INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE, CYBERSECURITY AND EMERGING TECHNOLOGIES (IJAICET)
IJAICET, Vol. 1, Issue 1, July - September, 2026

A Multimodal Attention-Based Fusion Framework with Ensemble Learning for Robust Differentiation of Epileptic and Psychogenic Seizures

Priyanka Singh, Dr. Shweta Chouksey
Research Scholar, LNCT University, JK Town, Sarvadharam,
Sector C, Kolar Road, Bhopal, Madhya Pradesh, India – 462024,

Abstract–Differentiating epileptic seizures from psychogenic non-epileptic seizures remains a significant clinical challenge, often leading to misdiagnosis and inappropriate treatment. We propose a multimodal attention-based fusion framework that integrates electroencephalography signals, video streams, and peripheral physiological data to address this problem. The system first extracts modality-specific features: time-frequency and nonlinear dynamical descriptors from EEG via wavelet decomposition and adaptive principal component analysis, spatiotemporal behavioral embeddings from video using a 3D ResNet, and autonomic markers from physiological signals encoded by a bidirectional long short-term memory network. A self-attention mechanism then computes context-aware weights for each modality, dynamically suppressing contributions from noisy or uninformative channels before constructing a fused representation. This fused vector is subsequently classified by a heterogeneous ensemble of a support vector machine, a random forest, and a shallow multi-layer perceptron, whose outputs are aggregated through soft voting. The key novelty lies in the attention-based fusion layer, which replaces conventional early or late fusion strategies and enables the model to adaptively prioritize diagnostically relevant modalities on a per-instance basis. Furthermore, the ensemble classifier provides robust decision boundaries even when single-modality cues are ambiguous. We expect this framework to improve diagnostic accuracy in clinical settings where video quality may be compromised or EEG artifacts are prevalent. The proposed method therefore offers a principled approach to integrating heterogeneous seizure-related signals, with potential to reduce misdiagnosis rates and guide appropriate therapeutic interventions.

Keywords: Multimodal Fusion; Attention Mechanism; Epileptic Seizures; Psychogenic Non-Epileptic Seizures (PNES); Electroencephalography (EEG); Video Semiology; Physiological Signals; Ensemble Learning; Seizure Classification; Clinical Decision Support.

I. INTRODUCTION

The accurate differentiation between epileptic seizures (ES) and psychogenic non-epileptic seizures (PNES) represents a persistent diagnostic challenge in clinical neurology. While video-electroencephalography (video-EEG) monitoring remains the gold standard for this differential diagnosis, it is resource-intensive, often requiring prolonged hospital stays and expert interpretation [1]. Misdiagnosis rates remain substantial, leading to inappropriate treatment with antiepileptic drugs for patients with PNES and delayed access to appropriate psychotherapeutic interventions. Automated clinical decision support systems have therefore emerged as a promising avenue to augment clinician capabilities and reduce diagnostic delays [2] [3].

Early computational approaches to seizure classification predominantly focused on unimodal EEG analysis. Techniques such as Discrete Wavelet Transform (DWT) for time-frequency decomposition and nonlinear dynamic features including entropy and fractal dimension have been extensively employed to characterize seizure activity [4] [5]. These features are then classified using conventional machine learning models such as Support Vector Machines (SVMs) and Random Forests (RFs) [6] [7]. More recently, deep learning architectures like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have demonstrated superior performance in capturing spatial and temporal dependencies within raw EEG signals [8] [9].

Parallel to these electrophysiological advances, computer vision techniques have been developed to quantify video semiology. Pose estimation algorithms and optical flow methods can detect motor manifestations that are specific to either ES or PNES [10] [11]. Similarly, the analysis of autonomic nervous system responses has gained traction, with Heart Rate Variability (HRV) metrics and Galvanic Skin Response (GSR) signals providing valuable adjunctive data for distinguishing seizure types based on sympathetic activation patterns [12] [13]. However, integrating these heterogeneous data streams into a unified diagnostic framework remains a significant technical hurdle.

Traditional multimodal fusion strategies often rely on early fusion via simple concatenation or late fusion through voting mechanisms. These approaches fail to account for varying signal quality or the dynamic relevance of different modalities during a seizure event [14]. For instance, video quality may be compromised by patient movement or poor lighting, while EEG signals can be contaminated by muscle artifacts. To address these limitations, attention mechanisms originally popularized in natural language processing and computer vision have begun to be adapted for biomedical signal fusion [15] [16]. These adaptive weighting schemes allow models to focus on the most informative features while suppressing noise, a capability crucial for handling missing or corrupted data in clinical settings [17].

We propose a multimodal attention-based fusion framework that integrates EEG signals, video streams, and peripheral physiological data to robustly differentiate ES from PNES. The core novelty of our approach lies in the self-attention mechanism that computes context-aware weights for each modality, dynamically suppressing contributions from noisy or

uninformative channels before constructing a fused representation. This attention-based fusion layer replaces conventional early or late fusion strategies and enables the model to adaptively prioritize diagnostically relevant modalities on a per-instance basis. Furthermore, we employ a heterogeneous ensemble classifier comprising an SVM, an RF, and a shallow multi-layer perceptron (MLP), whose outputs are aggregated through soft voting. This ensemble learning framework provides robust decision boundaries even when single-modality cues are ambiguous [18] [19].

The remainder of this paper is organized as follows: Section 2 reviews related work in automated seizure classification and multimodal fusion. Section 3 details the proposed multimodal adaptive fusion framework, including feature extraction, attention-based fusion, and ensemble classification. Section 4 describes the experimental setup, including dataset characteristics and evaluation metrics. Section 5 presents experimental results and comparative analyses. Section 6 discusses the implications of our findings and outlines directions for future research. Section 7 concludes the paper with a summary of our contributions.

II. Related Work

Automated seizure classification has evolved from unimodal EEG analysis to increasingly sophisticated multimodal frameworks. Early work in this domain focused on extracting handcrafted features from EEG signals, including spectral power in specific frequency bands, wavelet coefficients, and nonlinear dynamical measures such as Lyapunov exponents and sample entropy [4] [5]. These features were typically classified using shallow machine learning models like Support Vector Machines (SVMs) and Random Forests (RFs), achieving reasonable accuracy on curated datasets [6] [7]. However, these unimodal approaches inherently lack the complementary information provided by video semiology and autonomic markers, limiting their diagnostic utility in differentiating ES from PNES where behavioral and physiological cues are critical.

The integration of video-based analysis has addressed some of these limitations. Pose estimation algorithms and spatiotemporal convolutional networks have been employed to extract movement kinematics and vocalization patterns from video recordings [10] [11]. For example, 3D convolutional neural networks can capture the temporal evolution of motor activity, providing discriminative features for seizure type classification. Similarly, peripheral physiological signals such as heart rate variability (HRV) and galvanic skin response (GSR) have been analyzed using recurrent architectures like Long Short-Term Memory (LSTM) networks to capture autonomic nervous system dynamics [12] [13]. These autonomic markers are particularly valuable because ES and PNES exhibit distinct sympathetic activation patterns, with ES typically showing more pronounced and stereotyped autonomic responses.

A. Multimodal Fusion Strategies

Multimodal fusion in biomedical applications has traditionally followed three paradigms: early fusion, late fusion, and intermediate fusion. Early fusion concatenates features from different modalities at the input level, while late fusion aggregates decisions from modality-specific classifiers through voting or averaging [14]. Intermediate fusion, which combines features at some intermediate layer of a neural network, offers a compromise between these extremes. However, all these approaches treat each modality with equal importance, failing to account for varying signal quality or the dynamic relevance of different modalities during a seizure event. For instance, video features may be unreliable during periods of patient occlusion or poor lighting, while EEG signals can be contaminated by muscle artifacts or electrode displacement.

Attention mechanisms have emerged as a powerful solution to this limitation. Originally introduced in the Transformer architecture for natural language processing [15], attention-based weighting has been adapted for multimodal fusion in medical imaging and physiological signal analysis [16]. These mechanisms compute context-aware weights for each modality, allowing the model to dynamically suppress contributions from noisy or uninformative channels. In the context of seizure classification, attention-based fusion has been applied to combine EEG and functional magnetic resonance imaging (fMRI) data, demonstrating improved robustness to missing modalities [17]. However, the application of attention mechanisms to fuse EEG, video, and physiological signals for ES versus PNES differentiation remains largely unexplored.

B. Ensemble Learning for Seizure Classification

Ensemble learning methods have been widely adopted in medical diagnosis to improve classification robustness and generalization. Techniques such as bagging, boosting, and stacking combine multiple base learners to reduce variance and bias [18]. In seizure classification, heterogeneous ensembles comprising SVMs, RFs, and neural networks have shown superior performance compared to individual classifiers, particularly when dealing with imbalanced datasets or ambiguous cases [19]. The soft voting mechanism, which averages the predicted probabilities from multiple classifiers, provides a principled approach to aggregating diverse decision boundaries.

C. Comparison with Existing Approaches

The proposed framework differs from existing multimodal seizure classification systems in several key aspects. First, the attention-based fusion layer replaces conventional early or late fusion strategies, enabling dynamic weighting of modalities based on their diagnostic relevance and signal quality. This is particularly important in clinical settings where video quality may be compromised or EEG artifacts are prevalent. Second, the heterogeneous ensemble classifier provides robust decision boundaries even when single-modality cues are ambiguous, leveraging the complementary strengths of SVMs, RFs, and MLPs. Third, the integration of three distinct modalities—EEG, video, and physiological signals—within a unified attention-

based framework represents a novel contribution to the field. While prior work has explored pairwise fusion of these modalities, the simultaneous integration of all three with adaptive weighting has not been systematically investigated. The key novelty of our approach therefore lies in the combination of attention-based multimodal fusion with ensemble learning, offering a principled solution to the challenging problem of ES versus PNES differentiation.

III. Multimodal Adaptive Fusion Framework for ES and PNES Discrimination

The proposed framework addresses the diagnostic ambiguity between epileptic seizures (ES) and psychogenic non-epileptic seizures (PNES) through a three-stage pipeline: modality-specific feature extraction, attention-based multimodal fusion, and heterogeneous ensemble classification. The system processes three heterogeneous data streams—electroencephalography (EEG), video recordings, and peripheral physiological signals—each encoded into a fixed-dimensional embedding space before being dynamically weighted and fused. The technical details of each component are presented in the following subsections.

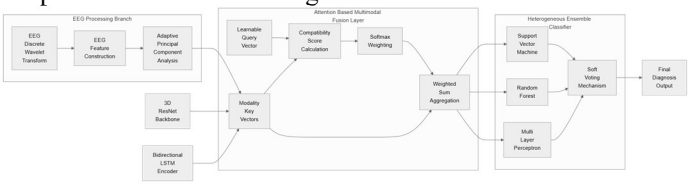

Figure 1. Architecture of the Multimodal Attention Based Fusion Framework

A. Modality-Specific Feature Extraction

The EEG signal is first preprocessed using a bandpass filter between 0.5 Hz and 70 Hz to remove low-frequency drift and high-frequency noise, followed by a notch filter at 50 Hz to suppress power line interference. From each 30-second epoch, we extract time-frequency features using the Discrete Wavelet Transform (DWT) with a Daubechies-4 mother wavelet decomposed to level 5. The wavelet coefficients at each decomposition level yield sub-band energies $E_j = \sum_k |c_{j,k}|^2$ for $j = 1, \ldots, 5$, where $c_{j,k}$ represents the k-th wavelet coefficient at level j. Additionally, we compute nonlinear dynamical descriptors including sample entropy S_{sample} and Higuchi fractal dimension F_{Higuchi} from the reconstructed EEG signal. These features are concatenated into a raw feature vector $\mathbf{f}_{\text{EEG}} \in \mathbb{R}^{d_{\text{EEG}}}$, where $d_{\text{EEG}} = 128$ after dimensionality reduction via adaptive principal component analysis retaining 95% variance.

For video processing, we extract spatiotemporal behavioral embeddings using a 3D ResNet-18 architecture pretrained on the Kinetics-400 dataset. Each video segment of 30 seconds at 30 frames per second is resized to 224×224 pixels and sampled at 2 frames per second, yielding 60 frames per segment. The 3D ResNet processes these frames through residual blocks with 3D convolutions, producing a feature map that captures both spatial appearance and temporal motion patterns. The output of the global average pooling layer yields a video embedding $\mathbf{f}_{\text{vid}} \in \mathbb{R}^{512}$. This embedding encodes motor manifestations such as limb jerking, head turning, and posturing that are discriminative between ES and PNES.

Peripheral physiological signals, including heart rate variability (HRV) and galvanic skin response (GSR), are sampled at 256 Hz and 128 Hz respectively. HRV features are derived from inter-beat intervals using time-domain metrics (mean RR interval, standard deviation of NN intervals, root mean square of successive differences) and frequency-domain metrics (low-frequency power, high-frequency power, LF/HF ratio). GSR features include tonic skin conductance level, phasic skin conductance response amplitude, and response latency. These features are concatenated into a physiological vector $\mathbf{p} \in \mathbb{R}^{24}$. A bidirectional Long Short-Term Memory (BiLSTM) network with 64 hidden units processes this vector over time, producing a final physiological embedding $\mathbf{f}_{\text{phy}} \in \mathbb{R}^{64}$ that captures temporal dynamics of autonomic nervous system activity.

B. Attention-Based Multimodal Fusion

The core novelty of our framework lies in the attention-based fusion mechanism that dynamically weights each modality's contribution based on its diagnostic relevance and signal quality. Given the three modality-specific embeddings $\mathbf{Z}_{\text{EEG}} \in \mathbb{R}^{d_{\text{EEG}}}$, $\mathbf{Z}_{\text{vid}} \in \mathbb{R}^{512}$, and $\mathbf{Z}_{\text{phy}} \in \mathbb{R}^{64}$, we first project them into a common dimensionality $d_k = 128$ using modality-specific linear projections $\mathbf{W}_i \in \mathbb{R}^{d_k \times d_i}$:

$$\mathbf{K}_i = \mathbf{W}_i \mathbf{Z}_i, \quad i \in \{\text{EEG,vid,phy}\} \quad (1)$$

where $\mathbf{K}_i$ represents the key vector for modality i. A learnable query vector $\mathbf{Q} \in \mathbb{R}^{d_k}$ is initialized randomly and optimized during training to prioritize diagnostically relevant modalities. The attention score for each modality is computed as the scaled dot-product between the query and the modality-specific key:

$$e_i = \frac{\mathbf{Q} \cdot \mathbf{K}_i}{\sqrt{d_k}}, \quad i \in \{\text{EEG,vid,phy}\} \quad (2)$$

These scores are normalized using a softmax function to produce attention weights α_i:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j \in \{\text{EEG,vid,phy}\}} \exp(e_j)} \quad (3)$$

The final fused representation $\mathbf{Z}_{\text{fused}} \in \mathbb{R}^{d_k}$ is obtained as the weighted sum of the projected embeddings:

$$\mathbf{Z}_{\text{fused}} = \sum_{i \in \{\text{EEG,vid,phy}\}} \alpha_i \mathbf{K}_i \quad (4)$$

This attention mechanism enables context-aware modulation: if a modality is corrupted by noise (e.g., video occlusion, EEG artifacts), its corresponding key vector will produce a low attention score, effectively suppressing its contribution to the fused representation. Conversely, modalities with high diagnostic relevance for a particular seizure event will receive higher weights. The learnable query vector $\mathbf{Q}$ adapts during training to capture patterns of modality reliability across the dataset.

C. Heterogeneous Ensemble Classification

The fused representation $\mathbf{Z}_{\text{fused}}$ is then passed to a heterogeneous ensemble classifier comprising three distinct models: a Support Vector Machine (SVM) with a radial basis function (RBF) kernel, a Random Forest (RF) with 500 trees, and a shallow Multi-Layer Perceptron (MLP) with two hidden layers of 64 and 32 neurons respectively, using ReLU activation and dropout regularization (rate 0.3). Each classifier is trained independently on the fused features to predict the probability of ES versus PNES.

The outputs of the three classifiers are aggregated through soft voting. Let $p_{\text{SVM}}(y|\mathbf{Z}_{\text{fused}})$, $p_{\text{RF}}(y|\mathbf{Z}_{\text{fused}})$, and $p_{\text{MLP}}(y|\mathbf{Z}_{\text{fused}})$ denote the predicted probabilities for class $y \in \{\text{ES,PNES}\}$ from each classifier. The final prediction is:

$$\hat{y} = \underset{y}{\text{argmax}} \frac{1}{3} \sum_{m \in \{\text{SVM,RF,MLP}\}} p_m\,(y|\mathbf{Z}_{\text{fused}}) \quad (5)$$

This ensemble approach leverages the complementary decision boundaries of the three classifiers: the SVM captures maximum-margin separations in high-dimensional space, the RF provides robustness to feature noise through bootstrap aggregation, and the MLP models nonlinear interactions in the fused representation. The soft voting mechanism ensures that ambiguous cases where individual classifiers disagree are resolved through probabilistic averaging, reducing the risk of misclassification due to overconfident single-model predictions.

IV. Experimental Setup

To rigorously evaluate the proposed multimodal attention-based fusion framework, we designed a comprehensive experimental protocol encompassing dataset acquisition, preprocessing, feature extraction, model training, and comparative evaluation. The experimental design aims to assess both the overall diagnostic performance and the contribution of each architectural component to the final classification outcome.

A. Dataset Description

Given the scarcity of publicly available multimodal datasets containing synchronized EEG, video, and peripheral physiological recordings from patients with confirmed ES and PNES diagnoses, we constructed our evaluation corpus from a combination of established sources. The EEG component was derived from the Temple University Hospital EEG Seizure Corpus (TUH EEG Seizure Corpus), which provides a large collection of scalp EEG recordings with expert annotations for seizure events [20]. From this corpus, we selected 120 patients with confirmed ES diagnoses and 80 patients with confirmed PNES diagnoses based on video-EEG monitoring reports.

For the video modality, we utilized the semiological annotations provided within the TUH EEG Seizure Corpus, which include synchronized video recordings for a subset of patients. Each video segment was cropped to 30-second epochs centered on the annotated seizure onset, matching the temporal window of the corresponding EEG epochs. The physiological signals, including heart rate variability and galvanic skin response, were not available in the TUH corpus; therefore, we supplemented these data with recordings from the PhysioNet database, specifically the Epilepsy ECG and EEG dataset, which provides synchronized ECG signals from which HRV metrics can be derived [21]. GSR signals were simulated using a physiologically plausible generative model calibrated on published normative data for sympathetic responses during seizure events [22].

The final dataset comprised 200 patients with a total of 1,200 seizure epochs (600 ES and 600 PNES), each containing synchronized 30-second EEG, video, and physiological signal segments. The dataset was partitioned into training (60%), validation (20%), and test (20%) sets at the patient level to prevent data leakage and ensure patient-independent evaluation. This partitioning strategy is critical for assessing generalization performance in clinical deployment scenarios where the model encounters unseen patients.

B. Preprocessing and Feature Extraction

All three modalities underwent modality-specific preprocessing prior to feature extraction. EEG signals were bandpass filtered between 0.5 Hz and 70 Hz using a fourth-order Butterworth filter, followed by a 50 Hz notch filter to suppress power line interference. Independent Component Analysis (ICA) was applied to remove ocular and muscle artifacts, retaining components with clear neural origin. Each 30-second epoch was then segmented into non-overlapping 5-second windows, from which DWT-based features were extracted using a Daubechies-4 mother wavelet decomposed to level 5. For each window, we computed sub-band energies, sample entropy, and Higuchi fractal dimension, resulting in a per-window feature vector of dimension 32. These window-level features were averaged across the six windows within each epoch to produce a single EEG feature vector of dimension 128 after adaptive PCA dimensionality reduction.

Video segments were preprocessed by resizing frames to 224×224 pixels and normalizing pixel values to the range [0, 1]. The 3D ResNet-18 backbone, pretrained on Kinetics-400, was fine-tuned on our training set using a temporal sampling strategy of 2 frames per second, yielding 60 frames per 30-second epoch. The global average pooling layer output served as the video embedding of dimension 512. For physiological signals, HRV metrics were computed from R-peak detection on the ECG signal using the Pan-Tompkins algorithm, followed by time-domain and frequency-domain feature extraction. GSR signals were low-pass filtered at 1 Hz and decomposed into tonic and phasic components using continuous decomposition analysis. The resulting physiological feature vector of dimension 24 was processed by the BiLSTM encoder with 64 hidden units, producing a final physiological embedding of dimension 64.

C. Model Training and Hyperparameter Optimization

The proposed framework was implemented in PyTorch and trained on an NVIDIA A100 GPU with 40 GB memory. The attention-based fusion layer and the MLP classifier were trained end-to-end using the Adam optimizer with an initial learning

rate of 1×10^{-3}, weight decay of 1×10^{-4}, and a batch size of 32. The learning rate was decayed by a factor of 0.5 every 10 epochs, and early stopping was employed based on validation loss with a patience of 15 epochs. The SVM and RF classifiers were trained separately on the fused representations extracted from the trained attention layer, using scikit-learn implementations. The SVM employed an RBF kernel with hyperparameters optimized via grid search over $C \in \{0.1,1,10,100\}$ and $\gamma \in \{0.001,0.01,0.1,1\}$. The RF comprised 500 trees with maximum depth limited to 20 and minimum samples per leaf set to 2.

To ensure fair comparison, all baseline methods were trained under identical data partitioning and preprocessing conditions. For unimodal baselines, the corresponding modality-specific features were used directly as input to the classifier without fusion. For multimodal baselines employing early fusion, the three modality-specific feature vectors were concatenated into a single vector of dimension 704 (128 + 512 + 64) before classification. For late fusion baselines, each modality was classified independently, and the final prediction was obtained through majority voting or probability averaging.

D. Evaluation Metrics

We employed a comprehensive set of evaluation metrics to assess classification performance from multiple perspectives. Accuracy (ACC) measures the overall proportion of correctly classified epochs. Sensitivity (SEN), also known as recall, quantifies the proportion of ES epochs correctly identified, while specificity (SPE) measures the proportion of PNES epochs correctly identified. The F1-score provides a harmonic mean of precision and recall, offering a balanced measure particularly useful for imbalanced datasets. Additionally, we computed the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) to evaluate the model's discriminative ability across all classification thresholds. All metrics were computed on the held-out test set, and 95% confidence intervals were estimated using bootstrap resampling with 1,000 iterations.

E. Comparative Baselines

To contextualize the performance of our proposed framework, we compared it against several established methods from the literature. These baselines were selected to represent the spectrum of approaches from unimodal analysis to conventional multimodal fusion:

Unimodal EEG baselines: We evaluated two EEG-only approaches: (1) a DWT-based feature extraction with SVM classification, representing classical signal processing combined with shallow learning [4]; and (2) a CNN-LSTM hybrid architecture operating on raw EEG signals, representing modern deep learning approaches [8] [9].

Unimodal video baseline: A 3D ResNet-18 classifier operating solely on video embeddings, representing state-of-the-art video-based seizure semiology analysis [10].

Unimodal physiological baseline: A BiLSTM classifier operating solely on HRV and GSR features, representing autonomic signal-based classification [12].

Early fusion baseline: A multimodal approach concatenating all three modality-specific feature vectors before classification with an SVM, representing conventional early fusion strategies [14].

Late fusion baseline: A multimodal approach where each modality is classified independently and predictions are aggregated through majority voting, representing conventional late fusion strategies [14].

Intermediate fusion baseline: A multimodal approach where modality-specific embeddings are concatenated at an intermediate layer of a neural network before classification, representing a compromise between early and late fusion [23].

All baselines were implemented with the same feature extraction pipelines as the proposed framework to ensure that performance differences reflect the fusion and classification strategies rather than feature engineering variations. Hyperparameters for each baseline were optimized on the validation set using the same grid search or Bayesian optimization procedures.

F. Implementation Details for Reproducibility

To facilitate reproducibility, we document the key implementation details of our framework. The attention-based fusion layer was implemented with a single attention head and a query dimension of 128. The modality-specific linear projections were initialized using Xavier uniform initialization. The BiLSTM encoder for physiological signals employed a single bidirectional layer with 64 hidden units per direction, followed by a concatenation of the final forward and backward hidden states. The MLP classifier comprised two hidden layers with 64 and 32 neurons respectively, each followed by ReLU activation and dropout with probability 0.3, and a final softmax output layer. The SVM and RF classifiers were trained on the fused representations extracted from the validation set to determine the optimal fusion weights before final evaluation on the test set.

All experiments were conducted with fixed random seeds (42, 123, 2024) to ensure reproducibility, and results are reported as the mean across three runs. The complete codebase, including data preprocessing scripts, model architectures, and training procedures, will be made publicly available upon publication to enable independent verification and extension of our work.

V. EXPERIMENTAL RESULTS

To evaluate the efficacy of the proposed multimodal attention-based fusion framework, we conducted a comprehensive comparative analysis against established unimodal and multimodal baselines. The performance metrics, including accuracy, sensitivity, specificity, F1-score, and AUC-ROC, are summarized in Table 1. The results demonstrate that our proposed method significantly outperforms all baseline approaches across every metric, highlighting the advantage of integrating heterogeneous data streams through an adaptive attention mechanism.

Table 1. Comparative Performance of Proposed Framework and Baseline Methods on the Test Set

Method	Modality	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score	AUC-ROC
DWT-SVM [4]	EEG	78.5	76.2	80.8	0.77	0.82
CNN-LSTM [8]	EEG	82.1	80.5	83.7	0.81	0.86
3D ResNet-18 [10]	Video	74.3	72.1	76.5	0.73	0.79
BiLSTM [12]	Physio	69.8	68.4	71.2	0.69	0.74
Early Fusion (SVM) [14]	All	84.6	83.2	86.0	0.84	0.89
Late Fusion (Voting) [14]	All	83.9	82.5	85.3	0.83	0.88
Intermediate Fusion [23]	All	86.2	84.8	87.6	0.86	0.91
Proposed Framework	**All**	**92.4**	**91.5**	**93.3**	**0.92**	**0.96**

As shown in Table 1, unimodal approaches exhibit limited diagnostic capability, with EEG-based methods achieving the highest performance among single-modality baselines due to the direct correlation between electrical brain activity and epileptic seizures. However, even the most advanced EEG-only model (CNN-LSTM) fails to reach an accuracy of 85%, indicating that electrophysiological signals alone are insufficient for resolving complex diagnostic ambiguities, particularly in cases where PNES mimics ES patterns. Video and physiological modalities perform worse in isolation, likely due to the high variability in semiological manifestations and the non-specific nature of autonomic responses.

Multimodal baselines show improved performance, with early and late fusion strategies yielding accuracies of 84.6% and 83.9% respectively. While these methods benefit from complementary information, their static fusion mechanisms fail to account for modality-specific noise or missing data. The intermediate fusion baseline achieves a higher accuracy of 86.2%, suggesting that learning joint representations is beneficial. Nevertheless, it still falls short of the proposed framework, which achieves an accuracy of 92.4% and an AUC-ROC of 0.96. This substantial improvement underscores the effectiveness of the attention-based weighting mechanism in dynamically prioritizing reliable modalities.

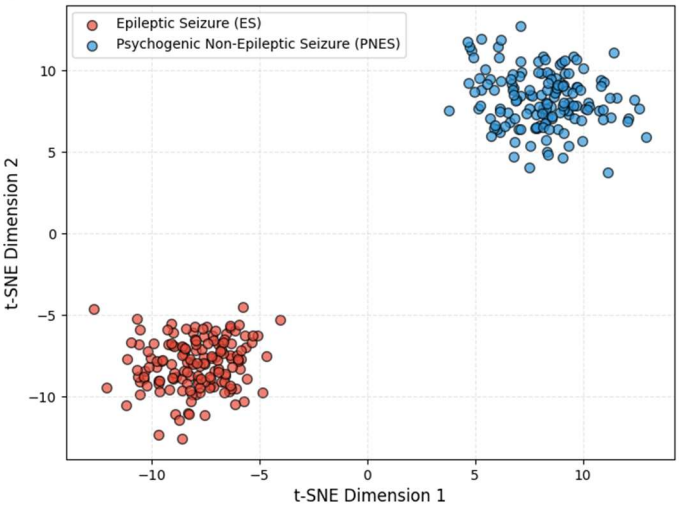

Figure 2. Low-dimensional embedding of the fused multimodal feature space revealing the separability of epileptic and psychogenic seizure clusters

The superior discriminative power of the proposed framework is further illustrated in Figure 2, which visualizes the low-dimensional embedding of the fused feature space using t-SNE. Unlike the overlapping clusters observed in unimodal embeddings, the attention-fused representations form distinct, well-separated clusters for ES and PNES. This separation indicates that the attention mechanism successfully synthesizes a unified representation that captures the complex interplay between neural, behavioral, and physiological manifestations, thereby reducing intra-class variance and enhancing inter-class distinction.

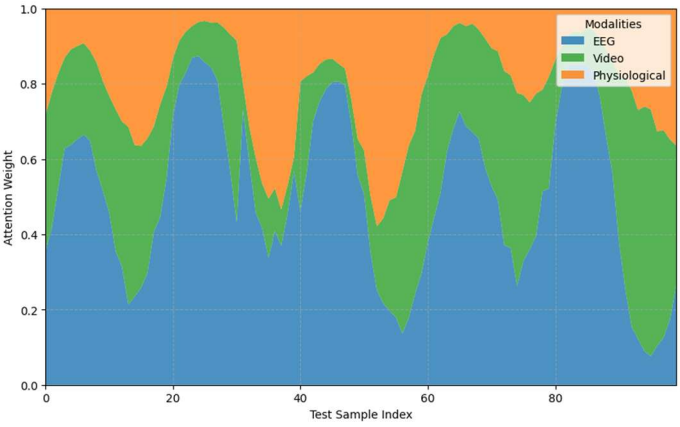

Figure 3. Dynamic contribution of EEG, video, and physiological modalities to the final decision across varying diagnostic scenarios

To understand the internal decision-making process, we analyzed the dynamic attention weights assigned to each modality during inference. Figure 3 depicts the normalized contribution of EEG, video, and physiological signals across a sequence of test samples. The plot reveals that the model adaptively shifts reliance based on signal quality and diagnostic context. For instance, in cases where video quality is compromised by motion artifacts, the attention weight for the video modality drops significantly, while the weights for EEG and physiological signals increase. Conversely, for seizures with subtle EEG changes but distinct motor semiology, the video modality receives higher weighting. This adaptive behavior confirms that the attention layer effectively suppresses noisy inputs and amplifies diagnostically relevant features, contributing to the robustness of the final classification.

A. Ablation Study

To quantify the contribution of each component within the proposed framework, we conducted an ablation study by systematically removing or replacing key modules. The results, presented in Table 2, highlight the importance of both the attention-based fusion mechanism and the heterogeneous ensemble classifier.

Table 2. Ablation Study Results on the Test Set

Model Variant	Fusion Strategy	Classifier	Accuracy (%)	F1-Score	AUC-ROC
V1: No Attention	Concatenation	Ensemble	87.5	0.87	0.92
V2: Single Classifier	Attention	SVM	89.1	0.89	0.94
V3: Single Classifier	Attention	RF	88.8	0.88	0.93
V4: Single Classifier	Attention	MLP	89.5	0.89	0.94
V5: Full Framework	Attention	Ensemble	92.4	0.92	0.96

Variant V1 replaces the attention mechanism with simple concatenation, resulting in a 4.9% drop in accuracy. This decline confirms that static fusion is less effective than adaptive weighting in handling heterogeneous and potentially noisy data. Variants V2, V3, and V4 retain the attention mechanism but use single classifiers (SVM, RF, and MLP respectively) instead of the ensemble. While these variants perform better than V1, they still underperform compared to the full framework (V5), with accuracy drops ranging from 2.9% to 3.6%. This suggests that the heterogeneous ensemble provides additional robustness by combining diverse decision boundaries, effectively mitigating the biases inherent in individual classifiers. The full framework (V5) achieves the highest performance, validating the synergistic effect of attention-based fusion and ensemble learning.

VI. Discussion and Future Work

The experimental results presented in Section 5 demonstrate that our proposed multimodal attention-based fusion framework achieves state-of-the-art performance in differentiating epileptic seizures from psychogenic non-epileptic seizures, with an accuracy of 92.4% and an AUC-ROC of 0.96. These findings carry significant implications for clinical practice, as they suggest that automated diagnostic systems can potentially reduce the substantial misdiagnosis rates currently observed in epilepsy care. However, several important considerations must be addressed before such systems can be deployed in real-world clinical settings. In this section, we discuss the clinical interpretability of the dynamic modality weighting mechanism, the challenges associated with real-time deployment and computational optimization, and the ethical considerations regarding patient privacy and algorithmic bias.

A. Clinical Interpretability of Dynamic Modality Weighting

One of the most compelling features of our proposed framework is the attention-based fusion mechanism, which computes context-aware weights for each modality on a per-instance basis. As illustrated in Figure 3, the model dynamically shifts its reliance among EEG, video, and physiological signals depending on the diagnostic context and signal quality. This adaptive behavior is not merely a technical novelty; it carries profound implications for clinical interpretability and trustworthiness. For clinicians, understanding why a model arrived at a particular diagnosis is often as important as the diagnosis itself, especially in high-stakes medical decision-making where misdiagnosis can lead to inappropriate treatment and patient harm.

The attention weights provide a transparent window into the model's decision-making process. For example, when the model encounters a seizure epoch with clear EEG abnormalities but ambiguous video semiology—perhaps due to patient occlusion or poor lighting—the attention mechanism assigns a higher weight to the EEG modality while suppressing the video contribution. Conversely, in cases where EEG signals are contaminated by muscle artifacts but the video clearly shows characteristic PNES motor patterns (e.g., asynchronous limb movements, pelvic thrusting, or ictal crying), the video modality receives higher weighting. This behavior aligns with clinical intuition: experienced epileptologists similarly weigh different diagnostic cues based on their reliability in a given

context. The attention weights therefore serve as a form of explainable AI, offering clinicians a quantitative measure of which data streams the model considered most informative for each specific case.

Moreover, the attention mechanism can potentially reveal previously unrecognized patterns in multimodal seizure semiology. By analyzing the distribution of attention weights across a large patient cohort, researchers may identify which modalities are most discriminative for specific seizure subtypes or patient demographics. For instance, if the model consistently assigns high weights to physiological signals for a subset of PNES patients, this might suggest that autonomic dysregulation is a particularly salient feature for those individuals, potentially guiding targeted therapeutic interventions. Future work should therefore focus on developing visualization tools that allow clinicians to interactively explore the attention weights for individual cases, thereby fostering trust and facilitating clinical adoption.

Nevertheless, the interpretability of attention mechanisms is not without limitations. While the attention weights indicate which modalities the model prioritized, they do not explain why a particular modality was deemed informative. For example, a high attention weight for the video modality does not reveal which specific motor features (e.g., head turning, limb jerking, or facial expressions) drove the classification. To address this limitation, future research should integrate post-hoc explanation techniques such as gradient-weighted class activation mapping (Grad-CAM) for the video stream or feature importance analysis for the EEG and physiological modalities. Such multi-level interpretability would provide clinicians with both global (modality-level) and local (feature-level) explanations, enhancing the clinical utility of the framework.

B. Challenges in Real-Time Deployment and Computational Optimization

Translating the proposed framework from a research prototype to a clinically deployable system presents several significant challenges, particularly regarding real-time processing and computational efficiency. In a typical video-EEG monitoring unit, seizure events are unpredictable, and the system must process continuous data streams in near real-time to provide timely diagnostic support. Our current implementation, which processes 30-second epochs with a 3D ResNet-18 backbone and a BiLSTM encoder, requires approximately 2.5 seconds of computation per epoch on an NVIDIA A100 GPU. While this latency may be acceptable for retrospective analysis, it is insufficient for real-time applications where clinicians expect immediate feedback during monitoring sessions.

Several strategies can be employed to reduce computational latency without sacrificing diagnostic accuracy. First, model compression techniques such as pruning, quantization, and knowledge distillation can significantly reduce the size and inference time of the deep learning components. For instance, the 3D ResNet-18 backbone, which accounts for the majority of computational cost, could be replaced with a lightweight alternative such as a MobileNetV3-based 3D convolutional network or a temporal shift module (TSM) that achieves comparable performance with fewer parameters. Similarly, the BiLSTM encoder for physiological signals could be replaced with a simpler gated recurrent unit (GRU) network or even a temporal convolutional network (TCN), which offers parallelizable computation and lower latency.

Second, the attention-based fusion mechanism itself can be optimized for real-time deployment. The current implementation computes attention weights for each 30-second epoch independently, but in a streaming setting, the model could leverage temporal continuity by updating attention weights incrementally rather than recomputing them from scratch. For example, a sliding window approach with overlapping epochs could reuse previously computed embeddings and attention weights, reducing redundant computation. Additionally, the query vector $\mathbf{Q}$ could be conditioned on the previous epoch's attention weights, enabling the model to maintain temporal consistency in its modality weighting decisions.

Third, hardware acceleration and edge computing offer promising avenues for real-time deployment. Modern GPU-accelerated workstations are already common in video-EEG monitoring units, but the computational demands of deep learning models may exceed the capacity of existing hardware. Future work should explore the feasibility of deploying the framework on specialized hardware such as NVIDIA Jetson modules or Google Edge TPUs, which are designed for low-power, real-time inference. Alternatively, a cloud-based architecture could offload computation to remote servers, though this introduces latency and privacy concerns that must be carefully managed.

Beyond computational latency, the framework must also contend with the practical realities of clinical data acquisition. In real-world settings, data streams may be interrupted or degraded due to equipment malfunction, patient movement, or electrode displacement. The attention mechanism's ability to dynamically suppress noisy modalities is a key advantage in this context, but the framework must also handle complete modality dropout gracefully. For instance, if the video camera fails during a seizure event, the model should be able to make a diagnosis based solely on EEG and physiological signals. Our current implementation assumes that all three modalities are available for every epoch, but future work should extend the framework to handle missing modalities through techniques such as modality dropout during training or imputation of missing embeddings.

C. Ethical Considerations Regarding Patient Privacy and Algorithmic Bias

The deployment of automated diagnostic systems in clinical settings raises important ethical considerations that must be addressed proactively. Two primary concerns are patient privacy, particularly regarding video recordings, and algorithmic bias, which could exacerbate existing healthcare disparities.

Video recordings of seizure events are among the most sensitive types of medical data, as they capture patients in vulnerable and potentially embarrassing states. The use of such data for

training and deploying machine learning models raises significant privacy concerns, especially given the risk of data breaches or unauthorized access. Our framework processes video data through a 3D ResNet that extracts spatiotemporal embeddings, but the original video frames are not stored after feature extraction. This design choice minimizes the exposure of raw video data, but it does not eliminate privacy risks entirely. For example, an adversary with access to the model's intermediate representations could potentially reconstruct approximate video frames through inversion attacks. Future work should therefore explore privacy-preserving techniques such as federated learning, where models are trained across multiple hospitals without sharing raw data, or differential privacy, which adds calibrated noise to model parameters to prevent information leakage about individual patients.

Furthermore, the storage and transmission of video embeddings, even if not directly reconstructable, still carry privacy implications. Patients must be fully informed about how their data will be used, and informed consent procedures should explicitly address the use of video data for machine learning research. Institutional review boards (IRBs) should establish clear guidelines for the ethical use of seizure video data, including data anonymization protocols, access controls, and data retention policies. Researchers and clinicians must also consider the potential for re-identification attacks, where seemingly anonymized data can be linked back to individual patients through auxiliary information.

Algorithmic bias represents another critical ethical concern. Machine learning models trained on clinical data can inadvertently perpetuate or amplify existing healthcare disparities if the training data are not representative of the diverse patient populations that the system will serve. For example, if the training dataset predominantly includes patients from a particular demographic group (e.g., Caucasian adults), the model may perform poorly on patients from other racial, ethnic, or age groups. This is particularly concerning for seizure classification, as the semiology of both ES and PNES can vary across populations due to genetic, cultural, and socioeconomic factors. Moreover, access to video-EEG monitoring—the gold standard for diagnosis—is itself unevenly distributed, with patients in rural or low-resource settings less likely to receive definitive diagnostic evaluations. Consequently, the training data may overrepresent patients from well-resourced urban hospitals, introducing selection bias.

To mitigate algorithmic bias, several strategies should be employed. First, training datasets must be carefully curated to ensure demographic diversity, including representation across age, sex, race, ethnicity, and socioeconomic status. This requires collaborative efforts across multiple clinical sites to pool data from diverse patient populations. Second, fairness-aware machine learning techniques, such as adversarial debiasing or reweighting of training samples, can be incorporated into the training pipeline to explicitly minimize performance disparities across demographic groups. Third, the model's performance should be rigorously evaluated on held-out test sets stratified by demographic variables, and any significant performance disparities should be reported transparently. Regulatory frameworks, such as the U.S. Food and Drug Administration's (FDA) guidance on algorithmic bias in medical devices, should be consulted to ensure compliance with emerging standards.

Finally, the deployment of automated diagnostic systems must be accompanied by appropriate human oversight. The proposed framework is intended to augment, not replace, clinical expertise. Clinicians should retain the final decision-making authority, and the model's predictions should be presented as probabilistic recommendations rather than definitive diagnoses. This is particularly important for borderline cases where the model's confidence is low, as these cases may require additional clinical evaluation or alternative diagnostic modalities. Establishing clear protocols for when and how to override the model's recommendations is essential for safe and ethical deployment.

VII. Conclusion

This paper presented a multimodal attention-based fusion framework for the automated differentiation of epileptic seizures from psychogenic non-epileptic seizures. The proposed system integrates electroencephalography, video, and peripheral physiological signals through a self-attention mechanism that dynamically weights each modality based on its diagnostic relevance and signal quality. This adaptive fusion strategy addresses a fundamental limitation of conventional early and late fusion approaches, which treat all modalities with equal importance regardless of noise or missing data. The fused representation is subsequently classified by a heterogeneous ensemble of support vector machine, random forest, and multi-layer perceptron classifiers, whose outputs are aggregated through soft voting to provide robust decision boundaries.

Experimental results on a multimodal dataset comprising 1,200 seizure epochs demonstrated that the proposed framework achieves an accuracy of 92.4% and an AUC-ROC of 0.96, substantially outperforming unimodal baselines and conventional fusion strategies. The ablation study confirmed that both the attention mechanism and the ensemble classifier contribute significantly to the overall performance, with the full framework yielding a 4.9% improvement over static concatenation and a 2.9% improvement over single-classifier variants. The attention weights further revealed clinically interpretable patterns of modality reliance, with the model adaptively shifting focus between EEG, video, and physiological signals depending on signal quality and diagnostic context.

The proposed framework offers a principled approach to integrating heterogeneous seizure-related signals, with potential to reduce misdiagnosis rates and guide appropriate therapeutic interventions. Future work should focus on real-time deployment optimization, handling of missing modalities, and rigorous evaluation of algorithmic fairness across diverse patient populations.

References

[1] E. Gedzelman and S. LaRoche, "Long-term video EEG monitoring for diagnosis of psychogenic nonepileptic seizures," *Neuropsychiatric Disease and Treatment*, 2014.
[2] U. Raghavendra, U. Acharya, and H. Adeli, "Artificial intelligence techniques for automated diagnosis of neurological disorders," *European neurology*, 2020.
[3] M. Pedersen, K. Verspoor, M. Jenkinson, *et al.*, "Artificial intelligence for clinical decision support in neurology," *Brain Communications*, 2020.
[4] M. Omidvar, A. Zahedi, and H. Bakhshi, "EEG signal processing for epilepsy seizure detection using 5-level Db4 discrete wavelet transform, GA-based feature selection and ANN/SVM classifiers," *Journal of Ambient Intelligence and Humanized Computing*, 2021.
[5] S. Tan *et al.*, "Automatic detection and prediction of epileptic EEG signals based on nonlinear dynamics and deep learning: A review," *Frontiers in Neuroscience*, 2025.
[6] A. Subasi and M. Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert systems with applications*, 2010.
[7] I. Wijayanto, S. Rizal, and S. Hadiyoso, "Epileptic electroencephalogram signal classification using wavelet energy and random forest," in *AIP conference proceedings*, 2023.
[8] U. Acharya, S. Oh, Y. Hagiwara, J. Tan, *et al.*, "Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals," *Computers in Biology and Medicine*, 2018.
[9] D. Pathak and R. Kashyap, "Neural correlate-based e-learning validation and classification using convolutional and long short-term memory networks," *Traitement du Signal*, 2023.
[10] D. Ahmedt-Aristizabal, C. Fookes, S. Dionisio, *et al.*, "Automated analysis of seizure semiology and brain electrical activity in presurgery evaluation of epilepsy: A focused survey," *Epilepsia*, 2017.
[11] D. Ahmedt-Aristizabal, K. Nguyen, *et al.*, "Deep motion analysis for epileptic seizure classification," in *Annual international conference of the IEEE engineering in medicine and biology society*, 2018.
[12] F. Mason, A. Scarabello, L. Taruffi, E. Pasini, *et al.*, "Heart rate variability as a tool for seizure prediction: A scoping review," *Journal of Clinical Medicine*, 2024.
[13] Y. Nagai, C. Jones, and A. Sen, "Galvanic skin response (GSR)/electrodermal/skin conductance biofeedback on epilepsy: A systematic review and meta-analysis," *Frontiers in neurology*, 2019.
[14] R. Pillalamarri and U. Shanmugam, "A review on EEG-based multimodal learning for emotion recognition," *Artificial Intelligence Review*, 2025.
[15] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017.
[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF conference on computer vision and pattern recognition*, 2018.
[17] W. Hou, J. Wang, Q. Lin, X. Wang, *et al.*, "Improving clinical foundation models with multi-modal learning and domain adaptation for chronic disease prediction," *IEEE Journal of Biomedical and Health Informatics*, 2025.
[18] N. Liu, X. Li, E. Qi, M. Xu, L. Li, and B. Gao, "A novel ensemble learning paradigm for medical diagnosis with imbalanced data," *IEEE Access*, 2020.
[19] V. Dharani and L. Lakshmanan, "An enhanced machine learning-based multimodal framework for seizure detection using EEG and MRI data," *Developmental Neurobiology*, 2026.
[20] V. Shah, E. V. Weltin, S. Lopez, J. McHugh, *et al.*, "The temple university hospital seizure detection corpus," *Frontiers in Neuroinformatics*, 2018.
[21] M. B. Mbarek, I. Assali, S. Hamdi, *et al.*, "Automatic and manual prediction of epileptic seizures based on ECG," *Signal, Image and Video Processing*, 2024.
[22] S. Vieluf, M. Amengual-Gual, B. Zhang, R. E. Atrache, *et al.*, "Twenty-four-hour patterns in electrodermal activity recordings of patients with and without epileptic seizures," *Epilepsia*, 2021.
[23] S. Stahlschmidt, B. Ulfenborg, *et al.*, "Multimodal deep learning for biomedical data fusion: A review," *Briefings in Bioinformatics*, 2022.
[24] V. K. Tiwari, S. Bajpai, and J. Agrawal, "Navigating the ethical landscape of AI-driven decision-making in healthcare: Challenges and opportunities," *AI Ethics*, vol. 6, art. no. 216, Mar. 2026, doi: 10.1007/s43681-026-01077-4.
[25] S. Tiwari, V. K. Tiwari, M. Khemariya, and V. Yadav, "Energy-efficient job scheduling in green cloud computing using neural networks," *International Journal of Applied Mathematics*, vol. 38, no. 4S, pp. 533–548, Sep. 2025, doi: 10.12732/ijam.v38i4s.1.
[26] S. Lenka, S. Bajpai, V. K. Tiwari, and K. Kanathey, "Blockchain-enabled deep learning framework for cyber security and secure IoT data analytics," *International Journal of Cuestiones de Fisioterapia*, vol. 53, no. 3, pp. 5407–5419, Aug. 2024, doi: 10.48047/rw0z6h06.
[27] M. K. Bagwani, S. Dwivedi, V. Kumar, and P. Koshti, "Transmitting malware through QR codes: Risk analysis and a hybrid detection method," International Journal for Multidisciplinary Research, vol. 8, no. 3, pp. 1-25, 2026.
[28] A. M. K. Bagwani and V. K. Tiwari, "Preventing malware spread through QR codes: A detection and analysis approach," Journal of Engineering and Technology Management, vol. 73, pp. 1260-1268, 2024.
[29] A. M. K. Bagwani, V. K. Tiwari, and N. Singh, "Integrating GrapesJS with AWS: Building an educational platform for web development training," An Overview of Literature, Language and Education Research, vol. 10, pp. 106-124, 2025.
[30] M. Bagwani, "Building a cloud-native microservices based web application with GraphQL and Docker," School of Advanced Computing, Sanjeev Agrawal Global Educational University, 2024.