

Article

CMEpiNet: Complex-Valued Multimodal Epilepsy Detection Network Model

Tianyi Su ^{1,†}, Haiyan Zhu ^{2,†}, Shuai Chen ^{3,*} and Haifeng Wang ^{3,*} 

¹ Department of Electrical and Information Engineering, Shandong University of Science and Technology, Jinan 250031, China; sutianyi@sdust.edu.cn

² School of Physical Education and Health Linyi, Linyi University, Linyi 276012, China; zhuhaiyan@lyu.edu.cn

³ School of Information Science and Engineering, Linyi University, Linyi 276012, China

* Correspondence: shuaichen@lyu.edu.cn (S.C.); wanghaifeng@lyu.edu.cn (H.W.)

† These authors contributed equally to this work.

Abstract

Existing seizure detection methods cannot fully exploit the spatiotemporal features of multimodal signals. They also fail to capture deep associations among cross-modal features. This limits their ability to learn unified representations of spatiotemporal dependencies. This work proposes CMEpiNet (Complex-valued Multimodal Epilepsy detection Network model) to address this issue. CMEpiNet first uses complex-valued convolutions for feature extraction. It explicitly models phase synchronization, phase shifts, and cross-frequency coupling. Thus, EEG, ECG, and EMG features are represented in the complex-valued domain. During feature fusion, CMEpiNet uses a two-level semantic alignment-based fusion method. It applies cross-modal consistency constraints in a shared alignment space. It also performs distribution-level alignment in an epilepsy-related semantic latent space. These operations ensure the consistency of multimodal features in the global semantic structure. Finally, CMEpiNet uses a spatial attention-guided 3D convolutional classifier. The classifier jointly models the temporal, feature, and modality dimensions. Experimental results on the SeizureIT2 dataset show that CMEpiNet improves seizure detection sensitivity, reduces the false alarm rate, and maintains stable performance under perturbations.

Keywords: feature fusion; machine learning; deep learning; epilepsy detection; classification

1. Introduction

Epilepsy is a common chronic neurological disorder characterized by seizures that are inherently unpredictable [1]. Globally, approximately 30% of epilepsy patients experience uncontrolled seizures despite adherence to medication, and consequently face an elevated risk of sudden unexpected death (SUDEP) [2]. Electroencephalography (EEG) reflects the brain electrophysiological activity and has become a key clinical tool for observing seizures [3]. However, manual EEG interpretation is inherently subjective and dependent on scarce human expertise [4]. This manual process is not only labor-intensive but also error-prone; the need for prolonged review introduces fatigue, which fundamentally limits diagnostic accuracy and consistency. Therefore, AI-supported automatic epilepsy diagnosis has become an important solution [5].

Current automated seizure detection methods primarily rely on single-modality EEG signals. EEG signals mainly reflect cerebral electrical activity. Therefore, they provide limited information on interactions with other physiological changes, such as abnormal increases in heart rate and alterations in muscle activity. However, other physiological



Academic Editor: Alan Jović

Received: 1 May 2026

Revised: 25 June 2026

Accepted: 27 June 2026

Published: 2 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

signals can contribute to early warning of epileptic seizures. Therefore, single-modality detection methods have clear limitations. Researchers use multimodal fusion to fully exploit physiological information from EEG, ECG, and EMG. Therefore, multimodal fusion-based seizure detection has become a new research trend. Multimodal signals can characterize epileptic seizure processes from multiple dimensions. These dimensions include neural electrical activity, cardiovascular responses, and muscle movements. They provide more comprehensive and complementary evidence for modeling complex physiological events. However, multimodal physiological signals differ significantly in noise characteristics, signal quality, and time–frequency structures. Under wearable acquisition conditions, they are susceptible to motion artifacts and poor electrode contact. Achieving robust multimodal feature fusion within a unified framework remains challenging, with two key issues remaining. (i) Existing studies often ignore deep oscillatory features related to phase and time–frequency structures in multimodal signals. This limits their ability to extract discriminative signal features. (ii) Cross-modal feature fusion lacks effective modeling of semantic consistency and distribution differences among modalities. This reduces model robustness under signal quality fluctuations and temporal asynchrony across modalities. To address the above issues, this work proposes CMEpiNet, a multimodal seizure detection model. CMEpiNet provides a multimodal feature optimization scheme that integrates complex-valued modeling, semantic alignment, and three-dimensional structural modeling. First, the complex-valued multimodal feature extraction module extracts time–domain features, phase features, and time–frequency structural features from analytic signals and complex-valued time–frequency spectra. This module fully uses the complementary information in amplitude, phase, and frequency-band energy distributions of multimodal signals. Second, the two-level semantic alignment fusion module performs cross-modal consistency constraints and shared latent space alignment. It reduces semantic and distributional biases among different modalities. Thus, the model obtains a more compact and consistent multimodal fusion representation. Finally, the classifier uses three-dimensional convolution and spatial attention. It jointly models the temporal, feature, and modality dimensions of the fused three-dimensional feature block. The main contributions are as follows.

(1) A complex-valued multimodal feature extraction method. This method uses complex-valued convolutions to explicitly model the coupling between signal amplitude and phase. It overcomes the limitation of traditional multimodal feature extraction methods that ignore phase information. It builds a dual-branch structure for time-domain-phase features and time-frequency features. In a unified complex-valued domain, it captures oscillatory features such as phase synchronization, phase shifts, and cross-frequency coupling. It unifies the representation forms of EEG, ECG, and EMG. It also extracts highly discriminative signal features.

(2) A two-level semantic alignment fusion method. This method constrains the multimodal fusion process from two aspects: representation-level consistency and distribution-level consistency. In the shared alignment space, it uses window-level soft matching and reliability gating to reduce semantic bias caused by temporal asynchrony and quality fluctuations across modalities. In the seizure-related semantic latent space, it aligns the global statistical structures of multimodal features. This reduces cross-modal distribution shifts. As a result, the model obtains stable cross-modal fusion representations under signal quality degradation and temporal asynchrony across modalities.

The rest of this paper is organized as follows. First, we introduce the multimodal seizure signal data and the corresponding preprocessing methods. Then, we describe the overall architecture and key module designs of the proposed complex-valued multimodal seizure detection model, CMEpiNet. Next, we evaluate the model through several com-

parative experiments and ablation studies. These experiments verify the effectiveness and robustness of each module in multimodal seizure detection. Finally, we present the conclusion and future work.

2. Related Work

Research on automated epilepsy detection has evolved from handcrafted rule-based pipelines to data-driven learning frameworks, driven largely by differences in feature construction, temporal modeling, and multimodal fusion mechanisms. Existing studies can be broadly grouped into four categories: statistical and nonlinear signal-based methods, classical machine-learning-based methods, deep-learning-based temporal and spatial modeling methods, and multimodal seizure detection methods.

(1) Statistical signal-based epilepsy detection methods

Early research on automatic EEG-based seizure detection was hindered by the challenge of objectively quantifying the complex and dynamic nature of brain signals. Conventional approaches operated in two stages: first, statistical properties were extracted from the time, frequency, or time–frequency domain; then, these were used to construct feature indicators with clear physical or statistical meaning. Seizure events were subsequently identified via thresholding or rule-based matching. Hjorth [6] introduced time-domain statistical parameters such as Activity, Mobility, and Complexity to characterize EEG energy level, temporal scale, and structural complexity, laying the foundation for subsequent statistical feature modeling. To address the limited spectral resolution of conventional Fourier analysis within short windows, Jansen et al. [7] systematically compared multiple autoregressive spectral estimation methods and found that AR modeling based on the Burg algorithm can yield higher-resolution spectral estimates under short-time-window settings. Gotman et al. [8] applied rhythmic statistical features to online seizure-trigger detection in long-term EEG and reduced missed detections by incorporating adaptive background comparison. Qu et al. [9] proposed patient-specific seizure-onset templates and a nearest-neighbor decision strategy to reduce false alarms.

Driven by a deeper understanding of EEG non-stationarity, research shifted from conventional time or frequency statistics toward time–frequency analysis, offering a more direct characterization of the transient structure of epileptic discharges. Adeli et al. [10] studied 3-Hz spike-and-wave complexes and systematically demonstrated that wavelet decomposition can localize transient spikes and slow-wave components simultaneously across multiple scales. This provides a tool better matched to multi-scale time–frequency representation of epileptic activity than Fourier spectra. Building on the time–frequency representation paradigm, researchers began using wavelet subband energies and related statistics as detection features and feeding them into learning-based classifiers. For example, Subasi [11] input DWT subband features into a neural network to improve detection robustness, representing a typical two-stage workflow of time–frequency feature extraction followed by statistical learning. Faust et al. summarized the broad use of wavelet methods in denoising, feature extraction, and classification, and observed that DWT—particularly db4—is the most commonly adopted approach in epilepsy EEG studies [12].

Second-order statistics are often insufficient for epileptic EEG. Non-Gaussianity is not fully captured. Nonlinear interactions are also missed. Later studies therefore incorporated complexity measures and higher-order statistics. Discriminative information is then enhanced. Kannathal et al. [13] systematically compared spectral entropy and embedding entropy and reported that seizure EEG tends to exhibit reduced complexity, while different entropy measures vary significantly in discriminative power. Chua et al. [14] further used bispectrum magnitude and its entropy-based features to characterize nonlinear phase coupling, improving discrimination among normal, pre-ictal, and ictal

states within GMM/SVM frameworks. Acharya et al. [15] combined embedding entropy (ApEn, SampEn) with bispectrum-based phase entropy to perform three-class classification under a multi-classifier setting, validating the effectiveness of low-dimensional nonlinear features. Beyond entropy and higher-order spectral features, another line of work has explicitly investigated the nonlinear and sequential dynamics of epileptic activity. Yang et al. [16] used recurrence plots and recurrence quantification analysis on stereo-electroencephalography (SEEG) recordings to characterize dynamical differences among interictal, preictal, and ictal stages, as well as between epileptogenic and non-epileptogenic regions. Their results showed that epileptogenic channels exhibited more deterministic and recurrent structures, and that synchronization between epileptogenic channels was strengthened during seizures. Farahi et al. [17] further used recurrence-based nonlinear analysis to identify seizure dynamic trajectories from EEG recordings and to trace the sequence of regions involved during seizure evolution. These studies demonstrate that epileptic seizures exhibit rich nonlinear and sequential dynamical structures. However, they mainly rely on handcrafted recurrence-based descriptors and dynamical interpretation, rather than end-to-end multimodal representation learning for seizure detection.

Overall, handcrafted statistical and nonlinear-dynamical methods provide interpretable descriptions of epileptic EEG, including spectral changes, complexity variations, phase coupling, recurrence structures, and seizure-state transitions. However, these methods usually depend on predefined descriptors and fixed analysis assumptions. The extracted features are not jointly optimized with the downstream detection objective, and most studies focus on EEG or SEEG alone. Therefore, although they provide important evidence that seizures are dynamic processes, their integration with multimodal deep representation learning remains limited.

(2) Classical machine-learning-based epilepsy detection methods

Building on statistical signal processing, researchers gradually combined handcrafted EEG features with classical machine learning models. This strategy reduces the instability introduced by manually tuned thresholds and improves the consistency and automation of detection decisions. Such methods typically follow a two-stage pipeline. First, signal processing and handcrafted rules are used to construct a fixed feature representation. Second, a supervised learning model is trained to discriminate seizure states. In these approaches, learning mainly operates at the decision layer, while the feature representation itself is not updated toward the task objective.

In terms of feature construction, researchers expanded the feature families developed in the statistical signal-processing stage. Polat et al. [18] combined FFT-based frequency-domain features with a decision-tree classifier to distinguish epileptic EEG from normal EEG, demonstrating the feasibility of conventional spectral features in a supervised learning setting. However, spectral features implicitly assume signal stationarity and thus have limited ability to characterize transient changes at seizure onset. To better capture time-varying frequency components during seizures, Tzallas et al. [19] extracted local energy-distribution features from time–frequency representations to describe evolving patterns on the time–frequency plane, achieving better detection performance than purely frequency-domain analysis. Because second-order statistics are often insufficient to represent complex dynamics, distribution-based and nonlinear feature representations were also introduced. Orhan et al. [20] clustered wavelet coefficients and computed distributional statistics to better describe the global shape of coefficient patterns. Chua et al. [21] extracted higher order spectral (HOS) features from EEG signals to characterize their nonlinear and non-Gaussian properties, achieving class-specific feature ranges with high statistical significance ($p = 0.002$). Nicolaou and Georgiou [22] employed permutation entropy as a low-complexity feature for automated epileptic seizure detection using an SVM classifier.

In classifier design and system optimization, related research has gradually shifted away from single discriminative models. Increased emphasis has been placed on system stability and engineering practicality. Subasi [23] proposed a seizure EEG classification framework based on discrete wavelet transform and a Mixture of Experts model. A gating mechanism coordinates multiple expert models. It adapts to different feature distributions. Sensitivity to feature heterogeneity is reduced. This limitation is common in single classifiers. To improve performance while retaining interpretability, Kabir et al. [24] introduced Logistic Model Trees, combining tree structures with logistic regression to obtain smoother decision boundaries. Mursalin et al. [25] evaluated multiple feature-classifier combinations from a system perspective and used feature-selection strategies to reduce the impact of redundant features on classification performance.

Overall, classical machine-learning-based epilepsy detection introduces supervised discrimination on top of statistical features, enabling a transition from heuristic thresholds to data-driven decisions. These methods still follow a modeling paradigm that assumes predefined feature representations. Feature construction is separated from classification decisions. Joint optimization of the feature representation is not possible during learning. This structural limitation makes it difficult to fully characterize the continuously evolving dynamics of seizures and leads to clear weaknesses in cross-patient generalization and adaptability to complex clinical scenarios.

(3) Deep-learning-based epilepsy detection methods

To overcome the limitations of handcrafted features in representational power and generalization, researchers have increasingly shifted toward deep-learning-driven modeling. The key idea of these methods is to jointly optimize representation learning and discriminative decision-making. Models can automatically learn multi-scale and high-dimensional discriminative representations from raw or minimally preprocessed EEG signals. Uncertainties caused by signal non-stationarity, inter-subject variability, and complex artifacts are mitigated to some extent.

For automatic feature representation and spatial-structure extraction, convolutional neural networks (CNNs) have become a widely used choice due to their strength in capturing local patterns. Thodoroff et al. [26] addressed the heavy reliance on handcrafted features in traditional methods and their insufficient robustness under complex background EEG by using deep networks to learn discriminative features directly from EEG, thereby reducing dependence on feature engineering. Acharya et al. [27] demonstrated that a deep CNN could classify seizure EEG signals without manual feature extraction, showing the potential of end-to-end learning for automated seizure detection. To better align with clinical visual inspection, Emami et al. [28] demonstrated that CNN applied to raw EEG plot images could detect seizures in a manner analogous to the visual inspection performed by clinical epileptologists. For real-time applications, Daoud and Bayoumi [29] emphasized deployability by using lightweight architectures and channel selection to reduce inference cost. However, CNNs are fundamentally built on local receptive fields and are better at capturing local morphology or texture-like features, while they are relatively less effective at modeling long-range evolution across time scales. Although image-based representations and lightweight designs improve usability, they can make models more sensitive to preprocessing strategies and capacity constraints, thereby reducing stability across scenarios.

To better exploit temporal dynamics, recurrent neural networks (RNNs) and their variants have been used to model EEG time evolution. Hussein et al. [30] employed a deep LSTM network to model temporal dependencies in EEG signals, enabling robust seizure detection under both ideal and artifact-contaminated conditions. As the research focus expanded from single-channel sequences to inter-channel interactions, brain-network-

structured features began to attract increasing attention. Mirowski et al. [31] highlighted the importance of cross-channel synchrony and dependency patterns in pre-seizure state discrimination, providing motivation for subsequent spatial-dependency modeling. Tang et al. [32] argued that conventional convolutions are limited in expressing non-Euclidean relationships and long-range spatial dependencies among channels and proposed graph neural networks (GNNs) to model electrode relationships, together with self-supervised learning to alleviate insufficient annotations. Li et al. [33] addressed the difficulty of jointly modeling spatiotemporal dependencies, spectral hierarchical structure, and uncertainty in the pre-seizure interval by proposing spatiotemporal-frequency hierarchical graph convolution combined with semi-supervised active learning, and further introduced an adaptive mechanism for estimating patient-specific pre-seizure intervals to reduce label noise. Nevertheless, graph-based methods are highly dependent on graph construction strategies and prior assumptions, and dynamic or hierarchical graphs markedly increase computational complexity, which makes cross-center transfer and real-time deployment challenging.

(4) Multimodal seizure detection methods

Multimodal physiological sensing has also become an important direction for seizure detection, especially in wearable and long-term monitoring scenarios. Chen et al. [34] reviewed seizure detection studies using portable and wearable multimodal signals, including EEG, ECG, EMG, accelerometry, electrodermal activity, and other non-cerebral physiological signals. Their review showed that multimodal systems can provide complementary information for ambulatory seizure monitoring. Yang et al. [35] proposed a multimodal artificial intelligence system combining EEG and ECG for seizure identification, and demonstrated improved out-of-distribution generalization across clinical datasets. Zhang et al. [36] evaluated wearable EEG, EMG, and accelerometry for tonic-clonic seizure detection using data from the SeizeIT2 study, and showed that combining wearable EEG with extracerebral modalities can substantially reduce false positive rates while retaining high sensitivity. These studies indicate that multimodal fusion is not limited to simple feature concatenation and has already been explored in epilepsy detection. Nevertheless, many existing multimodal approaches are designed for specific seizure types, modality combinations, or decision-level fusion settings. The joint modeling of phase-aware time-frequency dynamics, cross-modal temporal asynchrony, modality reliability, and high-level semantic consistency remains insufficiently explored.

Although deep learning has advanced epilepsy detection from fixed feature representations to jointly learned representations and has greatly improved automation, it still faces several challenges. These challenges include substantial dependence on labeled data. Stability is reduced by ambiguous boundaries of the pre-seizure interval. Generalization is degraded by distribution shifts across patients and medical centers. Overall, despite substantial progress from rule-based approaches to machine learning and deep learning, existing epilepsy detection methods still suffer from the following limitations: (1) Many EEG-based studies focus mainly on amplitude variations in the time or frequency domain. Temporal dependencies between signals are often overlooked. Sequential dynamical characteristics are not extracted sufficiently. The ability to characterize and recognize early pathological changes is therefore weakened. (2) Existing multimodal fusion methods often rely on simple feature concatenation and lack effective modeling and alignment of semantic consistency across modalities during representation learning. As a result, cross-modal complementarity is not fully exploited. Redundant information is not effectively suppressed. Multimodal epilepsy detection models are therefore obtained with limited robustness and weak generalization.

3. Materials and Methods

3.1. Model Architecture

EEG, ECG, and EMG in multimodal epilepsy detection differ substantially in sampling characteristics, noise levels, and representation forms. As a result, it is difficult to align and fuse multimodal information within a unified feature space. This section introduces the multimodal detection model CMepiNet (Complex-valued Multimodal Epilepsy Network), as shown in Figure 1. The model consists of a Complex-valued Multimodal Feature Extractor (CMFE), a Dual-Level Semantic Alignment Fusion module (DLSAF), and a 3D convolutional neural network with a spatial attention mechanism (3DCSA). The main working mechanism of CMepiNet is described as follows:

(1) Feature extraction module. We adopt a unified Complex-valued Multimodal Feature Extractor (CMFE) for all modalities. For each modality, CMFE builds an isomorphic dual-branch complex-valued convolutional network. The time-domain-phase branch takes the analytic signal as input and uses 1D complex-valued convolutions to extract sequential oscillatory features from the joint amplitude-phase perspective. The time-frequency branch takes the complex-valued time-frequency spectrum as input and applies 2D complex-valued convolutions to capture band-energy distributions and cross-band coupling patterns. The complex-domain features produced by the two branches are then fused along the feature dimension, forming a modality-level representation that contains both time-domain-phase information and explicit frequency-domain information.

(2) Feature fusion module. Under constraints of cross-modal consistency and shared latent-space alignment, this module models a unified representation of multimodal features. First, the CMFE features of each modality are mapped into an alignment space with a unified dimensionality. For each time segment, a cross-modal consistency constraint is imposed on the EEG, ECG, and EMG features, so that multimodal representations originating from the same epileptic physiological event remain semantically consistent in the feature space. Next, the aligned modality features are projected by their respective projection networks into a shared latent seizure space, and a distribution-level alignment loss is used to reduce the global distribution discrepancies among modalities within this latent space. Finally, the latent representations of the three modalities are stacked along the modality dimension to form a 3D fused feature block that jointly includes the time, feature, and modality dimensions. This approach fully leverages the complementary information across modalities in amplitude-phase structure and time-frequency characteristics.

(3) 3DCSA classification module. A 3D convolutional neural network with an embedded spatial attention mechanism is used to deeply process the 3D fused feature blocks in the shared latent space. The feature block obtained by stacking the latent representations of EEG, ECG, and EMG is denoted as $\mathcal{F} \in \mathbb{R}^{T \times d_z \times n}$, where the three axes correspond to the time, feature, and modality dimensions. Using 3D convolutions, the network jointly models these three dimensions and progressively extracts both local and global structures of multimodal epileptic patterns. A spatial attention mechanism is then introduced to adaptively learn importance weights for each spatial location in the 3D convolutional feature maps, i.e., different time segments, latent feature channels, and modality combinations. Regions that are more relevant to seizure-state discrimination receive higher responses. Noise and redundant information are suppressed. Finally, the weighted high-dimensional features are flattened and fed into a fully connected layer and a Softmax classifier to output the class probabilities of different epilepsy states.

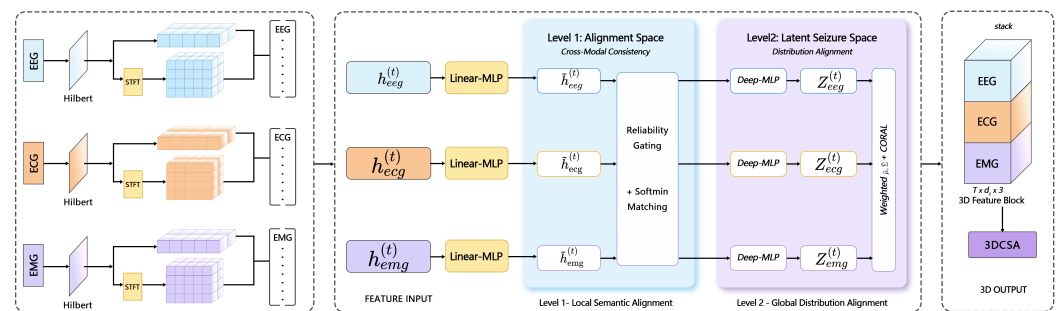


Figure 1. Overall framework of CMEpiNet. EEG, ECG, and EMG signals are first encoded by the shared complex-valued feature extractor (CMFE), then aligned across modalities by the dual-level semantic alignment fusion module (DLSAF). The aligned features are stacked into a $T \times d \times 3$ feature block and classified as seizure or non-seizure by the 3D spatial-attention classifier (3DCSA).

3.2. Data Preprocessing

To construct input samples suitable for model training, this work applies a unified preprocessing procedure to the SeizeIT2 dataset. The model is trained on the training subset, which includes subjects sub-001 to sub-096. The recordings consist of dual-channel EEG, single-channel ECG, and single-channel EMG. All signals were sampled at 256 Hz. First, signals from the three modalities are filtered. A 0.5–60 Hz band-pass filter is applied to remove DC drift and high-frequency noise. A 49.5–50.5 Hz notch filter is applied to suppress power-line interference. Then, segments labeled as impaired signal quality (IMPD) are excluded from the dataset. These segments are not used to construct training samples. This helps ensure the validity of the training data. A sliding-window strategy is used to segment the continuous signals. The window length is set to 2 s, corresponding to 512 samples. For seizure segments, dense sliding windows with a step size of 0.5 s are used to cover the onset and offset stages of seizures. For background segments, sliding windows with a step size of 1 s are used. The segmented windows are further organized into temporal sequences. Each input sequence contains $T = 30$ consecutive windows. For seizure sequences, a step size of 5 windows is used for sequence sampling. For background sequences, a step size of 15 windows is used. If a time gap longer than 3 s exists between adjacent windows, the sequence is split at the gap. The resulting parts are then treated as independent sequences. This process avoids interference from cross-gap sequences during temporal modeling. Class balance is then handled for the training samples. Background sequences are uniformly sampled across patients. The total number of background sequences is controlled at five times the number of seizure sequences. This alleviates the class imbalance between seizure and background samples. The final training samples are randomly shuffled at the patient level. Finally, window-level signal normalization is performed. For each window, z-score normalization is applied to each channel separately. This process reduces amplitude differences across subjects, modalities, and channels.

Through the above preprocessing procedure, training samples with stable signal quality, relatively balanced classes, and suitability for multimodal temporal modeling are obtained.

3.3. Complex-Valued Multimodal Feature Extractor (CMFE)

Multimodal seizure-related signals show clear heterogeneity in noise statistics and signal structures. Features based only on amplitude or spectral distributions cannot effectively characterize seizure-related changes across different modalities within a unified framework. Existing studies have shown that abnormal changes can be observed before seizure onset. These changes appear in inter-regional phase relationships. They also appear in the temporal organization of neural activity. Such changes occur earlier than prominent discharge

events. These findings indicate that phase information plays an important role in seizure prediction. In addition, frequency-domain and time–frequency features complement the temporal information reflected by phase by describing energy organization and dynamic evolution, thereby enabling a more comprehensive representation of the pre-seizure state.

Existing seizure detection methods cannot fully exploit deep oscillatory features in multimodal signals. These features include phase characteristics, time–frequency structures, temporal dependencies, and spatially coordinated changes. This limits the model’s ability to accurately represent complex spatiotemporal dynamics during seizures. To address this limitation, this section proposes a complex-valued multimodal feature extraction module, termed the Complex-valued Multimodal Feature Extractor (CMFE), which consists of a time-domain-phase branch and a time–frequency branch. Both branches are centered on complex-valued convolutions. By jointly processing the real and imaginary parts in the complex domain, the model can preserve phase variations as well as frequency- and amplitude-related information during feature extraction, thereby providing a more complete description of seizure-related dynamics in multimodal signals. This module also helps establish a consistent feature description across modalities, laying the foundation for joint modeling of seizure-related patterns.

CMFE adopts a parallel processing architecture to deeply fuse phase and amplitude-structure information in the complex domain. For each modality, CMFE includes two parallel complex-valued convolutional pathways. The time-domain-phase branch takes the analytic signal as input and focuses on capturing the dynamic characteristics of amplitude and instantaneous phase from the time series. The time–frequency branch takes the complex-valued time–frequency spectrum as input and focuses on extracting band-structure information and cross-band coupling features from the time–frequency plane. Both pathways perform convolution in the complex domain, with one emphasizing sequential phase-related features and the other emphasizing explicit frequency-domain structural features. Their outputs are then efficiently fused along the feature dimension, producing a unified complex-domain feature vector for the modality.

As shown in Figure 2, CMFE processes each modality’s input time series with two parallel branches—a time-domain/phase complex convolution branch and a time–frequency complex convolution branch, to generate a time-domain feature vector and a time–frequency feature vector, respectively. These vectors are concatenated along the feature dimension to form the CMFE output vector f_{cmfe} .

(1) Time-domain-phase complex convolution branch

As shown in the upper part of Figure 2, this branch takes each preprocessed single-channel time series (EEG, ECG, EMG) as input. A Hilbert transform is first applied to construct the corresponding analytic signal:

$$x_a(t) = x(t) + i\mathcal{H}[x(t)], \quad (1)$$

where $x(t)$ is the original continuous-time signal and $\mathcal{H}[x(t)]$ denotes its Hilbert transform. In practical implementations, the signal is discretely sampled at the sampling rate F_s . The physical time corresponding to the n -th sample is

$$t_n = \frac{n}{F_s}, \quad n = 0, 1, \dots, N - 1. \quad (2)$$

Accordingly, the discrete analytic signal $x_a[n]$ is defined as the value of $x_a(t)$ at the sampling instants t_n , i.e.,

$$x_a[n] = x[n] + i\mathcal{H}\{x[n]\}. \quad (3)$$

After constructing the analytic signal via the Hilbert transform, $x[n]$ serves as the real part and $\mathcal{H}\{x[n]\}$ serves as the imaginary part. This unified complex representation preserves both amplitude and phase information.

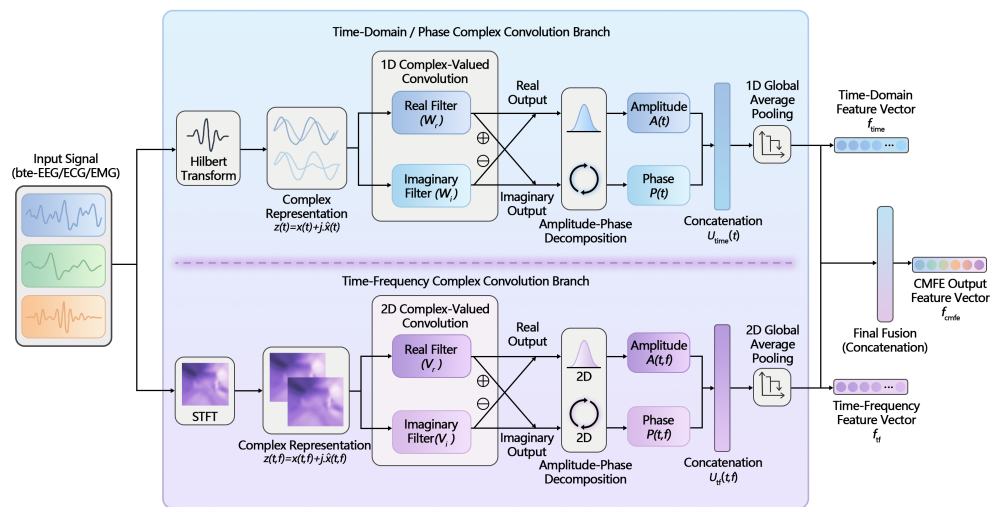


Figure 2. Architecture of the CMFE module. For each modality, the analytic signal is processed by two parallel complex-valued branches: a 1D time-domain phase branch and a 2D STFT-based time-frequency branch. The resulting amplitude and phase features are pooled and concatenated to form the complex-domain feature vector for that modality.

As shown in the middle of the upper part of Figure 2, the analytic signal $x_a[n]$ is then fed into a 1D complex-valued convolutional network. The network is composed of three cascaded 1D complex convolution layers, which progressively model oscillatory frequency characteristics and phase variation patterns within a local temporal range.

The complex-valued convolution kernel of the l -th layer is denoted as

$$\mathbf{W}^{(l)} = \mathbf{W}_r^{(l)} + i\mathbf{W}_i^{(l)}, \quad l \in \{1, 2, 3\}. \quad (4)$$

Here, l is the layer index, and $\mathbf{W}_r^{(l)}$ and $\mathbf{W}_i^{(l)}$ represent the learnable real and imaginary parameters of the l -th complex kernel, respectively.

To enable each kernel to simultaneously select frequency components and model phase variations within a local temporal window, we parameterize the 1D complex kernel of each layer as a Gaussian-windowed complex exponential:

$$w_k^{(l)}(t) = \exp\left(-\frac{t^2}{2(\sigma_k^{(l)})^2}\right) \exp\left(i\left(2\pi f_k^{(l)}t + \phi_k^{(l)}\right)\right). \quad (5)$$

Here, k is the kernel index; $f_k^{(l)}$ is the center frequency of the k -th kernel in layer l ; $\sigma_k^{(l)}$ controls the temporal support of the Gaussian window; and $\phi_k^{(l)}$ is the phase offset parameter. In the discrete implementation, we set $t = n/F_s$, discretize the kernel, and use it for 1D discrete complex convolution. After expansion, the real and imaginary parts are:

$$w_{k,r}^{(l)}[n] = \exp\left(-\frac{(n/F_s)^2}{2(\sigma_k^{(l)})^2}\right) \cos\left(2\pi f_k^{(l)} \frac{n}{F_s} + \phi_k^{(l)}\right), \quad (6)$$

$$w_{k,i}^{(l)}[n] = \exp\left(-\frac{(n/F_s)^2}{2(\sigma_k^{(l)})^2}\right) \sin\left(2\pi f_k^{(l)} \frac{n}{F_s} + \phi_k^{(l)}\right). \quad (7)$$

With this parameterization, EEG, ECG, and EMG can be modeled using a unified convolutional kernel form.

The forward propagation of the three complex convolution layers can be written recursively as

$$\mathbf{Z}^{(1)} = \mathbf{X} * \mathbf{W}^{(1)}, \quad \mathbf{Z}^{(2)} = \mathbf{Z}^{(1)} * \mathbf{W}^{(2)}, \quad \mathbf{Z}^{(3)} = \mathbf{Z}^{(2)} * \mathbf{W}^{(3)}, \quad (8)$$

where $*$ denotes 1D complex convolution, $\mathbf{X} = x_a[n]$ is the input analytic signal sequence, and $\mathbf{Z}^{(l)}$ is the feature map produced by the l -th layer. The third-layer output $\mathbf{Z}^{(3)}$ is a complex-valued feature map, whose real and imaginary parts are denoted by $\mathbf{Z}_r^{(3)}$ and $\mathbf{Z}_i^{(3)}$, respectively.

In the complex-valued feature map, the real part mainly characterizes amplitude-related features such as response strength within a local temporal range, energy distribution, and waveform shape. The imaginary part is more sensitive to changes in relative timing relationships across time points, and it can be used to describe phase variations, phase discontinuities, and changes in synchronization relationships.

To further reveal the amplitude and phase information contained in the third-layer complex feature map, as shown on the right side of the upper part of Figure 2, we perform amplitude–phase decomposition only on $\mathbf{Z}^{(3)}$ to obtain the amplitude and phase components:

$$A(t) = |\mathbf{Z}^{(3)}(t)|, \quad P(t) = \angle \mathbf{Z}^{(3)}(t). \quad (9)$$

Corresponding to “Amplitude $A(t)$ ” and “Phase $P(t)$ ” in the figure, the amplitude and phase can be computed from the real and imaginary parts of the third-layer output:

$$A(t) = \sqrt{\left(\mathbf{Z}_r^{(3)}(t)\right)^2 + \left(\mathbf{Z}_i^{(3)}(t)\right)^2}, \quad P(t) = \text{atan2}\left(\mathbf{Z}_i^{(3)}(t), \mathbf{Z}_r^{(3)}(t)\right). \quad (10)$$

Here, $\text{atan2}(\cdot, \cdot)$ is the two-argument arctangent function with quadrant correction.

Amplitude–phase decomposition is applied only to the third-layer output in order to preserve the representational capacity of complex convolution layers while avoiding unnecessary nonlinear decomposition in shallow layers.

The amplitude and phase components are then concatenated along the channel dimension to obtain a joint time-domain representation:

$$U_{\text{time}}(t) = [A(t), P(t)]. \quad (11)$$

Here, T is the number of time steps within the segment, and $[\cdot, \cdot]$ denotes concatenation along the channel dimension.

Next, 1D global average pooling (GAP) is applied to $U_{\text{time}}(t)$, i.e., a global average over the time dimension is computed for each channel, yielding a fixed-length time-domain feature vector:

$$f_{\text{time}} = \text{GAP}_{1\text{D}}(U_{\text{time}}), \quad f_c = \frac{1}{T} \sum_{n=1}^T U_{\text{time}}(c, n). \quad (12)$$

Here, $U_{\text{time}}(c, n)$ denotes the response at the c -th channel and the n -th time step in the joint representation. This operation aggregates amplitude and phase statistics over the entire segment along the time dimension.

Finally, the time-domain feature vector f_{time} is passed through a linear transformation and then normalized to obtain the time-domain–phase embedding vector for the segment:

$$h_t = \mathbf{W}_t f_{\text{time}} + \mathbf{b}_t. \quad (13)$$

Here, $\mathbf{W}_t$ and $\mathbf{b}_t$ are learnable parameters. The resulting h_t serves as the time-domain-phase feature vector of the current modality for this segment and is used for final fusion with the output of the other branch (the time-frequency branch).

(2) Time-frequency complex convolution branch and feature vector generation

As shown in the lower part of Figure 2, the time-frequency complex convolution branch takes the discrete analytic signal $x_a[n]$ (output from the time-domain-phase branch) as input. It first applies the short-time Fourier transform (STFT) to map the 1D complex sequence to a 2D complex-valued time-frequency map $X(f, \tau)$, thereby introducing structural information in the frequency and time-frequency domains. In the continuous-time domain, STFT is defined as

$$Z(f, \tau) = \int_{-\infty}^{+\infty} x_a(t) w(t - \tau) e^{-i2\pi ft} dt, \quad (14)$$

where $w(\cdot)$ is the window function, f is the frequency variable, and τ is the time parameter corresponding to the window position (or window center). $X(f, \tau) \in \mathbb{C}$ denotes the complex coefficient at the time-frequency point (f, τ) .

In practice, STFT is implemented in discrete time. Let $x_a[n]$ be the discrete analytic signal, with window length L , hop size H , and FFT size N_{fft} . The discrete form can be written as

$$Z(f, \tau) = \sum_{n=0}^{L-1} x_a[n + \tau H] w[n] e^{-i2\pi f n / N_{\text{fft}}}, \quad f = 0, \dots, F-1. \quad (15)$$

Here, f is the index of the discrete frequency bin, and τ is the time-frame index corresponding to the sliding STFT window along the time axis. After computing STFT for an input segment, we obtain a complex-valued time-frequency map $Z \in \mathbb{C}^{F \times T}$. Decomposing it into real and imaginary parts yields

$$Z = Z_r + iZ_i, \quad Z_r = \Re(Z) \in \mathbb{R}^{F \times T}, \quad Z_i = \Im(Z) \in \mathbb{R}^{F \times T}. \quad (16)$$

That is, the STFT output can be viewed as two real-valued matrices (the real-part matrix and the imaginary-part matrix) defined on the same (f, τ) grid. Furthermore, note that the analytic signal can be written as $x_a[n] = x[n] + i\hat{x}[n]$. Since STFT is a linear operator, we have

$$\text{STFT}\{x_a[n]\} = \text{STFT}\{x[n]\} + i \text{STFT}\{\hat{x}[n]\}. \quad (17)$$

Therefore, Z_r and Z_i correspond to the time-frequency projections of the original signal $x[n]$ and its Hilbert-orthogonal component $\hat{x}[n]$, respectively. This indicates that the branch does not apply STFT only to a “real-part signal”; rather, a single transform of the complex analytic signal simultaneously produces both real and imaginary 2D time-frequency maps.

The complex-valued time-frequency map $Z(t, f)$ is then used as a 2D complex input to a 2D complex-valued convolution module. This module contains two sets of learnable parameters for the real and imaginary parts of the complex kernels, corresponding to the “Real Filter” $\mathbf{V}_r$ and “Imaginary Filter” $\mathbf{V}_i$ in Figure 2. The complex-valued convolution kernel at layer l is

$$\mathbf{V}^{(l)} = \mathbf{V}_r^{(l)} + i\mathbf{V}_i^{(l)}, \quad l \in \{1, 2, 3\}. \quad (18)$$

Corresponding to the “Real Output” and “Imaginary Output” branches in the figure, given an input $\mathbf{U} = \mathbf{U}_r + i\mathbf{U}_i$ and a kernel $\mathbf{V} = \mathbf{V}_r + i\mathbf{V}_i$, the real and imaginary parts of the 2D complex convolution output $\mathbf{Y} = \mathbf{U} * \mathbf{V}$ are

$$\mathbf{Y}_r = \mathbf{U}_r * \mathbf{V}_r - \mathbf{U}_i * \mathbf{V}_i, \quad \mathbf{Y}_i = \mathbf{U}_r * \mathbf{V}_i + \mathbf{U}_i * \mathbf{V}_r, \quad (19)$$

where $*$ denotes 2D convolution.

To enable the kernel to characterize structural variations in both local frequency neighborhoods and local time neighborhoods on the time–frequency plane, each 2D complex-valued kernel is parameterized as a Gaussian-windowed complex exponential and is defined along both the frequency and time directions:

$$V_k^{(l)}(\Delta f, \Delta \tau) = \exp\left(-\frac{\Delta f^2}{2(\sigma_{f,k}^{(l)})^2} - \frac{\Delta \tau^2}{2(\sigma_{\tau,k}^{(l)})^2}\right) \exp\left(i\left(2\pi\nu_k^{(l)}\Delta f + 2\pi\mu_k^{(l)}\Delta \tau + \phi_k^{(l)}\right)\right). \quad (20)$$

Here, Δf and $\Delta \tau$ denote the local offsets along the frequency and time directions relative to the kernel center, defining the local receptive region on the time–frequency plane. The parameters $\sigma_{f,k}^{(l)}$ and $\sigma_{\tau,k}^{(l)}$ control the effective ranges of the k -th kernel in layer l along the frequency and time directions, respectively. Their values determine the smoothness and effective support size in each direction and thus affect the time–frequency resolution characteristics. The first exponential term forms a 2D Gaussian window that locally constrains the kernel response, focusing it on a limited frequency neighborhood and time neighborhood. The parameters $\nu_k^{(l)}$ and $\mu_k^{(l)}$ are the oscillation-center parameters of the complex exponential along the frequency and time directions, respectively, describing the kernel’s selectivity to different frequency-change and time-change patterns. $\phi_k^{(l)}$ is a phase offset term that adjusts phase alignment in the kernel response. With this Gaussian-windowed complex-exponential parameterization, each 2D complex kernel can jointly model local amplitude-variation structures and phase-evolution characteristics on the time–frequency plane, enabling a unified characterization of complex time–frequency patterns.

The forward propagation of the three 2D complex convolution layers can be written recursively as

$$\mathbf{Y}^{(1)} = \mathbf{Z} * \mathbf{V}^{(1)}, \quad \mathbf{Y}^{(2)} = \mathbf{Y}^{(1)} * \mathbf{V}^{(2)}, \quad \mathbf{Y}^{(3)} = \mathbf{Y}^{(2)} * \mathbf{V}^{(3)}. \quad (21)$$

$\mathbf{Y}^{(3)}$ is the complex-valued feature map output by the third layer. Its real part mainly reflects the response strength and energy variations of different frequency components within a local temporal range, whereas its imaginary part is more sensitive to phase variations and changes in relative timing relationships between frequency components.

As shown on the right side of the lower part of Figure 2, we perform amplitude–phase decomposition only on the third-layer complex output $\mathbf{Y}^{(3)}$ to obtain an amplitude map and a phase map:

$$A(t, f) = |\mathbf{Y}^{(3)}(t, f)|, \quad P(t, f) = \angle \mathbf{Y}^{(3)}(t, f). \quad (22)$$

The amplitude and phase features are concatenated along the channel dimension to form a joint time–frequency feature map:

$$U_{tf}(t, f) = [A(t, f), P(t, f)]. \quad (23)$$

Here, $[\cdot, \cdot]$ denotes concatenation along the channel dimension, which jointly represents amplitude and phase information while keeping the time–frequency resolution unchanged.

To obtain a fixed-length time–frequency feature vector, as shown in Figure 2, we apply 2D global average pooling to $U_{tf}(t, f)$, i.e., we compute a global average over the entire $T \times F$ time–frequency plane for each channel:

$$\mathbf{f}_{tf} = \text{GAP}_{2D}(U_{tf}(t, f)), \quad (f_{tf})_c = \frac{1}{TF} \sum_{t=1}^T \sum_{f=1}^F U_{tf}(c, t, f). \quad (24)$$

Here, $U_{tf}(c, t, f)$ denotes the response of the joint time–frequency feature at channel c and the time–frequency point (t, f) .

The time–frequency feature vector $\mathbf{f}_{tf}$ is then passed through a linear transformation and normalized to obtain the time–frequency embedding vector of the modality:

$$\mathbf{h}_{tf} = \mathbf{W}_{tf}\mathbf{f}_{tf} + \mathbf{b}_{tf}, \quad (25)$$

where $\mathbf{W}_{tf}$ and $\mathbf{b}_{tf}$ are learnable parameters.

Finally, as shown by “Final Fusion/Concatenation” on the right side of Figure 2, the time-domain–phase branch output $\mathbf{h}_t$ and the time–frequency branch output $\mathbf{h}_{tf}$ are concatenated along the feature dimension:

$$\mathbf{h} = [\mathbf{h}_t, \mathbf{h}_{tf}]. \quad (26)$$

Here, $[\cdot, \cdot]$ denotes vector concatenation. In this way, each modality obtains a unified representation that encodes both time-domain local response characteristics (amplitude–phase) and time–frequency structural variation information, providing a stable input for subsequent multimodal feature fusion and classification.

3.4. Feature Fusion: DLSAF

Multimodal features differ substantially in scale, spectral structure, and statistical characteristics. Simple feature concatenation limits the extraction of effective mutual information and reduces the overall signal-to-noise ratio. In addition, cross-modal response asynchrony and the dynamically non-stationary nature of signal quality further exacerbate the difficulty of feature alignment. Moreover, overly strong cross-modal consistency constraints may excessively compress complementary information across modalities and may also introduce misleading noise.

To address these issues, this section proposes a Dual-Level Semantic Alignment Fusion module (DLSAF), as shown in Figure 3. DLSAF consists of two components: a cross-modal consistency constraint in an alignment space and distribution alignment in a shared latent space. Specifically, EEG, ECG, and EMG features are first projected into a unified alignment space through learnable Linear-MLP mappings. Within this space, a window-level softmin soft matching mechanism is introduced to accommodate cross-modal asynchrony. In addition, a modality-reliability gating mechanism adaptively weights the constraint strength across different time segments. This design suppresses the negative impact of low-quality modalities on the fused representation while improving robustness, yielding the cross-modal consistency loss L_{cm} . Next, the aligned modality representations are fed into a Deep-MLP to map them into a shared epilepsy-semantic latent space. Within this space, distribution-level alignment is performed based on reliability-weighted first-order statistics (mean) and second-order statistics (covariance), resulting in the distribution alignment loss L_{dist} . Finally, the latent representation sequences of the three modalities are stacked along the modality dimension to construct a three-dimensional feature block (3D block), which serves as the input to the subsequent 3D module for joint modeling along the time-modality-feature dimensions.

Let $h_{eeg}^{(t)}$, $h_{ecg}^{(t)}$, and $h_{emg}^{(t)}$ denote the EEG, ECG, and EMG modality feature vectors, respectively, obtained from the feature extraction module at the t -th time segment.

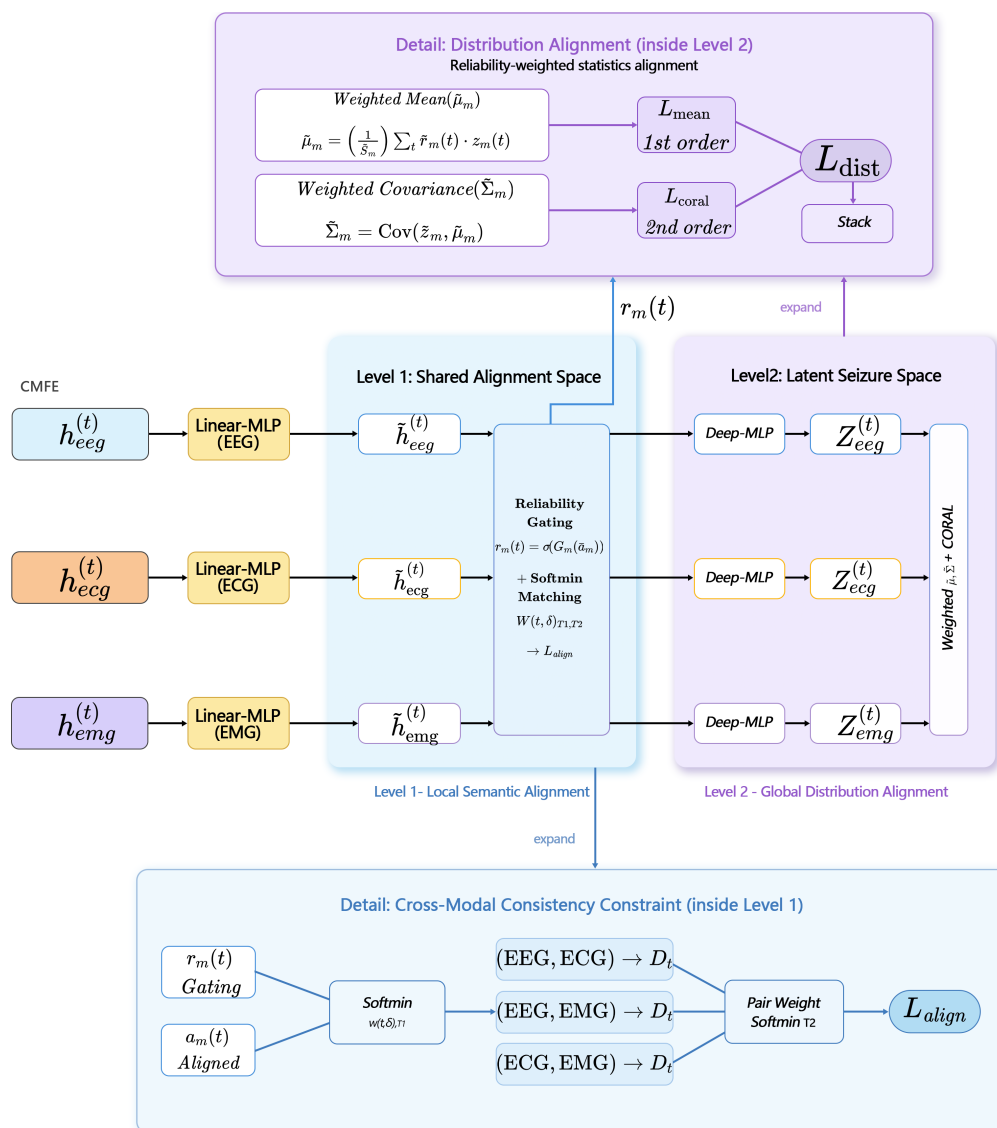


Figure 3. Architecture of the DLSAF module. The first level enforces cross-modal consistency in a shared alignment space through reliability gating and window-level softmin matching (L_{cm}). The second level aligns the mean and covariance of modality-specific features in a latent seizure space (L_{dist}).

(1) Cross-Modal Consistency-Constrained Feature Fusion in the Alignment Space

To align high-level semantic representations of different modalities along the temporal dimension, the features extracted from each modality at each time segment are first mapped into a shared alignment space of the same dimensionality. As illustrated by the Linear-MLP module in Figure 3, considering that the original feature dimensions and distributions may differ across modalities, a set of learnable linear projections is employed to map each modality into an alignment space of dimension d_h , formulated as:

$$\tilde{h}_{eeg}^{(t)} = W_{eeg} h_{eeg}^{(t)} + b_{eeg}, \quad \tilde{h}_{ecg}^{(t)} = W_{ecg} h_{ecg}^{(t)} + b_{ecg}, \quad \tilde{h}_{emg}^{(t)} = W_{emg} h_{emg}^{(t)} + b_{emg}.$$

Here, $W_{(\cdot)} \in \mathbb{R}^{d_h \times d}$ and $b_{(\cdot)} \in \mathbb{R}^{d_h}$ are learnable parameters, and $\tilde{h}_{(\cdot)}^{(t)} \in \mathbb{R}^{d_h}$ denotes the modality representation in the shared alignment space.

Considering that different modalities may exhibit temporal delays and asynchrony when reflecting the same epileptic state, and that signal quality dynamically fluctuates due to artifacts, missing data, and environmental noise, we adopt a window-level soft

matching–reliability gating–modality-pair adaptive weighting strategy for consistency regularization within the alignment space. As shown in the Reliable Window Consistency Branch in Figure 3, we first define the temporal window set:

$$\mathcal{K}(t) = \{k \mid 1 \leq k \leq T, |k - t| \leq \Delta\}, \quad (27)$$

where Δ denotes the window radius (measured in segments).

For any modality $m \in \{\text{eeg}, \text{ecg}, \text{emg}\}$, a reliability weight $r_m^{(t)} \in (0, 1)$ is estimated at each time segment t through a learnable gating mechanism:

$$r_m^{(t)} = \sigma\left(g_m\left(\tilde{h}_m^{(t)}\right)\right), \quad (28)$$

where $g_m(\cdot)$ is implemented as a lightweight fully connected network and $\sigma(\cdot)$ denotes the Sigmoid function. This reliability weight is used to suppress the adverse influence of low-quality modalities or segments on the consistency constraint.

For each modality pair (m, n) , using $\tilde{h}_m^{(t)}$ as the anchor, candidate segments from modality n within the window $\mathcal{K}(t)$ are softly matched. The reliability gating is directly incorporated into the matching distance, yielding the reliability-weighted window matching distance at segment t :

$$D_t^{m,n} = \sum_{k \in \mathcal{K}(t)} a_{t,k}^{m,n} \underbrace{r_m^{(t)} r_n^{(k)}}_{\text{reliability gating}} \left\| \tilde{h}_m^{(t)} - \tilde{h}_n^{(k)} \right\|_2^2, \quad (m, n) \in \mathcal{P}, \quad (29)$$

where $\mathcal{P} = \{(\text{eeg}, \text{ecg}), (\text{eeg}, \text{emg}), (\text{ecg}, \text{emg})\}$ denotes the set of modality pairs. The coefficient $a_{t,k}^{m,n}$ is defined via a softmax operation within the window (controlled by temperature parameter τ):

$$a_{t,k}^{m,n} = \frac{\exp\left(-\left\| \tilde{h}_m^{(t)} - \tilde{h}_n^{(k)} \right\|_2^2 / \tau\right)}{\sum_{j \in \mathcal{K}(t)} \exp\left(-\left\| \tilde{h}_m^{(t)} - \tilde{h}_n^{(j)} \right\|_2^2 / \tau\right)}. \quad (30)$$

Here, $\tau > 0$ controls the sharpness of matching within the window. A smaller τ makes the softmax operation closer to hard minimum matching. This window-based soft matching allows cross-modal asynchrony: when the true semantic correspondence occurs at $k \neq t$ but satisfies $|k - t| \leq \Delta$, the constraint will focus on the most consistent candidate segment within the window, thereby suppressing gradient noise caused by forced mismatching.

Next, as illustrated by the softmax $\rightarrow p_t^{m,n}$ branch in Figure 3, different modality pairs may exhibit varying consistency levels at the same segment t . Therefore, a softmax normalization is further applied across modality pairs over the set $\mathcal{P}$ to obtain pairwise weights:

$$p_t^{m,n} = \frac{\exp(-\gamma D_t^{m,n})}{\sum_{(i,j) \in \mathcal{P}} \exp(-\gamma D_t^{i,j})}, \quad (m, n) \in \mathcal{P}, \quad (31)$$

where $\gamma > 0$ controls the sharpness of the weight distribution, assigning larger constraint weights to modality pairs with stronger consistency (smaller distance).

The cross-modal consistency loss at segment t is then defined as:

$$L_{cm}^{(t)} = \sum_{(m,n) \in \mathcal{P}} p_t^{m,n} D_t^{m,n}, \quad (32)$$

and the overall consistency loss across all segments within a sample is computed as:

$$L_{cm} = \frac{1}{T} \sum_{t=1}^T L_{cm}^{(t)}. \quad (33)$$

By minimizing L_{cm} , the model encourages different modalities to form a stable and coherent structure around the same epileptic semantic representation while allowing cross-modal asynchrony. Meanwhile, the reliability gating mechanism mitigates the negative impact of low-quality modalities on the fused representation, thereby improving multi-modal synergy and robustness.

After cross-modal consistency regularization, the temporally aligned feature sequences of each modality are expressed as:

$$\tilde{H}_{\text{eeg}} = \{\tilde{h}_{\text{eeg}}^{(t)}\}_{t=1}^T, \quad \tilde{H}_{\text{ecg}} = \{\tilde{h}_{\text{ecg}}^{(t)}\}_{t=1}^T, \quad \tilde{H}_{\text{emg}} = \{\tilde{h}_{\text{emg}}^{(t)}\}_{t=1}^T.$$

(2) Shared Latent-Space Alignment for Feature Fusion (Shared Latent Seizure Space)

In the alignment space feature fusion step above, the three modality features are mapped into the same alignment space, and a cross-modal consistency loss L_{cm} is constructed via window-level soft matching, reliability gating, and modality-pair adaptive weighting. This achieves segment-level local semantic alignment along the temporal dimension. However, because different modalities may still exhibit global shifts in statistical distributions and correlation structures, relying only on local consistency constraints can lead to different distributional patterns in the latent representation space. This may in turn degrade generalization across subjects and under varying noise conditions. To further improve modality invariance at the distribution level, as shown in the pipeline “Deep-MLP $\rightarrow$ weighted mean/covariance $\rightarrow$ dist $\rightarrow L_{\text{dist}}$ ” on the right side of Figure 3, we introduce a shared latent seizure space. In this space, we align the first-order statistics (mean) and second-order statistics (covariance) of modality-specific latent representations to obtain a more compact epilepsy-semantic representation with consistent distributions.

For each time segment $t \in \{1, \dots, T\}$, the modality feature $\tilde{h}_m^{(t)} \in \mathbb{R}^{d_h}$ in the alignment space is fed into the modality-specific Deep-MLP $f_m(\cdot)$ in Figure 3, yielding the shared latent-space representation $\mathbf{z}_m^{(t)} \in \mathbb{R}^{d_z}$:

$$\mathbf{z}_m^{(t)} = f_m(\tilde{h}_m^{(t)}), \quad m \in \{\text{eeg}, \text{ecg}, \text{emg}\}. \quad (34)$$

To ensure that global distribution estimation is dominated by high-quality segments, we reuse the reliability weights $r_m^{(t)} \in (0, 1)$ obtained in the alignment space feature fusion step above, and define the normalized weight sum as

$$S_m = \sum_{t=1}^T r_m^{(t)} + \varepsilon, \quad (35)$$

where $\varepsilon > 0$ is a numerical stability term. Then, as shown in the “weighted mean/weighted covariance” modules in Figure 3, we compute the reliability-weighted mean and covariance of modality m in the shared latent space. The weighted mean vector is defined as

$$\bar{\mathbf{z}}_m = \frac{1}{S_m} \sum_{t=1}^T r_m^{(t)} \mathbf{z}_m^{(t)}, \quad m \in \{\text{eeg}, \text{ecg}, \text{emg}\}. \quad (36)$$

To further align distributional shapes, we introduce the reliability-weighted covariance matrix:

$$\mathbf{C}_m = \frac{1}{S_m} \sum_{t=1}^T r_m^{(t)} \left(\mathbf{z}_m^{(t)} - \bar{\mathbf{z}}_m \right) \left(\mathbf{z}_m^{(t)} - \bar{\mathbf{z}}_m \right)^\top, \quad m \in \{\text{eeg}, \text{ecg}, \text{emg}\}. \quad (37)$$

Based on these statistics, as indicated by “dist $\rightarrow L_{dist}$ ” in Figure 3, we jointly characterize modality distribution shifts using differences in means and covariances. The first-order alignment loss is

$$L_\mu = \|\bar{\mathbf{z}}_{\text{eeg}} - \bar{\mathbf{z}}_{\text{ecg}}\|_2^2 + \|\bar{\mathbf{z}}_{\text{eeg}} - \bar{\mathbf{z}}_{\text{emg}}\|_2^2 + \|\bar{\mathbf{z}}_{\text{ecg}} - \bar{\mathbf{z}}_{\text{emg}}\|_2^2. \quad (38)$$

The second-order (CORAL) alignment loss is

$$L_{cor} = \|\mathbf{C}_{\text{eeg}} - \mathbf{C}_{\text{ecg}}\|_F^2 + \|\mathbf{C}_{\text{eeg}} - \mathbf{C}_{\text{emg}}\|_F^2 + \|\mathbf{C}_{\text{ecg}} - \mathbf{C}_{\text{emg}}\|_F^2, \quad (39)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Finally, the distribution-level alignment loss is defined as

$$L_{dist} = L_\mu + \beta L_{cor}, \quad (40)$$

where $\beta \geq 0$ balances the relative strength between mean alignment and covariance alignment. With this distribution alignment mechanism, while ensuring that reliable segments dominate the estimation, the model can simultaneously reduce global center shifts and distribution-shape differences across modalities, thereby producing a more compact and modality-invariant epilepsy-semantic representation.

After shared latent-space alignment, as shown in the stacking module in Figure 3, the latent representation sequence of modality m is denoted as $\mathbf{Z}_m = \{\mathbf{z}_m^{(t)}\}_{t=1}^T$. To facilitate subsequent joint modeling across time–feature–modality by the 3D module, the latent representations of the three modalities are stacked along the modality dimension to form a 3D feature block:

$$\mathcal{F} = \text{Stack}(\mathbf{Z}_{\text{eeg}}, \mathbf{Z}_{\text{ecg}}, \mathbf{Z}_{\text{emg}}) \in \mathbb{R}^{T \times d_z \times n}, \quad n = 3. \quad (41)$$

This feature block serves as the input to the subsequent 3D convolution/attention classification module.

Finally, to enable end-to-end optimization for the overall task, we denote the prediction loss as L_{cls} . The total training loss is:

$$L = L_{cls} + \lambda_1 L_{cm} + \lambda_2 L_{dist} = L_{cls} + \lambda_1 L_{cm} + \lambda_2 (L_\mu + \beta L_{cor}), \quad (42)$$

where L_{cm} is the cross-modal consistency loss defined in the alignment space feature fusion step above, and $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ balance the contributions of local consistency regularization and distribution-level alignment regularization in the overall objective.

3.5. Classifier (3DCSA)

After applying the cross-modal consistency constraint and the shared latent-space alignment, the latent representations of EEG, ECG, and EMG from different time segments are stacked to form a 3D fused feature block:

$$\mathcal{F} \in \mathbb{R}^{T \times d_z \times n}, \quad (43)$$

where T denotes the number of time segments, d_z is the feature dimensionality in the shared latent space, and n is the number of modalities (in this study, $n = 3$ for EEG, ECG,

and EMG). This 3D block encodes multimodal epilepsy-semantic information along the time, feature, and modality dimensions and serves as the input to the subsequent classifier.

To fully exploit the joint structure of $\mathcal{F}$ across these three dimensions, we design a multimodal feature classifier based on a 3D convolutional neural network with a spatial attention mechanism, termed 3D Convolution with Spatial Attention (3DCSA), as illustrated in Figure 4. Specifically, $\mathcal{F}$ is treated as a 3D feature volume of size $T \times d_z \times n$, and it is further expanded to a single-channel tensor along the channel dimension:

$$\mathbf{X} \in \mathbb{R}^{T \times d_z \times n \times 1}, \quad (44)$$

which is used as the input to the 3D convolutional network.

3DCSA is composed of multiple stacked 3D convolutional blocks. Each block includes a 3D convolution layer, a nonlinear activation function, and a 3D pooling layer. The 3D convolution kernels slide simultaneously along the time, feature, and modality dimensions, enabling the network to capture (i) dynamic changes across time segments, (ii) coupling relationships among latent features, and (iii) correlation patterns across modalities. The subsequent 3D pooling operations progressively reduce the volume size while preserving higher-level semantic representations. After multiple layers of 3D convolution and pooling, a high-dimensional feature map is obtained:

$$\mathbf{G} \in \mathbb{R}^{H \times W \times D \times C}, \quad (45)$$

where H , W , and D correspond to the downsampled time, feature, and modality/spatial dimensions, respectively, and C denotes the number of channels.

To further enhance the model's ability to focus on key regions, 3DCSA introduces a spatial attention mechanism on the high-level feature map. A $1 \times 1 \times 1$ 3D convolution is used to compress channels in $\mathbf{G}$, producing a single-channel attention feature map $\mathbf{M}$. A Sigmoid activation is then applied to generate the spatial attention weights:

$$\mathbf{S} = \sigma(\mathbf{M}), \quad (46)$$

where each element in $\mathbf{S}$ lies in $[0, 1]$ and indicates the importance of the corresponding spatial location (i.e., a specific combination of time segment, latent feature channel, and modality). The attention map is multiplied element-wise with the original feature map:

$$\mathbf{G}' = \mathbf{G} \odot \mathbf{S}, \quad (47)$$

thereby amplifying regions that are more relevant to epilepsy-state discrimination while suppressing irrelevant or noisy regions.

Finally, the weighted feature map $\mathbf{G}'$ is flattened into a vector and fed into a fully connected layer and a Softmax classifier to output the class probabilities for different epilepsy states. By combining the complex-domain features extracted by CMFE, the fused representations obtained via cross-modal consistency and shared latent-space alignment, and the joint time–feature–modality modeling capability of 3DCSA, the model achieves higher detection accuracy and improved robustness on the SeizeIT2 [37] dataset.

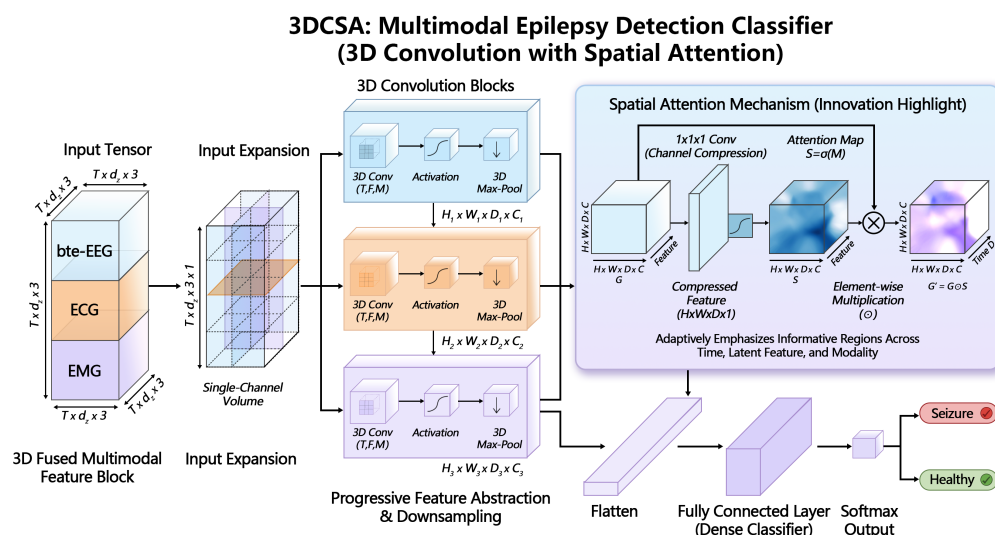


Figure 4. Architecture of the 3DCSA classifier. Stacked 3D convolutional blocks jointly model the temporal, feature, and modality dimensions of the fused feature block. A spatial attention map then reweights informative regions before a fully connected Softmax layer outputs the seizure probability.

4. Experimental and Results

4.1. Dataset

All experiments in this study were conducted on the public SeizeIT2 [37] benchmark dataset. SeizeIT2 is a multicenter wearable multimodal epilepsy dataset. Its public subset contains 125 patients with focal epilepsy from five epilepsy monitoring centers in Europe. The dataset contains approximately 11,640 h of wearable monitoring data and 886 focal seizure events. The recordings were collected in both home and hospital monitoring environments and provide multiple modalities, including behind-the-ear EEG, ECG, EMG, and motion signals. Therefore, SeizeIT2 is suitable for automatic seizure-event detection under long-term monitoring conditions.

To avoid data leakage, the dataset was divided at the subject level before sliding-window segmentation and temporal sample construction. We adopted the subject-independent holdout split recommended by SeizeIT2. The training set contained subjects from sub-001 to sub-096. The test set contained subjects from sub-098 to sub-125. Subject sub-097 was excluded because its multimodal recordings were incomplete and could not be used to construct the EEG, ECG, and EMG inputs required by this study. All recordings, window segments, and temporal samples from the same subject were assigned to the same data subset. Therefore, no subject appeared in both the training and test sets.

The experiments used three signal modalities: dual-channel EEG, single-channel ECG, and single-channel EMG. This study focuses on seizure-event detection rather than patient-level epilepsy diagnosis. Therefore, the model did not determine whether a subject had epilepsy by majority voting over all segments from that subject. Instead, it first generated window-level seizure probabilities. These probabilities were then converted into predicted seizure events through a unified post-processing procedure and matched with expert-annotated seizure events along the time axis.

4.2. Evaluation Metrics

The evaluation uses five metrics to assess model performance: sensitivity (Sens), false alarms per hour (FA/h), HMS, area under the Sens-FA/h curve (AUSF), and area under the receiver operating characteristic curve (AUROC). Sens, FA/h, and AUSF measure seizure-event-level performance, whereas AUROC measures window-level performance. Sens

reflects the seizure detection ability of the model. FA/h reflects the false alarm level during non-seizure periods. HMS provides a single-threshold composite score that summarizes the trade-off between sensitivity and false alarms. AUSF evaluates the overall trade-off between sensitivity and false alarms under different thresholds. AUROC evaluates the overall discriminative ability of the model between seizure and non-seizure samples.

In the experiments, the evaluation segments raw signals using a sliding window with a length of 2 s and a step size of 1 s. The signal sampling rate is 256 Hz. Therefore, each window contains 512 samples. For the i -th window, the model outputs a seizure probability $p_i \in [0, 1]$, forming a window-level probability sequence $\{p_i\}_{i=1}^N$. The evaluation computes AUROC directly from the window-level prediction probabilities and their corresponding labels. This computation does not involve threshold selection or post-processing.

For Sens, FA/h, and AUSF, the post-processing procedure first converts the window-level probability sequence into a list of predicted seizure events. The event-level matching procedure then matches the predicted seizure events with expert-annotated seizure events along the time axis. This study adopts the any-overlap criterion for event-level matching. Let the predicted event be $\hat{e}_j = [\hat{t}_j^s, \hat{t}_j^e]$, and let the true seizure event be $e_k = [t_k^s, t_k^e]$. Here, t denotes the temporal position of an event on the recording timeline. Superscripts s and e denote the start time and end time of an event, respectively. Subscripts j and k denote the indices of the predicted event and the true event, respectively. $\hat{e}_j$ denotes the model prediction, and e_k denotes the expert annotation.

If the two events satisfy

$$\min(\hat{t}_j^e, t_k^e) - \max(\hat{t}_j^s, t_k^s) > 0,$$

the evaluation regards the predicted event as matching the true event and counts it as a true positive (TP). If no predicted event covers a true seizure event, the evaluation counts the true seizure event as a false negative (FN). Sensitivity is defined as

$$\text{Sens} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

If a predicted event does not overlap with any true seizure event, the evaluation counts it as a false positive (FP). Let the total recording duration be T_{rec} , measured in seconds. FA/h is defined as

$$\text{FA/h} = \frac{\text{FP} \times 3600}{T_{\text{rec}}}.$$

Sens and FA/h characterize the seizure detection ability and false alarm control ability of the model, respectively. Together, they form key indicators for evaluating the clinical usability of the model. To provide an additional single-threshold summary of the trade-off between sensitivity and false alarm burden, this study also reports HMS as a supplementary composite metric. Following the definition introduced in the SeizIT2 seizure detection challenge, HMS is defined as

$$\text{HMS} = \text{Sens}(\%) - 0.4 \times \text{FA/h}.$$

A higher HMS indicates a better balance between seizure detection sensitivity and false alarm control at the fixed decision threshold.

AUSF is calculated from the Sens–FA/h curve. First, the detection threshold is uniformly divided into 51 values within the interval $[0, 1]$, denoted as $\{\theta_m\}_{m=0}^{50}$. The model first outputs the seizure probability sequence for each test recording. Then, each threshold θ_m serves as the decision threshold to convert the probability sequence into a binary prediction sequence, followed by unified post-processing. The evaluation then obtains the Sens–FA/h

curve for each test recording. Next, the curves of all recordings are interpolated onto a unified horizontal axis $x \in [0, 200]$. The evaluation averages the interpolated curves over all recordings containing seizures and obtains the mean sensitivity function $\bar{s}(x)$. AUSF is defined as

$$\text{AUSF} = \frac{1}{200} \int_0^{200} \bar{s}(x) dx.$$

Its value ranges from 0 to 1. A higher AUSF indicates that the model achieves a better balance between sensitivity and false alarm control under different thresholds. The main results are reported at the fixed threshold $\theta = 0.5$, whereas AUSF evaluates the overall performance of the model under different thresholds.

The post-processing procedure contains three steps. The boundary parameter is set to $\Delta = 10$ s. First, the procedure binarizes the window-level probability sequence according to the threshold:

$$b_i = \mathbb{1}[p_i > \theta].$$

Then, the procedure fills and merges segments whose gaps between adjacent predicted events are shorter than $0.2\Delta = 2$ s. This step reduces event fragmentation caused by short interruptions. Finally, the procedure removes isolated predicted events whose durations are shorter than $0.8\Delta = 8$ s. This step suppresses false detections caused by transient noise.

Before post-processing, the evaluation controls the signal quality of each EEG window. First, it calculates the root mean square (RMS) amplitude of each channel. If the RMS of any channel falls below $13 \mu\text{V}$ or exceeds $150 \mu\text{V}$ [37], the procedure sets the prediction probability of the corresponding window to zero. This strategy reduces false detections caused by low-quality signals and improves the reliability of event-level detection. Because the seizure-window prevalence in the test set is extremely low (about 0.096%, i.e., roughly 1:1041), the window-level AUROC is only a weak proxy for event-level clinical utility. We therefore regard AUROC as an auxiliary window-level diagnostic, and use the event-level metrics (Sens, FA/h, AUSE, and HMS) as the primary criteria for clinical usability.

4.3. Training Configuration

This section sets the following training strategies and hyperparameters to ensure stable training and reproducible results. The model uses the Adam optimizer with an initial learning rate of $\eta_0 = 1 \times 10^{-3}$ and a weight decay of $\lambda_w = 1 \times 10^{-4}$. The learning rate follows the StepLR strategy. Specifically, the learning rate is multiplied by a decay factor of 0.5 every 10 epochs, with a lower bound of $\eta_{\min} = 1 \times 10^{-5}$. The batch size is set to 32, and the maximum number of training epochs is set to 100. The gradient clipping threshold is set to 1.0. The random seed is fixed at 42.

The total loss consists of three components:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{cm}} + \lambda_2 \mathcal{L}_{\text{dist}}.$$

Here, $\mathcal{L}_{\text{cls}}$ denotes the Focal Loss with $\gamma = 2.0$. The dynamic positive-sample weight is defined as

$$w_{\text{pos}} = \frac{n_{\text{neg}}}{n_{\text{pos}}}.$$

$\mathcal{L}_{\text{cm}}$ denotes the cross-modal consistency loss, and $\mathcal{L}_{\text{dist}}$ denotes the distribution alignment loss. The loss weights are set to $\lambda_1 = \lambda_2 = 50$. The hidden dimension is set to $d_h = 128$, and the alignment-space dimension is set to $d_z = 64$. The dropout rates of the convolutional layer, alignment layer, and classification layer are set to 0.1, 0.3, and 0.5, respectively.

4.4. Comparison with Baseline Methods

To evaluate the performance of the proposed method, CMEpiNet was compared with the official SVM and ChronoNet benchmark results reported in the SeizeIT2 dataset paper. The SVM baseline represents a traditional feature-engineering-based machine-learning method. This baseline was adapted from the behind-the-ear EEG seizure detection framework proposed by Vandecasteele et al. [38]. In the official SeizeIT2 benchmark, the EEG signals were first filtered and then segmented into 2 s windows with 50% overlap. Each window segment was represented by 42 handcrafted features and classified using a support vector machine with a radial basis function kernel. ChronoNet represents the deep-learning baseline. It is a convolutional-recurrent neural network originally proposed by Roy et al. [39] for abnormal EEG identification. ChronoNet takes EEG temporal segments as input and combines one-dimensional convolutional layers with gated recurrent units to model temporal patterns in EEG signals. The SVM and ChronoNet results used in this study were taken from the official SeizeIT2 benchmark, rather than being independently reimplemented or reoptimized in this work. In addition, InceptionTime [40] was included as a typical deep multivariate time-series classification baseline and was evaluated under the same SeizeIT2-based protocol. InceptionTime is a convolutional time-series classification network composed of stacked Inception modules with multi-scale temporal convolution kernels and residual connections. In this study, we used a single InceptionTime network rather than the ensemble version. The EEG, ECG, and EMG inputs were concatenated along the channel dimension to form a four-channel multivariate time series for each 2 s window. To ensure a fair comparison with CMEpiNet, InceptionTime was evaluated using the same subject-independent SeizeIT2 split, input modalities, preprocessing pipeline, and event-level evaluation protocol. The training protocol was also aligned with that of CMEpiNet, including Focal Loss, Adam optimization, batch size, learning-rate schedule, gradient clipping, early stopping, random seed, and validation-based model selection. Table 1 reports the quantitative results of the compared methods on the SeizeIT2 test set. Figure 5 shows the threshold-dependent Sens-FA/h curve of CMEpiNet on the SeizeIT2 test set, which is used to analyze the operating characteristics of the proposed model under different decision thresholds. The quantitative and visual comparisons among the compared methods are provided in Table 1 and Figure 6. Figure 6 compares the Sens and FA/h values of the compared methods on the SeizeIT2 test set.

Table 1. Performance comparison with baseline methods on the SeizeIT2 test set. Sens, FA/h, and HMS are reported at a classification threshold of 0.5, whereas AUROC and AUSF are threshold-independent summary metrics.

Method	Sens (%)	FA/h	HMS	AUROC	AUSF
SVM [38]	71.10	11.00	66.70	0.6995	0.7229
ChronoNet [39]	84.20	100.50	44.00	0.7921	0.7953
InceptionTime [40]	58.60	9.18	54.93	0.6933	0.7566
CMEpiNet	86.44	21.41	77.88	0.6631	0.7611
95% CI	(75.49–94.77)	(17.88–25.35)	(66.92–86.29)	(0.6072–0.7125)	(0.6485–0.8577)

Note: Values in parentheses indicate 95% confidence intervals for CMEpiNet estimated by subject-level bootstrap resampling with 2000 resamples. Confidence intervals were not estimated for benchmark methods for which subject-level predictions were unavailable.

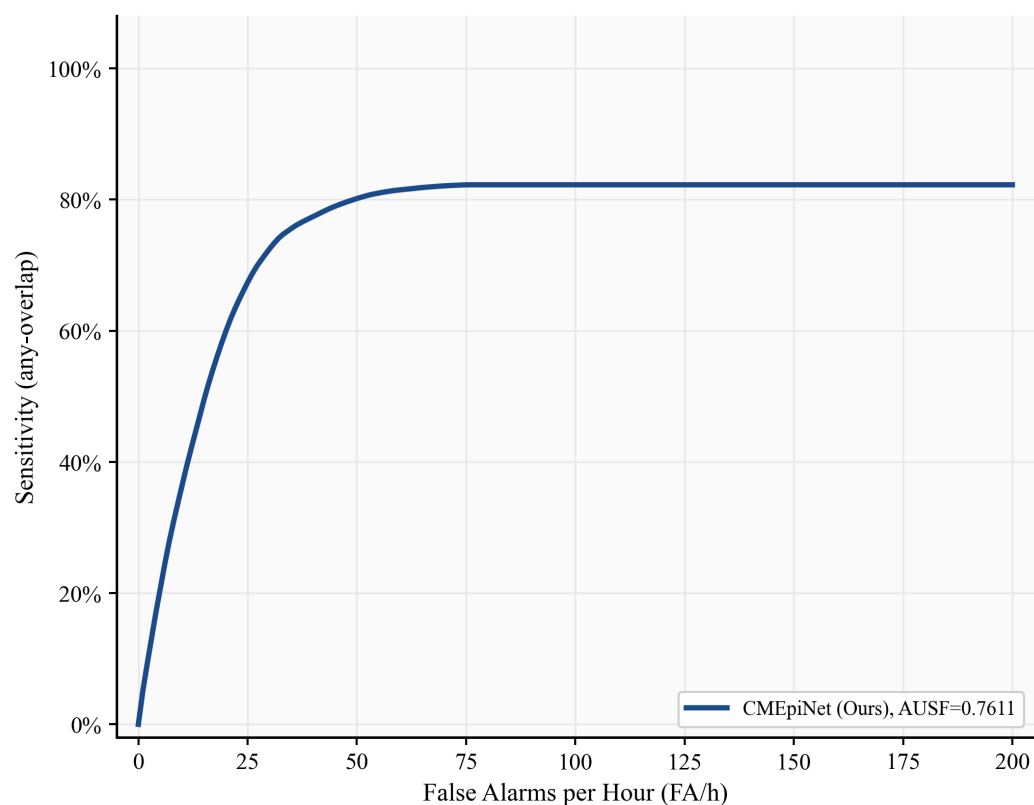


Figure 5. Threshold-dependent sensitivity–FA/h curve of CMEpiNet on the SeizeIT2 test set.

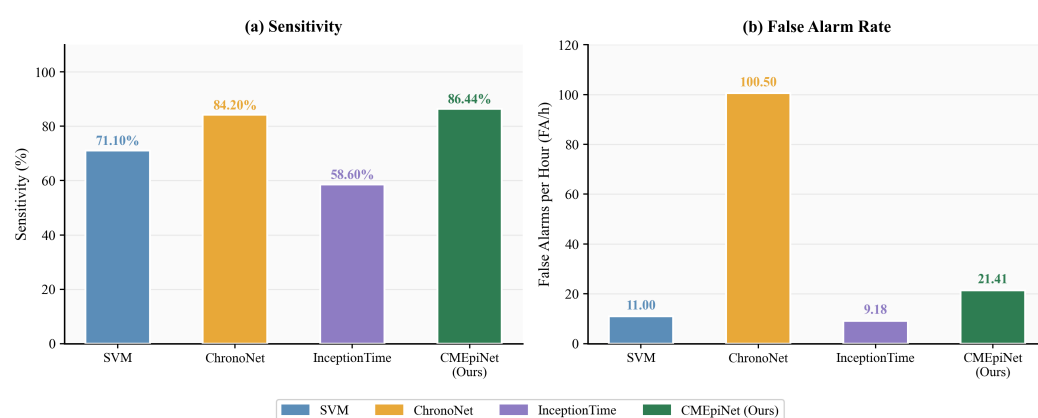


Figure 6. Comparison of sensitivity and FA/h on the SeizeIT2 test set.

Compared with SVM, CMEpiNet improves sensitivity from 71.10% to 86.44%, with an increase of 15.34%. It also improves AUSF from 0.7229 to 0.7611, with an increase of 3.82%. SVM achieves a low FA/h of 11.00 and therefore shows a certain advantage in false alarm control. However, its sensitivity is low, indicating a more serious missed-detection problem. In contrast, CMEpiNet significantly improves seizure detection ability while maintaining an acceptable false alarm level. The improvement in AUSF further indicates that this advantage is not limited to a single threshold but remains across a broader range of operating points. HMS also increased from 66.70 for SVM to 77.88 for CMEpiNet, further indicating that CMEpiNet achieves a better overall trade-off between sensitivity and false-alarm burden at the fixed decision threshold.

ChronoNet achieves a sensitivity of 84.20% and an FA/h of 100.50, indicating an extremely high false alarm level. This result means that the model triggers more than 100 false alarms per hour on average, which cannot meet the requirements of practical

clinical or home-monitoring scenarios. In contrast, CMepiNet improves sensitivity from 84.20% to 86.44%, with an increase of 2.24%. It also reduces FA/h from 100.50 to 21.41. These results show that CMepiNet improves seizure detection ability while substantially improving false alarm control. HMS also increased from 44.00 for ChronoNet to 77.88 for CMepiNet, indicating that CMepiNet achieves a more favorable balance between seizure detection performance and false-alarm control at the fixed operating point.

Compared with InceptionTime, CMepiNet improves sensitivity from 58.60% to 86.44%, corresponding to an increase of 27.84 percentage points. InceptionTime achieves a low FA/h of 9.18, indicating better false-alarm control, but its substantially lower sensitivity suggests a higher risk of missed seizures in this task. Although CMepiNet has a higher FA/h than InceptionTime, its HMS increases from 54.93 to 77.88. This result indicates that CMepiNet achieves substantially higher seizure detection sensitivity with a moderate increase in false alarms, leading to a better overall trade-off at the fixed decision threshold.

4.5. Sensitivity Analysis by Seizure Type

To examine whether the overall sensitivity masked subtype-specific detection deficits, we performed a seizure-type stratified analysis (Table 2). CMepiNet detected all focal to bilateral tonic-clonic (FBTC) seizures (55/55, 100.00%), and achieved sensitivities of 84.17% (101/120) for focal impaired awareness (FIA) seizures and 78.95% (45/57) for focal aware (FA) seizures. Four additional focal seizures were not further subclassified, of which three were detected. Because this category contained only four events, we report it for completeness without further statistical inference.

The subtype-specific results suggest a clinically plausible gradient in detection difficulty. Sensitivity was highest for FBTC seizures and decreased for FIA and FA seizures, consistent with the weaker motor and autonomic manifestations of less severe focal seizures. This pattern is also consistent with wearable multimodal detection, where FA seizures may provide weaker ECG and EMG evidence than FBTC seizures. Nevertheless, the per-type results do not suggest a complete failure in any major subtype: all clinically high-risk FBTC seizures were detected, and the more subtle FA seizures still reached a sensitivity of 78.95%.

This analysis also helps interpret the observed window-level AUROC. The missed detections were mainly from FA and FIA seizures, which may show weaker cross-modal signatures within individual 2 s windows. Their lower window-level separability may contribute to a weaker overall probability ranking and thus to the non-superior AUROC of CMepiNet. In contrast, by leveraging multimodal fusion and event-level post-processing, CMepiNet still detected a substantial proportion of seizure events. These results support our interpretation that the main advantage of CMepiNet lies in event-level seizure detection rather than window-level probability ranking.

Table 2. Sensitivity of CMepiNet by seizure type on the SeizelT2 test set.

Seizure Type	TP	FN	Sens (%)
FBTC (focal to bilateral tonic-clonic)	55	0	100.00
FIA (focal impaired awareness)	101	19	84.17
FA (focal aware)	45	12	78.95
Focal (unspecified) ^a	3	1	75.00
Overall	204	32	86.44

^a Only four events were available in this category; the result is listed for completeness and was not used for statistical inference.

4.6. Ablation Study

The ablation experiments on the SeizeIT2 dataset verify the effectiveness of the CMFE and DLSAF modules. For this purpose, the experiments construct three ablation variants. Table 3 reports the results of all variants on the SeizeIT2 test set. Figure 7 compares Sensitivity and FA/h.

Variant 1, RealConv+Linear, uses a three-layer real-valued Conv1d branch to extract features and uses linear projection for modality mapping. It does not introduce any cross-modal alignment loss.

Variant 2, CMFE+Linear, uses the dual-branch complex-valued CMFE to extract multimodal features and uses linear projection for modality mapping. It does not introduce the cross-modal soft-window consistency constraint.

Variant 3, CMFE+Linear+ $\mathcal{L}_{cm}$, uses the dual-branch complex-valued CMFE to extract multimodal features and introduces the cross-modal soft-window consistency constraint.

CMEpiNet further introduces shared latent-space alignment based on Variant 3 and forms the complete dual-level semantic alignment fusion mechanism.

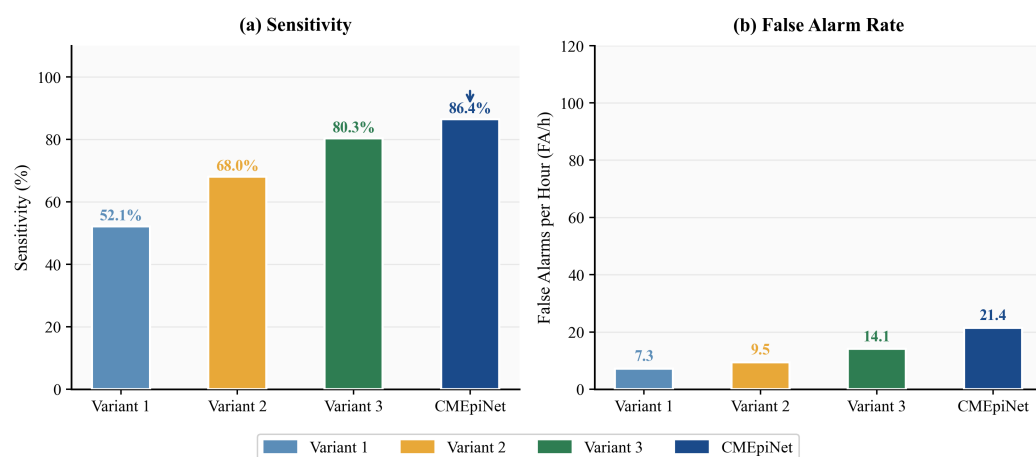


Figure 7. Ablation study of CMEpiNet on the SeizeIT2 test set in terms of sensitivity and FA/h.

Table 3. Ablation study results on the SeizeIT2 test set.

Model	Sens (%)	FA/h	AUROC	AUSF
Variant 1	52.15	7.26	0.6962	0.7597
Variant 2	68.04	9.46	0.6933	0.7786
Variant 3	80.26	14.13	0.6936	0.7620
CMEpiNet	86.44	21.41	0.6631	0.7611

After introducing the dual-branch complex-valued CMFE, sensitivity increases from 52.15% to 68.04%. AUSF increases from 0.7597 to 0.7786. These results show that complex-valued feature extraction significantly enhances the seizure detection ability of the model. The time-domain phase branch constructs analytic signals through the Hilbert transform and jointly models instantaneous amplitude and phase information. The time–frequency branch computes the STFT using the original signal and its Hilbert transform, thereby preserving a more complete spectral phase structure in the time–frequency plane. These designs enhance the model’s ability to represent seizure-related time–frequency patterns.

After introducing the cross-modal consistency constraint based on Variant 2, sensitivity increases from 68.04% to 80.26%. This result shows that the cross-modal consistency constraint further improves the seizure detection rate. This module first uses reliability gating to estimate the credibility of each modality at local time steps. It then uses the soft-

window matching loss to learn temporally consistent cross-modal representations. This mechanism allows the model to use information from other modalities for compensation when a single modality is disturbed or ambiguous, thereby reducing missed detections. Meanwhile, AUSF decreases from 0.7786 to 0.7620, and FA/h increases from 9.46 to 14.13. Although Level-1 alignment improves the model's sensitivity to subtle seizure patterns, it also increases false activation in some non-seizure segments. This result reflects the inherent trade-off between sensitivity and false alarms.

After further introducing shared latent-space alignment based on Variant 3, sensitivity increases from 80.26% to 86.44%, with an increase of 6.18%. This result shows that global distribution alignment further enhances the seizure detection ability of the model. This module imposes alignment constraints on the mean vectors and covariance matrices of the three modalities in the feature space, namely $\mathcal{L}_{dist}$. This constraint encourages the semantic representations of EEG, ECG, and EMG to become more consistent in the overall distribution, thereby reducing representation shifts caused by modality heterogeneity. Although the difference in AUSF between Variant 3 and the full model is small, with values of 0.7620 and 0.7611, sensitivity continues to improve. This indicates that shared latent-space alignment makes a substantive contribution to seizure detection ability.

Based on the ablation results in Table 3, the following conclusions can be drawn. As the core modules are gradually introduced, the model sensitivity continuously improves, reaching 52.15%, 68.04%, 80.26%, and 86.44%, respectively. From the basic baseline to the full model, sensitivity increases by 34.29%, indicating that each core module effectively improves seizure detection ability. Meanwhile, FA/h also increases from 7.26 to 9.46, 14.13, and 21.41. This indicates that the model produces more false alarms while improving sensitivity. As the feature representation ability becomes stronger, the model becomes more sensitive to weak seizure patterns and also produces a higher false alarm level. This result reflects the trade-off between improved detection ability and increased false alarms, and it should not be simply interpreted as performance degradation. On the other hand, the AUSF values of all ablation variants are close to or higher than 0.76. This indicates that the performance changes do not only reflect the movement of a single-threshold operating point but also maintain relatively stable comprehensive performance across thresholds.

4.7. Robustness Experiments

To evaluate the robustness of CMepiNet under input perturbations, we conducted robustness experiments. The baseline model corresponded to Variant 1 in the preceding ablation study, namely RealConv+Linear. This model uses a three-layer real-valued 1D convolutional branch for multimodal feature extraction and linear projection for modality mapping. RealConv+Linear does not include the complex-valued multimodal feature extraction module (CMFE), the Dual-Level Semantic Alignment Fusion module (DL-SAF), or the 3D Convolution with Spatial Attention (3DCSA) classifier incorporated in CMepiNet. Specifically, it lacks the cross-modal consistency constraint and the distribution-level alignment in the shared latent space introduced by DLSAF. Therefore, this comparison was used to analyze the robustness gains contributed by these core modules under input perturbations. AUROC and AUSF were used to evaluate model robustness under perturbation conditions.

The performance change caused by perturbation is defined as the difference between the metric value under the baseline condition and that under the perturbation condition for the same model:

$$\Delta M = M_{\text{baseline}} - M_{\text{perturbed}},$$

where M denotes AUROC or AUSF. If $\Delta M > 0$, the performance decreases after perturbation. If $\Delta M = 0$, the metric remains unchanged before and after perturbation. If $\Delta M < 0$, the metric after perturbation is slightly higher than the baseline value.

Noise perturbation is applied after window segmentation. The experiments randomly select 20% and 30% of the windows and replace the selected EEG, ECG, and EMG windows with Gaussian noise of equal power. The noise intensity of each channel is determined by the standard deviation of the original window in that channel. This setting simulates different degrees of signal quality degradation.

Temporal-shift perturbation is applied before window segmentation. To simulate acquisition delay or temporal asynchrony between auxiliary modalities and EEG, the experiments use EEG as the temporal reference and apply temporal shifts only to ECG and EMG. The shift magnitudes are set to 128, 256, and 512 samples. Since the sampling rate is 256 Hz, these three shift magnitudes correspond to temporal offsets of 0.5 s, 1.0 s, and 2.0 s, respectively.

Table 4 reports the performance degradation under each perturbation condition.

Table 4. Performance degradation under different perturbation conditions.

Perturbation Condition	CMEpiNet		Baseline Model	
	AUROC	AUSF	AUROC	AUSF
20% window noise replacement	0.0019	0.0000	0.0068	0.0135
30% window noise replacement	0.0018	0.0000	0.0091	0.0314
0.5 s auxiliary-modality temporal shift	0.0012	0.0000	−0.0004	0.0061
1.0 s auxiliary-modality temporal shift	0.0011	0.0000	0.0004	0.0091
2.0 s auxiliary-modality temporal shift	0.0013	0.0000	0.0008	0.0128

CMEpiNet shows small AUROC degradation under all five perturbation conditions, ranging from 0.0011 to 0.0019. Its AUSF remains unchanged from the baseline value. The baseline model shows a more obvious AUROC decrease under noise perturbation. Its AUSF also shows different degrees of degradation under both noise perturbation and temporal-shift perturbation.

4.7.1. Analysis of Noise Perturbation Results

Under noise perturbation, the AUROC of CMEpiNet decreases from the baseline value of 0.6631 to 0.6612 and 0.6613, corresponding to degradation values of 0.0019 and 0.0018, respectively. The AUSF of CMEpiNet remains 0.7611 under both 20% and 30% window noise replacement, corresponding to a degradation value of 0.0000 in both cases.

The AUROC of the baseline model decreases from the baseline value of 0.6962 to 0.6894 and 0.6871, corresponding to degradation values of 0.0068 and 0.0091, respectively. Its AUSF decreases from the baseline value of 0.7597 to 0.7462 and 0.7283, corresponding to degradation values of 0.0135 and 0.0314, respectively.

Under random window noise replacement, CMEpiNet shows a smaller performance decrease relative to its own baseline. Under the 30% window noise replacement condition, the AUSF degradation of the baseline model reaches 0.0314. This result shows that CMEpiNet has more stable event-level detection performance under signal quality degradation.

4.7.2. Analysis of Temporal-Shift Perturbation Results

Under temporal-shift perturbation, the AUROC of CMEpiNet changes from the baseline value of 0.6631 to 0.6619, 0.6620, and 0.6618, corresponding to degradation values of 0.0012, 0.0011, and 0.0013, respectively. The AUSF of CMEpiNet remains 0.7611 under the 0.5 s, 1.0 s, and 2.0 s auxiliary-modality temporal-shift conditions.

The AUROC of the baseline model changes from the baseline value of 0.6962 to 0.6966, 0.6958, and 0.6954, corresponding to degradation values of -0.0004 , 0.0004 , and 0.0008 , respectively. Under the 0.5 s auxiliary-modality temporal-shift condition, the AUROC of the baseline model is slightly higher than the baseline value, and therefore the degradation value is negative. The AUSF of the baseline model decreases from the baseline value of 0.7597 to 0.7536, 0.7506, and 0.7469, corresponding to degradation values of 0.0061, 0.0091, and 0.0128, respectively.

Both models maintain relatively stable AUROC under auxiliary-modality temporal shifts. However, the AUSF of the baseline model decreases as the temporal-shift magnitude increases. The AUSF of CMEpiNet shows no visible degradation under the three temporal-shift conditions. This indicates that its event-level detection performance has stronger stability under temporal asynchrony of auxiliary modalities.

4.7.3. Robustness Analysis

Figure 8 shows the trends of AUROC and AUSF for the two models under different perturbed test sets. Figure 8a shows that the AUROC curve of CMEpiNet remains nearly horizontal under all perturbation conditions, with values consistently maintained between 0.6612 and 0.6631. This indicates that perturbations have little influence on its window-level discriminative ability. The baseline model shows a more obvious decrease under noise perturbation. Figure 8b shows that the AUSF of CMEpiNet remains 0.7611 under all conditions, and its curve stays stable. In contrast, the AUSF of the baseline model decreases under both noise perturbation and temporal-shift perturbation. This figure indicates that CMEpiNet achieves more stable performance under different perturbation conditions, especially in terms of AUSF.

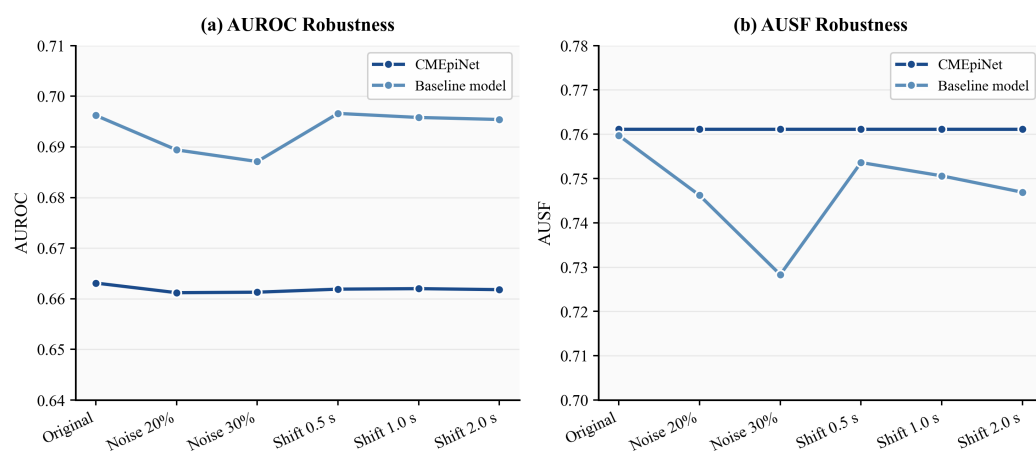


Figure 8. AUROC and AUSF trends of CMEpiNet and the baseline model under different perturbation conditions.

Figure 9 shows the performance degradation of the two models relative to their own baselines under window-level Gaussian noise replacement. The degradation value is defined as the baseline performance minus the perturbed performance. Therefore, a larger value indicates that the model is more strongly affected by perturbation. Figure 9a shows that under 20% and 30% window noise replacement, the AUROC degradation values of CMEpiNet are 0.0019 and 0.0018, whereas those of the baseline model are 0.0068 and 0.0091. Figure 9b shows that the AUSF degradation values of CMEpiNet are 0.0000 under both noise conditions. In contrast, the degradation values of the baseline model are 0.0135 and 0.0314. These results show that CMEpiNet exhibits a smaller relative performance decrease under signal quality degradation and achieves more stable robustness in terms of AUSF.

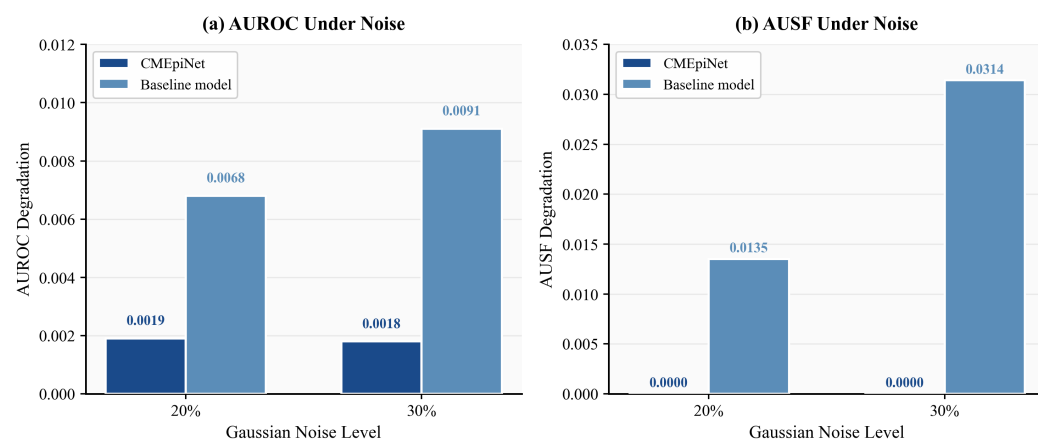


Figure 9. Performance degradation under window-level Gaussian noise replacement.

Figure 10 shows the performance degradation of the two models relative to their own baselines under auxiliary-modality temporal shifts. In the temporal-shift experiments, EEG serves as the reference, and temporal offsets of 0.5 s, 1.0 s, and 2.0 s are applied to ECG and EMG. Figure 10a shows that both models maintain small AUROC degradation. Figure 10b shows that the AUSE degradation of CMEpiNet remains 0.0000 under all three temporal-shift conditions, whereas the AUSE degradation values of the baseline model are 0.0061, 0.0091, and 0.0128. These values increase gradually with the shift magnitude. This result indicates that CMEpiNet has stronger event-level stability against temporal asynchrony of auxiliary modalities.

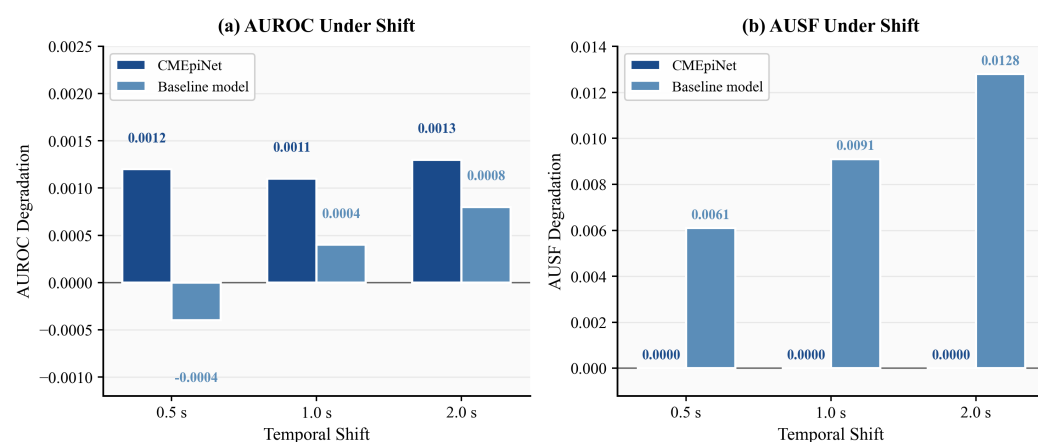


Figure 10. Performance degradation under auxiliary-modality temporal shifts.

This section analyzes model stability under perturbation conditions using AUROC and AUSE. AUROC reflects the window-level binary discriminative ability, whereas AUSE reflects the comprehensive trade-off between event-level detection performance and false alarm control under multiple thresholds. Therefore, these two metrics provide the main basis for robustness analysis.

In summary, CMEpiNet shows smaller relative performance degradation under window noise replacement and auxiliary-modality temporal shifts. It shows particularly stronger stability in terms of AUSE. This indicates that CMEpiNet has good robustness in the seizure detection task.

4.8. Computational Complexity Analysis

CMEpiNet remains compact in terms of model size, with 1.12 M parameters. It requires 12.3 GFLOPs to process a 60 s multimodal input, corresponding to 411 MFLOPs per 2 s

window. On an RTX 4060 GPU, the average inference time is 0.17 ms per window under batch processing and 0.71 ms per window under single-window streaming, corresponding to real-time factors of 9×10^{-5} and 3.6×10^{-4} , respectively. As summarized in Table 5, the real-time factor remains far below 1 in both settings, indicating that CMepiNet supports faster-than-real-time inference. These results suggest that the proposed model remains computationally efficient despite its complex-valued multimodal design and has potential for long-term wearable seizure monitoring.

Table 5. Computational complexity of CMepiNet. RTF denotes the real-time factor (processing time divided by signal duration); a value below 1 indicates faster-than-real-time processing.

Metric	Value
Parameters	1.12 M
FLOPs (per 60 s input)	12.3 G
FLOPs (per 2 s window)	411 M
Inference time (batch, per window)	0.17 ms
Inference time (streaming, per window)	0.71 ms
Throughput (batch)	5771 windows/s
RTF (batch/streaming)	$9 \times 10^{-5}/3.6 \times 10^{-4}$

5. Conclusions

This study proposes a new multimodal seizure detection model, CMepiNet, to achieve high-accuracy seizure detection using multisource physiological signals. The complex-valued multidomain feature extraction module extracts time-domain-phase features and time-frequency features from raw EEG, ECG, and EMG signals. This process provides a more comprehensive multiscale feature representation. The cross-modal alignment and fusion mechanism then reduces distribution differences among different modalities and strengthens seizure-related semantic consistency across modalities. This improves the quality of cross-modal feature fusion. Finally, the 3D spatial-modality attention classifier identifies the fused multimodal features and captures key dynamic seizure patterns from the time-feature-modality structure.

Experiments on the public SeizureIT2 multimodal epilepsy dataset show that CMepiNet achieves better performance than baseline methods in seizure sensitivity and false alarm control. The ablation experiments verify the effectiveness of each core module. The robustness experiments further show that the model maintains stable detection performance under signal quality degradation and temporal asynchrony of auxiliary modalities.

The performance improvement of CMepiNet mainly comes from three aspects. First, the complex-valued multidomain feature extraction module (CMFE) effectively enhances single-modality feature representation. Second, the cross-modal consistency constraint and latent-space alignment mechanism in DLSAF reduce structural differences among multimodal data. Third, the 3DCSA classifier efficiently captures multimodal dependencies in a three-dimensional space and improves the model's ability to represent complex seizure-related dynamic structures.

This study has several limitations. First, all experiments were conducted on a single dataset, SeizureIT2. Although SeizureIT2 is a large-scale and clinically representative wearable epilepsy dataset, the cross-dataset generalizability of CMepiNet has not yet been established on an independent external cohort. We used SeizureIT2 because it is one of the few public datasets that provide synchronized wearable EEG, ECG, and EMG recordings, which are required by the proposed multimodal framework. Commonly used public epilepsy datasets, such as CHB-MIT and EPILEPSIAE, do not provide this configuration. Evaluating CMepiNet on these datasets would require removing the ECG and EMG branches, thereby

reducing the task to single-modality EEG detection and preventing a like-for-like external validation of the proposed multimodal model.

Second, although the present evaluation includes overall and seizure-type-specific analyses, detailed per-patient performance analysis remains limited. The subject-level bootstrap confidence intervals partly quantify inter-patient variability, but they do not replace individualized evaluation. Future work should further explore personalized modeling and per-patient threshold adaptation to improve robustness across patients.

Overall, CMEpiNet shows promising event-level detection performance and robustness within the SeizeIT2 evaluation setting. However, the current FA/h rate should still be interpreted as a research-stage result rather than evidence of immediate readiness for clinical alarm deployment. Although CMEpiNet improves seizure-event sensitivity and achieves a better trade-off between sensitivity and false-alarm burden, false alarms remain an important limitation for long-term wearable monitoring. In addition, the proposed cross-modal alignment and three-dimensional attention modeling framework has good generality. It can be extended to other multimodal analysis scenarios, such as intelligent transportation and smart education, and provides a new technical pathway for complex event recognition.

Author Contributions: Writing—original draft preparation, T.S.; resources, H.Z.; writing—review and editing, S.C.; project administration, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Innovation Capacity Improvement Project for Technology-based SMEs in Shandong Province, grant number 2023TSGC0449; the Youth Innovation Team Development Plan for Universities in Shandong Province, grant number 2021QCY003; the College Students' Innovative Entrepreneurial Training Plan Program Project, grant number 202510424091.

Data Availability Statement: The data presented in this study are openly available in OpenNeuro at <https://doi.org/10.18112/openneuro.ds005873.v1.1.0>, reference number ds005873.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, Q.; Cao, W.; Zhang, A. Multi-stream feature fusion of vision transformer and CNN for precise epileptic seizure detection from EEG signals. *J. Transl. Med.* **2025**, *23*, 871. [[CrossRef](#)] [[PubMed](#)]
2. Eilbes, M.; Gallo, A.; Osmani, W.A.; Bittencourt-Silva, P.; Manis, A.D.; Hodges, M.R. Unique features of seizure-induced cardiorespiratory failure in SS Kcnj16^{-/-} rats: Implications for sudden unexpected death in epilepsy. *Epilepsia* **2025**, *66*, 2578–2591. [[PubMed](#)]
3. Smith, S.J. EEG in the diagnosis, classification, and management of patients with epilepsy. *J. Neurol. Neurosurg. Psychiatry* **2005**, *76*, ii2–ii7. [[CrossRef](#)] [[PubMed](#)]
4. Oliva, J.T.; Rosa, J.L.G. Classification for EEG report generation and epilepsy detection. *Neurocomputing* **2019**, *335*, 81–95. [[CrossRef](#)]
5. Liu, G.; Xiao, R.; Xu, L.; Cai, J. Minireview of epilepsy detection techniques based on electroencephalogram signals. *Front. Syst. Neurosci.* **2021**, *15*, 685387. [[CrossRef](#)] [[PubMed](#)]
6. Hjorth, B. EEG analysis based on time domain properties. *Electroencephalogr. Clin. Neurophysiol.* **1970**, *29*, 306–310. [[CrossRef](#)] [[PubMed](#)]
7. Jansen, B.H.; Bourne, J.R.; Ward, J.W. Autoregressive estimation of short segment spectra for computerized EEG analysis. *IEEE Trans. Biomed. Eng.* **1981**, *9*, 630–638. [[CrossRef](#)]
8. Gotman, J. Automatic recognition of epileptic seizures in the EEG. *Electroencephalogr. Clin. Neurophysiol.* **1982**, *54*, 530–540. [[CrossRef](#)] [[PubMed](#)]
9. Qu, H.; Gotman, J. A patient-specific algorithm for the detection of seizure onset in long-term EEG monitoring: Possible use as a warning device. *IEEE Trans. Biomed. Eng.* **1997**, *44*, 115–122. [[PubMed](#)]
10. Adeli, H.; Zhou, Z.; Dadmehr, N. Analysis of EEG records in an epileptic patient using wavelet transform. *J. Neurosci. Methods* **2003**, *123*, 69–87. [[CrossRef](#)] [[PubMed](#)]
11. Subasi, A. Epileptic seizure detection using dynamic wavelet network. *Expert Syst. Appl.* **2005**, *29*, 343–355. [[CrossRef](#)]

12. Faust, O.; Acharya, U.R.; Adeli, H.; Adeli, A. Wavelet-based EEG processing for computer-aided seizure detection and epilepsy diagnosis. *Seizure* **2015**, *26*, 56–64. [[CrossRef](#)] [[PubMed](#)]
13. Kannathal, N.; Choo, M.L.; Acharya, U.R.; Sadasivan, P. Entropies for detection of epilepsy in EEG. *Comput. Methods Programs Biomed.* **2005**, *80*, 187–194. [[CrossRef](#)] [[PubMed](#)]
14. Chua, K.C.; Chandran, V.; Acharya, U.R.; Lim, C.M. Application of higher order spectra to identify epileptic EEG. *J. Med. Syst.* **2011**, *35*, 1563–1571. [[PubMed](#)]
15. Acharya, U.R.; Molinari, F.; Sree, S.V.; Chattopadhyay, S.; Ng, K.H.; Suri, J.S. Automated diagnosis of epileptic EEG using entropies. *Biomed. Signal Process. Control* **2012**, *7*, 401–408. [[CrossRef](#)]
16. Yang, C.; Luan, G.; Liu, Z.; Wang, Q. Dynamical analysis of epileptic characteristics based on recurrence quantification of SEEG recordings. *Phys. A Stat. Mech. Its Appl.* **2019**, *523*, 507–515. [[CrossRef](#)]
17. Farahi, M.; Sajadi, S.S.; Karbasi, F.; Fesharaki, S.S.H.; Habibabadi, J.M.; Haidari, M.R.; Jafari, A.H. A Nonlinear Method to Identify Seizure Dynamic Trajectory Based on Variance of Recurrence Rate in Human Epilepsy Patients Using EEG. *J. Med. Signals Sens.* **2025**, *15*, 19. [[CrossRef](#)] [[PubMed](#)]
18. Polat, K.; Güneş, S. Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform. *Appl. Math. Comput.* **2007**, *187*, 1017–1026. [[CrossRef](#)]
19. Tzallas, A.T.; Tsipouras, M.G.; Fotiadis, D.I. Epileptic seizure detection in EEGs using time–frequency analysis. *IEEE Trans. Inf. Technol. Biomed.* **2009**, *13*, 703–710. [[CrossRef](#)] [[PubMed](#)]
20. Orhan, U.; Hekim, M.; Ozer, M. EEG signals classification using the K-means clustering and a multilayer perceptron neural network model. *Expert Syst. Appl.* **2011**, *38*, 13475–13481. [[CrossRef](#)]
21. Chua, K.; Chandran, V.; Rajendra Acharya, U.; Lim, C. Analysis of epileptic EEG signals using higher order spectra. *J. Med. Eng. Technol.* **2009**, *33*, 42–50. [[CrossRef](#)] [[PubMed](#)]
22. Nicolaou, N.; Georgiou, J. Detection of epileptic electroencephalogram based on permutation entropy and support vector machines. *Expert Syst. Appl.* **2012**, *39*, 202–209. [[CrossRef](#)]
23. Subasi, A. EEG signal classification using wavelet feature extraction and a mixture of expert model. *Expert Syst. Appl.* **2007**, *32*, 1084–1093. [[CrossRef](#)]
24. Kabir, E.; Siuly; Zhang, Y. Epileptic seizure detection from EEG signals using logistic model trees. *Brain Inform.* **2016**, *3*, 93–100. [[CrossRef](#)] [[PubMed](#)]
25. Mursalin, M.; Zhang, Y.; Chen, Y.; Chawla, N.V. Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier. *Neurocomputing* **2017**, *241*, 204–214. [[CrossRef](#)]
26. Thodoroff, P.; Pineau, J.; Lim, A. Learning robust features using deep learning for automatic seizure detection. In *Proceedings of the Machine Learning for Healthcare Conference*; PMLR: Cambridge, MA, USA, 2016; pp. 178–190.
27. Acharya, U.R.; Oh, S.L.; Hagiwara, Y.; Tan, J.H.; Adeli, H. Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Comput. Biol. Med.* **2018**, *100*, 270–278. [[CrossRef](#)] [[PubMed](#)]
28. Emami, A.; Kunii, N.; Matsuo, T.; Shinozaki, T.; Kawai, K.; Takahashi, H. Seizure detection by convolutional neural network-based analysis of scalp electroencephalography plot images. *Neuroimage Clin.* **2019**, *22*, 101684. [[CrossRef](#)] [[PubMed](#)]
29. Daoud, H.; Bayoumi, M.A. Efficient epileptic seizure prediction based on deep learning. *IEEE Trans. Biomed. Circuits Syst.* **2019**, *13*, 804–813. [[CrossRef](#)] [[PubMed](#)]
30. Hussein, R.; Palangi, H.; Ward, R.K.; Wang, Z.J. Optimized deep neural network architecture for robust detection of epileptic seizures using EEG signals. *Clin. Neurophysiol.* **2019**, *130*, 25–37. [[CrossRef](#)] [[PubMed](#)]
31. Mirowski, P.; Madhavan, D.; LeCun, Y.; Kuzniecky, R. Classification of patterns of EEG synchronization for seizure prediction. *Clin. Neurophysiol.* **2009**, *120*, 1927–1940. [[CrossRef](#)] [[PubMed](#)]
32. Tang, S.; Dunnmon, J.A.; Saab, K.; Zhang, X.; Huang, Q.; Dubost, F.; Rubin, D.L.; Lee-Messer, C. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. *arXiv* **2021**, arXiv:2104.08336.
33. Li, Y.; Liu, Y.; Guo, Y.Z.; Liao, X.F.; Hu, B.; Yu, T. Spatio-temporal-spectral hierarchical graph convolutional network with semisupervised active learning for patient-specific seizure prediction. *IEEE Trans. Cybern.* **2021**, *52*, 12189–12204.
34. Chen, F.; Chen, I.; Zafar, M.; Sinha, S.R.; Hu, X. Seizures detection using multimodal signals: A scoping review. *Physiol. Meas.* **2022**, *43*, 07TR01. [[CrossRef](#)]
35. Yang, Y.; Truong, N.D.; Eshraghian, J.K.; Maher, C.; Nikpour, A.; Kavehei, O. A multimodal AI system for out-of-distribution generalization of seizure identification. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 3529–3538. [[CrossRef](#)] [[PubMed](#)]
36. Zhang, J.; Swinnen, L.; Chatzichristos, C.; Broux, V.; Proost, R.; Jansen, K.; Mahler, B.; Zabler, N.; Epitashvilli, N.; Duempelmann, M.; et al. Multimodal wearable EEG, EMG and accelerometry measurements improve the accuracy of tonic-clonic seizure detection. *Physiol. Meas.* **2024**, *45*, 065005. [[CrossRef](#)]
37. Bhagubai, M.; Chatzichristos, C.; Swinnen, L.; Macea, J.; Zhang, J.; Lagae, L.; Jansen, K.; Schulze-Bonhage, A.; Sales, F.; Mahler, B.; et al. SeizelT2: Wearable dataset of patients with focal Epilepsy. *Sci. Data* **2025**, *12*, 1228. [[CrossRef](#)]

38. Vandecasteele, K.; De Cooman, T.; Dan, J.; Cleeren, E.; Van Huffel, S.; Hunyadi, B.; Van Paesschen, W. Visual seizure annotation and automated seizure detection using behind-the-ear electroencephalographic channels. *Epilepsia* **2020**, *61*, 766–775. [[PubMed](#)]
39. Roy, S.; Kiral-Kornek, I.; Harrer, S. ChronoNet: A deep recurrent neural network for abnormal EEG identification. In *Proceedings of the Conference on Artificial Intelligence in Medicine in Europe*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 47–56.
40. Ismail Fawaz, H.; Lucas, B.; Forestier, G.; Pelletier, C.; Schmidt, D.F.; Weber, J.; Webb, G.I.; Idoumghar, L.; Muller, P.A.; Petitjean, F. Inceptiontime: Finding alexnet for time series classification. *Data Min. Knowl. Discov.* **2020**, *34*, 1936–1962.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.