



EEG-Based Depression Detection Using Welch Power Spectral Density Analysis and Transformer Learning

Pratima Singh¹, Alok Singh Chauhan^{1,2*}, Dhyanendra Jain³, Ashok Kumar Rai⁴, Palak Gupta⁵, Rajiv Chourasiya¹

¹School of Computer Applications and Technology, Galgotias University, Greater Noida 203201, India

²INTI International University, 71800 Nilai, Negeri Sembilan, Malaysia

³Department of CSE-AIML, ABES Engineering College, Ghaziabad, India

⁴School of Computer Science Engineering and Technology, Bennett University, Greater Noida, India

⁵School of Business, Galgotias University, Greater Noida 203201, India

*alok.chauhan@galgotiasuniversity.edu.in

Abstract. Depression diagnosis remains largely dependent on subjective clinical assessments, creating a strong need for objective and reproducible biomarkers. Electroencephalography (EEG) offers a non-invasive and cost-effective solution for capturing neural activity associated with depressive states. This study proposes a deep learning-based framework for automated depression detection using EEG Power Spectral Density (PSD) features estimated through the Welch method. After preprocessing with bandpass filtering, notch filtering, and Independent Component Analysis (ICA)-based artifact removal, PSD features were extracted from the delta, theta, alpha, beta, and gamma bands. These spectral representations were then classified using Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Transformer-based architectures. Experimental results on EEG recordings from 90 subjects (60 depressed and 30 healthy controls) demonstrated that the Transformer model achieved a classification accuracy of 83.78%, with favorable convergence, generalization, and computational efficiency compared with the CNN and RNN models. The findings confirm that Welch-based spectral analysis combined with Transformer learning provides an effective and scalable framework for objective depression detection, with strong potential for clinical decision support and portable mental health monitoring systems.

Keywords: Electroencephalography, Depression Detection, Welch Method, Power Spectral Density, Convolutional Neural Networks, Recurrent Neural Networks, Transformer Learning

(Received 2026-04-21, Revised 2026-08-23, Accepted 2026-08-24, Available Online by 2026-10-12)

1. Introduction

Depression affects more than 280 million people worldwide and remains one of the most significant contributors to global disability [1], [2]. Conventional diagnosis primarily relies on subjective symptom-based assessments, which may introduce inter-clinician variability and diagnostic bias [3,4]. Consequently, there is an increasing demand for objective biomarkers that can support reliable depression screening. EEG has emerged as a promising modality due to its non-invasive nature, affordability, and ability to capture neural oscillatory dynamics linked to depressive disorders [5,6]. Previous studies have highlighted abnormalities in alpha asymmetry, theta power, and beta activity among depressed individuals [3,7]. In this work, EEG signals recorded under resting-state and task-specific conditions were preprocessed using a 0.5–45 Hz bandpass filter, a 50 Hz notch filter, and ICA-based artifact rejection. The Welch method was then employed to estimate PSD features across the standard EEG frequency bands [8–10].

Despite significant advances in EEG-based depression detection, most prior studies have focused either on handcrafted spectral biomarkers with classical machine learning or on CNN/RNN-based deep architectures that primarily capture local spatial or sequential dependencies. A clear research gap remains in effectively modeling global inter-band and inter-channel contextual relationships within Welch-derived EEG spectral representations while maintaining robustness against EEG noise variance [11–15]. EEG has emerged as a promising modality due to its non-invasive nature, affordability, and ability to capture neural oscillatory dynamics linked to depressive disorders. Previous studies have highlighted abnormalities in alpha asymmetry, theta power, and beta activity among depressed individuals. The major contribution of this study is the integration of robust Welch-based spectral feature extraction with advanced deep learning architectures, particularly Transformer encoders, for improved depression classification [16–18].

The literature on EEG-based depression detection spans clinical biomarker studies, spectral signal processing, compact CNN frameworks, explainable deep learning, and recent Transformer-based architectures. Foundational clinical and biomarker evidence is reported in [1–3,5] and [4]. Signal processing and PSD estimation principles are supported by [8,9] and [10]. CNN and EEGNet-based representation learning methods are detailed in [6,19,20] and [21], while advanced explainable and clinical AI pipelines are discussed in [22–24]. Recent Transformer and hybrid attention-based studies relevant to this work are covered in [16,17,25] and [18]. Additional machine learning surveys, nonlinear feature studies, and broader depression-classification reviews are reported in [7, 26–30].

By explicitly grounding the proposed work across all these dimensions, the present study positions itself as an integration of clinically meaningful EEG biomarkers with robust spectral estimation and modern attention-based deep learning. Recent advances in EEG-based depression detection have shown that both handcrafted spectral biomarkers and deep learning representations contribute significantly to classification performance [15,25,31]. Traditional machine learning approaches using frontal alpha asymmetry, entropy, and connectivity measures established the diagnostic value of EEG biomarkers, while more recent CNN, RNN, graph neural network, and Transformer architectures have substantially improved predictive accuracy. However, many studies rely on curated datasets with limited subject diversity, reducing clinical generalizability [32–34]. This work addresses that gap by combining robust Welch-based PSD estimation with Transformer learning on multi-subject EEG recordings.

The novelty of the present work lies in combining variance-reduced Welch PSD estimation with a Transformer self-attention framework that learns long-range dependencies among delta, theta, alpha, beta, and gamma band powers. Unlike conventional CNN and RNN approaches, the proposed model captures broader contextual interactions across spectral tokens, leading to stronger generalization and more clinically meaningful depression biomarkers. This integration provides a scalable and interpretable pathway for objective EEG-based depression screening in real-world mental health applications.

1.1 Contributions of this Study

The main contributions of this work are as follows:

- An end-to-end EEG-based depression detection system that covers the entire process of signal pre-processing, ICA-based artifact removal, Welch PSD estimation, and classification using Transformer.
- A condensed Transformer architecture for learning low-dimensional welch-derived EEG spectral features, which can be learned with a few trainable parameters.
- Representing EEG spectral band powers as ordered feature tokens enables the Transformer self-attention mechanism to capture the relationships between various frequency bands in the EEG signal.
- Evaluation of CNN, RNN and Transformer models comparing their performance, with the same preprocessing pipeline and training conditions.
- Computationally efficient demonstration of a framework that could be deployed in a portable EEG-based depression screening system in the future.

2. Methods

The proposed framework follows a five-stage pipeline supported by established EEG signal processing principles and modern deep learning workflows [8,19,20]:

1. EEG data acquisition using standard 10–20 electrode placement.
2. Signal preprocessing involves filtering, denoising, and epoching.
3. Feature extraction using Welch-based PSD estimation.
4. Deep learning classification using CNN, RNN, and Transformer models and Performance evaluation
5. Hyperparameter optimization for robust generalization.

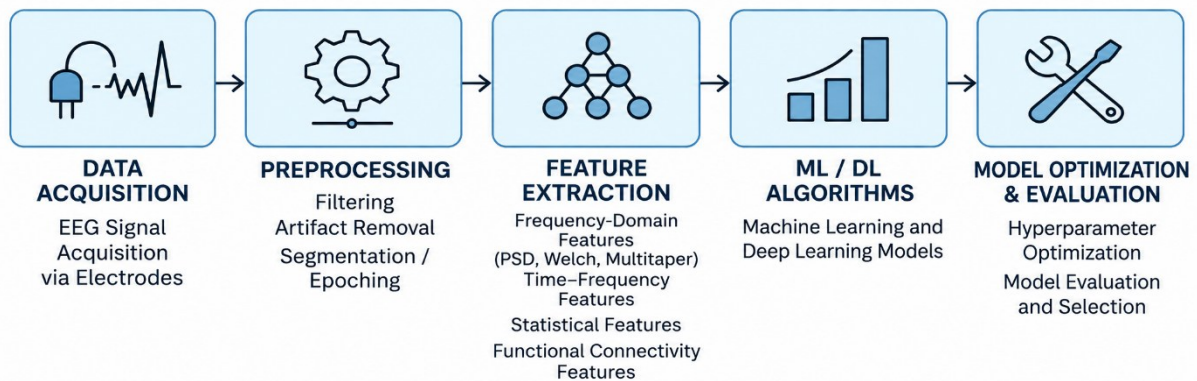


Figure 1. EEG-based ML/DL Analysis Pipeline

As shown in figure 1, we employed the Welch method [2–4,7] for feature extraction as it provides robust and smooth power spectral density (PSD) estimates by reducing noise and variance through segmentation, windowing, and averaging as shown. This makes it particularly effective for noisy, non-stationary EEG signals, enhancing the reliability of depression-related biomarkers and improving classification performance.

The Welch method was chosen because it is a method that can give stable power spectral densities (PSD) estimates by averaging modified periodograms and thus decreases the spectral variance. The continuous EEG recordings were initially divided into 1-second epochs without any overlap (256 samples at 256 Hz sampling frequency). Each epoch was first windowed using a Hamming window and then FFT-based spectral estimation was performed. These PSD values were then used to calculate the average power in 5 frequency bands: delta, theta, alpha, beta, and gamma that were used to construct the feature vector for classification.

Figure

The proposed framework for depression detection begins with EEG data acquisition from subjects under resting or task-specific conditions. The recorded signals undergo interference removal through bandpass filtering with a low cutoff of 0.5 Hz and a high cutoff of 45 Hz, along with a notch filter at 50

Hz to eliminate power line noise. Artifact removal is performed to discard non-cerebral signal components caused by eye blinks, muscle movements, and residual line noise. Independent Component Analysis (ICA) is applied to raw EEG data to isolate and remove ocular and muscular artifacts, thereby enhancing signal quality. In the feature extraction stage, Power Spectral Density (PSD) features are computed for standard EEG frequency bands—delta, theta, alpha, beta, and gamma—using the Welch method for improved spectral resolution and variance reduction. The Fast Fourier Transform (FFT) is applied to quantify frequency-specific energy distribution, with particular emphasis on alpha, beta, gamma, and theta bands, which have known associations with depressive states. The EEG data set is available in public domain at fig share which consists of 90 subjects which may be Healthy or may have Major Depressive Disorder. It can also have situations like eyes closed, eyes open or doing some tasks. The ready data after extracted spectral features are fed into a Spectro-Temporal Transformer or Time Series Transformer such as ViT or CNN-RNN is used for classification, enabling the identification of depressed versus non-depressed mental states.

3. Results and Discussion

The EEG dataset consisted of 90 subjects, including 60 clinically depressed participants and 30 healthy controls. Recordings were collected using the international 10–20 electrode system at a sampling frequency of 256 Hz. In the original source experiments, the raw depressed-subject recordings were first organized in CSV format with electrode-wise amplitude columns and later converted into EDF representation for standardized biomedical processing. This dual representation improved interoperability across preprocessing and visualization workflows. Preprocessing included bandpass filtering, notch filtering, ICA-based artifact rejection, normalization, and segmentation into fixed-length epochs.

The continuous EEG stream was divided into 1-second non-overlapping epochs, which improved spectral stability during Welch PSD estimation. Welch PSD estimation was then applied to derive delta, theta, alpha, beta, and gamma band powers. Particular emphasis was placed on alpha, theta, beta, and gamma alterations because these biomarkers repeatedly showed discriminative relevance in the source experimental observations. The CNN, RNN, and Transformer models were trained for 50 epochs with a batch size of 16 using the Adam optimizer and binary cross-entropy loss. Performance was evaluated using accuracy, precision, recall, F1-score, and loss convergence.

3.1 Data Splitting and Leakage Prevention

The dataset was split at the subject level and not the EEG epoch level to obtain an unbiased assessment of the proposed depression classification framework. EEG data were collected from 90 subjects (60 depressed and 30 healthy control) and subsequently split into mutually exclusive training, validation and test sets. Each EEG epoch from a given subject was only included in a single partition so that data from the same participant was not included in more than one partition. This approach of separating the subjects per each problem type was done in order to avoid data leakage and let the actual classification performance be representative of the model's learned ability to solve problems for subjects not previously encountered.

The data set was split into a 70% training set, 15% validation set and 15% testing set. The training subset was used to optimize the parameters of the model, the validation subset was used to select the hyperparameters and to stop the model training early, and the independent test subset was used only to evaluate the final performance of the model. 5-fold subject-wise cross validation was done to further assess the strength of the proposed model. These subjects were randomly split into five folds as nearly equally as possible whilst preserving the disease class distribution. For each iteration, the subjects in the evaluation set were not used in the training of the models. The whole process was repeated over all five folds, and the performance values were averaged over the folds to evaluate the stability and generalization ability of the proposed model. The validation loss was monitored, the hyperparameters

were tuned and early stopping was performed on the validation set, which served as a monitoring set for the model training. The independent test set was not seen at all during the model development process, but only at the end for final performance evaluation.

To minimize leakage of subject data by subject, splitting was done prior to the generation of the features at the epoch-level and model training; thus, an epoch of features for a given participant were kept together in the same data partition.

3.2 Data Representation and Preprocessing

The raw data for depressed subjects was initially stored in CSV format as shown in Figure 2, where each column represented the amplitude values recorded from the electrodes at different time points (time, Fp1, F3, C3, P3, O1, F7, T7, P7, Fz, Fp2, F4, C4, P4, O2, F8, T8, P8, Cz, Pz) as shown in figure 3.

time	Fp1	F3	C3	P3	O1	F7	T7	P7	Fz	Fp2	F4	C4	P4	O2	F8	T8	P8	Cz	Pz
0	4	-2	-2	-1	-2	-7	0	-2	3	5	0	3	7	0	5	6	3	0	3
0.003906	2	-4	-1	0	0	-8	-1	-2	0	3	-3	4	8	2	8	7	4	0	6
0.007813	0	-5	-2	1	1	-5	-3	-2	-2	0	-5	3	8	3	9	8	5	-2	6
0.011719	-2	-5	-2	1	1	-2	-4	-2	-4	-2	-6	2	6	4	7	6	5	-3	5
0.015625	-4	-5	-2	1	1	0	-5	-2	-5	-3	-5	2	5	5	3	3	6	-3	3
0.019531	-4	-4	-2	1	1	0	-5	-1	-5	-3	-4	4	4	5	0	0	7	-2	1
0.023438	-2	-3	-1	1	2	-1	-4	0	-4	-1	-3	5	4	4	0	-3	8	-1	0
0.027344	0	-2	0	1	3	-2	-2	1	-3	1	-2	6	3	4	0	-3	8	0	0
0.03125	1	-1	0	1	3	-1	0	1	-2	3	-2	6	3	2	1	-2	7	1	-1
0.035156	2	-1	-1	1	2	2	2	1	-2	3	-2	5	2	1	3	1	6	1	0
0.039063	3	-1	-1	0	1	5	4	1	-1	1	-1	2	1	0	3	4	4	1	0
0.042969	3	-1	-1	-2	0	5	5	0	-1	-2	1	0	1	-1	3	8	3	0	0
0.046875	2	-1	0	-2	0	1	4	-1	-1	-4	4	-1	2	-2	2	10	2	-1	1
0.050781	1	-2	1	-2	-1	-4	3	-1	-1	-6	6	-1	2	-2	0	9	1	-1	1
0.054688	0	-3	2	-1	0	-8	2	0	-2	-6	6	0	1	-2	0	7	0	-1	2
0.058594	-2	-3	4	0	1	-10	2	2	-3	-5	3	1	0	-1	0	3	-1	-1	2
0.0625	-3	-2	4	2	3	-9	1	4	-3	-3	-2	0	-2	0	0	0	-2	0	1
0.066406	-3	-2	4	3	4	-5	0	5	-3	-1	-7	-1	-3	0	0	-1	-2	0	1
0.070313	-2	-1	2	3	5	-3	-2	4	-1	1	-9	-2	-3	1	-1	-1	-2	1	0
0.074219	-1	-1	0	1	4	-1	-4	2	1	3	-8	-2	-1	1	-2	-2	0	2	0

Figure 2. Raw CSV representation of depressed subject EEG

The same dataset, when converted into European Data Format (EDF), is illustrated in Figure 3.

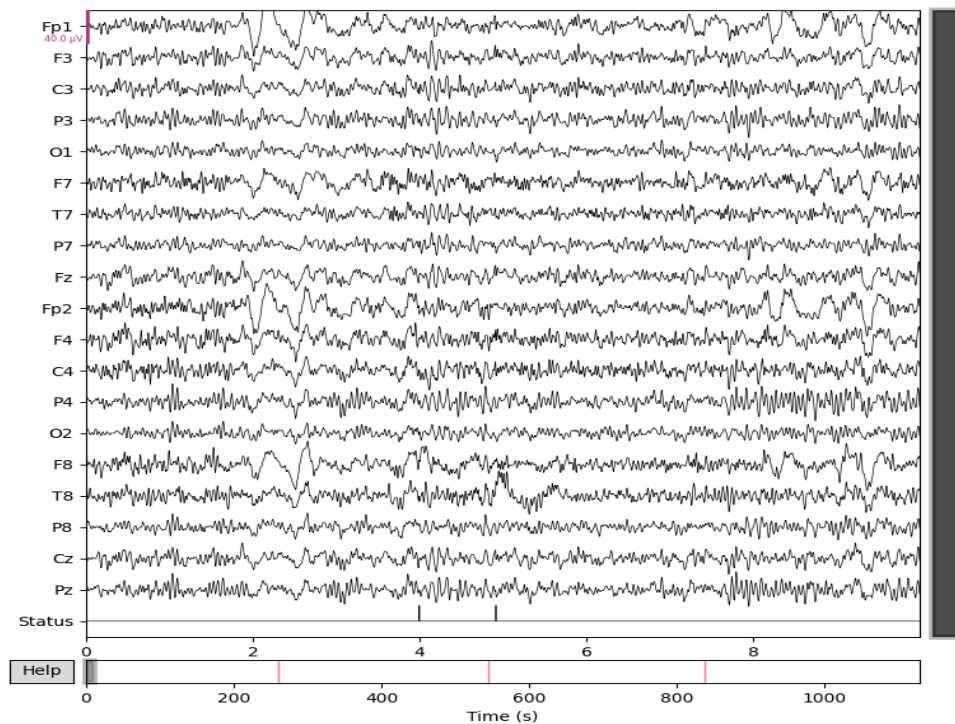


Figure 3. EDF representation of depressed subject EEG

EEG recordings were acquired under resting-state conditions with both eyes-open and eyes-closed trials. To remove noise and artifacts, a band-pass filter with lower and upper cutoff frequencies of 0.5 Hz and 45 Hz, respectively, was applied. Additionally, a notch filter at 50 Hz was employed to suppress power-line interference. The continuous EEG data was segmented into non-overlapping epochs of 1 second, resulting in a total of five epochs for each subject used in subsequent feature extraction. Frequency-domain features were derived using Power Spectral Density (PSD) estimation. The features were extracted in the frequency domain using Welch Power Spectral Density (PSD) estimation method. The EEG data was windowed with a Hamming window, and each window was Fourier transformed to generate the PSD. To reduce spectral variance and improve robustness to EEG noise, PSD estimates were computed for the windowed epochs, and the resulting spectral information was used to calculate the average band powers. The average band powers were then divided into five frequency bands: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–45 Hz) and these frequency bands were used to feed the deep learning models.

Figure 4 represents the average band power (Delta, Theta, Alpha, Beta, Gamma) computed from our EEG CSV file of a healthy participant.

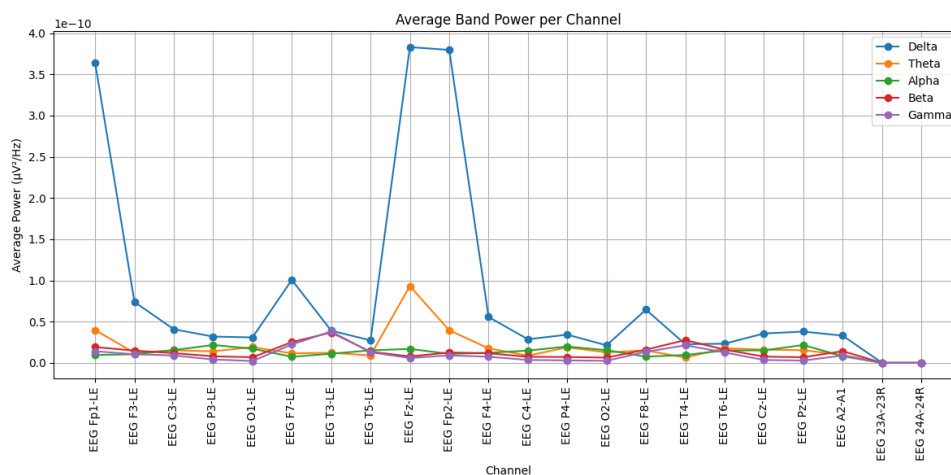


Figure 4. Average band power from csv file of a healthy participant

Analysis of the PSD values revealed distinct differences between depressed and healthy groups. Specifically, delta-band power was consistently elevated in depressed subjects, whereas beta- and gamma-band powers were reduced compared to healthy controls. These results are illustrated in Figure 5 which presents the PSD comparison between depressed and healthy subjects across frequency bands.

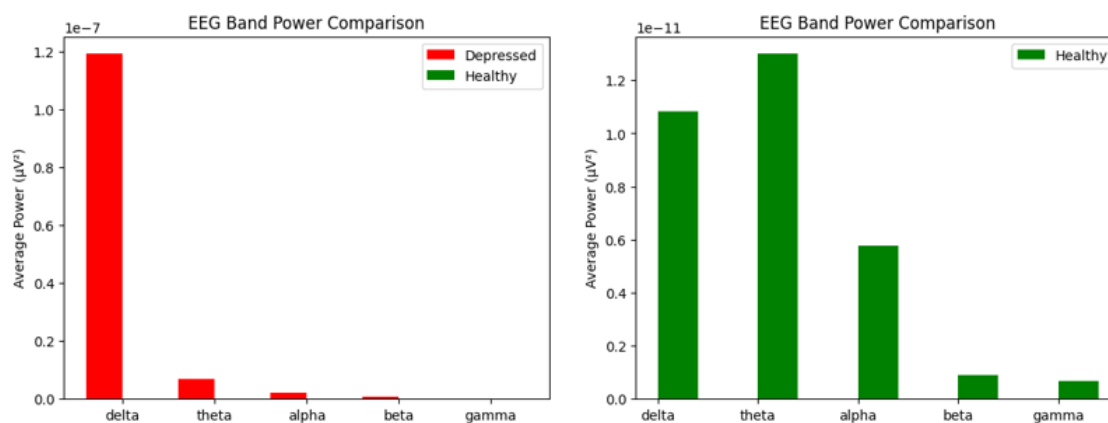


Figure 5. PSD comparison between depressed and healthy groups

3.3. Classification Experiment Setup

In this study, power spectral density (PSD) features were used as the primary input for classification experiments. To evaluate the effectiveness of deep learning models in distinguishing between depressed and healthy subjects, three architectures were explored: Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Transformer-based models.

3.3.1 Convolutional Neural Network (CNN) Architecture

The CNN was designed to capture the spatial-spectral patterns embedded in the PSD features. The features derived from the Welch-derived EEG Power Spectral Density (PSD) were used to train a Convolutional Neural Network (CNN) to learn discriminative spatial-spectral representations automatically. The input was the PSD features extracted from all the five EEG frequency bands (delta, theta, alpha, beta, gamma) in a two-dimensional feature representation, suitable for convolutional operations in CNN. The proposed CNN architecture (see Fig. 6) involved two convolutional blocks, which are a series of convolutional and max-pooling layers to progressively learn higher-level feature representations for the reduction of the spatial dimensions. Non-linearity was introduced in the network after each convolutional layer with the help of rectified linear unit (ReLU) activation. Dropout after all of the pooling layers was used to decrease overfitting and enhance the generalization of the model. The extracted feature maps were then flattened and fed into a fully connected dense layer along with an output neuron with sigmoid activation function for binary classification of depressed and healthy subjects. The network was trained with the Adam optimizer and binary cross-entropy loss, 16 as a batch size and 50 training epochs. The final CNN that was trained had 3809 trainable parameters, which makes it a light CNN and its complexity is low, but it still achieves competitive classification results.

Layer (type)	Output Shape	Param #
conv2d_2 (Conv2D)	(None, 5, 1, 16)	64
max_pooling2d_2 (MaxPooling2D)	(None, 2, 1, 16)	0
dropout_3 (Dropout)	(None, 2, 1, 16)	0
conv2d_3 (Conv2D)	(None, 2, 1, 32)	1,568
max_pooling2d_3 (MaxPooling2D)	(None, 1, 1, 32)	0
dropout_4 (Dropout)	(None, 1, 1, 32)	0
flatten_1 (Flatten)	(None, 32)	0
dense_2 (Dense)	(None, 64)	2,112
dropout_5 (Dropout)	(None, 64)	0
dense_3 (Dense)	(None, 1)	65

Total params: 3,809 (14.88 KB)
Trainable params: 3,809 (14.88 KB)
Non-trainable params: 0 (0.00 B)

Figure 6. Parameters and detailed architecture of the proposed CNN model

In CNN as shown in Figure 7 shows how training accuracy steadily increased, while validation accuracy plateaued at ~0.8378 after epoch 18.

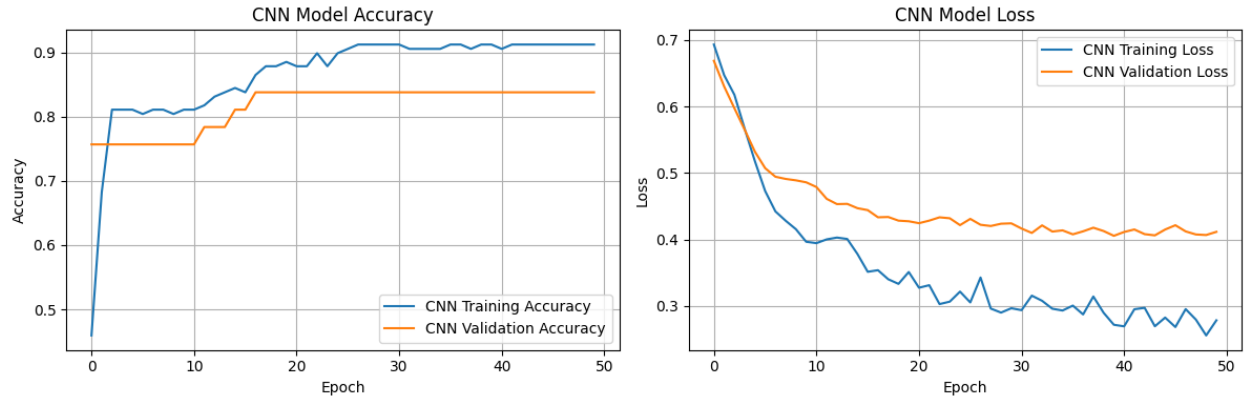


Figure 7. Accuracy and loss per epoch in CNN model

3.3.2 Recurrent Neural Network (RNN) Architecture

The proposed RNN model is a sequential deep learning architecture designed for binary classification of EEG signals. A Long Short-Term Memory (LSTM) recurrent neural network (RNN) was employed to model the temporal dependency that exists in PSD features derived from EEG signals. The sequential EEG spectral features extracted by Welch PSD were the input to the model. To represent the long-term temporal relationships between the spectral features, an LSTM layer with 64 hidden units was used. Figure 8 shows the structure of the LSTM-based RNN model that has been developed to classify the EEG depression signal.

Layer (type)	Output Shape	Param #
simple_rnn (SimpleRNN)	(None, 32)	1,216
dropout (Dropout)	(None, 32)	0
dense (Dense)	(None, 16)	528
dropout_1 (Dropout)	(None, 16)	0
dense_1 (Dense)	(None, 1)	17

Total params: 1,761 (6.88 KB)

Trainable params: 1,761 (6.88 KB)

Non-trainable params: 0 (0.00 B)

Figure 8. Parameters and detailed architecture of the proposed RNN model

A dropout layer with dropout rate of 0.3 was added to the LSTM layer to improve the generalization and reduce overfitting. The obtained temporal representations were then fed into a fully connected dense layer of 64 ReLU neurons. Another dropout layer (0.3) was added prior to the last classification layer. The output layer contained only one neuron with a sigmoid activation function, whose output was the probability of an EEG sample being depressed or healthy. The Adam optimizer, binary cross-entropy loss, batch size 16, and 50 epochs of training were used to train the model. To avoid overfitting and keep the best model, early stopping was used, which was based on validating loss.

Figure 9 shows the training and validation accuracy and loss graph for the LSTM-based RNN model through 50 epochs. It is evident from the figure that the model shows stable convergence with increasing training accuracy and reducing loss.

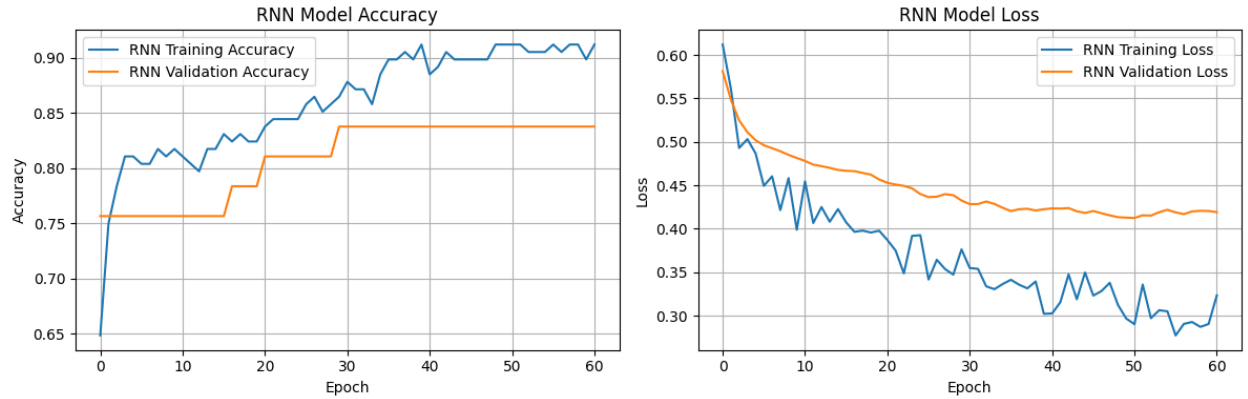


Figure 9. Accuracy and Loss per Epoch in RNN

As shown in Figure 10, Training accuracy first surpasses validation accuracy at Epoch 6 (0.8007 > 0.7568). Validation accuracy increases in steps: 0.7568 → 0.7838 → 0.8108 → 0.8378. and Training accuracy steadily rises and peaks at 0.9140 at Epoch 50.

Figure 10 below shows the comparative Accuracy and loss measure of the CNN and RNN models.

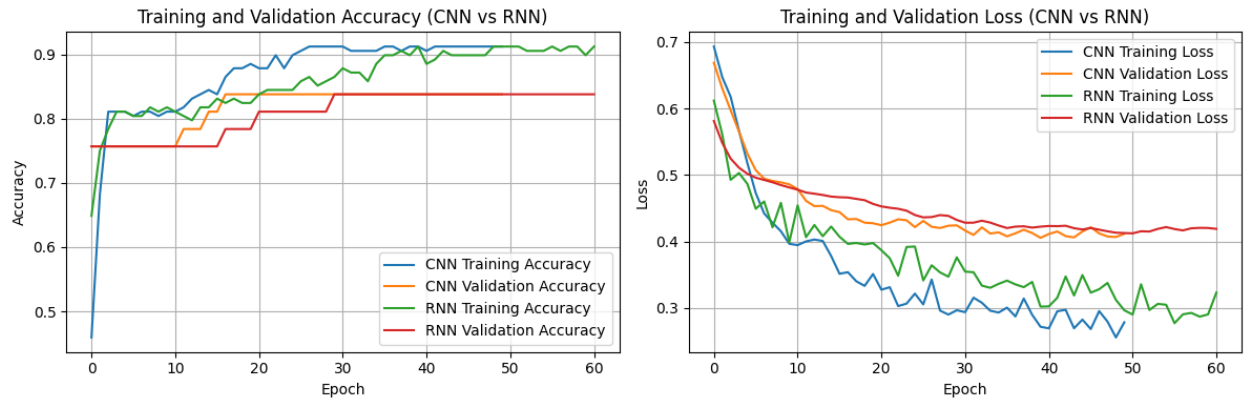


Figure 10. Accuracy and Loss per epoch in RNN vs CNN

3.3.3. Transformer Model Architecture

The proposed Transformer model (see Fig. 11) was designed to classify EEG signals on the basis of spectral features extracted by Welch. Five normalized band-power features that were extracted from each EEG epoch were used as the input to the model: the band powers for delta, theta, alpha, beta, and gamma. The five spectral features were used in an ordered fashion and embedded into a 32-dimensional space using a learnable embedding layer. Positional embeddings were added which were learnt to maintain the relative order of the spectral bands prior to feature processing.

The embedded feature sequence was processed by a Transformer encoder consisting of a single encoder block with two-head multi-head self-attention. The encoder used a position-wise feed-forward network with an embedding dimension of 32, residual connections and layer normalization. Self-attention was used to acquire local interactions between different frequency bands of EEG data and maintain global feature interactions. The output of the encoder was flattened and then fed into a fully connected layer with 20 neurons and a ReLU activation function. The dropout layer was used to prevent overfitting by dropping out 0.1 of the weight, which results in better generalization. Finally, a sigmoid activation function in the output layer resulted in the classification of depression probability.

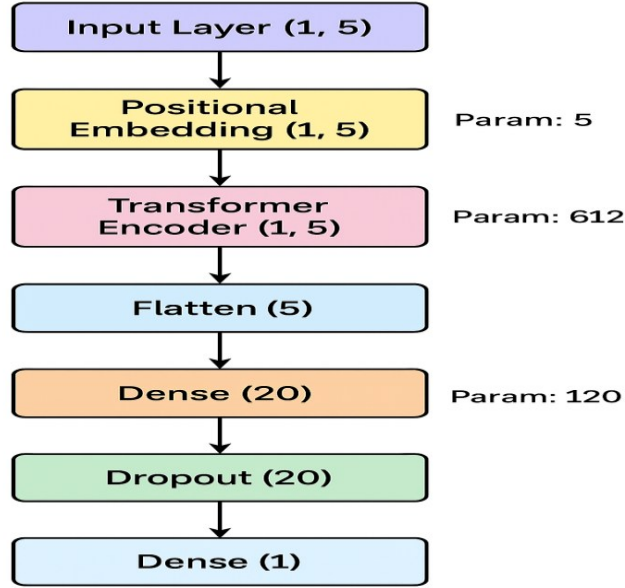


Figure 11. Architecture of the proposed Transformer model

The Transformer processing pipeline is made up of four key components: (i) embedding of the features in the spectrum, (ii) positional encoding, (iii) multi-head self-attention and feed-forward learning, and (iv) a dense classifier, with a sigmoid output layer.

Self-attention is used for Transformer architecture to better learn the global dependency among the EEG spectral features compared to the convolutional or recurrent neural networks that are mainly used for learning local spatial or sequential dependency. This is especially ideal for modelling interactions between multiple bands of the EEG signal linked to depression.

The hyperparameters of the Transformer model as shown in Table 1 were chosen empirically to find a balance between accuracy and computational complexity. Only a single encoder layer with two attention heads was enough to model the dependency between the five EEG spectral bands with a light architecture (758 trainable parameters). The number of embedding dimensions was set to 32 while 0.1 dropout rate was sufficient to prevent overfitting. With the learning rate set to 0.001, the Adam optimizer converged to a stable solution during training.

Table 1. Transformer Hyperparameter Configuration

Hyperparameter	Value
Input Features	Five EEG band powers (δ , θ , α , β , γ)
Sequence Length	5
Embedding Dimension	32
Transformer Encoder Layers	1
Attention Heads	2
Feed-Forward Dimension	32
Activation	ReLU
Dropout	0.1
Output Activation	Sigmoid
Optimizer	Adam
Learning Rate	0.001
Batch Size	16

Epochs	50
Loss Function	Binary Cross-Entropy

The proposed Transformer architecture only has 758 trainable parameters, which is far less than the number of parameters used in the traditional deep Transformer architecture. With fewer parameters, the proposed approach can be trained and tested more efficiently, with competitive accuracy in depression classification, which is suitable for real-time and resource-constrained applications. Figure 12 shows the detailed specifications of the model used.

Layer (type)	Output Shape	Param #
input_layer_2 (InputLayer)	(None, 1, 5)	0
positional_embedding_1 (PositionalEmbedding)	(None, 1, 5)	5
transformer_encoder_1 (TransformerEncoder)	(None, 1, 5)	612
flatten (Flatten)	(None, 5)	0
dense_6 (Dense)	(None, 20)	120
dropout_7 (Dropout)	(None, 20)	0
dense_7 (Dense)	(None, 1)	21

Total params: 758 (2.96 KB)
Trainable params: 758 (2.96 KB)
Non-trainable params: 0 (0.00 B)

Figure 12. Parameters and detailed architecture of the proposed Transformer model

Figure 13 shows the training and validation accuracy and loss curves of the proposed Transformer model for 50 epochs. The model exhibits stability with consistent improvement in training performance and good validation accuracy.

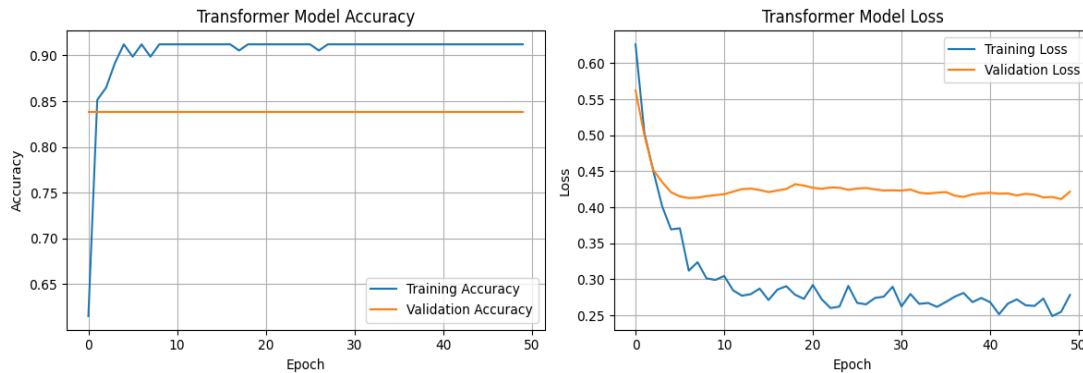


Figure 13. Accuracy and Loss of Transformer model

Finally, a Transformer architecture as shown in Figure 12 was employed to exploit attention mechanisms for capturing global dependencies across EEG channels. The PSD features were structured as input sequences, where each channel-frequency component was treated as a token embedding.

A lightweight encoder architecture was used in the Transformer model to learn the contextual dependency of the Welch derived EEG spectral features. The five band-powered features were normalized and then fed into a 32-dimensional feature representation and passed through one Transformer encoder block, two attention heads, and a feed-forward dimension of 32. Regularization with a dropout rate = 0.1 was used. The encoded representation was fed into a fully connected layer of

20 ReLU activated neurons followed by a single sigmoid output neuron for a binary classification. The model has 758 trainable parameters and was trained for 50 epochs with a learning rate of 0.001, a batch size of 16, and the loss function was binary cross-entropy. This model was optimized using the Adam optimizer with binary cross-entropy loss and trained under the same experimental settings (50 epochs, batch size 16). The Transformer's self-attention mechanism enabled the network to learn global interactions across EEG channels more effectively than CNN or RNN.

The experimental analysis, as done in Figure 13 revealed several important observations regarding the performance of the Transformer Encoder model:

- The Transformer Encoder model achieved a test accuracy of approximately 83.78%, demonstrating its effectiveness in capturing discriminative features from EEG signals for depression classification.
- During the training phase, the validation accuracy stabilized at around 83.78% from the very first epoch, whereas the training accuracy continued to improve across subsequent epochs. This indicates that the model was able to generalize early, while still optimizing its fit to the training data as shown in Figure 13.
- The training loss exhibited a consistent decreasing trend across epochs, signifying effective learning. In contrast, the validation loss remained relatively stable, which suggests that the model maintained a balanced generalization capability without significant overfitting as shown in Figure 13.
- Thus, the plot comparing the training and validation accuracy and loss for the CNN, RNN, and Transformer models has been generated in Figure 14.

These findings highlight the suitability of the Transformer-based approach for EEG-based depression detection, while also indicating the potential scope for further optimization to improve generalization performance.

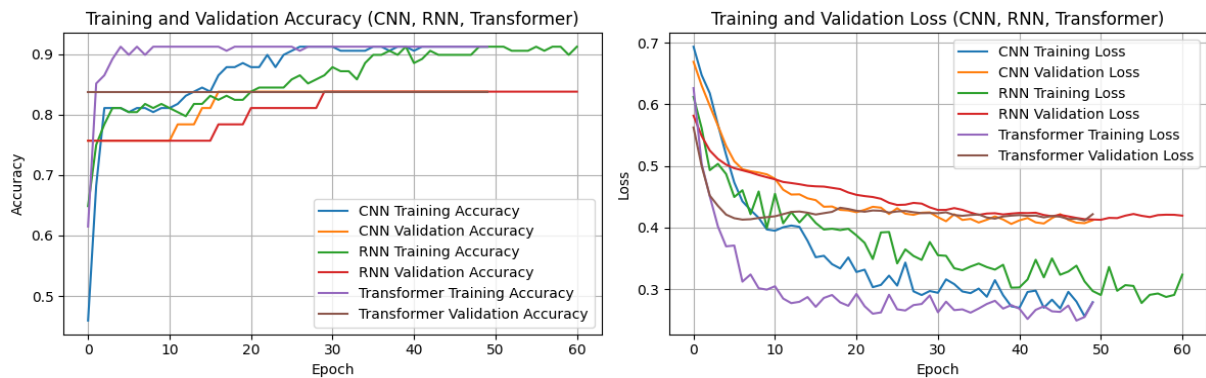


Figure 14. All three models' Accuracy and Loss in each epoch

Three deep learning models were evaluated on the extracted EEG spectral features. The CNN model successfully captured spatial-spectral EEG patterns and achieved a peak validation accuracy of 83.78% after 50 epochs. While it demonstrated stable convergence, its ability to model temporal dependencies was limited. The RNN model improved temporal learning through sequential PSD modeling and achieved comparable validation accuracy. However, the training-validation gap indicated mild overfitting, especially in later epochs. The Transformer encoder achieved the best overall performance (83.78%) with faster convergence and stronger generalization. Its self-attention mechanism effectively captured inter-band and inter-channel dependencies that were less efficiently modeled by CNN and RNN architectures. The comparative results confirm that Transformer-based learning is better suited for EEG depression detection tasks where contextual relationships between spectral bands are critical.

Convergence was achieved by the CNN with stable validation accuracy from around 18 epochs, indicating the successful learning of the spatial patterns of discriminants without overfitting. The LSTM showed continuous improvement in training accuracy owing to the recurrence memory which helped

the model in learning the sequential pattern of the EEG spectrum. Validation loss being stable shows satisfactory generalization ability. Rapid convergence of the Transformer can be attributed to the multi-head self-attention that captured the global relationships between the EEG spectral bands effectively. Small training and validation accuracy gap shows good generalization. The superior performance of the Transformer can be attributed to its self-attention mechanism, which efficiently models long-range interdependencies among EEG spectral bands. Unlike CNNs, which mainly capture local spatial patterns, and RNNs, which may suffer from vanishing gradient issues in longer dependencies, Transformers preserve global contextual relationships more effectively. From a clinical perspective, the proposed framework offers a scalable pathway for objective depression screening, especially in telemedicine and portable EEG settings.

There are some limitations in this study. The evaluation was performed on a relatively small public EEG dataset that potentially could restrict the generalizability of the reported performance. In addition, the Welch method was used only for deriving spectral band-power features. The proposed framework will be tested on larger, multi-center datasets in the future and will be extended with more temporal and connectivity-based EEG features for further enhancements in classification performance.

3.4 Engineering Significance

The proposed framework is practical in engineering considerations. The use of Welch PSD provides a better estimate of the spectrum with low variance and yet remains relatively simple to compute. The Transformer is a lightweight architecture that has relatively few trainable parameters, which is important for computational efficiency, as compared with other significant deep learning architectures. For this reason, the proposed framework is suitable to be implemented in portable EEG equipment as well as in embedded clinical decision support systems with limited computational resources.

Table 2 shows a comparison of the engineering properties of the CNN, RNN based on LSTM, and our proposed Transformer model. While CNN was able to model local spatial features and LSTM learned the time series relationships, our Transformer learns global relationships between the EEG spectral bands using the self-attention mechanism. Moreover, having only 758 trainable parameters makes it lightweight.

Table 2. Engineering Comparison of Deep Learning Models

Criterion	CNN	LSTM-RNN	Proposed Transformer
Local Feature Learning	✓	Moderate	✓
Temporal Dependency Learning	✗	✓	✓
Global Context Modeling	✗	Limited	✓
Self-Attention	✗	✗	✓
Computational Efficiency	Moderate	Moderate	High
Real-Time Deployment	Moderate	Moderate	High

4. Conclusion

This study demonstrates that EEG spectral band power features derived through the Welch method can serve as reliable biomarkers for automated depression detection. Among the evaluated models, the Transformer encoder provided the most balanced performance by effectively learning both contextual and sequential dependencies in EEG PSD representations. The proposed framework offers strong potential for integration into portable EEG systems and real-time mental health monitoring applications. CNN captured spatial features but lagged in temporal modeling, RNN better handled sequential patterns but overfit, while the Transformer Encoder achieved the best performance ($\approx 83.78\%$ accuracy) with strong generalization and minimal overfitting, making it the most robust model for EEG-based depression detection. The framework introduces a powerful Welch-based spectral analysis while keeping a light-weight Transformer architecture to offer an effective and repeatable depression detection system from EEG signals. In addition to competitive classification accuracy, the proposed approach is

computationally efficient and scalable, enabling its future potential integration in portable EEG devices and clinical decision-support systems.

Overall, the comparative analysis suggests that while CNNs and RNNs capture spatial and temporal aspects of EEG signals, respectively, the Transformer Encoder offers a more balanced and effective representation, combining both sequential and contextual dependencies. This positions the Transformer as a promising architecture for EEG-based depression detection tasks. The proposed method can support early depression screening in hospitals, educational institutions, and workplace wellness systems. Its integration with portable EEG headsets may enable real-time, non-invasive, and affordable mental health monitoring. Future research should investigate multimodal fusion frameworks integrating EEG with ECG, HRV, and functional near-infrared spectroscopy signals. Explainable AI techniques such as SHAP and attention visualization can improve clinician trust.

Declaration of AI and AI assisted technologies in the writing process

During the preparation of this work the authors used Grammarly, Quillbot and ChatGPT to check grammar, spelling and summarize referenced papers. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] Ćukić M, López V, Pavón J. Classification of Depression Through Resting-State Electroencephalogram as a Novel Practice in Psychiatry: Review. *J Med Internet Res* 2020;22:e19548. <https://doi.org/10.2196/19548>.
- [2] Kennis M, Gerritsen L, Van Dalen M, et al. Prospective biomarkers of major depressive disorder: a systematic review and meta-analysis. *Mol Psychiatry* 2020;25:321–38. <https://doi.org/10.1038/s41380-019-0585-z>.
- [3] Thibodeau R, Jorgensen RS, Kim S. Depression, anxiety, and resting frontal EEG asymmetry: A meta-analytic review. *Journal of Abnormal Psychology* 2006;115:715–29. <https://doi.org/10.1037/0021-843X.115.4.715>.
- [4] Nosek BA, Hawkins CB, Frazier RS. Implicit social cognition: from measures to mechanisms. *Trends in Cognitive Sciences* 2011;15:152–9. <https://doi.org/10.1016/j.tics.2011.01.005>.
- [5] Wu C-T, Dillon DG, Hsu H-C, et al. Depression Detection Using Relative EEG Power Induced by Emotionally Positive Images and a Conformal Kernel Support Vector Machine. *Applied Sciences* 2018;8:1244. <https://doi.org/10.3390/app8081244>.
- [6] Acharya UR, Oh SL, Hagiwara Y, et al. Automated EEG-based screening of depression using deep convolutional neural network. *Computer Methods and Programs in Biomedicine* 2018;161:103–13. <https://doi.org/10.1016/j.cmpb.2018.04.012>.
- [7] Liu W, Jia K, Wang Z, et al. A Depression Prediction Algorithm Based on Spatiotemporal Feature of EEG Signal. *Brain Sciences* 2022;12:630. <https://doi.org/10.3390/brainsci12050630>.
- [8] Welch P. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans Audio Electroacoust* 1967;15:70–3. <https://doi.org/10.1109/TAU.1967.1161901>.
- [9] Pfister H. Discrete-time signal processing. Lecture Note, Pfister Ee Duke Edu/Courses/Ece485/Dtsp Pdf 2017;4:2. <https://pfister.ee.duke.edu/courses/ece485/dtsp.pdf>. Accessed 5 April 2026.
- [10] Stoica P, Moses RL. Spectral analysis of signals. vol. 452. Pearson Prentice Hall Upper Saddle River, NJ; 2005. <https://doi.org/10.1016/B978-013113956-5.X5000-7>.
- [11] Wang Y, Liu F, Yang L. EEG-Based Depression Recognition Using Intrinsic Time-scale Decomposition and Temporal Convolution Network. *The Fifth International Conference on*

- Biological Information and Biomedical Engineering, Hangzhou China: ACM; 2021, p. 1–6. <https://doi.org/10.1145/3469678.3469683>.
- [12] Liu W, Jia K, Wang Z. Graph-based EEG approach for depression prediction: integrating time-frequency complexity and spatial topology. *Front Neurosci* 2024;18:1367212. <https://doi.org/10.3389/fnins.2024.1367212>.
 - [13] Wang B, Kang Y, Huo D, et al. EEG diagnosis of depression based on multi-channel data fusion and clipping augmentation and convolutional neural network. *Front Physiol* 2022;13:1029298. <https://doi.org/10.3389/fphys.2022.1029298>.
 - [14] Liu B, Chang H, Peng K, et al. An End-to-End Depression Recognition Method Based on EEGNet. *Front Psychiatry* 2022;13:864393. <https://doi.org/10.3389/fpsyt.2022.864393>.
 - [15] Li X, La R, Wang Y, et al. A Deep Learning Approach for Mild Depression Recognition Based on Functional Connectivity Using Electroencephalography. *Front Neurosci* 2020;14:192. <https://doi.org/10.3389/fnins.2020.00192>.
 - [16] Umair M, Ahmad J, Alasbali N, et al. Decentralized EEG-based detection of major depressive disorder via transformer architectures and split learning. *Front Comput Neurosci* 2025;19:1569828. <https://doi.org/10.3389/fncom.2025.1569828>.
 - [17] Mouazen B, Benali A, Chebchoub NT, et al. Enhancing EEG-Based Emotion Detection with Hybrid Models: Insights from DEAP Dataset Applications. *Sensors* 2025;25:1827. <https://doi.org/10.3390/s25061827>.
 - [18] Tan Q, Miao M. Detection of Depression in EEG Signals Based on Convolutional Transformer and Adaptive Transfer Learning. In: Shi Z, Witbrock M, Tian Q, editors. *Intelligence Science V*, vol. 720, Cham: Springer Nature Switzerland; 2025, p. 30–9. https://doi.org/10.1007/978-3-031-71253-1_3.
 - [19] Lawhern VJ, Solon AJ, Waytowich NR, et al. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering* 2018;15:056013. <https://doi.org/10.1088/1741-2552/aace8c>.
 - [20] Schirrmester RT, Springenberg JT, Fiederer LDJ, et al. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping* 2017;38:5391–420. <https://doi.org/10.1002/hbm.23730>.
 - [21] Song X, Yan D, Zhao L, et al. LSDD-EEGNet: An efficient end-to-end framework for EEG-based depression detection. *Biomedical Signal Processing and Control* 2022;75:103612. <https://doi.org/10.1016/j.bspc.2022.103612>.
 - [22] Qayyum A, Razzak I, Tanveer M, et al. High-Density Electroencephalography and Speech Signal Based Deep Framework for Clinical Depression Diagnosis. *IEEE/ACM Trans Comput Biol and Bioinf* 2023;20:2587–97. <https://doi.org/10.1109/TCBB.2023.3257175>.
 - [23] Tasci G, Loh HW, Barua PD, et al. Automated accurate detection of depression using twin Pascal’s triangles lattice pattern with EEG Signals. *Knowledge-Based Systems* 2023;260:110190. <https://doi.org/10.1016/j.knosys.2022.110190>.
 - [24] Dehghani M, Asayesh V, Nikjeh MT, et al. Investigation Capability of EEG-Based Non-Linear Features in Depression Detection. *Fbt* 2025. <https://doi.org/10.18502/fbt.v12i4.19809>.
 - [25] Liu X, Qu S, Luo G, et al. A generalized depression recognition framework based on cross-center and cross-task EEG signals. *Biomedical Signal Processing and Control* 2025;104:107492. <https://doi.org/10.1016/j.bspc.2025.107492>.
 - [26] Lau ZJ, Pham T, Chen SHA, et al. Brain entropy, fractal dimensions and predictability: A review of complexity measures for EEG in healthy and neuropsychiatric populations. *Eur J of Neuroscience* 2022;56:5047–69. <https://doi.org/10.1111/ejn.15800>.
 - [27] Cukic M, Pokrajac D, Stokic M, et al. EEG machine learning with Higuchi fractal dimension and Sample Entropy as features for successful detection of depression 2018. <https://doi.org/10.48550/ARXIV.1803.05985>.
 - [28] Avots E, Jermakovs K, Bachmann M, et al. Ensemble approach for detection of depression using EEG features 2021. <https://doi.org/10.48550/ARXIV.2103.08467>.

- [29] Radenković MČ, Lopez VL. Machine Learning Approaches for Detecting the Depression from Resting-State Electroencephalogram (EEG): A Review Study 2019. <https://doi.org/10.48550/ARXIV.1909.03115>.
- [30] Thoduparambil PP, Dominic A, Varghese SM. EEG-based deep learning model for the automatic detection of clinical depression. *Phys Eng Sci Med* 2020;43:1349–60. <https://doi.org/10.1007/s13246-020-00938-4>.
- [31] Shen A, Sun J, Chen X, et al. A data-centric and interpretable EEG framework for depression severity grading using SHAP-based insights. *J NeuroEngineering Rehabil* 2025;22:116. <https://doi.org/10.1186/s12984-025-01645-5>.
- [32] Kwok AMH, Cheong J, Kalkan S, et al. Machine Learning Fairness for Depression Detection Using EEG Data. 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI), Houston, TX, USA: IEEE; 2025, p. 1–5. <https://doi.org/10.1109/ISBI60581.2025.10980986>.
- [33] Bagherzadeh S, Norouzi MR, Ghasri A, et al. Automated depression detection via cloud based EEG analysis with transfer learning and synchrosqueezed wavelet transform. *Sci Rep* 2025;15:18008. <https://doi.org/10.1038/s41598-025-02452-7>.
- [34] Krisna JIT, Luthfiarta A, Cahya LD, et al. Comparing Optimizer Strategies For Enhancing Emotion Classification In IndoBERT Models. *ASSET* 2024;6:0240203. <https://doi.org/10.26877/asset.v6i2.18228>.