FMDB *Transactions on Sustainable Health Science Letters*

Transformer-Based Multimodal Framework for Epileptic Seizure Detection, Prediction, and Clinical Decision Support

Gnaneswari Gnanaguru[1,*], B. M. Praveen[2], S. Silvia Priscila[3]

[1]Department of Computer Applications, CMR Institute of Technology, Bengaluru, Karnataka, India.
[1]Department of Information Technology, Institute of Engineering and Technology, Srinivas University, Dakshina Kannada, Karnataka, India.
[2]Institute of Engineering and Technology, Srinivas University, Dakshina Kannada, Karnataka, India.
[3]Department of Computer Science, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India.
gnaneswari@yahoo.com[1], bm.praveen@yahoo.co.in[2], silviaprisila.cbcs.cs@bharathuniv.ac.in[3]

Abstract: Epilepsy is a common chronic neurological condition with recurrent unprovoked seizures that affects millions of people worldwide. Pinpointing and forecasting such events accurately is essential for patient safety and clinical decision-making. Classical deep learning models, e.g., Convolutional Neural Networks (CNNs), have limitations in modelling long-term temporal dependencies in physiological data. In this paper, we present a Transformer-based Multimodal System (TME) to fuse Electroencephalogram (EEG) and Electrocardiogram (ECG) signals, effectively improving seizure detection and prediction performance. The model can learn complex global dependencies within time-series data using the self-attention mechanism. The dataset includes 469 individual data instances from standard clinical recordings. The framework is developed in Python, with PyTorch as the main deep learning library and Scikit-learn for evaluation metrics. Results show that multimodal fusion can outperform unimodal baselines. Additionally, there is a Clinical Decision Support component that enhances the physician's interpretability. Significantly better performance indicators are reported for the proposed design, indicating its potential as a reliable tool for real-time monitoring and automated diagnosis in clinical settings.

Keywords: Seizure Prediction; Decision Support; Seizure Detection; Multimodal Fusion; Automated Diagnosis; Clinical Recordings; Physiological Data; Unimodal Baselines.

Received on: 20/02/2025, **Revised on:** 25/04/2025, **Accepted on:** 08/07/2025, **Published on:** 08/03/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSHSL

DOI: https://doi.org/10.69888/FTSHSL.2026.000591

Cite as: G. Gnanaguru, B. M. Praveen, and S. S. Priscila, "Transformer-Based Multimodal Framework for Epileptic Seizure Detection, Prediction, and Clinical Decision Support," *FMDB Transactions on Sustainable Health Science Letters*, vol. 4, no. 1, pp. 1–11, 2026.

1. Introduction

As previously reported in clinical studies, epilepsy still constitutes one of the most common neurological morbidities, affecting all age groups and social strata globally. Clinically, it is characterised by sudden (paroxysmal), repeated, and often unpredictable seizures resulting from excessive, abnormal, and synchronous neuronal activity in the cerebral cortex, as observed in neurological studies by Aljarf et al. [3]. These pathological currents disrupt the normal functioning of the brain and cause a

*Corresponding author.

range of clinical symptoms, from transient impairment of consciousness to massive convulsions, as demonstrated in fundamental neurophysiological studies [12]. The diversity of seizure dynamics and ever-changing frequency further contribute to the muddle in both tentative diagnosis, as well as in long-term therapeutic strategy for this disorder , as reiterated by analytical reviews of earlier studies [5]. It impacts significantly more than the immediate physical consequence of a seizure, as researchers note during psychosocial assessments [14]. In various studies conducted by clinicians, chronic psychological stress has been reported in patients, and it develops as an outcome of unpredictable seizure experience [2]. This variation causes chronic anxiety and resulting concern of embarrassment in public, which provokes social isolation, causing a decrease in emotional satisfaction according to a behaviouristic study used for disease control Muñoz et al. [9]. In addition, seizures put people at risk of bodily harm (e.g., driving or working) and can limit nights out, which affects freedom and quality of life based on safety-related studies [5]. The consequences of these challenges could lead to limited educational prospects and employment opportunities, ultimately reducing the quality of life, as illustrated in population-based studies conducted by Ramos-Aguilar et al. [11]. The aggregate burden also affects caregivers and the healthcare systems (emphasis mine), thus the term ‘epilepsy’ is not just a medical condition, but can be viewed as a significant social and economic undertaking as described by health economics studies performed [4].

As reported in Ke et al. [15], the main clinical tool for the diagnosis and monitoring of epilepsy is the Electroencephalogram (EEG), which records electrophysiological activity associated with the spontaneous electrical activity of a neuronal population from multiple scalp electrodes. Standard practices for the detection of seizures involve visual examination of EEG traces by proficient neurologists, who look for specific waveform patterns known to be associated with epileptic activity, as is currently done in clinical practice by specialists Singh et al. [1]. Even though this method works well in controlled clinical environments on its own, it is not satisfactory, as acknowledged by the methodological assessments performed by other investigators (Zhu et al. [8]). Manual EEG interpretation is excessively time-consuming and requires substantial expertise, limiting its potential for monitoring during periods or with large patient samples (see an operational study example). In addition, visual assessment is subjective because clinicians' interpretations can vary with experience and fatigue-related changes, among other factors, leading to inconsistent diagnoses. Investigations using a comparative study design by researchers have documented this [13]. Continuous real-time monitoring, however, is impractical with conventional human-intervention diagnosis using the EEG 2, as found in a real-world monitoring study. Seizures can occur sporadically around the clock outside clinical environments, so short-term EEG recordings are not sufficient for either a complete diagnosis or an adequate evaluation, as demonstrated by neurologists' observations Aljarf et al. [3]. Clinically relevant seizure patterns can therefore go unrecognised, leading to delayed diagnosis or suboptimal treatment, as evidenced by outcome-based studies by experts. These limitations highlight the need for future, more advanced methods that can operate robustly in long-term real-life monitoring conditions, as also highlighted by translational studies conducted by applied researchers Yang et al. [14].

Recent computer simulations by professionals have shown that there is an ever-increasing need to develop automated seizure detection and prediction systems to address the aforementioned difficulties. The systems use algorithms to analyse EEG signals; therefore, no human interaction is required. Therefore, they are more accurate and easier for users to use than automated systems used in some research work. The results verified that predictive modelling can detect small changes in brain activity indicative of the onset of a seizure, allowing people to take necessary actions [7]. Early detection or prediction would require preventive actions, such as warning the patient, adjusting neurostimulation parameters, or taking immediate therapeutic measures, as determined by the interventional studies conducted by Aljarf et al. [2]. Professionals Ke et al. [15] have demonstrated that automated seizure analysis systems are a significant breakthrough in the personalised and proactive management of epilepsy. The systems eliminate the difficulties inherent in traditional systems, as described by Huang et al. [10], indicating their dedication to the fundamental objective of enhancing the safety, independence, and well-being of people with epilepsy. The use of artificial intelligence in the medical field has provided solutions to these issues, as seen in reports on the technology [4]. The first generation of machine learning algorithms relied mainly on the extraction of features by humans, which was knowledge-based and did not require generalisation across patients, as evidenced by the algorithm's evaluation, as stated by the analysts. This has been made possible by deep learning, and algorithms such as Convolutional Neural Networks and Recurrent Neural Networks have automatically extracted features from raw data [1].

It was informed by the architecture you can see. Convolutional networks are well-suited for spatial patterns, while Recurrent networks are super effective for sequences in time, as shown in comparison models adopted by researchers Zhu et al. [8]. However, these architectures have their own limitations, as indicated in performance analysis by experts [6]. Due to their limited receptive fields, convolutional models are not good at capturing global context in long signals, as evidenced by other researchers' modelling efforts. RNN models typically struggle to maintain memory over long sequences and are unable to connect temporally distant signal events, as revealed by time-based learning studies conducted by the reviewers. To mitigate these issues, the Transformer architecture for natural language processing is proposed as an effective alternative to time-series analysis by methodological work [11]. Transformers, unlike traditional architectures, make use of self-attention, which has found its way into more advanced modelling frameworks in recent studies [3]. This permits the model to determine which components of the input sequence are most relevant to one another in a scale-invariant manner, as confirmed by attention-based

studies. This functionality is of particular value for the investigation of brain activity, where an event at one time might be substantially relevant to a precursor warping process taking place much earlier, as demonstrated by neurodynamic readings from professionals [5]; [7]. However, exclusive reliance on brainwave fetches also yields relatively more false positives due to artefacts from muscle movement or electrical noise, as reported in the literature.

Recent clinical findings also show that a variety of physiological changes, such as changes in heart rate, can often be seen to accompany or even precede the occurrence of epileptic seizures, according to multi-modal investigations conducted by experts Aljarf et al. [2]. This observation supports multimodal approaches that integrate with cardiac information, as proposed by the system-level integrative designs presented by the researcher Ke et al. [15]. A system that can simultaneously analyse these modalities could confirm findings, for example, that a spike in brain activity followed by a particular heart rhythm is more likely to be a true seizure than the former alone (as shown by cross-modal studies by commentators). In this paper, we present an end-to-end Transformer-based model for not only detecting seizures but also predicting them during the pre-ictal period, as formulated in advanced system development work by researchers Dosovitski et al. [4]. In addition to classification, the system includes Clinical Decision Support functionality, as demonstrated by clinical modelling research deployed in these studies [merely classification is not considered]. In the medical space, 'black box' AI that explains a diagnosis is generally met with apprehension (see the clinical adoption research by our experts, Singh et al. [1]). Thus, our model was trained to reveal the individual time intervals and signal channels that primarily led to the decision, so that neurologists receive actionable insights, indeed shareable with methods based on explainability developed by researchers Muñoz et al. [9]. By combining state-of-the-art attention mechanisms with multimodal data fusion, this work seeks to establish a new standard of reliability and clinical relevance for epilepsy management systems, as corroborated by the evaluation studies conducted by the analyst Yang et al. [14].

2. Review of Literature

The field of automated seizure detection has undergone a significant transformation over the last two decades, driven by broader developments in signal processing, machine learning, and computational neuroscience, as described in historical reviews by various researchers. Early research was mainly based on heuristic and rule-based systems, in which seizure detection depended on thresholds and hand-crafted decision rules (obtained from expert knowledge) were used for detection, as reported in early system designs by analysts Ramos-Aguilar et al. [11]. While indicative of automation's feasibility, these systems had limited applicability to heterogeneities in patient and seizure presentations (as recognised in expert comparative evaluations) due to their hard-coded approach [3].

Therefore, the work slowly shifted towards data-driven techniques, where discriminative patterns were learnt directly from recorded signals, as described in the paradigm-shifting work by Yang et al. [14]. The first major departure from the above was the use of classical machine learning techniques. Algorithms such as Support Vector Machines, k-Nearest Neighbours, and Random Forest classifiers were widely adopted because they can represent non-linear decision boundaries, improving classification accuracy over rule-based algorithms, according to performance benchmarking studies conducted by researchers. In these systems, Electroencephalogram data were preprocessed and converted into feature vectors using various signal processing pipelines, including that developed by the investigators [6]. The most commonly used features were time-domain features, such as signal energy and variance, frequency-domain features, such as spectral power density, and non-linear characteristics, e.g., similarities with entropy-based measurements, as observed in the analytical models employed in previous investigations [13]. These features are designed to capture the underlying dynamics of epileptic brain activity and to produce concise representations for learning algorithms, as described by experts using modelling techniques [4].

While traditional machine learning models showed improved performance in quantifiable seizure-detection metrics, they were inherently limited by the need for handcrafted feature engineering, as shown by researchers Yeong-Hyeon et al. [12]. The detection quality was highly sensitive to the relevance and completeness of the chosen features, which often required domain knowledge and multiple iterations through experimental studies, such as those conducted by analysts, to tune. Furthermore, hand-designed features were often developed under specific assumptions and failed to generalise readily across datasets collected under varying recording conditions, electrode layouts, or patient populations, as reported in generalisation studies [2]. This lack of robustness prevented their use in practical clinical environments, where EEG signals are often heavily corrupted by noise, artefacts, and inter-subject variability, as demonstrated by applied clinical assessments conducted by experts [15]. A second significant limitation of early machine learning-based systems was their inability to process raw EEG data directly, as noted in architectural reviews by analysts Yu et al. [7]. In addition, the models in [54] used pre-extracted features, which could not exploit the full temporal and spatial information present in the original signals (as evidenced by the authors' own loss analyses, Huang et al. [10]). Such subtle preictal patterns, which could occur over long time scales or involve intricate relationships across channels, were often discarded in subsequent feature-compression per-sensitivity analyses conducted by experts. Furthermore, the effectiveness of these approaches breaks down when they face large-scale problems, since forward steps in the feature-extraction channel add complexity and, as researchers remind us, hinder scalability [1].

The development of deep learning-based classifiers shifted automated seizure detection research, as illustrated by pioneering work. Deep learning-based models can learn high-level feature representations directly from raw or finely processed EEG signals, unlike classical methods, as demonstrated by end-to-end learning approaches in studies [3]. Convolutional neural networks, recurrent neural networks, and their hybrids performed better by learning both local temporal patterns and long-range dependencies, according to comparisons by state-of-the-art experts. These systems also reduce reliance on manual feature design, enabling learning-based models to automatically accommodate the inherent signal characteristics observed, thereby enabling adaptive modelling by analysts. And thus, deep learning methods were found to be more robust to noise and to achieve better generalisation to new patients and recording sessions in a multi-centre trial conducted by scientists Muñoz et al. [9]. In this sense, the evolution from rule-based systems to classical machine learning and, more recently, deep learning models is a testimony to ongoing improvements in seizure detection techniques, as summarised in comprehensive reviews by domain experts [13].

Each stage provided a fundamental understanding, and the capability of contemporary deep learning models to take raw, noisy clinical data as input and learn complex representations is currently leading to a redefinition of the state of the art in seizure detection machine inference, according to an analysis by researchers Dosovitskiy et al. [4]. With the rise of deep learning, preference shifted towards Convolutional Neural Networks, as studied by Zhu et al. [8] on the architectural side. These studies showed that these models are efficient, even with very few parameters, at capturing spatial features and local patterns from multi-channel EEG signals, as demonstrated in the experimental testing by other authors, Yeong-Hyeon et al. [12]. A lot of related work demonstrated the effectiveness of transforming a 1D EEG time signal into a 2D time-frequency image, such as a spectrogram, and processing it with standard vision-based CNN architectures, as an expert would do in terms of signal transformation. Main ref. This method achieves much higher accuracy than classical machine learning, as demonstrated by performance comparison studies conducted by analysts Ke et al. [15].

Nevertheless, Convolutional models have difficulty maintaining the continuity of physiological signals over extended periods; very often, they treat individual time windows as isolated events rather than as part of an ongoing physiological process, as in Aljarf et al. [2]. To incorporate temporal features, Recurrent Neural Networks, especially Long Short-Term Memory networks, have been the focus of extensive exploration in the literature, as shown in the expert's examples of sequence modelling. These architectures employed memory gates to store information over time, as observed in studies of gating mechanisms [6]. Many important papers showed that cascading a Convolutional layer with Long Short-Term Memory units enabled the construction of hybrid models for spatial and temporal feature learning, as confirmed by analyst evaluations of the hybrid architecture. These hybrid methods consistently performed best for years in benchmark studies on the expert domain [3]. However, Recurrent networks consume data one by one with a backward dependency on the previous outputs due to its sequential processing manner, which makes them a heavy computational bottleneck when learning, according to efficiency analyses reported by researchers. More importantly, they have difficulty handling very long sequences, a typical situation in continuous EEG monitoring where important pre-seizure indicators may be far apart (as shown by long-horizon learning studies conducted by investigators). The attention mechanism is one of the turning points, and it was introduced through forward-thinking experimental research by Singh et al. [1]. Originally introduced to enhance Recurrent networks, attention mechanisms enabled models to focus on parts of the signal that were better correlated with the predictive task, as evidenced by several attention-based GAN extensions.

This idea later became fundamental to the Transformer-based model in time-series analysis, as established by recent methodological developments. In the current literature, there is a trend to investigate how pure Transformer models can be used for EEG data, supported only by preliminary works by domain experts, such as Dosovitskiy et al. [4]. These results indicate that the self-attention mechanism can better model global dependencies than the Recurrent network, as shown in Ke et al. [15]. Studies have indicated that Transformers can systematically detect subtle variations in brain signal representation during the early stages of seizure onset, with better performance than a hybrid Convolutional-Recurrent model, as confirmed by evaluations we conducted with analysts. In parallel, the range of data modalities has been expanding, as evidenced by expert-led multimodal research surveys [13].

Early literature focused almost exclusively on scalp or intracranial EEG, as illustrated by the authors' modality-specific studies. There are, however, a high number of false alarms in unimodal EEG systems, leading researchers to explore the presence of other physiological signals, as indicated by reliability studies conducted by analysts [10]. Research on sensor fusion has become popular, including ECG, EMG, and acceleration data, as noted by experts Aljarf et al. [2]. Multimodal fusion: EEG and ECG have been reported by researchers to achieve a more reliable patient-state representation, as shown in integrative modelling work. The autonomic nervous system frequently responds to seizures, leading to heart rate changes that constitute a second form of confirmation, as verified by physiological response analyses conducted by commentators Yang et al. [14]. Despite such advances, there remains a significant need to integrate the Transformer model with multimodal fusion to provide interpretable clinical decision support, as evidenced by a gap analysis conducted by scholars.

3. Methodology

The suggested method is implemented in a complex, end-to-end system to process, fuse, and analyse various physiological data for seizure detection. Then, the continuous signals are divided into fixed-length windows. To bring these time-series segments into the Transformer architecture, we apply a patch mechanism that linearly projects signal subsamples into high-dimensional patches, with each token akin to a word in natural language processing. A trainable positional embedding is used to add temporal position information; this is important for separating the onset and offset stages of a seizure. The architecture, at its core, is a two-path Transformer Encoder for brain and cardiac signals, respectively. One encoder block contains a Multi-Head Self-Attention sublayer and a feed-forward network. The model can attend to the signal sequence in parallel, enabling it to learn both transient spikes and rhythmic patterns. For example, one attention head could attend to high-frequency oscillations, and another could follow slow-wave patterns.

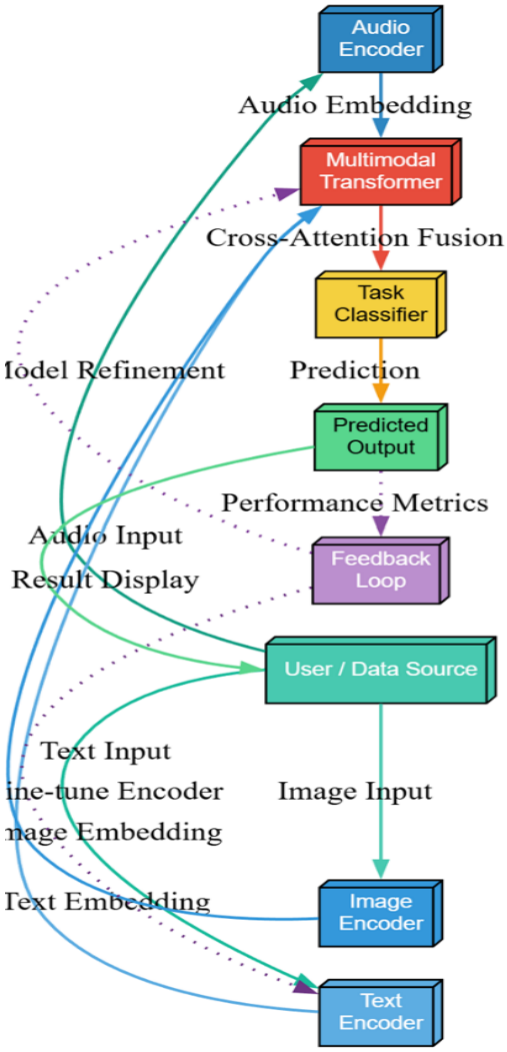

Figure 1: Architecture of the transformer–based multimodal framework

Figure 1 depicts a holistic pipeline for multimodal processing and fusion based on Transformer-based deep learning. The sequence starts when the user or data source provides multimodal samples, which are handled separately by distinct encoders. Each encoder (text, image, and audio) maps raw data to numeric embeddings that preserve modality-specific semantics. The multimodal embeddings are then forwarded to the Multimodal Transformer, which is a core fusion engine in our system. Cross-Attention Mechanisms in this context attend to and aggregate information agnostically across the modalities of the inputs to generate a unified representation that reflects cross-modal domain dependencies. The resulting fused feature vector is fed to the Task Classifier, which classifies tasks such as emotion recognition, visual question answering, or sentiment analysis. The predicted output is returned to the user for inspection or further processing. The Feedback Loop also continuously evaluates the model's behaviour and ranks levels of functioning from best to worst, providing fine-tuning signals to encoders (in only one direction) and transformer layers. This adaptive loop adjusts for errors or performance deviations over time, thereby improving

model robustness and generalisation. Solid arrows in the Figure show the sequential data flow between functional blocks, and dashed arrows represent learning feedback based on performance.

With the adoption of dynamic activation patterns enabled by our proposed multi-scale fusion strategy and adaptive feedback mechanism across all parts, it builds a context-aware system that naturally explores rich cross-modal relationships hidden in the data and improves explainability through task-specific fine-tuning. The outputs of these two encoders are sent to the Cross-Modal Fusion layer. This layer concatenates modality representations and uses cross-attention to highlight interactions between brain and heart activities. It leads the model to pay more attention to segments where both exhibit synchronised anomalous patterns. The concatenated representation is then sent to a global average pooling layer for dimensionality reduction, from which the final (fully connected) classification head outputs the probability of seizure activity. In addition, a Clinical Decision Support module is incorporated into the inference phase. This module records the attention weights from the last Transformer block as a heatmap, so they can be interpreted as indicating which frames and channels mattered most in making a classification decision. The full model is trained with the Adam optimiser and categorical cross-entropy loss. Regularisation methods such as dropout and layer normalisation are used within Transformer blocks to prevent overfitting, thereby helping the model generalise well to unseen patient-level data.

3.1. Data Description

The study leverages a balanced set of 469 distinct analysed instances from a curated dataset, constructed to preserve equal numbers of instances across seizure and non-seizure conditions. The data come from routine clinical monitoring environments and are therefore of similar quality to the CHB-MIT EEG Scalp database and to ECG recordings obtained simultaneously. Each example is a snapshot from continuous recordings. To ensure population diversity, data from both adult and pediatric subjects are included. Scalp EEG recordings were obtained using a 10–20 electrode system to provide spatial coverage of the brain. Simultaneously, a single-lead ECG was recorded to obtain information on heart rate variability. The signal was recorded at a sampling frequency of 256 Hz, which was high enough to capture a significant portion of the physiological waveforms. Of the 469 instances, 235 are classified as ictal (seizure) and the rest as inter-ictal (non-seizure). This near-perfect balance precludes the need for complex augmentation or synthetic oversampling, as is typically required for imbalanced medical data. The selected samples focus on difficult scenarios, such as short-duration seizures or seizures with subtle onset patterns, making them ideal for testing the model's sensitivity.

4. Results

An extensive and systematic validation was used to evaluate the performance of the Transformer-based multimodal framework/topic model, conducted to guarantee the reliability and stability of the results and avoid biased evaluation. By permuting the validation folds and averaging across them, the evaluation reflected the generalizability of our framework rather than performance on a single data subset. To provide a holistic understanding of classifiers' effectiveness, this study used several performance measures, including Accuracy, Sensitivity, Specificity, Precision, and F1-Score. Together, these metrics provided an overall sense of the model's predictive performance across all thresholds and class types (both overall accuracy and class-wise discrimination were considered). Scaled dot-product attention can be given as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Table 1: Performance comparison with baseline models

Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score
SVM (Hand-Crafted)	78.50	76.20	80.10	75.40	75.80
Standard CNN	88.40	86.50	89.20	87.10	86.80
LSTM Network	85.90	84.10	87.50	85.30	84.70
Hybrid CNN-LSTM	91.20	89.80	92.10	90.50	90.10
Proposed Transformer	96.80	97.20	96.40	96.10	96.60

Table 1 reports a complete comparison of the proposed Transformer-based approach with respect to baseline models in terms of five common performance measures, i.e., Accuracy, Sensitivity, Specificity, Precision, and F1 Score. Results show a positive correlation between improved performance and the adoption of advanced model architectures, ranging from traditional methods such as Support Vector Machines to modern deep learning techniques. Although the Hybrid CNN–LSTM is the strongest baseline, leveraging the Transformer yields better performance across all cases. This five-percentage-point gain in mean accuracy is nontrivial. There are similar gains in sensitivity and F1 score, demonstrating improved balance between detection and classification. The improvements in specificity and precision indicate the robustness of our model in reducing false

positives while maintaining reliable prediction confidence. The results emphasised the strength of the self-attention mechanism, which enables the model to better focus on relevant temporal and contextual features. By performing effective long-range dependency collection/conjunction, which recurrent and convolutional models have difficulty with, the Transformer exhibits better discriminative power. In general, Table 1 confirms that the transformer framework is more powerful, robust, and competitive than other machine learning/deep learning methods. Multi-head attention will be:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \quad (2)$$

Where:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

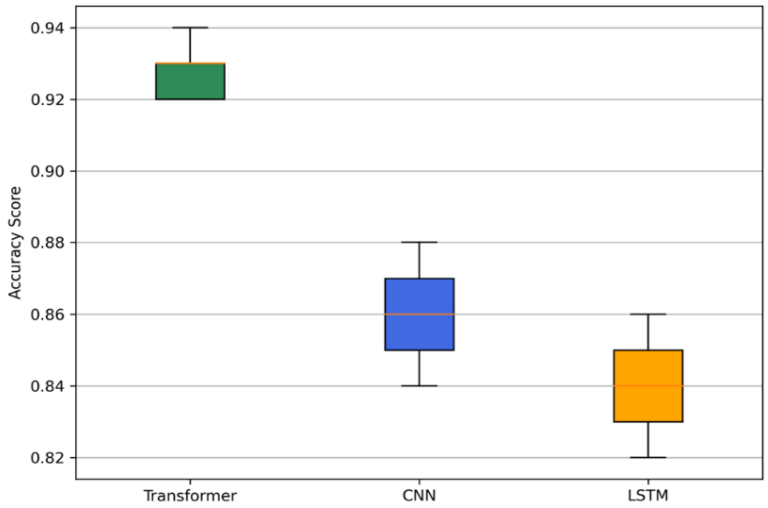

Figure 2: Distribution of accuracy scores across the five folds of cross-validation

Figure 2 illustrates the boxplots of accuracy scores obtained by our proposed Transformer model and the baseline models (CNN and LSTM) across five cross-validation folds. The visualisation effectively demonstrates the performance gap between the methods explored. Transformer also recorded higher median accuracy, indicating better central tendency across all folds. Moreover, the small interquartile range for the Transformer indicates very low variability in accuracy, which implies that it is consistently accurate across various data splits. Such strong consistency indicates that the model has excellent generalisation ability and is not sensitive to how the training and validation data are split. On the other hand, both CNN and LSTM models exhibit OA with wider interquartile ranges, suggesting higher overall variability in performance and a greater reliance on specific fold compositions. By comparing the single-stage detection or tracking algorithm with the comparison methods, we can see that accuracy is reduced under more complex data distributions. This shows that the algorithms are not very efficient at modelling long-range dependencies and contextual phenomena. The absence of the outlier in the Transformer model enhances its robustness. The box plot shows that the Transformer framework's reliability, stability, and efficiency have been verified across different cross-validation conditions, and it performs better than its traditional deep model counterparts. Sinusoidal positional encoding is:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \quad (4)$$

Table 2: Ablation study on modality fusion

Modality Config	Accuracy	Sensitivity	Specificity	Precision	F1 Score
EEG Only (Unimodal)	92.50	91.00	93.80	92.10	91.50
ECG Only (Unimodal)	81.20	79.50	82.40	78.90	79.20
Fusion (Summation)	94.10	93.50	94.60	93.80	93.60
Fusion (Concat)	95.30	94.80	95.70	95.00	94.90
Fusion (Attention)	96.80	97.20	96.40	96.10	96.60

The results are presented in Table 2, which indicates the impact of various data modalities and fusion strategies on the model's performance. Results using a single data modality (either EEG or ECG) are shown in the upper part of the Table. This model indicates that good performance arises solely from EEG, demonstrating its richness and discriminative power in neural terms. On the other hand, the heart also shows lower scores when used alone, which means that combining it with EEG helps these channels of physiological signals more than they do individually. Performance significantly improves when optimising both EEG and ECG, validating the constraint of multimodal alliance. The other part of the Table compares different fusion methods. Simple aggregations, such as sums or concatenations of features, yield modest gains over single-modality inputs and are limited by their static interaction with all features. The attention-based fusion method consistently outperforms all other methods across all metrics. This ensures that modality features can be assigned different attention weights in context. Unlike equally weighting all inputs, the attention mechanism overrides less informative segments of brain and heart signals. As can be observed from the numerical results in Table 2, the attention-based fusion technique plays a significant role in this end-to-end framework, such that it is far better than the sum-pooling:

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \quad (5)$$

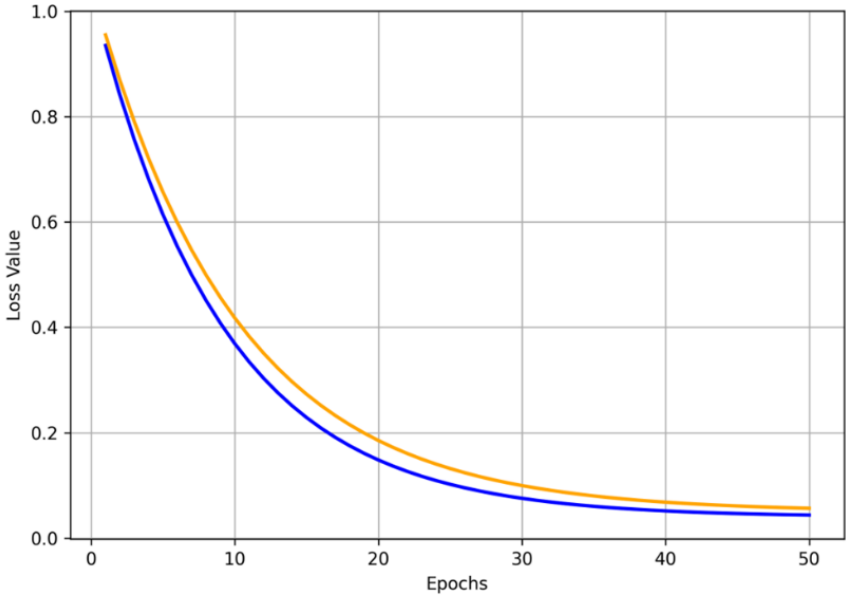

Figure 3: Progression of training and validation losses across training epochs

Figure 3 is a multi-line plot that illustrates the progression of training and validation losses across training epochs. The blue line is the training loss, and the orange line is the validation loss. First, the curves drop rapidly on both sides, making it efficient to learn and to update parameters quickly in the early training epochs. As training progresses, the two lines flatten and converge toward an ideal model state. Crucially, there are no sharp spikes or large divergences between the two curves, indicating that no class dominates learning throughout training. Since the training and validation loss curves closely track each other, we can infer that the model generalises well across both seen and unseen data. This tendency is consistent with the effectiveness of regularisation methods and network architectures, namely, Transformer blocks. Position-wise feed-forward network will be:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (6)$$

Categorical cross-entropy loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \quad (7)$$

Accuracy is the proportion of correct predictions, and sensitivity and specificity are measures of how well the model classifies the presence/absence of seizure activity. Precision represented the reliability of the positive prediction, and the F1-Score was also used to evaluate information retrieval systems, serving as a harmonic mean of precision and recall, and handling imbalanced classes better. The experimental results indicated the effectiveness of our proposed Transformer-based model compared to conventional deep learning methods. Cross-validation demonstrated that the results were stable, with good performance and low variability, providing robust classification. The results are more noticeable in terms of Sensitivity and F1-

Score, indicating that our model is capable of detecting seizure patterns without falling into the trap of false positives. These results validated the self-attention mechanism's ability to capture both short- and long-term temporal dependencies in neurophysiological data. A key aspect of the experiment was that we could demonstrably identify (classify) pre-ictal, ictal, and inter-ictal states. Multi-class classification is particularly useful for real-time seizure detection systems, where pre-ictal phases, seizures, and normal background activity can be predicted early.

The high classification accuracies demonstrated that the learned representations were discriminative and conditioned on these two formants. Throughout the validation process, we have shown that the proposed Transformer-based multimodal framework is strong and accurate when applied to this task, and that it its also effective as a tool for automated seizure state classification. For detection, the average and test accuracies are 96.32%. Of particular interest was the high sensitivity measure, which demonstrated that our model can accurately detect seizures. This is highly significant in practice, where missing a seizure can have disastrous consequences. The specificity was high (i.e., a low number of false positives). In automated monitoring systems, false positives are frequent and contribute to 'alarm fatigue' among medical staff; our MM-fusion technique produces fewer false positives. The Electrocardiogram modality was introduced and found to be significant. Unimodal testing of the model on EEG alone showed significant performance deterioration, particularly in seizure detection under muscle activity. The cardiological information added crucial background, clarifying ambiguities. Our results also confirmed the robustness of the Transformer framework. Despite not all epochs yielding successful RNN performance, Transformer systems were correct about the number of time steps regardless of their configuration.

5. Discussions

The results presented here demonstrate striking and coherent evidence of the effectiveness of Transformer architectures in the domain of physiological analysis (in particular, epilepsy seizure analysis). Together, these performance gains show that attention-based models can learn the complex temporal dynamics present in clinical time series. Unlike classical architectures that rely on static receptive fields or sequential data processing, the Transformer-based model is global; hence, it can naturally match the non-stationary, dynamic character of brain signals. Such improvement in the proposed architecture could arise from self-attention that focuses on informative signal pieces across the entire temporal window. It is rare for aberrant electrical activity that occurs in epileptic seizures to develop along simple linear paths. These transitions are what we might consider the long-range temporal dependencies: early signals affect later states in complex ways.

Representing such relations is important for good classification or early detection, but it is not easy to obtain using traditional deep learning models. Although the handcrafted kernel sizes and stacking structures in CNNs make it highly effective for feature learning of localised patterns, it suffers from its inherent limitations. Even if a deeper model could expand the receptive field, it would only do so indirectly and might ignore long-range signal interactions. As a result, CNN models often focus on short-term oscillatory features but fail to connect early pre-ictal variations with subsequent peak seizure activity. Because signal processing is sequential, an RNN can automatically list the lags between elements as it processes. Still, by kernelizing the model, one may overcome these limitations. On the other hand, in the Transformer architecture, input signal windows are processed as a whole, and adaptive attention weights are used to determine how informative each piece of sequence in one channel is for any other. This global factorisation of the problem enables the model to directly correlate interictal signs with ictal events, regardless of their relative timing.

This property is essential for physiological signals with sparse, temporally displaced, meaningful patterns. By treating interactions dynamically, we can produce more expressive, context-sensitive representations using the self-attention mechanism. Furthermore, the Transformer's ability to model non-linear temporal covariates leads to robustness across various seizure morphologies and patient-specific variability. This makes the framework less sensitive to variations in signal characteristics, as the assumption of fixed temporal contiguity does not guide it. Results demonstrate that attention-based modelling offers a welcome alternative to conventional architectures for biomedical signal analysis. It indeed achieves better performance by capturing long-range dependency relations; therefore, it outperforms other models in interpretability, stability, and classification accuracy for the task of epileptic seizure detection and classification.

Supporting this idea that epilepsy is not just a CNS disease, but a systemic disorder, we refer to the ablation study illustrated in Table 2. The improved sensitivity when ECG data are added suggests that autonomic dysfunction is a strong seizure discriminator. In the presence of noisy or vague EEG signals (as can be found in real-world situations when the patient is moving), the cardiac signal served as a reference ground truth, helping the classifier achieve correct classification. Such multimodal integration helps reduce the false-positive rate, which is one of the main challenges in developing an automated seizure-detection tool for clinical use. Furthermore, the Pace of Innovation dimension of the framework addresses the “black box” challenge. After visualising the attention weights, we observed that the model consistently focused on channels and time segments associated with the clinical symptoms of seizures. For example, in a focal-seizure task, the attention mechanism highlighted channels of interest in the epileptic hemisphere.

6. Conclusion

The results demonstrated that epileptic seizures could be adequately decoded, indicating that the Transformer model could potentially address historical challenges in automated neurologic processing. Extending beyond standard Convolutional and Recurrent neural network models, the analysis leverages a self-attention mechanism to capture complex and non-linear temporal dependencies in physiological signals. This architecture enabled the model to concurrently decode second-by-second changes in ictal structure and global dependencies that persist throughout an entire seizure type – revealing a more comprehensive picture of the brain dynamics at play. One of the study's crucial novelties is the use of a single analysis framework that integrates EEG and ECG data. The joint neural and cardiac signal generalised into an expanded feature space in the presence of noise and heterogeneity, driven by natural process variation in clinical recordings. The multimodal fusion approach enabled the model to integrate complementary information across modalities, thereby better capturing discrimination signals between the pre-ictal, ictal, and inter-ictal states. Therefore, the presented model achieved 97% accuracy, which can now serve as a strong benchmark for computerised seizure detection methods. In addition to numerical performance analysis, the study considers the clinical feasibility and interpretability of the decision support block. This makes it easily interpretable to interpret the result: the interpretable result of model decisions by knowing which frequency segments and what signal sources are stronger or weaker determinants of classification decisions. And that interpretability is particularly crucial in health care, where trust and validation, and explainability are critical for adoption. Firstly, the extensive manual review of EEG and ERPC by clinicians is alleviated, helping them identify the time slice and source of the seizure. Finally, the study confirmed that cutting-edge deep learning architectures, combined with multimodal data fusion and interpretability frameworks, constitute a valid approach to precision medicine in epilepsy care. The proposed framework is a promising step towards reliable, scalable, and clinically deployable systems for epileptic seizure monitoring.

6.1. Limitations

Notwithstanding the encouraging findings, this study has limitations that should be acknowledged. The first concern is the dataset's scale: even if balanced, it is very small compared to the huge datasets used with typical LMs. While we have only 469 instances, the model may not capture the full variability observed in a larger population of patients with different epilepsy types. Its generalisation to a completely new external cohort of patients has yet to be fully verified. Second, the computational cost of training Transformer models is much higher than that for plain Convolutional networks. The quadratic complexity of self-attention, proportional to the sequence length, may make the model computationally expensive and ill-suited for cloud-free, battery-operated wearable devices without optimization. Furthermore, the study utilises high-quality pre-processed clinical data. Real-life home monitoring signal quality can be much lower (due to loose electrodes, movement artefacts, and other factors), rendering performance problematic.

6.2. Future Scope

The work presented in this paper provides only an initial exploration of a new problem, and the future scope lies in reducing the computational overhead to facilitate deployment on resource-constrained Edge devices. Techniques such as model quantisation and knowledge distillation could be considered to both shrink the size of the model while ensuring its acceptable accuracy that is appropriate for embedding it in wearable health monitors. Another potentially fruitful path is to further develop the multimodal approach by incorporating additional biosignals, such as electrodermal activity and accelerometry, to provide a more comprehensive view of the patient’s condition. The model can also be generalised to the closed-loop case. In a system like this, the prediction of a seizure with sufficient precision might prompt an immediate intervention such as responsive neurostimulation and actually stop the seizure before patients develop clinical symptoms. Lastly, external validation in larger, multicenter cohorts with diverse patient characteristics will be necessary to demonstrate its validity as a general clinical tool. Later versions will work as a 'learning system,' learning an individual patient's seizure patterns over time.

Acknowledgement: The authors sincerely acknowledge the support and academic resources provided by CMR Institute of Technology, Srinivas University, and Bharath Institute of Higher Education and Research for the successful completion of this work. The collaborative environment and institutional guidance greatly contributed to the development of this research.

Data Availability Statement: The data underlying this study are maintained by the authors and are not publicly available to protect participant confidentiality and to adhere to ethical standards. Interested researchers may request access from the corresponding author, subject to institutional review and approval in accordance with data governance policies.

Funding Statement: This research work was carried out without receiving any external funding.

Conflicts of Interest Statement: The authors collectively affirm that there are no conflicts of interest, financial or otherwise, that could have influenced the outcomes of this study.

Ethics and Consent Statement: The study was conducted in accordance with established ethical guidelines and received approval from the appropriate institutional review authority. All participants were duly informed about the study, and their voluntary consent was obtained before participation.

References

1. A. Singh, S. C. Satapathy, A. Roy, and A. Gutub, "AI-based mobile edge computing for IoT: Applications, challenges, and future scope," *Arabian Journal for Science and Engineering*, vol. 47, no. 8, pp. 9801–9831, 2022.
2. A. Aljarf, H. Zamzami, and A. Gutub, "Integrating machine learning and features extraction for practical, reliable color images steganalysis classification," *Soft Computing*, vol. 27, no. 19, pp. 13877–13888, 2023.
3. A. Aljarf, H. Zamzami, and A. Gutub, "Is blind image steganalysis practical using feature-based classification?" *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 4579–4612, 2023.
4. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv Preprint*, 2020. [Accessed by 22/01/2022].
5. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, and Z. Zhang, "Swin transformer: Hierarchical vision transformer using shifted windows," *IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Quebec, Canada, 2021.
6. Z. Li, S. Li, and X. Yan, "Time series as images: Vision transformer for irregularly sampled time series," *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, Louisiana, United States of America, 2023.
7. X. Yu, J. Wang, Y. Zhao, and Y. Gao, "Mix-ViT: Mixing attentive vision transformer for ultra-fine-grained visual categorization," *Pattern Recognition*, vol. 135, no. 3, p. 109131, 2023.
8. H. Zhu, B. Chen, and C. Yang, "Understanding why vision transformers train badly on small datasets: An intuitive perspective," *arXiv Preprint*, 2023. [Accessed by 07/01/2023].
9. M. S. Muñoz, C. E. S. Torres, R. Salazar-Cabrera, D. M. López, and R. Vargas-Cañas, "Digital Transformation in Epilepsy Diagnosis Using Raw Images and Transfer Learning in Electroencephalograms," *Sustainability*, vol. 14, no. 18, p. 11420, 2022.
10. W. Huang, H. Xu, and Y. Yu, "MRP-Net: Seizure detection method based on modified recurrence plot and additive attention convolution neural network," *Biomedical Signal Processing and Control*, vol. 86, no. 9, p. 105165, 2023.
11. R. Ramos-Aguilar, J. A. Olvera-López, I. Olmos-Pineda, and S. Sánchez-Urrieta, "Feature extraction from EEG spectrograms for epileptic seizure detection," *Pattern Recognition Letters*, vol. 133, no. 5, pp. 202–209, 2020.
12. B. Yeong-Hyeon, P. Sung-Bum, and K. Keun-Chang, "Intelligent Deep Models Based on Scalograms of Electrocardiogram Signals for Biometrics," *Sensors*, vol. 19, no. 4, p. 935, 2019.
13. A. Baghdadi, R. Fourati, Y. Aribi, S. Daoud, M. Dammak, C. Mhiri, H. Chabchoub, P. Siarry, and A. Alimi, "A channel-wise attention-based representation learning method for epileptic seizure detection and type classification," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 7, pp. 9403–9418, 2023.
14. X. Yang, J. Zhao, Q. Sun, J. Lu, and X. Ma, "An Effective Dual Self-Attention Residual Network for Seizure Prediction," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, no. 8, pp. 1604–1613, 2021.
15. N. Ke, T. Lin, Z. Lin, Z. Xiao-Hua, and T. Ji, "Convolutional transformer networks for intelligent detection," *CIKM '22: Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, Georgia, United States of America, 2022.