An Ultra-Low Latency and High-Precision Stacked-CNN Model for Epileptic Seizure Prediction

Yishan Wu [1,*,†], Shiyue Su [2,†], Daolin Cui [3,†]

[1] Yanjing Medical College, Capital Medical University, Beijing, China, 101300

[2] College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China, 110179

[3] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, 150081

* Corresponding Author Email: w15022378142@163.com

† These authors contributed equally.

Abstract. Epilepsy, a prevalent and intricate chronic neurological disorder, poses significant challenges in clinical diagnosis and treatment, especially in promptly identifying pre-seizure periods. To comprehensively address the challenges of computational efficiency, storage cost, and real-time performance in multi-channel EEG analysis, this paper proposes a novel multimodal feature extraction framework specifically designed for single-channel electroencephalogram (EEG) signal modeling. The framework integrates traditional analytical techniques which include time and frequency analysis, short-time Fourier transform (STFT) and discrete wavelet transform (DWT) and addresses the inherent limitation of poor generalization in these methods. To mitigate this issue, we introduce nonlinear approaches such as local neighbor descriptive pattern (LNDP) and deep learning models like Transformer, thereby constructing a multimodal heterogeneous feature extraction architecture. Additionally, to overcome the limitation of long training times commonly observed in convolutional neural network (CNN) -based methods, we incorporate efficient models such as Gradient Boosting Decision Trees (GBDT) and propose a stacked convolutional neural network model, significantly improving prediction efficiency. To validate the generalization capability of the model, we transfer the trained system to the Children's Hospital Boston and the Massachusetts Institute of Technology (CHB-MIT) database. The results demonstrate that the proposed model achieves the testing accuracy of 95% with an inference time of 4 ms. Moreover, in transfer learning scenarios, the model attains a loss of approximately 7.56e-9 and processes 1128 data in merely 130 ms, which is approximately one twentieth of the time required by a multi-scale CNN model. These results highlight the model's potential to improve patients with epilepsy outcomes through timely and accurate seizure prediction.

Keywords: Ensemble Learning, Stacked-CNN Classifier, Transformer-based Features, Multi-feature Fusion.

1. Introduction

Epilepsy is a chronic brain disorder caused by abnormal discharge of brain neurons, affecting over 70 million people worldwide, making it one of the most common neurological diseases. The main clinical manifestations include recurrent seizures and transient disturbances of brain function [1, 2]. Therefore, during an epileptic seizure, rapidly distinguishing between the ictal and interictal periods is crucial for alleviating the patient's suffering.

Currently, automated detection methods for epilepsy predominantly utilize traditional machine learning, extracting features from electroencephalogram (EEG) signals across time, frequency, and time-frequency domains. Time-domain features encompass mean and variance, frequency-domain features include power spectrum and frequency band energy, and time-frequency domain features are derived via wavelet transform. After processing, these features are combined with classifiers like support vector machines (SVM) and random forests (RF) for seizure detection and prediction [3, 4]. Deep learning methods have also gained traction for automatically extracting features from raw EEG

signals or their time-frequency images. For example, convolutional neural networks (CNN) have shown excellent performance in this field [5, 6], while recurrent neural networks (RNN), particularly long short-term memory networks (LSTM) [7], address the vanishing gradient problem by retaining gradient values during training and backpropagating them. In the bidirectional long short-term memory (Bi-LSTM) architecture, each LSTM unit consists of two blocks that process the time series in opposite directions simultaneously, thus better capturing the forward and backward dependencies of the signals [8].

In feature extraction, the short-time Fourier transform (STFT) suffers from limitations due to fixed resolution [9], while discrete wavelet transform (DWT) offers improved temporal-frequency resolution at the cost of increased computational complexity. Conventional time-domain and frequency-domain analyses lack integrated localization capabilities, and the local neighborhood difference pattern (LNDP) disregards global contextual information. Moreover, Transformer models, though powerful, are computationally intensive. To overcome these limitations, we propose a collaborative framework that integrates multiple analytical methods. Specifically, wavelet transform enhances the resolution trade-off inherent in STFT, LNDP and Transformer complement each other in capturing both local and global features, and dimensionality reduction via wavelet and LNDP mitigates the computational load of the Transformer. This integrative approach enables adaptive resolution, hierarchical characterization (from biomarkers to distributed dynamics), and a balanced trade-off between computational efficiency and analytical accuracy, thereby enhancing the robustness and interpretability of EEG analysis beyond the capabilities of any single method.

In terms of classifiers, Convolutional Neural Networks (CNNs) have a high demand for large, labeled datasets to achieve optimal performance. Support Vector Classification (SVC) tends to struggle as the dataset size or dimensionality grows, while Gradient Boosting Decision Trees (GBDT) are highly sensitive to hyperparameters and require careful tuning to avoid overfitting [3]. Traditional Decision Trees (DT), though simple, often lack robustness, particularly in the presence of noisy data or when generalizing to unseen examples [10]. To address these issues, we employ ensemble learning [9], leveraging stacked generalization, to enhance generalization, stability, and accuracy while mitigating individual weaknesses.

This study presents a novel multi-modal framework for signal processing optimization, systematically integrating three complementary approaches: conventional analytical techniques (time-frequency analysis, STFT and DWT), LNDP operators and Transformer. Then, the features are trained and validated on different machine learning models (SVC, GBDT and DT), CNN and a stacked-CNN model. Given that existing research has seldom evaluated the performance on out-of-sample datasets [11], we utilize a transferring technique to enhance diagnostic precision and reducing the likelihood of misdiagnosis and missed diagnosis.

2. Materials and methods

2.1. Public dataset

The model is trained on two datasets: the Bonn Dataset and the Neurology & Sleep Centre Dataset. The Bonn Dataset comprises 500 single-channel EEG signals categorized into three groups: normal individuals (A and B: eyes open and closed), interictal epileptic patients (C and D), and ictal epileptic patients (E). The Neurology & Sleep Centre Dataset provides 1024 signals, which are divided into preictal period, interictal period and ictal period. It is available through the following link: https://www.researchgate.net/publication/308719109_EEG_Epilepsy_Datasets. In the "Training Model" column of the following table1.

We use the CHB-MIT dataset for transfer learning to verify the model's generalization ability and performance across different datasets [12, 13]. In the "Generalization Learning" column of the following table1.

Table 1. Number of data labels in the training model and generalization learning

Data Label	Training model Number	Generalization Learning Number
Interictal period	805	450
Preictal period	161	50
Ictal period	161	150

2.2. Data preprocessing workflow

To address discrepancies in sampling rates between the Bonn dataset and the Neurology & Sleep Centre (NSC) dataset, we resampled the data to ensure a consistent sampling frequency across all datasets. As the CHB-MIT dataset already features a uniform sampling rate, no additional processing was necessary in this regard. Subsequently, we applied a band-pass filter to eliminate noise by removing irrelevant frequency components, thereby improving the overall signal quality. Finally, to enhance prediction accuracy and reduce latency, we segmented the CHB-MIT dataset into 4-minute intervals, enabling more precise characterization of interictal and ictal states and thereby improving the overall performance of the model. Since all the signals from Bonn dataset and Neurology & Sleep Centre dataset have the same length, there is no need to enforce a standard length while loading the data. During the training of the Stacked-CNN classifier, features of both datasets are mixed for training and testing, with a training-to-testing ratio of 7:3. To ensure data consistency, the Min-Max normalization method is used, mapping the feature values to a range of 0 to 1. The normalization formula is:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Where, x_{norm} signifies the output of normalization, while x denoting the input. x_{max} and x_{min} represent the maximum and minimum in the values, respectively.

2.3. Multimodal heterogeneous feature extraction framework

Feature extraction in this study is implemented through a synergistic combination of multimodal methods, as presented in Figure 1.

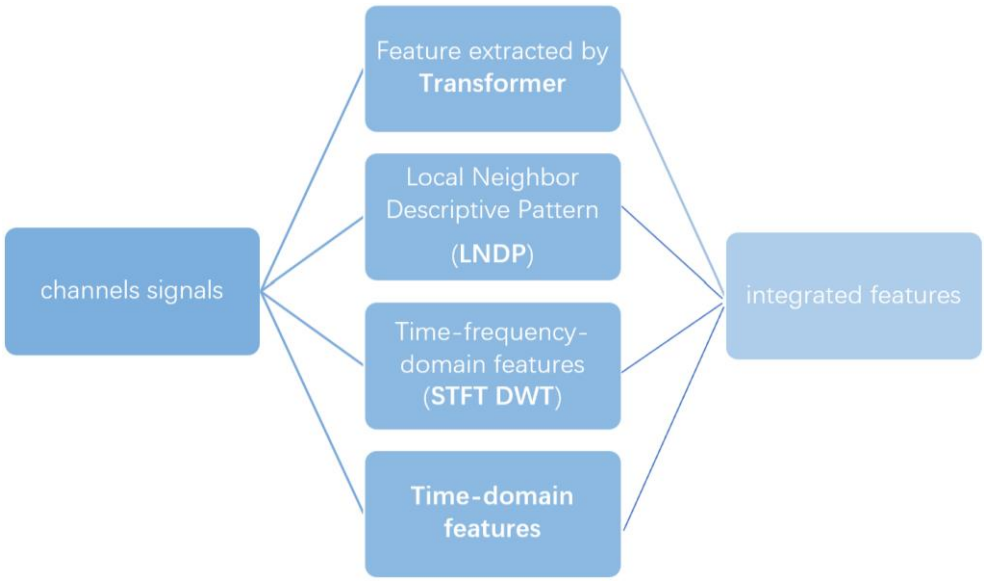

Figure 1. Feature Extraction Process for EEG Signals

2.3.1 Transformer feature extraction

The overall schematic diagram of the Transformer feature extraction model is shown in Figure 2. First, it processes raw input through an input layer, converting it to the required dimensions. It uses higher-dimensional representations to capture complex data patterns. The input then passes through

the Transformer encoder layer, which employs a multi-head self-attention mechanism and a feed-forward neural network. These components, stacked in multiple layers, extract higher-level features, with Dropout layers added to prevent overfitting.

Subsequently, a self-attention pooling layer weighs each step of the input sequence, performing a weighted average to generate a fixed-size feature representation. This allows the model to identify the most critical information within the sequence. To meet the Transformer's input requirements, the outputs from the above steps are integrated and mapped through a linear and fully connected layer. The final output is adjusted to reflect the global features of the entire sequence, combining Transformer's output with self-attention pooling layer's results.

To prevent overfitting, the model is set to evaluation mode, where dropout layers are activated to mask neurons randomly, reducing overfitting risks. Batch Normalization layers normalize data using global statistics, minimizing noise from dynamic changes. Gradient computation is also disabled to enhance computational efficiency and speed up inference [14-16].

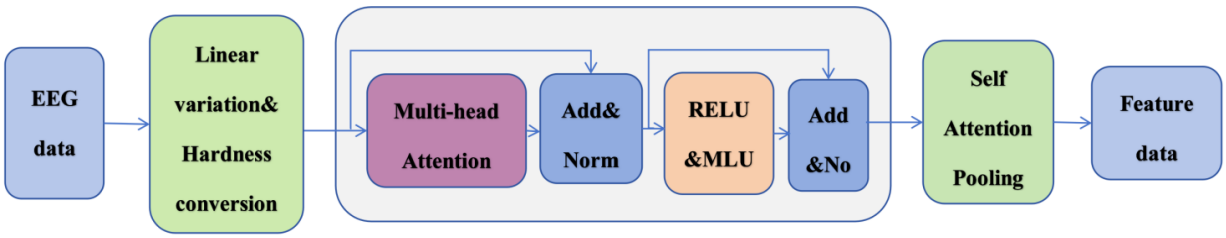

Figure 2. Transformer Structure

2.3.2 Raw Time-Domain feature calculation

Table 2. Signal Feature Formulas (Array Length = n)

Feature	Formula	
Interquartile Range	$\mathrm{IQR} = \mathrm{Q_3} - \mathrm{Q_1}$	(2)
Skewness	$\mathrm{skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - mean)^3}{std^3}$	(3)
Kurtosis	$\mathrm{kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - mean)^4}{std^4} - 3$	(4)
Log Sum	$\mathrm{log\_sum} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \ln(\lvert \mathrm{x_i} \rvert + 1 * 10^{-6})$	(5)

The results are particularly evident on the corresponding signals of each feature, and these features are also the fundamental characteristics of the signals.

2.3.3 Short-Time Fourier Transform (STFT) feature calculation

(1) Segment the signal x(t) and apply Fourier transform to obtain the STFT.

$$\mathrm{STFT(t,f)} = \sum_{\mathrm{n=0}}^{\mathrm{N-1}} \mathrm{x[n]} \times \mathrm{w[n - t_k]} \times \mathrm{e^{-j2\pi f_m n}} \quad (6)$$

Here, x[n] represents the discrete-time sequence of the input signal, w[n] is the short-time window function used to segment the signal, which is set to 256 in this study; t_k is the time offset, corresponding to the position of the short-time window, and f_m is the discrete frequency component, while $\mathrm{e^{-j2\pi fn}}$ is the frequency complex exponential in the Fourier transform. Through this step, the signal is transformed into the time-frequency domain, resulting in a complex matrix where time and frequency correspond one-to-one. As shown in Figure 3, (a) illustrates the magnitude distribution in the time-frequency domain for normal individuals, (b) corresponds to interictal epileptic patients, and (c) represents ictal epileptic patients. Specifically, in (a), the frequency components remain stable over time and are predominantly concentrated within the 0-20Hz range, encompassing delta, theta, and alpha waves, which is characteristics of normal physiological rhythms. In contrast, (b) displays

temporal instability in frequency distribution while remaining confined to the 0-20Hz range, reflecting intermittent abnormal activity. Meanwhile, (c) exhibits a broader spectral spread spanning 10-30Hz with sustained high-energy patterns, indicative of heightened neural synchronization associated with beta and gamma waves during seizure episodes.

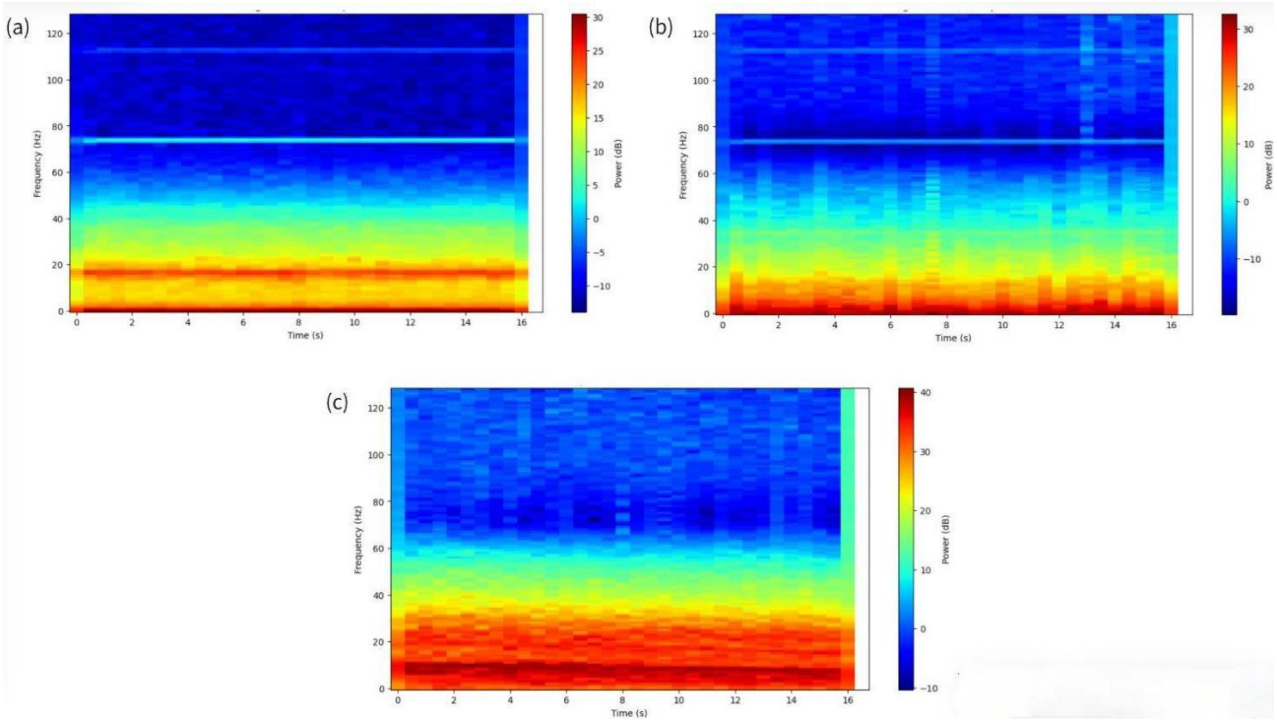

Figure 3. Average STFT power spectrum in the time-frequency domain

(2) Calculate the Power Spectral Density (PSD) Total Energy

$$E(f_m) = \sum_{t_k} STFT(t_k, f_m)^2 \tag{7}$$

The total energy at frequency f_m, denoted as $E(f_m)$, is obtained. In this step, the energy corresponding to each frequency is summed along the time dimension, resulting in the total energy corresponding to f_m.

(3) Extract the energy of the first ten frequency bands.

$$FEARURE[i] = E(f_i) \tag{8}$$

The performance of the STFT frequency bands on each signal is shown in Figure 4. It illustrates that most of signals have dominant energy in lower bands (1-3), which suggests strong low-frequency components, while several signals (451-551) including abnormal energy in higher bands (7-10) indicates noise interference.

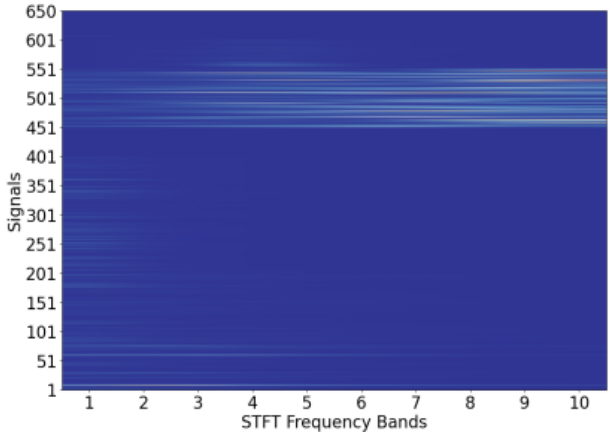

Figure 4. STFT energy distribution of the first 10 frequency bands

2.3.4 Discrete Wavelet Transform (DWT) feature calculation

(1) Discrete Wavelet Decomposition

$$x[n] = A_4 + D_4 + D_3 + D_2 + D_1 \tag{9}$$

$$A_j[k] = \sum_n x[n] * \emptyset_{j,k}[n] \tag{10}$$

$$D_j[k] = \sum_n x[n] * \psi_{j,k}[n] \tag{11}$$

This step uses the Daubechies-4 wavelet basis to decompose the signal up to the fourth level. The decomposition results are: $A_j[k]$— approximation coefficients, representing the low-frequency part; $D_j[k]$ — detail coefficients, representing the high-frequency part.

(2) Calculate the energy of each set of coefficients.

$$E_i = \sum_{n=1}^{N_i} c[n]^2 \tag{12}$$

Herein, N_i is the length of the i group of coefficients and c stands for the coefficient.

2.3.5 Local Neighbor Descriptive Pattern (LNDP) feature calculation

The size of the sliding window is set as 8 and a local neighborhood matrix using the sliding window is constructed. Based on this, we calculate the local differences from the local neighborhood matrix and binarize the difference values.

$$b_{i,j} = \begin{cases} 1, d_{i,j} >= 0 \\ 0, d_{i,j} <= 0 \end{cases} \tag{13}$$

$b_{i,j}$ is the binarized differential value. Then, LNDP encoding value is calculated and the weight vector is defined as:

$$w = 2^{j-1} \tag{14}$$

$$\mathrm{LNDP} = \sum_{j=1}^{m-1} b_{i,j} * w \tag{15}$$

All the extracted features (including Transformer features, traditional features, and LNDP features) are blended as the input of the model.

2.4. Stacked-CNN classifier model

When constructing the Stacked-CNN model, we employed a multi-scale CNN [17] and a stacking ensemble framework [18]. For the ensemble component, we selected GBDT owing to its superior predicting performance among machine learning models.

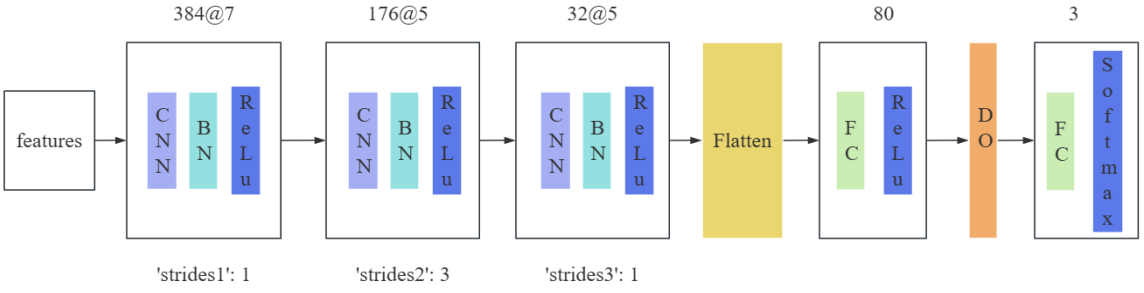

Figure 5. Multi-scale CNN Classifier Structure

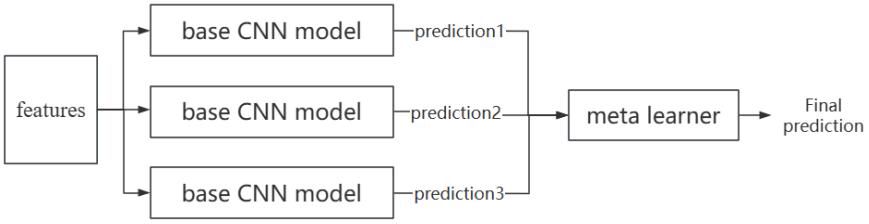

Figure 6. Stacked Ensemble Framework based on CNN models

As shown in Figure 5, regarding the multi-scale CNN, the network extracts spatial features from the input features using three filters. The multi-scale CNN extracts spatial features using three filters. The first layer uses 384 filters with a 7x7 kernel to capture local waveform features, providing a larger receptive field beneficial for EEG signal processing. The convolution operation here targets low-level feature extraction, aiding training speed and addressing the vanishing gradient problem. The formula is as follows:

$$y_{i,d}^{(1)} = \sum_{u=-2}^{2} w_{d,u}^{(1)} \cdot x_{i+u} + b_d^{(1)} \tag{16}$$

$y_{i,d}^{(1)}$ denote the value of the output feature map at position ii for the d filter in the first convolutional layer. The term $w_{d,u}^{(1)}$ represents the weight of the d filter at offset position u in the first layer. x_{i+u} indicates the value of the input signal at position i + u. Additionally, $b_d^{(1)}$ represents the bias associated with the d filter in the first layer.

The second layer employs 176 filters with a 5x5 kernel and a stride of 3, reducing spatial dimensions and boosting computational efficiency. The third layer uses 32 filters with a 5x5 kernel, further expanding the receptive field to enhance the capture of spatially distributed features across different scales. The separate formula is as follows:

$$y_{j,d}^{(2)} = \sum_{c=1}^{200} \sum_{u=-2}^{2} w_{d,u,c}^{(2)} \cdot x_{3j+u,c}^{(1)} + b_d^{(2)} \tag{17}$$

$$y_{k,d}^{(3)} = \sum_{c=1}^{145} \sum_{u=-2}^{2} w_{d,u,c}^{(3)} \cdot x_{k+u,c}^{(2)} + b_d^{(3)} \tag{18}$$

$y_{j,d}^{(2)}$ and $y_{k,d}^{(3)}$ represent the values at position ii for the d filter in the output feature maps of the 2nd and 3rd layers, respectively. $w_{d,u,c}^{(2)}$ and $w_{d,u,c}^{(3)}$ represent the weights of the d filter at the offset position u in the 2nd and 3rd layers, respectively. $x_{3j+u,c}^{(1)}$ and $x_{k+u,c}^{(2)}$ represent the values of the input signal at positions 3j+u and k + u respectively. $b_d^{(2)}$ and $b_d^{(3)}$ represent the biases for the d filter in the 2nd and 3rd layers, respectively.

In each convolutional layer, Batch Normalization is introduced to improve the model's generalization ability and nonlinear expression capacity, stabilizing the distribution of intermediate layer features. To ensure that the model's output has biological significance, we also use the Relu activation function to enhance the network's nonlinear characteristics. The formulas are as follows:

$$\widehat{y^{(l)}} = \gamma^{(l)} \cdot \frac{y^{(l)} - \mu^{(l)}}{\sqrt{\sigma^{(l)2} + \epsilon}} + \beta^{(l)} \tag{19}$$

$$a^{(l)} = \max\left(0, \widehat{y^{(l)}}\right) \tag{20}$$

After entering the Flatten section, the Adam optimizer is used, and the branch cross-entropy loss function is employed to calculate the loss. The specific loss calculation formula is as follows:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{3} y_{i,c} \log\left(p_{i,c}\right) \tag{21}$$

After the stacked convolutional layers, a fully connected layer is introduced to integrate all the extracted features and map them to the stacked ensemble classifier. The stacked ensemble classifier comprises three base learners, all of which are CNN models demonstrating optimal validation performance. GBDT is then employed to extract nonlinear features.

Additionally, a SoftMax activation function outputs class probabilities, with a Dropout layer to prevent overfitting.

As illustrated in Figure 6, this approach employs multiple models, with the integrated features dataset used to train each individual model. First, base learners are trained to predict target class labels, which are of three types as previously mentioned. In the subsequent step, all predictions are fed into a meta-model for further fitting, with the utilized labels remaining consistent.

2.5. Performance measurement

The new performance evaluation standards we have adopted are as follows:

$$[\text{Accuracy} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} I\left(y_{\text{test}}^{(i)} = \widehat{y_{\text{test}}^{(i)}}\right)] \tag{22}$$

$$[\text{Log Loss} = -\frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \sum_{c=1}^{C} y_{\text{test},c}^{(i)} \log\left(\widehat{p_{\text{test},c}^{(i)}}\right)] \tag{23}$$

3. Results

3.1. Model iteration visualization

After the annotations during the onset period were extracted as the iteration unit for each epoch, the accuracy remained stable during the first eight epochs but exhibited a sharp increase between epochs 9 and 10, as shown in Figure 7. A plausible explanation for this phenomenon is that adaptive optimizers (this study adopted Adam) inherently accumulate historical gradient information. Resetting their internal states at epoch 9 may have redirected parameter updates, thereby enabling the model to converge toward a more optimal solution. The loss shows a declining trend and reaches the minimum at epoch 9, as shown in Figure 8.

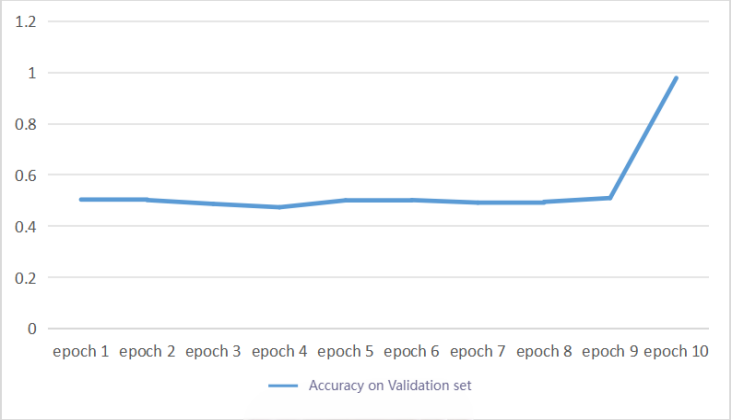

Figure 7. Accuracy vs. Epochs

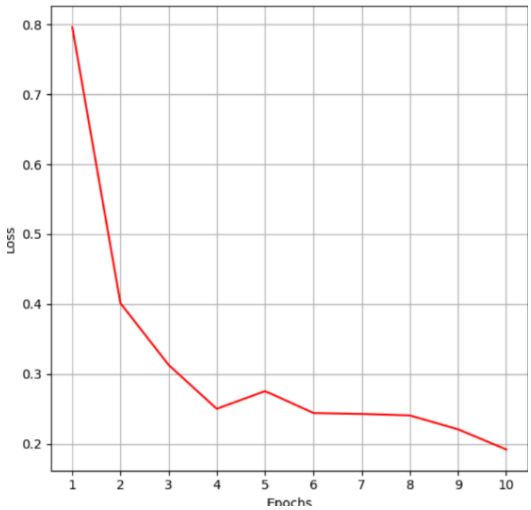

Figure 8. Loss vs. Epochs

The fluctuations in EEG signals across multiple epochs are displayed, with a focus on capturing changes in brain electrical activity over time that may be related to specific events of interest. The heatmap at the top provides an overview of global EEG activity across epochs, while the plot at the bottom offers a time-domain view of the signal amplitude variations over time in Figure 9.

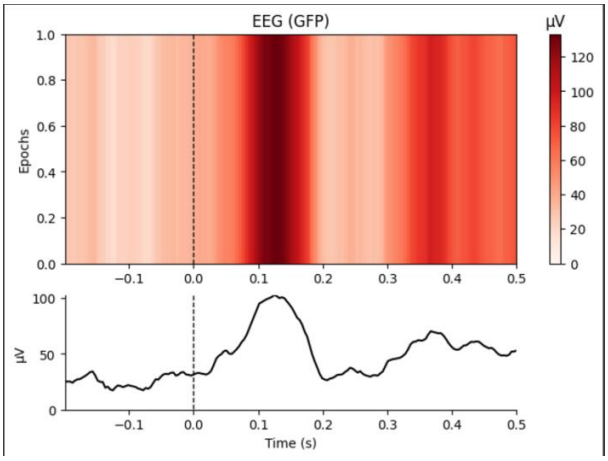

Figure 9. EEG Signal Variation Across Epochs

3.2. Ablation study of the module

3.2.1 Ablation of the feature extraction module

In this study, several classical feature extraction methods were sequentially reproduced, and the features extracted by each method were input into the same pre-trained classifier for validation. The resulting classification accuracy and loss values are presented as table 3:

Table 3. Assessment of Feature Effectiveness

Feature Extraction Method	Accuracy	Log Loss
Raw Time Domain	0.91	0.31
STFT	0.92	0.33
DWT	0.70	0.80
LNDP	0.83	0.78
TRANSFORMER	0.94	0.24
Proposed	**0.95**	**0.17**

As shown in the table, the individual use of features extracted from the raw time domain, STFT, DWT, LNDP or Transformer results in sub-optimal performance, with classification accuracy all falling below 0.95. In contrast, the integrated feature set demonstrates better performance, highlighting the advantage of combining multiple feature representations.

3.2.2 Ablation of the classifier module

Initially, classical feature extraction methods are employed for validation. Subsequently, ensemble learning techniques are applied to combine multiple models in order to identify high-performing meta-models. The resulting performance is presented as table 4:

Table 4. Model Performance Evaluation

Model	Training Acc	Training Loss	Testing Acc	Testing Loss
SVM[3, 4]	0.93	0.18	0.90	0.26
DT[10]	1.00	0.00	0.90	8.87
GBDT[3]	1.00	0.00	0.94	0.27
CNN[17]	1.00	0.00	0.93	0.18
Stacked-CNN	1.00	0.00	**0.95**	**0.17**

As shown in the table, the classification accuracies of Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting Decision Tree (GBDT), and Convolutional Neural Network (CNN) on the testing set are all below 0.95, performing worse than the proposed Stacked-CNN model on the testing set.

3.3. Temporal prediction performance of the model

In clinical applications, the timely detection of epileptic seizures is critically important. Successful early prediction can significantly reduce neurological and physical harm to patients. Our proposed

model achieves an average prediction time of 4 ms on both datasets. The prediction time is defined as the computational duration required by the final layer of the model to generate the output feature prediction. With the comparison of some classical approaches, the results are summarized as table 5:

Table 5. Model Prediction Time

Model	Inference Time (ms)
SVM[3,4]	1.8
DT[10]	5.1
GBDT[3]	99.2
CNN[17]	31.4
Stacked-CNN	4

Through comparative analysis, it is evident that CNN training significantly improves classification accuracy; however, it comes at the cost of longer computation time. By introducing a meta-model constructed with GBDT, we achieved substantial improvements in computational efficiency while maintaining high accuracy, making the approach more suitable for clinical application.

3.4. Model performance across CHB-MIT datasets for transfer learning

The Bonn and Neurology & Sleep Centre datasets have relatively small sample sizes, making them susceptible to individual-specific variability that may affect overall model performance. Moreover, since the training and validation sets originate from the same larger dataset, this setup cannot effectively demonstrate the model's generalization capability.

Given that continuous recording and specific seizures annotation are essential for training a robust model, we transferred the knowledge of stacked-CNN model to a patient-specific model with the CHB-MIT dataset [19]. We selected one case (Chb01) with the detection in about 41 hour totally. To simulate a real-world situation and investigate our model's robustness, time-series data from the interictal and preictal periods in the CHB-MIT dataset are randomly sampled, with 4-minute duration samples processed, which yields a total of 1,128 samples.

The model employed the integrated feature extraction strategy, combined with a Stacked-CNN classifier, in which the CNN model serves as the base learner to ensure high accuracy, while additional meta-model is integrated to improve computational efficiency. The model's parameters are fine-tuned, including 100 filters, a stride of 3 with a kernel size of 3*3 in the first layer, 25 filters, a stride of 2 with a kernel size of 5*5 in the second layer and 30 filters with a kernel size of 4*4 in the third layer. The model achieved a prediction loss of approximately 7.56×10^{-9}, indicating that the model possesses strong self-generalization capabilities.

As seen in table 6, we compare the Gradient Boosting classifier model, which achieves the best performance among the machine learning models. It is quite evident that our proposed model significantly outperforms other existing models in terms of classification, accuracy and prediction loss. Additionally, the stacked-CNN model demonstrates a remarkable reduction in inference time, requiring only roughly 130 ms per inference cycle compared to the roughly 2318 ms needed by the standard CNN architecture, highlighting the proposed model's superior computational efficiency.

Table 6. Model Performance Evaluation in MIT dataset

Model	Training Loss	Testing Loss	Inference Time(ms)
GBDT[3]	1e-3	1e-3	489.98
CNN[17]	0.18	0.17	2318.41
Stacked-CNN	7.56e-9	**7.56e-9**	**130.19**

4. Conclusion

This study innovates by integrating multiple feature types—combining conventional time and frequency domain features with complex representations from deep learning algorithms like Transformers—enhancing model accuracy. Our model achieves 100% training accuracy, 95% testing

accuracy, and a fast 4ms inference time. Furthermore, in transfer learning scenarios, the model attains 7e-9 loss, indicating improved predictive performance with larger datasets.

References

[1] Thijs R D, Surges R, O'Brien T J, et al. Epilepsy in adults [J]. The Lancet, 2019, 393 (10172): 689-701.

[2] Arnold S T, Dodson W E. Epilepsy in children [J]. Bailliere's Clinical Neurology, 1996, 5 (4): 783-802.

[3] Song Y, Zhang J. Discriminating preictal and interictal brain states in intracranial EEG by sample entropy and extreme learning machine [J]. Journal of Neuroscience Methods, 2016, 257: 45-54.

[4] Bou Assi E, Nguyen D K, Rihana S, et al. Towards accurate prediction of epileptic seizures: A review [J]. Biomedical Signal Processing and Control, 2017, 34: 144-157.

[5] Abou Jaoude M, Jing J, Sun H, et al. Detection of mesial temporal lobe epileptiform discharges on intracranial electrodes using deep learning [J]. Clinical Neurophysiology, 2020, 131 (1): 133-141.

[6] Cheng C, Liu Y, You B, et al. Multilevel Feature Learning Method for Accurate Interictal Epileptiform Spike Detection [J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2022, 30: 2506-2516.

[7] Daoud H, Bayoumi M A. Efficient Epileptic Seizure Prediction Based on Deep Learning [J]. IEEE Transactions on Biomedical Circuits and Systems, 2019, 13 (5): 804-813.

[8] Yin X, Fang W, Liu Z, et al. A novel multi-scale CNN and Bi-LSTM arbitration dense network model for low-rate DDoS attack detection [J]. Scientific Reports, 2024, 14 (1): 5111.

[9] Cao J, Zhu J, Hu W, et al. Epileptic Signal Classification with Deep EEG Features by Stacked CNNs [J]. IEEE Transactions on Cognitive and Developmental Systems, 2020, 12 (4): 709-722.

[10] Siddiqui M.K., Morales-Menendez R., Huang X, et al. A review of epileptic seizure detection using machine learning classifiers [J]. Brain Inf, 2020, 7, 5.

[11] Nhu D, Janmohamed M, Shakhatreh L, et al. Automated Interictal Epileptiform Discharge Detection from Scalp EEG Using Scalable Time-series Classification Approaches [J]. INTERNATIONAL JOURNAL OF NEURAL SYSTEMS, 2023, 33 (01).

[12] Andrzejak R G, Lehnertz K, Mormann F, et al. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state [J]. Physical Review E, 2001, 64 (6): 061907.

[13] Shoeb A H. Application of machine learning to epileptic seizure onset detection and treatment [D]. Massachusetts Institute of Technology, 2009.

[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need [C] // Advances in Neural Information Processing Systems: Systems. 2017, 30.

[15] Pattnaik S, Dash M, Sabut S K. DWT-based feature extraction and classification for motor imaginary EEG signals [C] // 2016 International Conference on Systems in Medicine and Biology (ICSMB). IEEE, 2016: 186-201.

[16] Lee J, Lee I, Kang J. Self-Attention Graph Pooling [C] // Proceedings of the 36th International Conference on Machine Learning. 2019: 3734-3743.

[17] Ullah I, Hussain M, Qazi E ul H, et al. An automated system for epilepsy detection using EEG brain signals based on deep learning approach [J]. EXPERT SYSTEMS WITH APPLICATIONS, 2018, 107: 61-71.

[18] AkyolKemal. Stacking ensemble based deep neural networks modeling for effective epileptic seizure detection [J]. Expert Systems with Applications, 2020, 148.

[19] Hu S, Liu J, Yang R, et al. Exploring the Applicability of Transfer Learning and Feature Engineering in Epilepsy Prediction Using Hybrid Transformer Model[J]. IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING, 2023, 31: 1321-1332.