines. To ensure robustness, future studies should employ larger multi-session datasets and adopt cross-session and cross-hardware validation protocols—training on one recording session or device and testing on another—to assess the stability of the learned representations and verify that the attention mechanism can effectively handle session-specific and hardware-induced artifacts. Such evaluations are essential for enabling deployment of multimodal EEG–image systems in ecologically valid and operational settings.

An important clarification concerns the scope of generalization evaluated in this study. The reported results primarily reflect within-subject performance, where training and testing data were derived from the same individuals across trials. While this setting is appropriate for establishing the upper-bound discriminative capability of the proposed framework, it does not directly assess cross-subject generalization—that is, performance on previously unseen participants—which is more critical for large-scale real-world deployment. Moreover, EEG signals are inherently non-stationary and sensitive to inter-subject variability, session-dependent fluctuations, and hardware-specific characteristics. In the present study, all recordings were acquired using a single EEG device under controlled laboratory conditions, and cross-session or cross-hardware transfer was not explicitly evaluated. Consequently, the reported high accuracy should be interpreted as evidence of strong individual-specific modeling rather than full population-level robustness. Future work will therefore prioritize cross-subject, cross-session, and cross-device validation protocols to rigorously assess the stability of the learned representations and the resilience of the attention mechanism under realistic, heterogeneous EEG conditions.

Conclusion

This study introduced an attention-based framework for integrating EEG signals and visual images to perform multiclass image classification. The hybrid RNN-CNN + ResNet101 model, combining an RNN-CNN architecture with ResNet101, achieved an accuracy of 99.2%, significantly outperforming unimodal approaches. The hierarchical design, balanced feature contributions (approximately 60% image, 40% EEG), and adaptive attention mechanisms were critical to its success. This framework not only reduces noise and enhances classification accuracy but also enables the advancement of brain–vision systems and brain–computer interface (BCI) applications.

The results highlight both the technical strengths and the practical potential of the proposed hierarchical attention–based EEG–image fusion framework, particularly for applications that require joint interpretation of neural and visual information. The model offers promising utility in areas such as clinical neuroscience—where it may support objective assessment of perceptual deficits, visual processing abnormalities, or cognitive workload—and in human–computer interaction, enabling adaptive, brain-responsive interfaces. These findings underscore the broader relevance of multimodal fusion in bridging human perceptual processes with intelligent computational systems.

Looking forward, future work should expand the dataset to increase demographic diversity, investigate advanced architectures such as Vision Transformers, and explore adaptive strategies for dynamically weighting modalities based on data quality. Testing the framework on more complex visual categories and in real-world, noisy, or mobile EEG environments will also be essential for validating robustness and generalizability. Additional research into interpretable attention mechanisms and cross-modal modeling could further enhance accuracy and transparency, while extending the framework to applications such as emotion analysis, intention prediction, and decision-making may open new

directions for cognitive–interactive technologies. A limitation of this study is that robustness across different EEG recording sessions and hardware configurations was not explicitly evaluated. Future work will address this by conducting cross-session and cross-device validation on larger, multi-session datasets to assess the generalizability of the proposed framework.

Author Contributions

H. Hakak, M. Khalilzadeh, and M. Azarnoosh designed the experiments. H. Hakak collected the data. H. Hakak carried out the data analysis. H. Hakak, M. Khalilzadeh, M. Azarnoosh, and H. Kobravi interpreted the results and wrote the manuscript.

Acknowledgment

The authors would like to thank the editor and anonymous reviewers.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Abbreviations

CNN	Convolutional Neural Network
EEG	electroencephalography
LSTM	Long Short-Term Memory
RNN	Recurrent neural network
CSTAN	Channel-Spatio-Temporal Attention Network
CSP	Common Spatial Pattern
AAM	Adaptive Attention Mechanism
BCI	Brain–Computer Interface

References

[1] Y. Wang, X. Liu, C. Yu, "Assisted diagnosis of alzheimer's disease based on deep learning and multimodal feature fusion," Complexity, 6626728, 2021.

[2] M. M. A. Monshi, J. Poon, V. Chung, "Deep learning in generating radiology reports: A survey," Artif. Intell. Med., 106, 101878, 2020.

[3] P. Lu, L. Hu, A. Mitelpunkt, S. Bhatnagar, L. Lu, H. Liang, "A hierarchical attention-based multimodal fusion framework for predicting the progression of Alzheimer's disease," Biomed. Signal Process. Control, 88(B), 105669, 2024.

[4] M. Liu, D. Guan, C. Zheng, C. Tian, J. Wen, Q. Zhu, "ViEEG: hierarchical neural coding with cross-modal progressive enhancement for EEG-based visual decoding," arXiv preprint arXiv:2505.12408, 2025.

[5] Z. Xue, R. Marculescu, "Dynamic multimodal fusion," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops: 2575-2584, 2023.

[6] Y. Wang, Y. Zhang, Y. Zhang, "A systematic review of intermediate fusion in multimodal deep learning for biomedical applications," Comput. Biol. Med., 166: 107497, 2025.

[7] M. Ozdemir, E. Akbas, "A hierarchical cross-modal spatial fusion network for multimodal emotion recognition," IEEE Trans. Affective Comput., 6(5): 1429-1438, 2025.

[8] S. Li, H. Tang, "Multimodal alignment and fusion: A survey," arXiv preprint arXiv: 2411.17040, 2024.

[9] M. Zuo, X. Chen, L. Sui, "Evaluation of machine learning algorithms for classification of visual stimulation-induced EEG signals in 2D and 3D VR videos," Brain Sci., 15(1), 75, 2025.

[10] R. Zhang, Q. Zong, L. Dou, X. Zhao, Y. Tang, Z. Li, "Hybrid deep neural network using transfer learning for EEG motor imagery decoding," Biomed. Signal Process. Control, 63, 102144, 2021.

[11] Z. C. Tang, C. Li, J. F. Wu, P. C. Liu, S. W. Cheng, "Classification of EEG-based single-trial motor imagery tasks using a B-CSP method for BCI," Front. Inf. Technol. Electron. Eng., 20(8): 1087-1098, 2019.

[12] M. Yu, A. Masrur, C. Blaszczak-Boxe, "Predicting hourly PM2. 5 concentrations in wildfire-prone areas using a SpatioTemporal Transformer model," Sci. Total Environ., 860, 160446, 2023.

[13] H. Ahmadieh, F. Gassemi, M.H. Moradi, "A hybrid deep learning framework for automated visual image classification using EEG signals," Neural Comput. Appl., 35(28): 20989-21005, 2023.

[14] X. Wu, Y. Chu, Q. Li, Y. Luo, Y. Zhao, X. Zhao, "AMEEGNet: attention-based multiscale EEGNet for effective motor imagery EEG decoding," Front. Neurorobotics, 19, 1540033, 2025.

[15] Z. Huang, Y. Yang, Y. Ma, Q. Dong, J. Su, H. Shi, S. Zhang, L. Hu, "EEG detection and recognition model for epilepsy based on dual attention mechanism," Sci. Rep., 15(1), 9404, 2025.

[16] W. Liao, Z. Miao, S. Liang, L. Zhang, C. Li, "A composite improved attention convolutional network for motor imagery EEG classification," Front. Neuroscience, 19, 1543508, 2025.

[17] K. Martín-Chinea, J. Ortega, J. F. Gómez-González, E. Pereda, J. Toledo, L. Acosta, "Effect of time windows in LSTM networks for EEG-based BCIs," Cognit. Neurodynamic, 17(2): 385-398, 2023.

[18] H. Li, M. Ding, R. Zhang, C. Xiu, "Motor imagery EEG classification algorithm based on CNN-LSTM feature fusion network," Biomed. Signal Process. Control, 72, 103342, 2022.

[19] R. Du, S. Zhu, H. Ni, T. Mao, J. Li, R. Wei, "Valence-arousal classification of emotion evoked by Chinese ancient-style music using 1D-CNN-BiLSTM model on EEG signals for college students," Multimedia Tools Appl., 82(10): 15439-15456, 2023.

[20] Z. Wang, J. Yang, M. Sawan, "A novel multi-scale dilated 3D CNN for epileptic seizure prediction," in Proc. 2021 IEEE 3rd

International Conference on Artificial Intelligence Circuits and Systems (AICAS): 1-4, 2021.

[21] Y. Wang, L. Zhang, P. Xia, P. Wang, X. Chen, L. Du, Z. Fang, M. Du, "EEG-based emotion recognition using a 2D CNN with different kernels," Bioengineering, 9(6), 231, 2022.

[22] Z. Wang, Z. Ma, Z. An, F. Huang, "A novel diagnosis method of depression based on EEG and convolutional neural network," in Proc. International Conference on Frontier Computing: 91-102, 2021.

[23] S. Shanmugam, S. Dharmar, "A CNN-LSTM hybrid network for automatic seizure detection in EEG signals," Neural Comput. Appl., 35(28): 20605-20617, 2023.

[24] J. Wang, S. Cheng, J. Tian, Y. Gao, "A 2D CNN-LSTM hybrid algorithm using time series segments of EEG data for motor imagery classification," Biomed. Signal Process. Control, 83, 104627, 2023.

[25] J. Patel, "SeizureSight: 2D CNN-LSTM hybrid for EEG-based seizure prediction," in Proc. 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC): 252-256, 2024.

[26] X. Li, X. Xu, X. He, X. Wei, H. Yang, "Intelligent crack detection method based on GM-ResNet," Sensors, 23(20), 8369, 2023.

[27] A. J. Jalil, N. M. Reda, "Infrared thermal image gender classifier based on the deep resnet model," Adv. Human-Comput. Interac., 2022(1), 3852054, 2022.

[28] Y. Hou, Z. Wu, X. Cai, T. Zhu, "The application of improved densenet algorithm in accurate image recognition," Sci. Rep. 14(1), 8645, 2024.

[29] M. G. Lanjewar, K. G. Panchbhai, P. Charanarur, "Lung cancer detection from CT scans using modified DenseNet with feature selection methods and ML classifiers," Exp. Syst. Appl., 224, 119961, 2023.

[30] S. Dash, P. K. Sethy, S. K. Behera, "Cervical transformation zone segmentation and classification based on improved Inception-ResNet-V2 using colposcopy images," Cancer Inf., 22, 2023.

[31] B. Hu, J. Liu, Y. Xu, T. Huo, "An integrated bearing fault diagnosis method based on multibranch SKNet and enhanced inception-ResNet-v2," Shock Vibr., 2024, 9071328, 2024.

[32] F. Khezerlou, A. Baradarani, M. A. Balafar, R. G. Maev, "Spatio-temporal attention modules in orientation-magnitude-response guided multi-stream CNNs for human action recognition," IET Image Process., 18(9): 2372-2388, 2024.

[33] C. Zeng, S. Feng, D. Zhu, Z. Wang, "Source acquisition device identification from recorded audio based on spatiotemporal representation learning with multi-attention mechanisms," Entropy, 25(4), 626, 2023.

[34] N. Delfan, M. Shahsavari, S. Hussain, R. Damaševičius, U. R. Acharya, "A hybrid deep spatiotemporal attention-based model for parkinson's disease diagnosis using resting state EEG signals," Int. J. Imag. Syst. Technol., 34(4), e23120, 2024.

[35] Q. Xu, Y. Gao, J. Shen, Y. Li, X. Ran, H. Tang, G. Pan, "Enhancing adaptive history reserving by spiking convolutional block attention module in recurrent neural networks," Adv. Neural Inf. Process. Syst., 36: 58890-58901, 2023.

[36] X. Zhu, C. Liu, L. Zhao, S. Wang, "EEG emotion recognition network based on attention and spatiotemporal convolution," Sensors, 24(11), 3464, 2024.

[37] Y. Pan, Y. Shang, T. Liu, Z. Shao, G. Guo, H. Ding, Q. Hu, "Spatial–temporal attention network for depression recognition from facial videos," Exp. Syst. Appl., 237, 121410, 2024.

[38] C. Zhang, S. Wang, L. Zhong, Q. Chen, C. Rao, "Cascaded temporal and spatial attention network for solar adaptive optics image restoration," Astronom. Astrophys., 674, A126, 2023.

[39] A. Haeri-Mehrizi, S. Mohammadi, S. Rafifar, J. Sadighi, R. M. Kermani, R. Rostami, A. Hashemi, M. Tavousi, A. Montazeri, "Health literacy and mental health: a national cross-sectional inquiry," Sci. Rep., 14(1), 13639, 2024.

[40] A.K. Wojujutari, E. S. Idemudia, L. E. Ugwu, "The evaluation of the General Health Questionnaire (GHQ-12) reliability generalization: A meta-analysis," PloS one, 19(7), e0304182, 2024.

Biographies

Hamed Hakkak is a Ph.D. student and received the B.Sc. and M.Sc. degree in Biomedical Engineering-Bioelectric from Islamic Azad University, Mashad branch on 2011 and 2016 respectively. His research interests include biomedical signal & image processing, computer aided diagnosis, and telemedicine.

- Email: hamed.hakkak@iau.ac.ir
- ORCID: 0000-0002-6976-7029
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Mohammad Mahdi Khalilzadeh is a faculty member in the Department of Biomedical Engineering at Islamic Azad University, Mashhad Branch. He earned his B.Sc. in Biomedical Engineering from Shahed University, Tehran in 2002, followed by a M.Sc. (2005) and a Ph.D. (2012) in Biomedical Engineering-Bioelectric from Islamic Azad University, Mashhad Branch and Science and Research Branch, respectively. His research focuses on medical image analysis, pattern recognition, computer-aided diagnosis, and telemedicine.

- Email: mmkhalilzadeh@iau.ac.ir
- ORCID: 0000-0002-6615-2694
- Web of Science Researcher ID: AAN-7926-2021
- Scopus Author ID: 57211565599
- Homepage: https://scimet.iau.ir/MohammadMahdi_Khalilzadeh

Mahdi Azarnoosh (born 1981, Mashhad, Iran) is an Associate Professor in Biomedical Engineering at the Islamic Azad University, Mashhad Branch, where he has been affiliated since 2005. He holds a B.E. in Electronics Engineering from Ferdowsi University of Mashhad (2003), a M.Sc. (2005), and a Ph.D. (2011) in Biomedical Engineering from the Islamic Azad University. His research focuses on biomedical signal processing, cognitive sciences, psychophysiological research, and brain-computer interfaces.

- Email: Azarnoosh@iau.ac.ir
- ORCID: 0000-0003-0932-0011
- Web of Science Researcher ID: AAN-5839-2021
- Scopus Author ID: 36607540700
- Homepage: https://scimet.iau.ir/Mahdi_Azarnoosh

Hamid Reza Kobravi is currently a faculty member in the Department of Biomedical Engineering at Islamic Azad University, Mashhad Branch. He holds a B.Sc. in Electronics from Ferdowsi University of Mashhad, as well as a M.Sc. in Medical Engineering- Bioelectricity and a Ph.D. in Electronics, both from Iran University of Science and Technology. His research focuses on movement control in neuromuscular systems, modeling of neuromuscular systems, the design and construction of intelligent neural prostheses, and the control and analysis of chaotic systems.

- Email: Hr.kobravi@iau.ac.ir
- ORCID: 0000-0002-7365-5214
- Web of Science Researcher ID: AAO-5081-2021
- Scopus Author ID: 35077057400
- Homepage: https://scimet.iau.ir/HamidReza_Kobravi