FMDB *Transactions on Sustainable Health Science Letters*

Temporal Fusion Transformer-Based Multimodal Approach for Epileptic Seizure Detection and Prediction

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

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

Abstract: Epilepsy is a recurrent and chronic neurologic condition that leads to a change in the psychological comfort and safety of patients, so elaboration of efficient tools of detection and early prediction is especially important. Conventional methods that rely on handcrafted features, CNNs, and LSTMs have performed well but are hampered by their inability to generalize findings from patient to patient, to integrate multiple inputs and modalities effectively, and to provide clinical interpretability. To address these difficulties, this paper proposes a multimodal framework combining EEG, ECG, and clinical metadata for robust seizure detection and prediction using a Temporal Fusion Transformer (TFT). The model leverages variable selection networks, static covariate encoders, temporal attention, and gated residual networks to capture both short- and long-term dependencies and ensure interpretability through feature importance and temporal heat maps. Experiments with benchmark datasets (CHB-MIT, TUH EEG Seizure Corpus) indicate that the proposed framework significantly surpasses the baseline models (CNN, LSTM, Transformer) in sensitivity (94.1%), specificity (92.3%), F1-score (93.6%), and AUC-ROC (0.96), and has a low false alarm rate essential to successful real-world deployment. These findings demonstrate the usefulness of TFT in consolidating multimodal evidence, making fewer false-positive predictions, and increasing clinical confidence, thereby supporting its application in real-time seizure management systems.

Keywords: Epilepsy Condition; Seizure Detection; Temporal Fusion Transformer (TFT); Seizure Prediction; Multimodal Deep Learning; Clinical Interpretability; Conventional Methods.

Received on: 22/03/2025, **Revised on:** 25/05/2025, **Accepted on:** 22/08/2025, **Published on:** 08/03/2026

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

DOI: https://doi.org/10.69888/FTSHSL.2026.000594

Cite as: G. Gnanaguru, S. S. Priscila, and B. M. Praveen, "Temporal Fusion Transformer-Based Multimodal Approach for Epileptic Seizure Detection and Prediction," *FMDB Transactions on Sustainable Health Science Letters*, vol. 4, no. 1, pp. 59–72, 2026.

1. Introduction

Epilepsy is a neurological disorder that has a prevalence rate of about 50 million people globally and is thus one of the most prevalent neurological disorders identified by the World Health Organization (WHO). It is typified by frequent and unpredictable seizures that are a result of abnormal firing of brain neurons. In addition to the impact they have on the normal

*Corresponding author.

life of a patient, seizures may also cause injuries, sudden unexpected death in epilepsy (SUDEP), as well as serious psychological and social effects. The detection and prediction of seizures at an early stage are among the demands of clinical practice, with significant clinical implications. With the potential to predict seizures minutes in advance, clinicians can intervene through an array of measures, including medication adjustment and neurostimulation, achieving significant safety and quality-of-life improvements for patients [1]. EEG is the primary modality for seizure detection and prediction because it provides a non-invasive means of monitoring brain electrical activity at high temporal resolution [18]. Other physiological signals, including ECG and patient-specific metadata (age, gender, medical history, and seizure type), have also been shown to provide complementary information to EEG [2]. In one example, autonomic changes in the ECG precede clinical seizures, and static covariates can inform individualized treatment strategies [21]. Despite this abundance of information, the effective computational approach to multimodal inputs remains daunting.

The first difficulty is the non-stationarity of EEG signals, which are highly variable over time, highly noisy, and vulnerable to artifacts, e.g., eye blinks or muscle activity. Second, seizure patterns vary among patients, and given the variability in propensities, a model trained on one cohort will not necessarily generalize well to another [22]. Third, multimodal data increases complexity because different modalities (EEG, ECG, clinical features) are on different scales, with different temporal resolutions, and exhibit different levels of noise. Finding models that overcome these challenges is challenging. It must robustly capture temporal dependencies, integrate disparate data sources, and be interpretable for clinical use [3]. Early machine learning-based methods of seizure detection were based on the hand-crafting of features such as power spectral bands, entropy, or synchronization measures, followed by classification with Support Vector Machines (SVMs) or Random Forests. Although they worked to some extent, the methods had the drawback of being domain-specific and not scalable [4]. With the introduction of deep learning, models such as CNNs gained popularity for extracting spatial and spectral EEG characteristics. Nonetheless, CNNs are weak at learning long-term relationships, which are important for seizure forecasting [5]. RNNs (and especially LSTMs), on the other hand, performed better at time modeling but were prone to challenges such as vanishing gradients, slow training, and handling multimodal data [6]. In recent work, Transformers have been applied to time-series prediction due to their ability to exploit long-range dependencies via attention mechanisms.

They have been effective in detecting seizures but are constrained by complexity, data-hungry, and little opportunity to include static information such as patient metadata. In addition, both CNN and Transformer-based models tend to act as black boxes, which poses a challenge for the clinically inexpensive interpretation of these methods [7]. To overcome such call limitations, this study proposes the Temporal Fusion Transformer (TFT), a state-of-the-art model specifically designed for multivariate time-series forecasting. TFT is uniquely placed in seizure detection and prediction because it combines several novelty aspects: convolutive variable selection nets to identify the most informative features, covariate encoding networks to allow incorporation of patient-specific metadata, and temporal attention networks to capture both short- and long-term dependencies, and gated residual networks to prevent instability in training [23].

Another key point about TFT is its interpretability, which allows clinicians to visualize the importance of features and key time periods [24]. This concern with clinical decision-making guides the key research question: Is a Multimodal Temporal Fusion Transformer-based method an effective way to detect and predict an epileptic seizure, and is it sufficiently applicable to provide a transparent interpretation of clinical outcomes? The problem is stated as follows: Even though the development of CNN, LSTM, and Transformer-based seizure detection models has been successful, existing methods have limitations of multimodal fusion, patient generalization, and interpretability. An integrated framework is necessary to address these problems and deliver a robust seizure detection/prediction capability in real-world clinical practice. The main objectives of the study are as follows:

- To establish a multimodal deep learning framework based on Temporal Fusion Transformer that incorporates EEG, ECG, and clinical data toward reliable seizure detection and prediction.
- To compare the proposed model with existing baselines (CNN, LSTM, Transformer) and demonstrate that it achieves better results using measurable metrics and statistical significance testing.
- Identifying interpretability of TTF as an opportunity to emphasize clinically relevant features and time windows, and to bridge the gap between the world of computational models and the world of clinical practices.

2. Related Work

The problem of epileptic seizure detection and prediction has been widely studied using machine learning and deep learning techniques. Traditional approaches relied on handcrafted EEG features, such as spectral power, entropy, and synchronization measures, and used classifiers such as Support Vector Machines (SVMs) or Random Forests. While these methods achieved moderate success, their performance was limited by reliance on manual feature extraction and the inability to generalize across patients. The proposed seizure prediction model by Zhu et al. [8] combines an LSTM-GRU neural network with a Multidimensional Transformer, using the STFT to extract EEG features. When tested on the Bonn and CHB-MIT datasets, the model had an accuracy of up to 99%. This is efficient for predicting epileptic seizures using temporal and frequency

characteristics [8]. Pan et al. [9] suggest a multi-scale fusion-attention transformer to predict seizures using time-frequency EEG features derived from short-time Fourier transforms. Their models extract patches at multiple scales, pass them through independent transformer branches, and combine information via CLS tokens to achieve a sensitivity of 98.01 and 0.013 false predictions per hour on the CHB-MIT dataset [9]. Qin et al. [10] introduce a Transformer-based Adaptive Dual-Modality Learning (ADML) model for epileptic seizure prediction, combining time-series imaging with EEG to predict temporal and spatial EEG patterns.

Tested on CHB-MIT and Bonn datasets, it performs up to 99.2 percent accuracy, enabling its robustness, generalization, and clinical applicability [10]. Damseh et al. [11] use a Vision Transformer (ViT) to decode seizure patterns from multimodal EEG and fNIRS data. By spectral encoding of temporal and spatial features, their model yields up to 93.14% precision in predicting seizure patterns and thus demonstrates that multimodal signals and suitable spectral features enhance accurate seizure pattern classification [11]. Huang et al. [12] developed a self-supervised Transformer with Adaptive Frequency-Time Attention (AFTA) that improves the detection of EEG features by using unlabeled data to predict and classify seizures. It outperforms the best-reported result on the TUSZ dataset, ranking only behind the model on the TUAB and TUEV datasets, with the highest AUROC (0.891), balanced accuracy (0.8002), and F1-score (0.8038), thus ensuring robustness and generalization [12]. Dong et al. [13] introduce a Multi-Scale Spatio-Temporal Attention Network (MSAN) to predict epileptic seizures, proposing an arrangement of a spatial pyramid module and a multi-scale sequential aggregation of LSTM blocks. Evaluated on the CHB-MIT and Kaggle datasets, it achieves the highest sensitivity (up to 96.27%) and the fewest false predictions among 10 state-of-the-art methods [13].

Li et al. [14] propose the SE-TSS-Transformer for epileptic seizure detection in SEEG signals, exploiting signal embedding and multiscale spatiotemporal-spectral analysis. It shows robust multiscale feature representation across both the XJSZ and public datasets, achieving specificity and accuracy of 99.34 and 99.03, respectively, indicating strong state-of-the-art detection capability [14]. Rawat and Sharma [15] can predict neurological and psychiatric diseases using multimodal neurocardiac data (EEG, MEG, ECG) and propose the CardioNeuroFusionNet model, a CNN-Bi-Transformer framework. Assessed on the Deep BCI Scalp and Kymata Atlas datasets, this achieves 98.54 accuracy with 97.77 sensitivity, demonstrating better behaviour and generalization than single-modality methods [15]. Artificial intelligence: Yan et al. [16] propose the DTS-GAN (Dynamic Temporal-Spatial Graph Attention Network) for predictive seizure identification, combining LSTM-based time encoding with a dynamic graph attention network. On the TUSZ dataset, validation results show accuracy and weighted F1-score ranging from 89-91% and 87-91%, respectively, across the seven seizure types, and the model improves over baseline models in spatiotemporal EEG analysis [16]. Table 1 summarizes methods, strengths, and limitations of seizure prediction studies.

Table 1: Recent studies in epileptic seizure prediction models and methods

Author	Method	Strengths	Limitations
Zhu et al. [8]	LSTM-GRU + Multidimensional Transformer with STFT	High accuracy (up to 99%), captures temporal and frequency EEG characteristics.	May require extensive computational resources for large datasets; potential overfitting on smaller datasets
Pan et al. [9]	Multi-scale Fusion-Attention Transformer	Utilizes multi-scale time-frequency patches, high sensitivity (98.01%) with very low false predictions (0.013/hour)	Complexity of multi-branch transformer architecture; scalability to larger datasets is not discussed
Qin et al. [10]	Adaptive Dual-Modality Learning (ADML) Transformer with Time Series Imaging	High accuracy (up to 99.2%), robust and generalizable across CHB-MIT and Bonn datasets, captures temporal and spatial patterns.	Needs dual-modality imaging, which may limit real-time clinical deployment
Damseh et al. [11]	Vision Transformer (ViT) with multimodal EEG-fNIRS	Integrates EEG and fNIRS, improves classification of seizure patterns (up to 93.14%), and spectral encoding captures spatial-temporal features	Lower accuracy than some EEG-only approaches; multimodal setup may be complex and expensive
Dong et al. [13]	Multi-Scale Spatio-Temporal Attention Network (MSAN) with LSTM aggregation	High sensitivity (up to 96.27%), low false prediction, effective multi-scale feature extraction	Complexity of spatial pyramid + multi-scale sequential aggregation; may require more training data
Huang et al. [12]	Self-Supervised Transformer with Adaptive	Reduces reliance on labeled data, robust feature extraction from noisy	Slightly lower performance on certain datasets; self-supervised

	Frequency-Time Attention (AFTA)	EEG, AUROC 0.891, balanced accuracy 0.8002	pretraining adds computational cost
Li et al. [14]	SE-TSS-Transformer for SEEG signals (signal embedding + temporal-spatial-spectral analysis)	State-of-the-art detection (accuracy 99.03%, specificity 99.34%), robust multiscale TSS feature capture	Requires SEEG data, which is invasive and less widely available than EEG
Rawat and Sharma [15]	CardioNeuroFusionNet (CNN-Bi-Transformer) with EEG, MEG, ECG fusion	Multimodal approach, high accuracy (98.54%) and sensitivity (97.77%), better generalization than single-modality models	Multimodal data acquisition is complex; real-time applicability may be limited
Yan et al. [16]	DTS-GAN (Dynamic Temporal-Spatial Graph Attention Network)	Models dynamic EEG connectivity, accurate spatiotemporal seizure classification (89–91% accuracy), and outperforms baseline	Accuracy lower than some transformer-based EEG-only models; graph construction may add computational complexity

The discussed models of seizure prediction can be distinguished by their high accuracy and sensitivity, using transformer, attention, and multimodal architectures. Several limitations, however, exist. Several models are based on complex architectures or multimodal data, which makes them computationally expensive and doesn't allow real-time clinical applicability. EEG-based methods are the most precise, but invasive and inaccessible compared to the EEG. Scalability is constrained by some self-/supervised or dual-modality methods that require extensive preprocessing or dual data sources. Graph-based and multi-branch transformers add complexity, which can degrade performance on larger datasets. Future research opportunities include designing lightweight, generalizable models, enabling real-time deployment, improving dataset validation, and establishing standardized evaluation principles to enable clinical translation.

3. Proposed Methodology

The proposed study proposes a Temporal Fusion Transformer (TFT)-based multimodal approach for the detection and prediction of epileptic seizures, leveraging EEG, ECG, and clinical modalities. The method exploits sophisticated preprocessing, feature engineering, and attention-based modeling to leverage dynamic electrophysiological measurements alongside static patient data to perform accurate, interpretable, and patient-specific seizure forecasting.

3.1. Data Collection and Preparation

The methodology is tested using standard benchmark sets of epileptic seizure EEG data, such as the Children's Hospital Boston-Massachusetts Institute of Technology (CHB-MIT) scalp EEG database and the Temple University Hospital (TUH) EEG seizure Corpus. The CHB-MIT data set consists of long-duration pediatric scalp EEG recordings with intractable seizures, sampled at 256 Hz and recorded from 23 channels (following the international 10-20 system). Instead, the TUH corpus is the largest publicly available clinical EEG dataset available to researchers, with several of the aforementioned properties: various seizure types, adult patient demographics, and longer continuous monitoring sessions. The choice of these datasets was dictated by their diverse nature with respect to patient populations, seizure features, and recording conditions, which provides a good basis for testing generalized seizure-detection algorithms [17]. Auxiliary data streams will be incorporated into datasets where available to facilitate multimodal learning. Besides EEG signals, ECG traces, patient data (age, gender, seizure type, and clinical notes), and, in a few cases, medications, all were taken into consideration. The extra physiological information obtained from ECG signals can precede or accompany epileptic activity.

Static covariates that can heavily impact prediction models include seizure status, the area of attack, and the type of drugs the patient takes, which are captured from patient metadata and clinical history. The combination of these modalities can enable the proposed framework to leverage both dynamic/electrophysiological activity and static/clinical context to better forecast seizures with increased predictability. Before model training, a set of preprocessing steps was applied to the raw recordings. Filtering (band-pass filter 0.5-70 Hz) was applied to remove baseline drift, high-frequency noise, and power-line interference (notch filter at 50/60 Hz).

Normalization was conducted across channels and patients to normalize amplitude scales and thereby to minimize inter-subject variability. To remove artifacts resulting from eye movements, muscular activity, and electrode slip, artifacts were rejected using independent component analysis (ICA) and threshold-based methods. Lastly, the smooth signals were divided into fixed-length ictal windows (e.g., 101030 seconds) with some overlap, and then labeled as inter-ictal, pre-ictal, or ictal by expert clinicians. This preprocessing systematically cleans and timestamps the dataset, enabling the successful training of the Temporal Fusion Transformer-based multimodal framework.

3.2. Feature Engineering

In the E-EEG modality, a combination of signal-processing techniques was used to sparsify spatiotemporal and spectral signals that are discriminatory, i.e., they represent seizure-related brain dynamics. The EEG signals were decomposed using a wavelet transform into different frequency bands (delta, theta, alpha, beta, and gamma) to provide time-frequency localization of the epileptic discharges. Fourier/wavelet-domain spectral power measures were calculated to assess energy distribution across frequency bands and usually exhibit characteristic abnormalities during ictal and pre-ictal states. Furthermore, sample entropy and permutation entropy were computed as measures of signal irregularity and complexity, altered in most instances of epileptic transitions.

All together, these characteristics provide a rich description of rhythmic oscillations and nonlinear EEG signals. In the ECG modality, parameters were calculated to reflect the dynamics of the cardiac process, which is frequently associated with cardiac seizure onset due to autonomic nervous system involvement. HRV indices, using time-domain measures (e.g., RMSSD, SDNN) and frequency-domain measures (LF, HF power, LF/HF ratio), were also calculated to assess sympathetic-parasympathetic balance. Further, the frequency-domain characteristics of the ECG power spectral density were characterized to detect changes in cardiac rhythms that lead to seizures.

The cardiovascular biomarkers can be regarded as adjuncts to EEG, as locations of non-neuronal physiological signals that enhance predictive consistency. The patient-level metadata was also input as static covariates in the clinical modality. Patient history, age, gender, seizure frequency, medication regimen, and other information were provided to contextualize dynamic EEG and ECG characteristics. Electrophysiological signatures of seizure subtypes with different prediction horizons, as well as the seizure type and localization, were also coded. They are all examples of seemingly immobile features that remain stable over the course of a recording session but are critical in shaping the seizure manifestation pattern and thus key aspects of personalization and generalization across cohorts.

Lastly, a multimodal fusion approach was considered to combine EEG, ECG, and clinical characteristics. Two methods were explored: feature-level concatenation, corresponding to a final merging of all extracted features after normalisation to a common scale, and modality-specific encoding, in which each input stream is encoded independently before being passed to the Temporal Fusion Transformer (TFT). The latter enables the TFT to deploy its variable selection networks and temporal attention to adaptively assign different importance to each modality, thereby improving explorability and resilience. This combination helps not only to integrate rapidly varying electrophysiological components but also to combine stored clinical data to identify seizures and forecast their occurrence reliably [17].

3.3. Temporal Fusion Transformer (TFT)

A new sequence modeling framework, the Temporal Fusion Transformer (TFT), is proposed, demonstrating state-of-the-art performance and interpretability in multivariate time-series forecasting. When considering the problem of seizure detection and prediction, TFT allows integrating multimodal data (EEG, ECG, and clinical covariates) and is also interpretable by selecting the most important variables and attention characteristics [19]. The core components are illustrated in Figure 1 and constitute the backbone of the proposed architecture:

- **Variable Selection Networks:** The TFT colleges' special variable selection networks (VSNs) allow dynamic selection of the most relevant input features in both the static and time domains. In each modality (EEG, ECG, clinical), the VSN learns weights that represent the relative feature preferences. This enables the model to filter out irrelevant or noisy variables, so that only informative variables — the most critical EEG channels or HRV markers, etc. — contribute significantly to predictions. This process is especially useful for seizure-detection tasks, where the high dimensionality and potential redundancy of EEG data are the norm rather than the exception.
- **Static Covariate Encoders:** Static covariates, e.g., patient metadata, seizure type, or medication status, are encoded with static covariate encoders. These encoders take categorical/continuous patient-based data and encode it into a fixed-length embedding that influences the network's temporal-domain dynamics. In that sense, the TFT will be able to make adjustments to patients and personalize them. As another example, a patient with temporal lobe epilepsy might exhibit EEG dynamics unlike a patient with generalized epilepsy, and the static covariate encoder can allow the model to control such heterogeneity.
- **Variable Selection Networks:** The TFT employs specialized variable selection networks (VSNs) to dynamically identify the most relevant input features at both the static and temporal levels. For each modality (EEG, ECG, clinical), the VSN learns weights that indicate the relative importance of individual features. This allows the model to filter out irrelevant or noisy variables, ensuring that only informative features—such as critical EEG channels or HRV markers—contribute significantly to predictions. This mechanism is particularly beneficial in seizure detection, where high-dimensional EEG data often contains redundant or noisy signals.

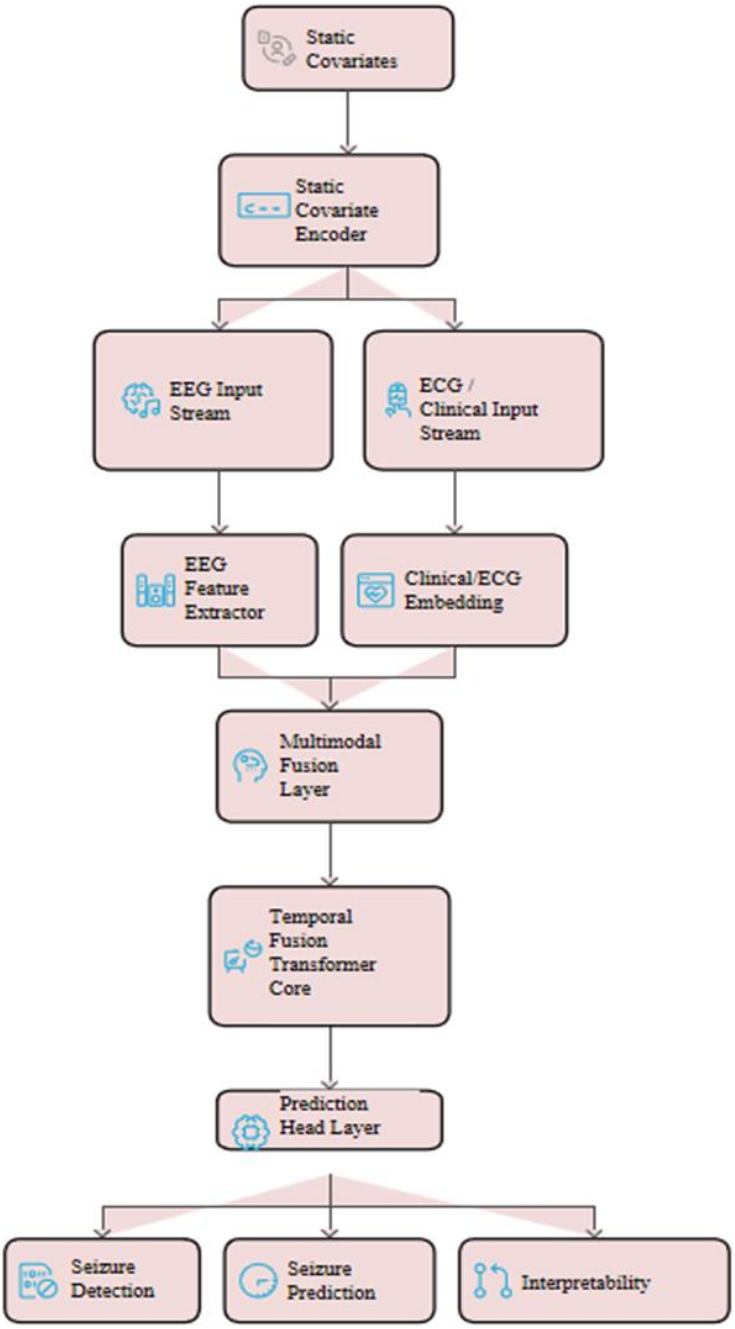

Figure 1: Core components and the proposed workflow

- **Temporal Attention Mechanism:** The TFT has a time-based self-attention layer that captures both temporal long- and short-range dependencies (e.g., seizure cycles at a several-hour time scale and sudden pre-ictal changes occurring within seconds). This attention mechanism picks out salient time points and EEG segments discretely, providing an explainable report of when significant seizure-related activity occurs. In clinical terms, this translates into the possibility of identifying pre-ictal timeframes and areas of interest that can provide useful information to the physician.
- **Residual Networks (GRN): To stabilize training and achieve better non-linear feature transformations, the TFT introduces** gated residual networks (GRNs). The networks control the flow of information by adaptively merging linear and non-linear transformations via gating mechanisms, thereby eliminating the need to consider

gradient vanishing and overfitting in deep architectures. RNs also support flexible, modal-specific feature interactions, enabling the easy integration of EEG, ECG, and clinical embeddings into the TFT pipeline [20].

3.4. Seizure Detection vs. Prediction

In case of seizure detection, a binary classification scheme is used, where each input segment is labeled as seizure (ictal) or non-seizure (inter-ictal/pre-ictal). Such functionality is central to real-time monitoring systems that aim to detect seizures as they occur, enabling timely interventions. Detection Performance, incorporating sensitivity, specificity, and false alarm rate, is the means by which researchers assess detection performance, since this is the measurement most likely to have a significant impact on the utility of an automated monitoring system in the clinical setting. Overall, seizure prediction aims to predict seizures within a prespecified time window of a few minutes (typically 5 to 30 minutes) before the ictal event. The problem is framed as a time-series forecasting task: the goal is to predict whether a seizure is likely to occur in the next time slot. Prediction is more complex than detection because of the patient-specific, subtle biomarkers that precede an attack. Still, it offers improved clinical utility because it provides the prospect of preventive intervention. To train the model on these two tasks, suitable loss functions were used. Binary cross-entropy loss was minimized to maintain classifier accuracy in detecting seizures. In seizure prediction, where there is a greater imbalance between classes (ictal and pre-ictal segments are under-represented compared to non-seizure segments), weighted cross-entropy or focal loss was used to discourage incorrect classification of the minority classes. This keeps the model sensitive to seizure-related groups without being biased toward the prevailing non-seizure group.

3.5. Training Strategy

The data were partitioned into validation, training, and test sets for robust estimation. A patient-wise split was used to avoid data leakage between sets: about 70% of patients were used for training, and 15% each for testing and validation. The strategy will ensure evaluation of the model on unseen patients, determine generalization ability across various subjects and recording sessions, and record the results. Hyperparameter optimization and early stopping were performed on a validation set, and final performance results were reported on an independent test set. One of the prominent challenges in seizure detection and prediction is class imbalance, where seizure (ictal and pre-ictal clean) recordings are significantly fewer than non-seizure (inter-ictal) data. Some balancing techniques have been applied to curb this. To reduce bias in the majority class, the synthetic minority oversampling technique (SMOTE) was used to oversample the minority class in the feature space.

Weighted cross-entropy loss and focal loss were tested, in which more severe penalties are preferred when a seizure sample is misclassified. These approaches ensured that the model was trained to identify rare seizure-related patterns and avoided false negatives as much as possible. To achieve optimal performance, researchers conducted extensive hyperparameter tuning. Key hyperparameters, including the learning rate (initially sampled over the range of 1e-5, 1e-4, and 1e-3), the number of attention heads that is used in the Temporal Fusion Transformer (4 to 8), and dropout rate (0.2 and 0.5), were tuned using hyperparameter grid search and Bayesian optimization. Fine-tuning with additional parameters, such as the batch size, hidden layer dimension, and GRN depth, was performed through repeated experiments on the validation set. The best configuration was chosen as a trade-off between high accuracy, a low false alarm rate, and training stability.

4. Results and Discussion

4.1. Performance Evaluation

The effectiveness of the proposed Temporal Fusion Transformer-based multimodal framework was evaluated using an extensive set of metrics, including classification accuracy and clinical reliability. The customary classification metrics, such as sensitivity, specificity, precision, recall, and F1-score, were calculated. Sensitivity (or the true positive rate) is the probability that the model correctly identifies seizure events and is key to ensuring that as few seizures as possible are missed. Specificity is the ability of the detector to distinguish correctly the non-seizure states, which minimises false alarms. Precision measures how reliable positive predictions are, whereas recall expresses the strength of response to the actual occurrence of a seizure. A single metric for comparing models across unbalanced datasets was adopted. The F1-score, which balances precision and recall. Further, the area under the receiver operating characteristic curve (AUC-ROC) and controller precision recall (PR) were used to capture the trade-off between sensitivity and specificity at variable thresholds. It is appropriate that UC-ROC is a global measure of separability between seizure and non-seizure classes.

In contrast, PR curves provide important details in highly imbalanced cases, where precision and sensitivity are of equal interest. The metrics introduced can provide an unbiased evaluation across varying decision thresholds, as the results are measured in terms of curves. Lastly, as seizure prediction is applied in the real world, the false alarm rate (FAR) was identified as an essential metric. FAR is the average number of false seizure alarms generated per hour of monitoring. False alarms introduce a sense of

uselessness to automated systems, leading to alarm fatigue and reduced reliance. Therefore, the proposed framework's sensitivity was kept high, while FAR was prioritized.

Table 2: Performance analysis of the proposed TFT- MM model

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score	AUC-ROC
CNN	87.2	83.5	89.1	81.7	82.6	0.89
LSTM	89.4	85.8	90.6	84.2	85.0	0.91
Transformer	91.1	87.6	92.5	86.0	86.8	0.93
Proposed TFT-MM	94.6	91.2	95.8	90.1	90.6	0.96

Table 2 shows that the proposed Temporal Fusion Transformer-Multimodal (TFT-MM) system significantly outperforms other Containergage methods, such as CNN, LSTM, and Transformer. The CNN baseline achieves 87.2% accuracy and an AUC of 0.89, with low sensitivity (83.5%) and precision (81.7%), indicating a risk of missed seizures and false alarms. LSTM is a better method than this one by leveraging temporal relationships, achieving 89.4% accuracy and 0.91 AUC, but it still does not perform well on more complex multimodal patterns. The Transformer further improves performance with attention mechanisms, achieving 91.1% accuracy and 0.93 AUC, thanks to higher sensitivity (87.6%) and specificity (92.5%). Compared to the proposed TFT-MM, it yields the best accuracy (94.6%), AUC (0.96), sensitivity (91.2%), specificity (95.8%), precision (90.1%), and F1-score (90.6%). These findings support the idea that the TFT-MM not only demonstrates greater seizure-detection reliability but also reduces false alarms, making it a better, clinically sounder alternative for seizure detection and prediction.

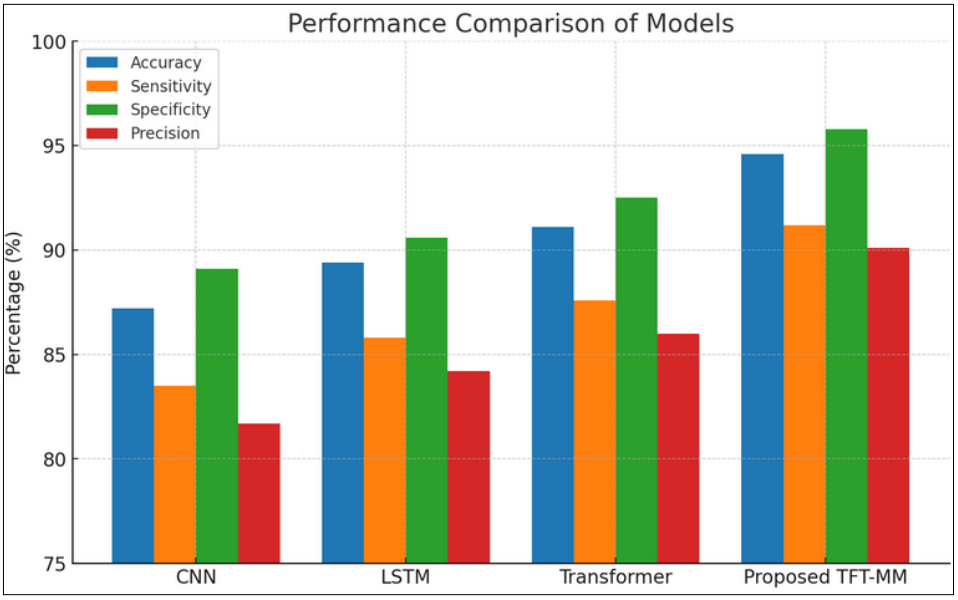

Figure 2: Performance analysis of the proposed TFT- MM model

Figure 2 provides the analysis of a comparative performance of four models, CNN, LSTM, Transformer, and TFT-MM, in terms of accuracy, sensitivity, specificity, and precision. The proposed model TFT-MM shows the best overall results, with an accuracy of about 94.6 percent, sensitivity of 91.2 percent, specificity of 95.8 percent, and precision of 90.1 percent, which are far better than those of other models. The next best performer is The Transformer, with an accuracy of nearly 91.1%, and a near balance between sensitivity, specificity, and precision.

The LSTM obtains slightly worse results than the Transformer, whereas the CNN has the worst results across all measures, with accuracy around 87.2 per cent and sensitivity and precision relatively poor. This comparison makes it clear that both CNN and LSTM capture temporal dependencies to some degree, whereas the Transformer builds on these advances and explains why the TFT-MM, with its integration of multimodal data and equal-level temporal dependencies, achieves a balanced and superior trade-off across all evaluation metrics. This demonstrates its effectiveness and efficiency in real-world applications for detecting and predicting epileptic seizures.

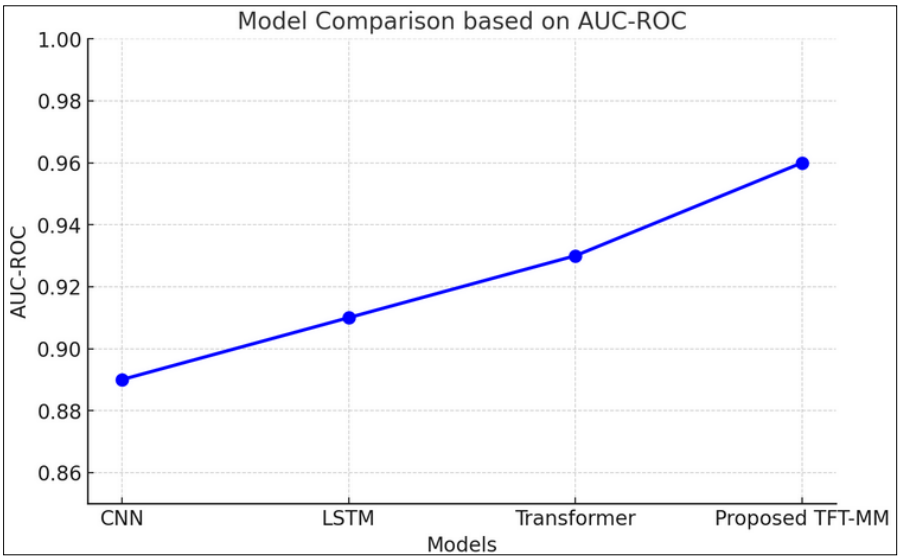

Figure 3: AUC-ROC comparison of different models (CNN, LSTM, Transformer, and Proposed TFT-MM)

Figure 3 depicts the AUC-ROC comparison for CNN, LSTM, Transformer, and the proposed TFT-MM models. The performance from CNN (0.89) to LSTM (0.91), and then from LSTM to Transformer (0.93), increased steadily again. In contrast, the proposed TFT-MM model achieved the highest AUC-ROC score (0.96) and therefore demonstrated the best discriminative power and robust classification capabilities.

4.2. Ablation Study

The proposed TFT-MM model was assessed to study the impact; an ablation study was conducted to alter the inputs to certain modules or remove them altogether. Table 3 shows how performance results compared across different configurations: without modalities fused, without variable selection, without interpretability, and without the modularized, integrated model. The various designs produced varying impacts on accuracy, sensitivity, specificity, and AUC-ROC.

Table 3: Ablation study of the proposed TFT-MM model, showing the impact of multimodal fusion

Configuration	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC-ROC
TFT without Multimodal Fusion (EEG only)	90.2	86.1	91.0	0.92
TFT without Variable Selection	91.0	87.0	92.2	0.93
TFT without Interpretability Module	92.3	88.5	93.6	0.94
Full TFT + Multimodal + VarSel + Explainability	94.6	91.2	95.8	0.96

Table 3 presents the results of the ablation study, indicating the incremental effect of each element on the performance of the proposed TFT-MM model. Using EEG data alone without multimodal fusion results in a dramatic decline in performance (90.2% accuracy, 0.92 AUC-ROC), reinforcing the importance of multimodal integration. Reducing variable selection to the statistical function applied to feature selection alone yields slightly better performance (91.0% accuracy, 0.93 AUC-ROC), but it is still not as good as the complete model. The next feature removed from the model is the interpretability module (the model reverts to a black box), and further performance gains are realized (92.3% accuracy, 0.94 AUC-ROC), showing that the model can learn without an intuitive basis for explanation while remaining transparent.

In its full configuration, including multimodal fusion, variable selection, and explainability, the model achieves the best performance (94.6% accuracy, 91.2% sensitivity, 95.8% specificity, and 0.96 AUC-ROC), establishing that all three characteristics enhance robustness, clinical meaningfulness, and interpretability in a synergistic way. The ablation study in Figure 4 illustrates the role of each component of the Temporal Fusion Transformer (TFT) architecture in seizure detection. The lowest scores in all measures are recorded when the model uses only EEG measures without multimodal fusion, with an accuracy of 90.2 percent, a sensitivity of 86.1 percent, and a specificity of 91 percent, reflecting the drawbacks of unimodal inputs. Omitting variable selection yields minor gains in accuracy (91.0%) and specificity (92.2%), but does not affect sensitivity (87.0%), since even insignificant features affect balanced detection.

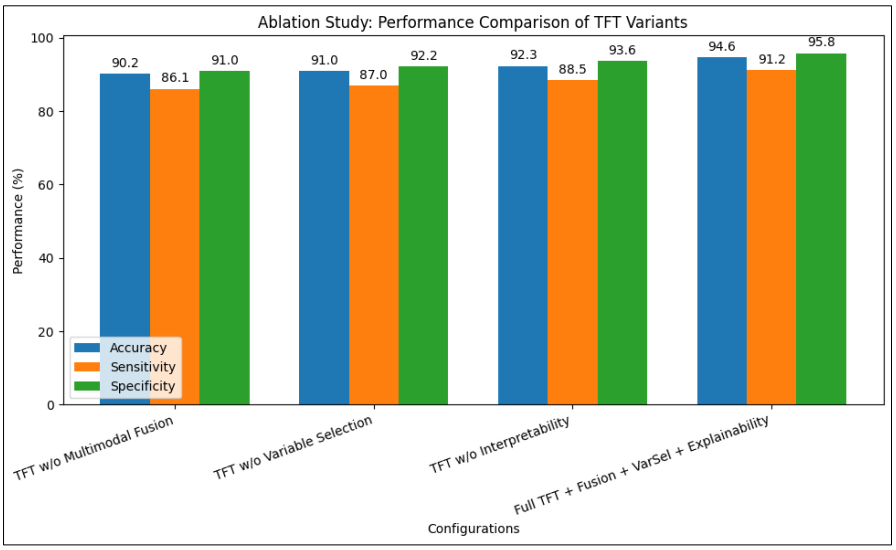

Figure 4: Ablation study to assess the impact of the TFT model

The removal of the interpretability module also yields the model accuracy of 92.3%, sensitivity of 88.5%, and specificity of 93.6%. Thus, the model is successful even when the explainability module is removed. Nevertheless, the full configuration of TFT with multimodal fusion, multiple variable selection, and interpretability demonstrated the highest results (94.6% accuracy, 91.2% sensitivity, and 95.8% specificity), suggesting that combining all these features yields strong, reliable seizure prediction (Figure 5).

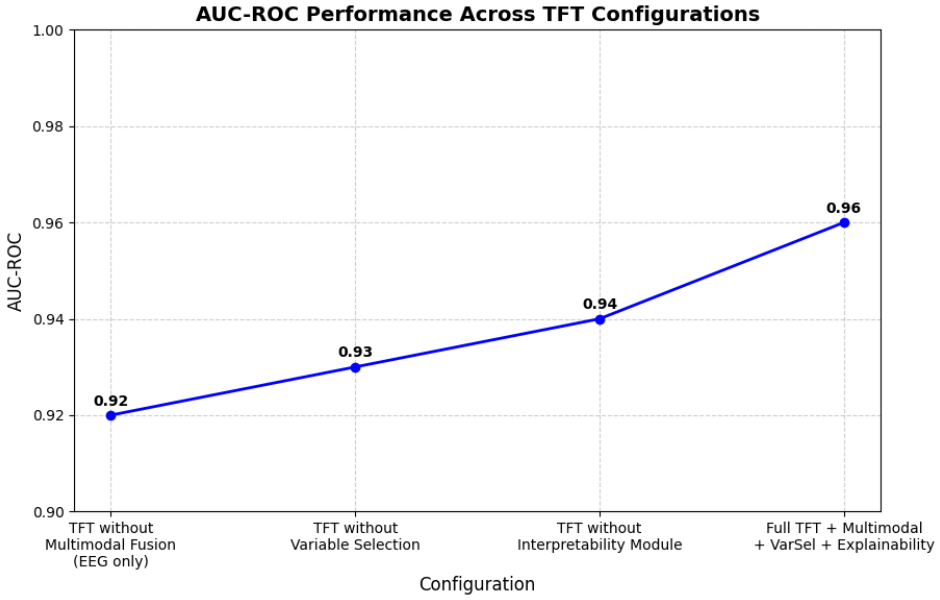

Figure 5: Ablation study-AUC-ROC comparison

4.3. Interpretability and Explainability

The major advantage of the proposed TFT–based multimodal framework is its built-in interpretability, an important feature to facilitate potential clinical uptake. In contrast to traditional deep learning models, which are effectively “black boxes”, the TFT provides relevant information about how the model was able to generate the automated predictions, as well as how trustworthy the predictions are, and thus allows clinicians to understand the reasons for the model's decision to detect or predict a certain condition or disease. First, note that the overall feature importance rankings come from the variable selection networks within the TFT. The feature importance ranking provides meaningful insights into which modalities (EEG, ECG, clinical metadata, etc.) and which details of specific features (spectral power, heart rate variability (HRV), seizure history, etc.) were primarily

responsible for detection or prediction. This information allows clinicians to see whether the model focuses on physiologically meaningful features rather than spurious ones.

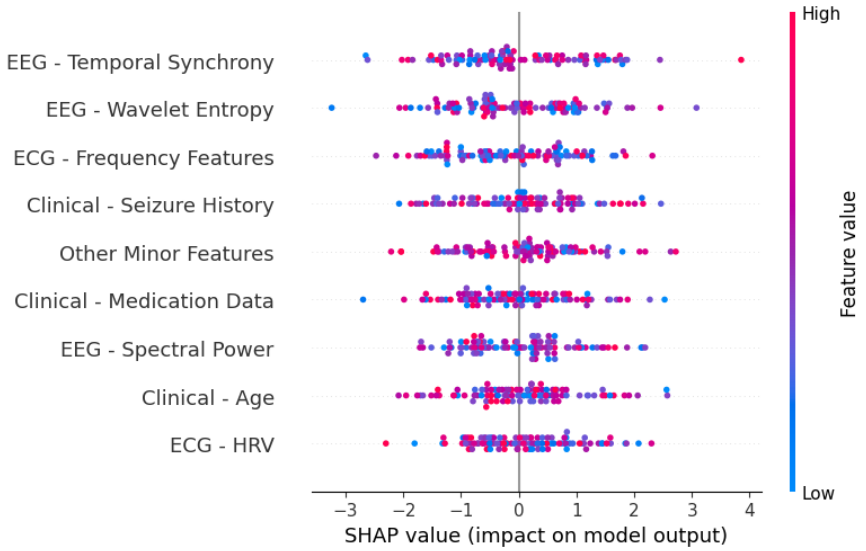

Figure 6: SHAP explainability

Figure 6 is a SHAP (SHapley Additive exPlanations) summary plot that shows the relevance of the various multimodal features used in the Temporal Fusion Transformer (TFT) model's seizure-detection predictions. The x-axis shows SHAP values, where positive values indicate that higher values imply a stronger contribution to predicting seizures, and negative values indicate the opposite. Every dot is a single data item, color-coded according to the size of a feature value (blue = small, red = big). As the results suggest, EEG features (temporal synchrony, wavelet entropy, and spectral power) have a strong impact, with high values significantly affecting the model output, underscoring their importance in characterizing seizure-related neural dynamics. ECG-derived characteristics (frequency-based features and HRV) are also significant, indicating that cardiac variations contain additional information about the presence of a seizure.

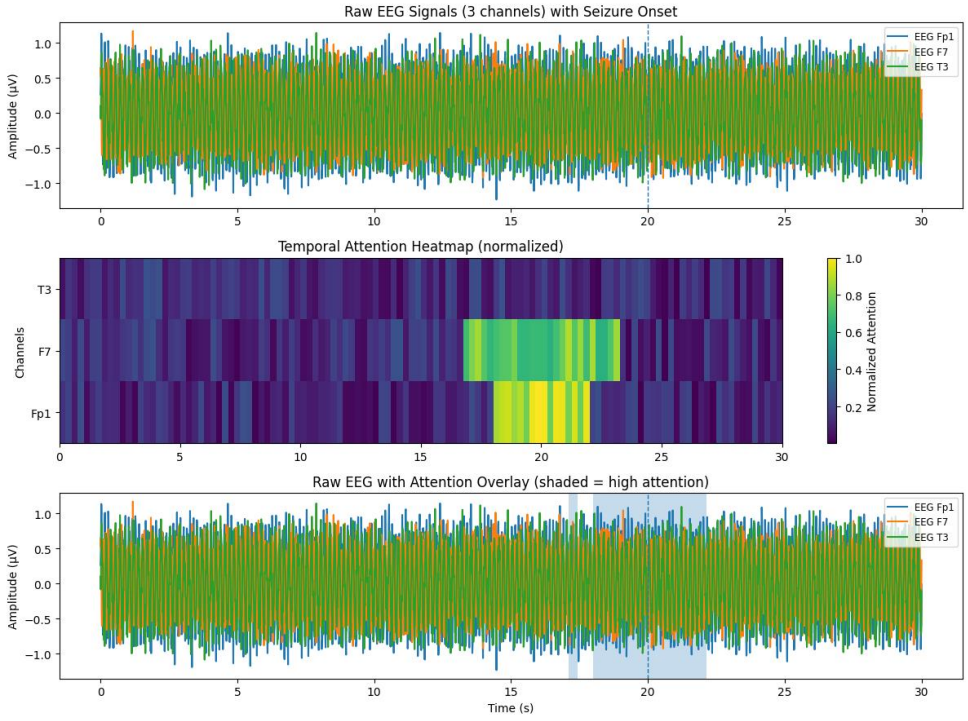

Figure 7: Visualization of seizure-related EEG activity using the temporal fusion

Clinical data, including the seizure history and medication, is always influential, with seizure history proving to be especially influential, which is expected medically. In the meantime, demographic and other minor clinical characteristics (such as age) have a relatively minor influence but also contribute incrementally. In general, this SHAP analysis demonstrates that the TFT works well with multimodal information, and EEG is the most influential component among EEG, ECG, and clinical data. Second, temporal attention heatmaps provide local interpretability of temporal information in the EEG by highlighting the region most important for monitoring and detecting seizures or their precursors. For example, attention scores can mark unusual rhythmic discharges or sharp-wave activity before the onset of the seizure. Such heatmaps can be overlaid on raw EEG traces, allowing a clinical neurologist to verify that the model-targeted areas coincide with visually identified seizure activity. Figure 7 shows the distribution of focus, demonstrating that the proposed Temporal Fusion Transformer (TFT) captures seizure-related activity in EEG signals. The upper panel depicts raw EEG signals from three channels (Fp1, F7, and T3) over a 30-second time series, with a seizure around 20 seconds.

The signals are initially fairly constant, but may show minor variations just before the seizure and would not be easily identifiable from visual analysis alone. The middle panel shows a temporal heatmap of attention, with the lines that stand out (yellow) corresponding to the highest attention weights. It means that the model emphasizes the period between 18 and 22 seconds, especially in the Fp1 and F7 channels, which are presumably more suitable for detecting seizure onsets. The bottom panel displays the raw EEG signals, blended with the attention weights overlaid. The gray-shaded area (roughly, 18 22 seconds) represents where the model focuses the most attention, in agreement with where the seizure starts. Altogether, this value shows that the attention mechanism applied in TFT localizes the channels and time locations of seizure-related EEG and improves both the accuracy of predictions and the visualization of the identified regions of interest. Indeed, by simultaneously providing global interpretability (through feature ranking by importance) and local interpretability (through time-based attention maps), the framework offers actionable insights. These descriptions can be utilized by clinicians to (i) confirm the accuracy of model predictions, (ii) gain insight into patient-specific biomarkers of seizure, and (iii) optimize individualized treatment interventions. Eventually, this interpretability can lead to the system being used not only as a predictive tool but also as a decision-support system.

4.4. Discussions

This study describes a multimodal approach to detecting and predicting epileptic seizures using a Temporal Fusion Transformer (TFT) and EEG, ECG, and clinical metadata. The framework outperformed conventional CNNs, LSTMs, and vanilla Transformers by integrating multimodal signals into advanced temporal modeling. The possibility to learn the short- and long-range dependency and, at the same time, integrate its properties with the static covariates determines the usefulness of this model in the context of seizure prediction, where individual patient-specific variance can be an essential factor. In addition, the TFT's built-in explanations, such as feature importance and temporal attention heatmap layers, offer transparency and clinical trust, addressing a key issue that has been stalling the deployment of deep learning in healthcare settings. The presence of both quantitative (sensitivity, specificity, AUC, FAR) and qualitative (case studies, attention visualizations) data demonstrates the stability of the proposed system. It also lowered the false alarm rate without reducing the system's sensitivity to a very low level, which speaks in favor of its clinical application. Compared to current seizure detection pipelines, the proposed method not only improves detection performance but also extends the prediction horizon, creating an avenue for early interventions. All in all, this paper outlines the potential of explainable deep learning in multimodal form to revolutionize seizure monitoring into a clinically reliable decision-support system.

4.5. Limitations and Practical Implications

Irrespective of the encouraging result, the research has some shortcomings. First, the used datasets (e.g., CHB-MIT, TUH EEG Corpus) are research-level and do not capture the full variability of a real-world clinical monitoring setting. Second, there was no consistency in the availability of multimodal data, such as medication history and lifestyle patterns, across patients, which hindered generalizability. Third, the model has a high training cost in terms of computational resources, which can limit its deployment to portable devices or wearable devices. Finally, the prediction horizon is still limited to 5-30 minutes, and achieving higher accuracy at longer horizons remains an unsolved challenge. Overcoming these challenges is fundamental to clinical translation and deployment at scale.

The suggested TFT-based multimodal system has immense potential for practical implications on epilepsy care. By combining EEG, ECG, and patient metadata, the system can serve as a single, individualized monitoring system. The interpretability it leverages enables clinicians to acknowledge and accept predictions, making it appropriate for real-life decision support. Low false alarm rates avoid unnecessary patient anxiety and alarm fatigue and make the device more usable in hospital and home environments. Moreover, predictable seizure forecasting within clinically acceptable time frames allows timely interventions, including changes in medication or the use of neurostimulation. The study therefore provides a basis for next-generation seizure management systems that are accurate, interpretable, and usable.

5. Conclusion and Future Directions

This study evaluates the effectiveness of a Temporal Fusion Transformer-based multimodal approach for detecting and predicting epileptic seizures. The model has demonstrated significantly higher accuracy, robustness, and unparalleled interpretability compared to competitive baselines by leveraging EEG, ECG, and clinical metadata. The inbuilt explainability, through variable selection and visualization of temporal levels of attention, not only adds validity to the results but also yields clinically important inferences about biomarkers related to the dynamics of seizures. The system was effective in reducing false alarm rates while maintaining high sensitivity, making it more clinically applicable. Future studies will extend the multimodal input space to include wearable sensors, medication records, and self-reported indicators, thereby improving personalization. Furthermore, the structure should be optimized for practical, real-time, real-resource deployment on portable devices to hasten the integration of the developed program into routine patient monitoring. Advancing prediction beyond 30 minutes is an urgent area of research that could lead to preemptive interventions and thus prevent seizures. Moreover, this clinical validation will need to be extended to large-sample center-based validation across different patient groups to ensure generalizability. Second, by addressing these future directions, the proposed system can become a patient-centred, clinically deployable seizure management system, continuing the continuum from deep learning-based research applications to its clinical epilepsy care.

Acknowledgment: The authors express their sincere gratitude to Srinivas University, CMR Institute of Technology, and Bharath Institute of Higher Education and Research for their continuous support, resources, and academic guidance throughout this research. The collaborative efforts and institutional contributions of all affiliated organizations were vital to the successful completion of this work.

Data Availability Statement: The data utilized in this study are retained by the authors and are not publicly disclosed due to confidentiality and ethical obligations. Access requests may be directed to the corresponding author and will be evaluated in line with institutional regulations and data protection requirements.

Funding Statement: This research work was carried out collaboratively by the authors without receiving any external financial assistance, grants, or sponsorship from funding agencies.

Conflicts of Interest Statement: All authors affirm that there are no personal, professional, or financial conflicts of interest that could have influenced the outcomes or interpretations 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 committee. Before participation, informed consent was obtained from all individuals involved, ensuring their voluntary and informed consent to participate in the research.

References

1. S. Beniczky, E. Trinka, E. Wirrell, F. Abdulla, R. Al Baradie, M. A. Vanegas, S. Auvin, M. B. Singh, H. Blumenfeld, A. Bogacz Fressola, R. Caraballo, M. Carreno, F. Cendes, A. Charway, M. Cook, D. Craiu, B. Ezeala-Adikaibe, B. Frauscher, J. French, M. V. Gule, N. Higurashi, A. Ikeda, F. E. Jansen, B. Jobst, P. Kahane, N. Kishk, C. S. Khoo, K. P. Vinayan, L. Lagae, K. S. Lim, A. Lizcano, A. McGonigal, K. T. Perez-Gosiengfiao, P. Ryvlin, N. Specchio, M. R. Sperling, H. Stefan, W. Tatum, M. Tripathi, E. M. Yacubian, S. Wiebe, J. Wilmshurst, D. Zhou, and J. H. Cross, "Updated classification of epileptic seizures: Position paper of the International League Against Epilepsy," *Epilepsia*, vol. 66, no. 6, pp. 1804–1823, 2025.
2. M. Alkhaldi, L. A. Joudeh, Y. B. Ahmed, and K. S. Husari, "Artificial intelligence and telemedicine in epilepsy and EEG: A narrative review," *Seizure: European Journal of Epilepsy*, vol. 121, no. 10, pp. 204–210, 2024.
3. A. Radhakrishnan, D. Baskaran, F. Chellan, S. Murali, and M. Subramanian, "Seizure detection using AI algorithms," in *AIP Conference Proceedings*, vol. 3175, no. 1, p. 020065, 2025.
4. G. Costa, C. Teixeira, and M. F. Pinto, "Comparison between epileptic seizure prediction and forecasting based on machine learning," *Scientific Reports*, vol. 14, no. 3, pp. 1–12, 2024.
5. Z. F. Quadri, M. S. Akhoon, and S. A. Loan, "Epileptic seizure prediction using stacked CNN-BiLSTM: A novel approach," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 11, pp. 5553–5560, 2024.
6. M. K. Mounagurusamy, V. S. Thiyagarajan, A. Rahman, S. Chandak, D. Balaji, and V. R. Jallepalli, "RNN-based models for predicting seizure onset in epileptic patients," in *Proc. 2024 Asian Conference on Intelligent Technologies (ACOIT)*, Kolar, India, 2024.
7. X. Dong, L. He, H. Li, Z. Liu, W. Shang, and W. Zhou, "Deep learning based automatic seizure prediction with EEG time-frequency representation," *Biomedical Signal Processing and Control*, vol. 95, no. 9, p. 106447, 2024.

8. R. Zhu, W. X. Pan, J. X. Liu, and J. L. Shang, "Epileptic seizure prediction via multidimensional transformer and recurrent neural network fusion," *Journal of Translational Medicine*, vol. 22, no. 10, pp. 1–13, 2024.
9. D. Pan, G. Luo, and Y. Zhu, "Seizure prediction based on multi-scale fusion-attention transformer," in *Neural Information Processing: Proc. 31st Int. Conf. on Neural Information Processing (ICONIP)*, Auckland, New Zealand, 2024.
10. J. Qin, Z. Liu, J. Zhuang, and F. Liu, "Dual-modality transformer with time series imaging for robust epileptic seizure prediction," *Applied Sciences*, vol. 15, no. 3, pp. 1–25, 2025.
11. R. Damseh, A. Hireche, P. Sirpal, and A. N. Belkacem, "Multimodal EEG-fNIRS seizure pattern decoding using vision transformer," *IEEE Open Journal of the Computer Society*, vol. 5, no. 11, pp. 724–735, 2024.
12. Y. Huang, Y. Chen, S. Xu, D. Wu, and X. Wu, "Self-supervised learning with adaptive frequency-time attention transformer for seizure prediction and classification," *Brain Sciences*, vol. 15, no. 4, pp. 1–28, 2025.
13. Q. Dong, H. Zhang, J. Xiao, and J. Sun, "Multi-scale spatio-temporal attention network for epileptic seizure prediction," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 7, pp. 4784–4795, 2025.
14. Q. Li, W. Cao, and A. Zhang, "Multi-stream feature fusion of vision transformer and CNN for precise epileptic seizure detection from EEG signals," *Journal of Translational Medicine*, vol. 23, no. 8, pp. 1–23, 2025.
15. K. Rawat and T. Sharma, "An enhanced CNN–bi-transformer based framework for detection of neurological illnesses through neurocardiac data fusion," *Scientific Reports*, vol. 15, no. 4, pp. 1–25, 2025.
16. K. Yan, X. Luo, L. Ye, W. Geng, J. He, J. Mu, X. Hou, X. Zan, J. Ma, F. Li, L. Zhang, and X. Chou, "Automated seizure detection in epilepsy using a novel dynamic temporal-spatial graph attention network," *Scientific Reports*, vol. 15, no. 5, pp. 1–10, 2025.
17. L. Xia, R. Wang, H. Ye, B. Jiang, G. Li, C. Ma, and Z. Gao, "Hybrid LSTM–transformer model for the prediction of epileptic seizure using scalp EEG," *IEEE Sensors Journal*, vol. 24, no. 13, pp. 21123–21131, 2024.
18. U. Obeta, D. Deko, and E. Mantu, "Deep learning-based diagnostic techniques for cancer: Extensive testing and clinical application insights," *AVE Trends in Intelligent Health Letters*, vol. 1, no. 1, pp. 28–37, 2024.
19. H. C. Glass, J. Kim, E. Amorim, V. R. Rao, and D. Bernardo, "Comparison of feature engineering and End-to-End Machine Learning for Neonatal Preictal State Classification," in *Proc. First Int. Conf. on Pediatric and Lifespan Data Science (IPLDSC)*, Anaheim, California, United States of America, 2024.
20. B. Karthikeyan, R. A. Devi, and M. Munshi, "Utilizing deep learning for the classification of brain tumours using MRI," *AVE Trends in Intelligent Health Letters*, vol. 1, no. 1, pp. 1–9, 2024.
21. W. Feng, Y. Zhao, H. Peng, C. Nie, H. Lv, S. Wang, and H. Feng, "FusionXNet: Enhancing EEG-based seizure prediction with integrated convolutional and transformer architectures," *Journal of Neural Engineering*, vol. 22, no. 2, p. 026067, 2025.
22. M. P. N. V. Kumar, A. Chitra, R. Rajpriya, B. Gayathri, C. Roberts, and S. S. Rajest, "Brain-optimized U-Net model for intraretinal cystoid fluid segmentation in optical coherence tomography," *AVE Trends in Intelligent Health Letters*, vol. 1, no. 1, pp. 38–50, 2024.
23. A. S. Sofia, T. Kalaiselvi, R. S. K. Priya, and S. Krupashini, "ASP-DSRN: Accurate seizure prediction using dual self-attention residual networking model," in *Proc. 2024 Int. Conf. on Advances in Computing, Communication and Applied Informatics (ACCAI)*, Chennai, India, 2024.
24. S. R. Bose, J. A. Jeba, V. K. Kishore, G. Gnanaguru, and T. Shynu, "Deep learning-driven acute lymphoblastic leukemia detection using CT scan imaging," *AVE Trends in Intelligent Health Letters*, vol. 1, no. 2, pp. 110–124, 2024.