Classification of Epilepsy Seizure Types Based on EEG Signals Using a Hybrid Deep Learning Architecture of Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM)

Cantik Zelda Herawan
Biomedical Engineering Study Program
School of Electrical Engineering
Telkom University, Main Campus
Bandung, Indonesia
Email: cantikzelda@student.telkomuniversity.ac.id

Inung Wijayanto
Telecommunication Engineering Study Program
School of Electrical Engineering
Telkom University, Main Campus
Bandung, Indonesia
Email: iwijayanto@telkomuniversity.ac.id

Annisa Humairani
Biomedical Engineering Study Program
School of Electrical Engineering
Telkom University, Main Campus
Bandung, Indonesia
Email: annisahumairani@telkomuniversity.ac.id

***Abstract*—Epilepsy is a neurological disorder characterized by recurrent seizures caused by abnormal electrical activity in the brain; therefore, electroencephalography (EEG)-based seizure-type classification is needed to support a more objective analysis process. This study develops an epileptic seizure-type classification system using a deep learning architecture that combines a Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM). The proposed system consists of preprocessing, Wavelet Packet Decomposition (WPD), statistical feature extraction, a two-stage hierarchical classification scheme, and Focal Loss to address class imbalance. The evaluation was conducted on the TUH EEG Seizure Corpus (TUSZ) version 2.0.3 using a 5-fold stratified cross-validation scheme. The experimental results show that the Hybrid CNN–BiLSTM achieved an accuracy of 92.00% and an F1-score of 92.00% in the Top 3-class scenario, an accuracy of 93.00% and an F1-score of 93.00% in the Specific 5-class scenario, and an accuracy of 90.56% and an F1-score of 91.00% in the seven-class hierarchical scenario. As further validation, the direct (flat) seven-class classification scenario without the hierarchical scheme achieved only 71.90% accuracy and a 72.00% F1-score, demonstrating that the two-stage hierarchical scheme plays a significant role in reducing ambiguity among similar seizure classes. These results indicate that the Hybrid CNN–BiLSTM architecture can effectively support EEG-based epileptic seizure-type classification with good performance.**

***Keywords*—Epilepsy, electroencephalography (EEG), Hybrid CNN–BiLSTM, hierarchical classification, seizure-type classification**

I. Introduction

Neurological disorders are among the leading causes of global disability burden [1]. Epilepsy, as one of the most common chronic neurological disorders, is characterized by recurrent seizures resulting from excessive electrical activity in brain neurons and affects more than 50 million people across various age groups [2]. Each seizure type has distinct clinical manifestations and requires different medical management, making accurate seizure-type identification a critical step in determining appropriate medical action. Electroencephalography (EEG) is the primary electrophysiological technique used to diagnose epilepsy; however, manual interpretation requires specialized clinical expertise, is time-consuming, and can be affected by inter-rater variability. Therefore, the development of automated and objective EEG-based seizure-type classification systems has become an important need to support the diagnostic process.

Various approaches have been developed to address this need. In conventional machine learning approaches, feature extraction from EEG signals has been widely used because it can informatively represent signal characteristics, such as EEG signal dynamics analysis [3] and non-linear feature extraction based on the Multiscale Hjorth Descriptor [4]. Berrich and Guennoun combined CNN and DNN as feature extractors before classification with SVM, obtaining an accuracy of approximately 96% on the Bonn dataset [5]. Advances in deep learning such as CNN, LSTM, and BiLSTM are increasingly utilized because they can automatically learn the spatial and temporal patterns of EEG signals without requiring manual feature engineering. Albaqami et al. proposed MP-SeizNet, a multi-path CNN–BiLSTM architecture with DWT, DTCWT, and WPD features, which achieved an F1-score of approximately 98.01% in classifying five specific seizure types using the TUSZ dataset [6]. Hermawan et al. developed a multi-representation approach using CNN, LSTM, and hybrid CNN–LSTM for epileptic seizure detection [7], while McCallan et al. conducted a systematic review of various deep learning architectures on the TUSZ dataset and reported that most models achieved accuracies above 95% in classifying up to six seizure classes, while also highlighting the challenges posed by montage and sampling-frequency heterogeneity in clinical EEG recordings [8].

Among the various conventional machine learning methods, Support Vector Machine (SVM) has been recorded as

one of the most widely used methods and has shown good performance in EEG-based epileptic seizure detection and classification [9]. The dominance of SVM in the literature makes it a relevant benchmark for assessing the feasibility of a new classification model. Therefore, this study includes SVM as a comparison baseline, not to compare two methods on equal footing, but to validate whether the performance of the proposed model has reached a comparable feasibility standard or surpasses the most commonly used method in this field.

Nevertheless, several issues remain inadequately addressed in prior research. First, some studies have not elaborated in detail on how they handle the heterogeneity of clinical EEG recordings arising from differences in montage configuration and sampling frequency, even though this has been highlighted as a distinct challenge [8]. Second, evaluation in several studies is generally conducted on a relatively balanced subset of classes [5], [6], so it does not yet reflect the substantial class imbalance in real clinical data, where non-specific seizure classes actually dominate. Third, most studies still perform multiclass classification directly without staged grouping, even though some seizure types, such as non-specific focal and generalized seizures, have similar characteristics and are prone to confusion. Fourth, most studies proposing deep learning architectures for seizure classification directly report model performance without including a conventional method baseline for comparison [6], [7], [10], making it difficult to assess that performance relative to simpler approaches.

Based on these gaps, this study develops a more systematic EEG signal processing pipeline, starting with preprocessing to standardize variations in montage and sampling frequency, followed by Wavelet Packet Decomposition (WPD) to decompose the signal into several frequency sub-bands. The statistical features resulting from this decomposition are used in the main classification model, a Hybrid CNN–BiLSTM adapted from the MP-SeizNet architecture [6], with Focal Loss and dynamic class weighting applied to handle the imbalance in seizure class distribution. This study also designs a two-stage hierarchical classification scheme that first separates non-specific and specific seizure groups before proceeding to final classification, allowing the model to focus more effectively on the characteristics of each class group. In addition, this study includes SVM as a comparison baseline. Unlike most prior studies that restrict evaluation to specific seizure classes, this study deliberately includes non-specific seizure classes (FNSZ and GNSZ) within the scope of seven-class classification, so that the resulting system more fully represents real clinical data conditions.

A summary of this study's position relative to reference studies is presented in Table I.

The main contributions of this study are summarized as follows: (1) developing a systematic EEG signal preprocessing pipeline to handle montage and sampling-frequency heterogeneity in a multi-source clinical dataset; (2) analyzing the effectiveness of Focal Loss and dynamic class weighting in reducing classification bias caused by seizure class imbalance; (3) designing and evaluating a two-stage hierarchical classification scheme compared to a direct (flat) classification scheme for classes with similar characteristics; and (4) measuring the performance of the Hybrid CNN–BiLSTM model relative to an SVM baseline in classifying seven epileptic seizure types encompassing both non-specific and specific classes.

II. Research Methodology

This study develops a processing pipeline starting from raw EEG signals, through preprocessing, feature extraction, and finally classification using two approaches, namely SVM and the Hybrid CNN–BiLSTM model, as illustrated in Fig. 1.

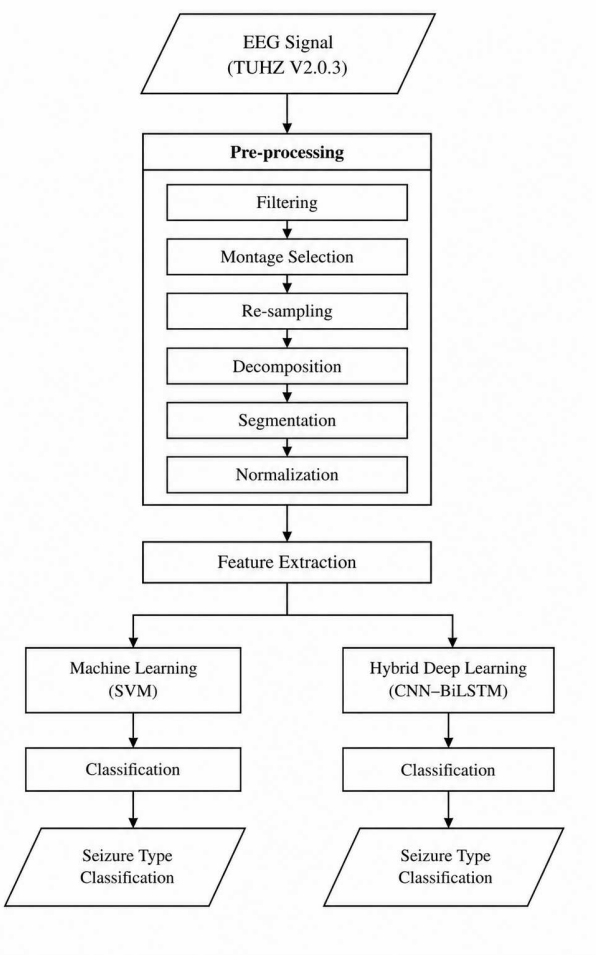

Fig. 1. Research flow diagram: preprocessing → feature extraction → classification (SVM and Hybrid CNN–BiLSTM) → seizure type.

A. Dataset

This study uses the TUH EEG Seizure Corpus (TUSZ) version 2.0.3 [11], a subset of the Temple University Hospital EEG Corpus that focuses on clinical EEG recordings with a high probability of containing seizure activity. The dataset consists of 7,366 EEG files from 675 subjects, stored in European Data Format (EDF) with seizure annotations provided in CSV format. Signals were recorded following the international 10–20 electrode system standard, with sampling frequencies varying between 250–1024 Hz depending on the recording

TABLE I
SUMMARY OF RELATED STUDIES ON EEG-BASED EPILEPTIC SEIZURE-TYPE CLASSIFICATION

Study	Year	Method	Result
Albaqami et al. [6]	2022	Multi-path CNN + BiLSTM with DWT, DTCWT, WPD features (TUSZ dataset)	5 specific seizure classes; F1-score $\approx$ 98.01% (seizure-wise)
McCallan et al. [8]	2023	Systematic review of WPD + various deep learning architectures (TUSZ dataset)	6 seizure classes; most models achieve >95% accuracy
Berrich & Guennoun [5]	2025	CNN–SVM and DNN–SVM with PCA and DWT (Bonn dataset)	5 seizure classes; accuracy $\approx$ 96%
Hermawan et al. [7]	2024	CNN, LSTM, and hybrid CNN–LSTM	5 seizure classes; accuracy $\approx$ 97%
This study	2026	Hybrid CNN–BiLSTM + WPD, Focal Loss, two-stage hierarchical classification (TUSZ v2.0.3 dataset)	7 seizure classes (including non-specific); F1-score 91.00%, accuracy 90.56% (seizure-wise)

device. All data were anonymized by the dataset provider prior to release for academic research.

Seizure types in TUSZ follow the ILAE 2017 nomenclature [12], [13] and include Focal Non-Specific Seizure (FNSZ), Generalized Non-Specific Seizure (GNSZ), Simple Partial Seizure (SPSZ), Complex Partial Seizure (CPSZ), Absence Seizure (ABSZ), Tonic Seizure (TNSZ), and Tonic–Clonic Seizure (TCSZ). Unlike some prior studies that evaluate only specific seizure classes, this study includes all seven classes, including FNSZ and GNSZ, which are the most dominant classes in real clinical data.

B. EEG Signal Preprocessing

The preprocessing stage aims to standardize the heterogeneity of clinical EEG recordings caused by differences in montage and sampling frequency. The montage configuration used is REF-01_tcp_ar (bipolar, 22 channels), which is the most dominant montage scheme in the dataset (74.8% of all files). A summary of the preprocessing specifications is presented in Table II.

TABLE II
SPECIFICATION OF EEG SIGNAL PREPROCESSING STAGES

Stage	Specification
Filtering	4th-order Butterworth bandpass filter, 0.5–40 Hz range
Montage selection	22 bipolar channels, REF-01_tcp_ar montage
Resampling	All recordings standardized to 250 Hz
Decomposition	Level-3 Wavelet Packet Decomposition (WPD), Daubechies 4 (db4) mother wavelet, producing 8 frequency sub-bands
Segmentation	2-second sliding window (62 samples at 250 Hz), 50% overlap
Normalization	Z-Score Normalization, computed independently per patient recording
Value clipping	Feature/signal clipping to $[-5.0; 5.0]$ across all deep learning architectures

The 0.5–40 Hz bandpass range was chosen to encompass the five main EEG wave bands (delta, theta, alpha, beta, gamma) while suppressing noise outside that range. The 250 Hz resampling target was determined as the lowest sampling frequency found in the dataset after montage standardization, so that all signals have a uniform temporal resolution without requiring upsampling interpolation, which risks introducing artificial information. The 2-second sliding-window duration was chosen because the shortest seizure duration in the data is around 1 second, so this window is considered sufficient to capture seizure information without cutting off important patterns at segment boundaries.

After the file-integrity verification, resampling, and WPD decomposition stages, 5,271 valid decomposed files were obtained. The segmentation process produced 2,285 seizure segments, with an imbalanced distribution across classes: FNSZ (56.0%), GNSZ (32.2%), CPSZ (6.1%), ABSZ (2.2%), SPSZ (2.1%), TNSZ (0.9%), and TCSZ (0.5%). After the sliding-window windowing process with 50% overlap, a total of 153,396 windows were obtained, with FNSZ and GNSZ together accounting for 88.1% of the overall data. A summary of the distribution is presented in Table III.

TABLE III
DISTRIBUTION OF WINDOW COUNTS PER SEIZURE TYPE AFTER PREPROCESSING

Seizure Type	Group	Windows	Pct.
FNSZ	Non-specific	79,585	51.9%
GNSZ	Non-specific	55,523	36.2%
CPSZ	Specific	14,645	9.5%
SPSZ	Specific	1,884	1.2%
TCSZ	Specific	930	0.6%
TNSZ	Specific	617	0.4%
ABSZ	Specific	212	0.1%
Total	–	**153,396**	**100%**

C. Feature Extraction

From the normalized WPD decomposition coefficients, six statistical features were computed for each channel and each frequency sub-band: energy, variance, standard deviation, skewness, kurtosis, and Shannon entropy. The combination of 22 channels, 8 sub-bands, and 6 statistical features produces a tabular feature representation of dimension $(22, 48)$. In addition to this tabular representation, this study also uses a raw signal representation consisting of approximation coefficients

from the decomposition, with dimension $(22, 62)$, i.e., 22 channels with 62 sequential time points per 2-second window. Both representations originate from the same data source but are used on different model processing paths: the tabular WPD features are used as input for the SVM model (after flattening into 1,056 features) and the convolutional path of the Hybrid CNN–BiLSTM, while the raw signal representation is used as input for the BiLSTM path of the Hybrid CNN–BiLSTM. Feature standardization (Standard Scaler) was computed separately and only from the training data in each fold to prevent data leakage.

D. Classification Model Architecture

1) Support Vector Machine (baseline): SVM is used as a comparison baseline with a Radial Basis Function (RBF) kernel, chosen for its ability to handle the complex and not always linear relationships among EEG features [9]. The flattened WPD statistical features (1,056 values) are used as the input vector.

2) Hybrid CNN–BiLSTM: The Hybrid CNN–BiLSTM architecture is adapted from MP-SeizNet [6], with adjustments to the type of input features, the attention mechanism configuration, the fusion scheme, and the class-imbalance handling strategy. The model consists of two parallel processing paths.

The first path is a 1D-CNN that processes the tabular WPD features of dimension $(22, 48)$ using multiscale convolution, i.e., three parallel convolution operations with 64 filters and different kernel sizes (3, 5, and 7), to capture frequency feature patterns at both local and regional scales across electrode channels [14]. The three outputs are concatenated into a 192-dimensional representation, followed by batch normalization, Leaky ReLU activation, and spatial dropout. A residual connection is applied to preserve information flow in the deeper network, and Global Average Pooling reduces the final representation to a 128-feature vector.

The second path is a BiLSTM that processes the raw signal representation of dimension $(22, 62)$ to capture temporal dependencies and amplitude change patterns in seizure signals [15]. This path uses a Bidirectional LSTM with 128 units per direction (256 total dimensions), followed by an attention mechanism [16] that computes a relevance score for each time step, normalizes it into attention weights, and sums the sequential representation in a weighted manner (Global Sum Pooling) into a single 256-dimensional context vector.

The final representations from the CNN path (128 dimensions) and the BiLSTM path with attention (256 dimensions) are concatenated into a single 384-dimensional fusion vector, passed through three stacked dense layers (256, 128, 64 units) with ReLU activation and dropout, before being classified using a final dense layer with softmax activation. An illustration of the architecture is shown in Fig. 2, and a summary of the configuration is presented in Table IV.

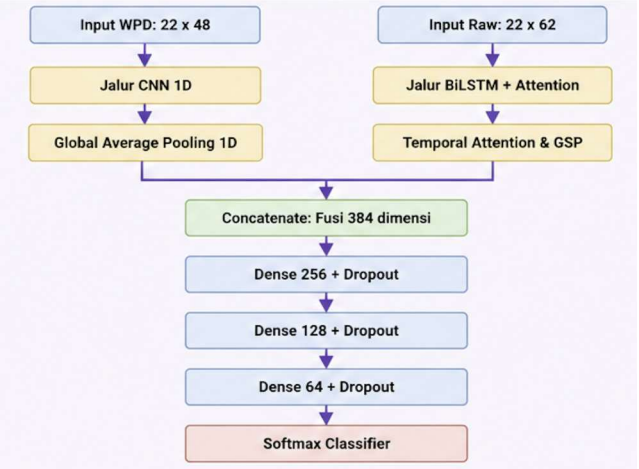

Fig. 2. Hybrid CNN–BiLSTM architecture with two parallel paths (multiscale 1D-CNN and BiLSTM+attention) and a fusion layer.

TABLE IV
SUMMARY OF THE HYBRID CNN–BILSTM ARCHITECTURE CONFIGURATION

Component	Specification
1D-CNN path	Input (22,48); parallel convolution with 64 filters, kernels 3/5/7; residual connection; Global Average Pooling → 128 features
BiLSTM + Attention path	Input (22,62); BiLSTM 128 units/direction (256 dim); attention + Global Sum Pooling → 256 features
Fusion	Concatenate 128 + 256 = 384 dimensions
Dense layers	256 → 128 → 64 units, ReLU activation, dropout at each layer
Final classification	Dense + softmax according to the number of target classes

E. Two-Stage Hierarchical Classification Scheme

The distribution of seizure classes in TUSZ is highly imbalanced: the non-specific group (FNSZ and GNSZ) accounts for up to 88.1% of the overall data, while the specific group (ABSZ, CPSZ, SPSZ, TCSZ, TNSZ) accounts for only a small remaining fraction. To address this, this study designs a two-stage hierarchical classification scheme as an alternative to direct (flat) seven-class classification, as illustrated in Fig. 3.

In the first stage, a sub-model acts as a binary gatekeeper classifier that separates the data into a non-specific group (class 0) and a specific group (class 1). If the first-stage output is the non-specific group, the data is routed to stage-2a sub-model, which performs binary classification of FNSZ vs. GNSZ. If the first-stage output is the specific group, the data is routed to stage-2b sub-model, which performs five-class classification (ABSZ, CPSZ, SPSZ, TCSZ, TNSZ). Routing at inference time uses the first-stage predictions rather than the true labels, so errors in the first stage can propagate to the subsequent stage — this scheme is deliberately maintained without manual correction so that the evaluation reflects realistic deployment conditions.

F. Class Imbalance Handling

The class-imbalance handling strategy is applied differently for each model, tailored to its computational characteristics.

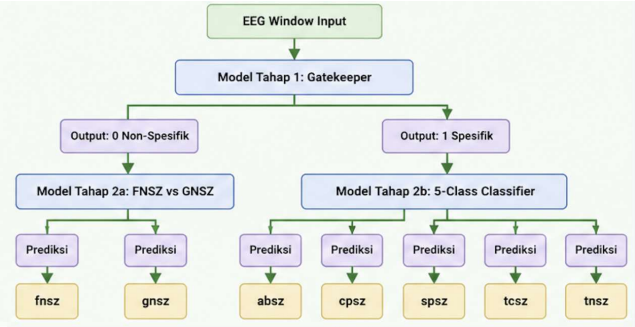

Fig. 3. Two-stage hierarchical classification flow: gatekeeper (non-specific vs. specific) → FNSZ/GNSZ or five specific classes.

For SVM, tiered Random Undersampling (RUS) is applied to each hierarchical classification sub-task, on the training data only, given the significantly increased computational complexity of training a kernel SVM as sample size grows. For the Hybrid CNN–BiLSTM, class imbalance is handled using a multiclass Focal Loss function ($\gamma = 1.0$) combined with a dynamic class-weighting scheme based on the square root of class frequency (Sqrt Class Weight), chosen to avoid overly extreme weighting and to maintain convergence stability. The Focal Loss function is formulated as:

$$L_{FL}(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t) \tag{1}$$

where p_t denotes the predicted probability for the true class, α_t is the class weighting factor, and γ is the focusing parameter. With this approach, the deep learning model can make use of the entire training dataset without needing to discard samples as in the SVM approach.

G. Training and Evaluation Configuration

The model was trained and evaluated using a 5-fold stratified cross-validation (seizure-wise) scheme, with the test-data indices kept identical between SVM and Hybrid CNN–BiLSTM in each fold to ensure a consistent comparison. The Hybrid CNN–BiLSTM model was trained using the Adam optimizer (initial learning rate 0.0001), a batch size of 128, a maximum of 150 epochs, gradient clipping at a norm of 1.0, mixed precision for GPU acceleration, early stopping (patience of 30 epochs), and ReduceLROnPlateau (patience of 10 epochs, factor 0.5).

Model performance was evaluated using confusion matrices along with accuracy, precision, recall, and weighted F1-score metrics [17], given that the substantial class imbalance makes accuracy alone insufficiently representative. Three evaluation scenarios were conducted: (1) direct classification on a dominant-class subset (Top 3: CPSZ, FNSZ, GNSZ) and a specific-class subset (Specific 5: ABSZ, CPSZ, SPSZ, TCSZ, TNSZ); (2) comparison of the two-stage hierarchical classification scheme against the flat seven-class scheme using an identical architecture; and (3) final performance evaluation of the seven classes with SVM as a comparison baseline.

III. Results and Discussion

A. Classification Results on Class Subsets

Table V summarizes the evaluation results for the two direct classification scenarios, namely the three-dominant-class subset (Top 3) and the five-specific-class subset (Specific 5). The Hybrid CNN–BiLSTM model consistently outperforms the SVM baseline in both scenarios. In the Top 3 scenario, Hybrid CNN–BiLSTM achieved an F1-score of 92.00% compared to 73.00% for SVM. In the Specific 5 scenario, Hybrid CNN–BiLSTM achieved an F1-score of 93.00% compared to 88.00% for SVM.

These results indicate that dominant (non-specific) and specific classes have different learning patterns, which is one of the reasons underlying the design of the two-stage hierarchical classification scheme evaluated in the following stage.

B. Comparison of Hierarchical, Flat, and SVM Baseline Schemes

Table VI compares seven-class classification performance across three approaches: Hybrid CNN–BiLSTM with the two-stage hierarchical scheme, Hybrid CNN–BiLSTM with the direct (flat) classification scheme, and the SVM baseline with the hierarchical scheme.

The two-stage hierarchical Hybrid CNN–BiLSTM scheme achieves an F1-score of 91.00%, significantly higher by 19 percentage points than the flat scheme with an identical architecture (72.00%), and also higher than the SVM baseline (72.00%). An interesting finding is that the flat CNN–BiLSTM model's performance is nearly on par with the SVM baseline under the hierarchical scheme, despite the two using very different modeling approaches. This indicates that the advantage of deep learning architectures does not automatically emerge merely from model complexity, but depends heavily on how the classification problem is structured.

Confusion-matrix analysis of the flat model shows that the most dominant classification errors occur between the non-specific classes: 3,895 GNSZ instances were predicted as FNSZ, and 1,051 FNSZ instances were predicted as GNSZ, in addition to 651 CPSZ instances confused with FNSZ. Under the hierarchical scheme, similar errors in the FNSZ–GNSZ pair decreased significantly to 926 and 797 instances, respectively. This pattern shows that without an initial separation between the non-specific and specific groups, the model struggles to form clear decision boundaries between FNSZ, GNSZ, and CPSZ, causing errors to spread more broadly compared to when classification is broken down into more targeted stages.

C. Per-Class Performance Analysis and Imbalance Handling

Table VII summarizes the precision, recall, and F1-score per class under the two-stage hierarchical scheme of the Hybrid CNN–BiLSTM model.

The highest F1-score was obtained for the FNSZ class (91.8%), while the lowest F1-score was obtained for the TCSZ class (72.0%), giving a gap of 19.8 percentage points between the majority class and the minority class with the

TABLE V
MODEL EVALUATION RESULTS FOR THE CLASS-SUBSET CLASSIFICATION SCENARIOS (WEIGHTED AVERAGE, 5-FOLD CV SEIZURE-WISE)

Scenario	Model	Acc.	Prec.	Rec.	F1
Top 3 (CPSZ, FNSZ, GNSZ)	**Hybrid CNN-BiLSTM**	**92.00%**	**92.00%**	**92.00%**	**92.00%**
Top 3 (CPSZ, FNSZ, GNSZ)	SVM baseline	71.00%	76.00%	71.00%	73.00%
Specific 5 (ABSZ, CPSZ, SPSZ, TCSZ, TNSZ)	**Hybrid CNN-BiLSTM**	**93.00%**	**94.00%**	**93.00%**	**93.00%**
Specific 5 (ABSZ, CPSZ, SPSZ, TCSZ, TNSZ)	SVM baseline	87.00%	89.00%	87.00%	88.00%

TABLE VI
COMPARISON OF SEVEN-CLASS EVALUATION RESULTS: HIERARCHICAL, FLAT, AND SVM BASELINE SCHEMES

Model	Acc.	Prec.	Rec.	F1
Hybrid CNN-BiLSTM (Hierarchical)	**90.56%**	**91.00%**	**91.00%**	**91.00%**
Hybrid CNN-BiLSTM (Flat)	71.90%	74.00%	72.00%	72.00%
SVM Baseline (Hierarchical)	70.00%	78.00%	70.00%	72.00%

TABLE VII
PER-CLASS F1-SCORE UNDER THE TWO-STAGE HIERARCHICAL SCHEME (HYBRID CNN–BILSTM)

Class	Group	Prec.	Rec.	F1
FNSZ	Majority	92.5%	91.2%	91.8%
GNSZ	Majority	89.6%	90.4%	90.0%
CPSZ	Specific	86.6%	89.7%	88.1%
SPSZ	Specific (minority)	87.4%	86.6%	87.0%
ABSZ	Specific (minority)	92.1%	92.1%	92.1%
TNSZ	Specific (minority)	85.5%	77.7%	81.4%
TCSZ	Specific (minority)	69.0%	75.3%	72.0%

lowest performance. This gap is considerably smaller than that of the SVM baseline, where the recall for the GNSZ class reached only 63.10% compared to 90.40% for the Hybrid CNN–BiLSTM on the same class. These results show that Focal Loss combined with dynamic class weighting (Sqrt Class Weight) in the Hybrid CNN–BiLSTM model is more effective at suppressing majority-class bias than the Random Undersampling scheme used for SVM, although the TCSZ and TNSZ classes still show relatively lower performance due to the very limited number of samples in the test data.

D. Comparison with Reference Studies

Table VIII compares the performance of the proposed model with the reference study MP-SeizNet [6], which forms the basis of the architectural adaptation in this study.

Although this study's F1-score (91.00%) is lower than that of the reference study (98.01%), this comparison should be interpreted while considering the difference in evaluation scope. The reference study evaluated only five specific seizure classes and did not include non-specific classes (FNSZ and GNSZ), even though, based on the distribution of the TUSZ dataset used, these two non-specific classes are actually the most frequently found seizure types in clinical data (more than 88% of the overall segmented data). This study deliberately includes FNSZ and GNSZ within the seven-class classification scope, so that the resulting system is able to distinguish all seizure types comprehensively, not only the minority classes that are naturally easier to separate due to their smaller and more homogeneous sample sizes. This increase in complexity due to the expanded scope is the primary cause of the F1-score gap, as demonstrated by the flat seven-class experiment, which showed a significant performance decline when FNSZ and GNSZ were included without a hierarchical scheme (Table VI). In other words, the performance gap does not indicate a weakness in the adapted architecture or method, but rather a logical consequence of an evaluation scope deliberately made more comprehensive and more representative of real clinical data conditions.

Technically, the addition of FNSZ and GNSZ increases classification complexity because both classes have very large numbers of windows and are characteristically close to one another in terms of temporal and spectral characteristics in the WPD-derived features, particularly when focal seizure patterns spread or when generalized activity appears across several channels. This makes the model's decision boundary more complex, since the model must not only distinguish relatively specific minority classes but also separate two majority classes with closely related feature representations — a pattern consistent with the FNSZ–GNSZ misclassification findings in Section III-B.

E. Study Limitations

Several limitations should be noted. First, the data split in this study was performed on a window basis using stratified k-fold, rather than on a patient basis (patient-wise), so the resulting F1-scores reflect performance under a seizure-wise/window-wise scenario and require further validation using a patient-wise split scheme to assess the model's generalization capability to new patients — as highlighted in various prior studies, patient-wise scenarios generally yield lower performance compared to seizure-wise scenarios. Second, the TCSZ and TNSZ classes still show relatively low performance due to very limited sample sizes, so more sophisticated data-balancing methods could potentially further improve performance for these classes. Third, ambiguity between FNSZ and

TABLE VIII
COMPARISON OF THE PROPOSED MODEL'S RESULTS WITH THE REFERENCE STUDY

Study/Model	Classes Classified	Evaluation Scenario	F1-Score	Accuracy
Albaqami et al. [6] (reference)	5 specific classes (ABSZ, CPSZ, SPSZ, TCSZ, TNSZ)	Seizure-wise (5-fold CV)	98.01%	Not reported
Hybrid CNN-BiLSTM Hierarchical (proposed)	7 classes (ABSZ, CPSZ, FNSZ, GNSZ, SPSZ, TCSZ, TNSZ)	Seizure-wise (5-fold CV)	91.00%	90.56%

GNSZ, although reduced through the hierarchical scheme, still leaves hundreds of misclassifications in the test data, which could potentially be improved through the addition of inter-channel spatial features or an additional attention mechanism at the gatekeeper stage.

IV. CONCLUSION

This study proposed a Hybrid CNN–BiLSTM model, adapted from the MP-SeizNet architecture, combined with Wavelet Packet Decomposition-based preprocessing, Focal Loss with dynamic class weighting, and a two-stage hierarchical classification scheme for classifying seven epileptic seizure types on the TUSZ version 2.0.3 dataset. Three main conclusions can be drawn from the results of this study.

First, the application of Focal Loss and dynamic class weighting (Sqrt Class Weight) in the Hybrid CNN–BiLSTM model successfully reduced bias toward the majority class, as shown by the GNSZ class recall reaching 90.40% compared to 63.10% for the SVM baseline, as well as a relatively smaller F1-score gap between majority and minority classes. Second, the two-stage hierarchical classification scheme proved significantly more effective than direct (flat) classification in classifying seven epileptic seizure types, with an F1-score of 91.00% compared to 72.00% for the flat scheme with an identical architecture. Third, the Hybrid CNN–BiLSTM model with the hierarchical scheme delivered the best performance across all test scenarios, achieving 90.56% accuracy and a 91.00% F1-score in seven-class classification, far higher than the SVM baseline (70.00% accuracy, 72.00% F1-score).

Although this study's F1-score is lower than that of the reference study by Albaqami et al. (98.01%), this study evaluates a broader class scope, namely seven classes including FNSZ and GNSZ, which were not evaluated in the reference study, so that the resulting system is more comprehensive and more representative of real clinical conditions in which non-specific seizures actually dominate the data. As a direction for future development, subsequent research is recommended to apply patient-wise validation to evaluate the model's generalization capability to new patients, to develop more effective data-balancing methods for minority classes such as TCSZ and TNSZ, to explore additional mechanisms for reducing classification ambiguity between FNSZ and GNSZ, and to develop frequency decomposition schemes that adapt to the characteristics of each seizure type.

REFERENCES

[1] World Health Organization, "Neurological disorders," 2025. [Online]. Available: https://www.who.int/publications/b/40350. [Accessed: Dec. 26, 2025].

[2] World Health Organization, "Epilepsy," 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/epilepsy. [Accessed: Oct. 2, 2025].

[3] S. Hadiyoso, I. Wijayanto, and A. Humairani, "Signal dynamics analysis for epileptic seizure classification on EEG signals," *Traitement du Signal*, vol. 38, no. 1, pp. 65–71, 2021.

[4] A. Rizal, S. Hadiyoso, I. Wijayanto, Triwiyanto, and S. Ziani, "Multiscale Hjorth descriptor on epileptic EEG classification," *J. Elect. Comput. Eng.*, vol. 2023, p. 4961637, 2023.

[5] Y. Berrich and Z. Guennoun, "EEG-based epilepsy detection using CNN-SVM and DNN-SVM with feature dimensionality reduction by PCA," *Sci. Rep.*, vol. 15, no. 14313, 2025.

[6] H. Albaqami, G. M. Hassan, and A. Datta, "MP-SeizNet: A multi-path CNN Bi-LSTM network for seizure-type classification using EEG," *Biomed. Signal Process. Control*, vol. 84, p. 104780, 2023.

[7] A. T. Hermawan, I. A. E. Zaeni, A. P. Wibawa, Gunawan, W. H. Hendrawan, and Y. Kristian, "A multi representation deep learning approach for epileptic seizure detection," *J. Robot. Control (JRC)*, vol. 5, no. 1, pp. 187–204, 2024.

[8] N. McCallan, S. Davidson, K. Y. Ng, P. Biglarbeigi, D. Finlay, and J. McLaughlin, "Epileptic multi-seizure type classification using electroencephalogram signals from the Temple University Hospital Seizure Corpus: A review," *Expert Syst. Appl.*, 2023.

[9] M. S. Farooq, A. Zulfiqar, and S. Riaz, "Epileptic seizure detection using machine learning: Taxonomy, opportunities, and challenges," *Diagnostics*, vol. 13, no. 6, p. 1058, 2023.

[10] D. S. Beu, A. Rizal, and I. Wijayanto, "Classification of epileptic EEG spectrogram images using CNN: A study on magnitude, phase, imaginary and complex representation," in *Proc. 2024 10th Int. Conf. Educ. Technol. (ICET)*, 2024, pp. 92–96.

[11] SzCORE Epilepsy Benchmarks, "TUH SZ: Subset of the Temple University Hospital EEG Seizure Corpus," 2025. [Online]. Available: https://epilepsybenchmarks.com/datasets/tuh/. [Accessed: Nov. 29, 2025].

[12] R. S. Fisher *et al.*, "Operational classification of seizure types by the International League Against Epilepsy: Position paper of the ILAE Commission for Classification and Terminology," *Epilepsia*, vol. 58, no. 4, pp. 522–530, 2017.

[13] R. S. Fisher *et al.*, "Epileptic seizures and epilepsy: Definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE)," *Epilepsia*, vol. 46, no. 4, pp. 470–472, 2005.

[14] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mech. Syst. Signal Process.*, 2021.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 2015.

[16] E. d. S. Farias and E. Palhares Júnior, "Application of attention mechanism with bidirectional long short-term memory (BiLSTM) and CNN for human conflict detection using computer vision," *arXiv preprint arXiv:2502.18555*, 2025.

[17] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informat.*, 2021.