The Journal of Neurobehavioral Sciences (JNBS)

https://dergipark.org.tr/en/pub/jnbs

ISSN: 2149-1909

e-ISSN: 2148-4325

Research Article

Compact Deep Learning for Major Depressive Disorder Classification from Resting-State EEG: Independent Cohort Validation and Channel-Level Explainability

Caglar Uyulan[*1]

[1]*Faculty of Engineering and Architecture, Department of Mechanical Engineering, İzmir Katip Çelebi University, İzmir, Türkiye, caglar.uyulan@ikcu.edu.tr*

***Correspondence:** Caglar Uyulan, Faculty of Engineering and Architecture, Department of Mechanical Engineering, İzmir Katip Çelebi University,İzmir, Türkiye; caglar.uyulan@ikcu.edu.tr*

How to cite this article:	Uyulan C. Compact Deep Learning for Major Depressive Disorder Classification from Resting-State EEG: Independent Cohort Validation and Channel-Level Explainability. J Neurobehav Sci 2026; 13 (2): (49-61).

https://doi.org/10.32739/jnbs.13.2.289

Received: 14.08.2026
Accepted: 22.08.2026
Published: 31.08.2026

Abstract

Aims: Major depressive disorder (MDD) is a prevalent psychiatric disorder, and objective neurophysiological measures may complement clinical assessment. This study assessed whether short resting-state EEG segments can distinguish MDD from healthy controls (HC) using two compact deep learning architectures and whether channel ablation provides interpretable spatial information. **Materials and Methods:** The retrospective dataset included 112 eyes-closed resting-state EEG recordings from 56 individuals with MDD and 56 HC. Ninety-two recordings were used for model development and internal evaluation, while 20 independent recordings (10 MDD, 10 HC) were reserved for external validation. Nineteen channels were standardised to 125 Hz and segmented into non-overlapping 1-s windows. Lightweight 1D Convolutional Neural Network (Light 1D-CNN) and CNN-MiniTransformer models were trained using identical inputs. Channel ablation quantified global and class-specific spatial importance. **Results:** Internally, Light 1D-CNN achieved 96.43% accuracy, 96.39% F1-score, and 0.9941 AUROC; CNN-MiniTransformer achieved 95.26%, 95.31%, and 0.9906, respectively. On external validation, Light 1D-CNN achieved 93.72% accuracy and 0.9785 AUROC, while CNN-MiniTransformer achieved 94.65% and 0.9872. P4 was the most influential MDD-related channel in both models, followed by Fz and F4, with parietal and frontal regions showing highest importance. Attention-based modelling showed greater external stability while preserving efficient temporal feature extraction. **Conclusion:** Compact deep learning models maintained strong segment-level discrimination on independent EEG recordings and converged on similar parietal-frontal importance patterns, supporting further participant-level, multicentre validation of interpretable EEG-based MDD classification.

Keywords: *Major depressive disorder; EEG; deep learning; explainable artificial intelligence; external validation.*

Ethics committee approval: There is no need for ethics committee approval.

Introduction

Major depressive disorder (MDD) remains a clinically important psychiatric condition for which diagnosis is based primarily on clinical assessment rather than a single objective biological test. Computational analysis of electroencephalography (EEG) has therefore been investigated as a complementary source of neurophysiological information for distinguishing MDD from healthy-control patterns. Early machine-learning work demonstrated that EEG-derived features can support MDD detection and emotion-related decoding.[1] More recent deep-learning studies have extended this direction using multi-task, end-to-end, and robust convolutional frameworks that learn discriminative representations directly from EEG data.[2-4]

EEG is attractive for this purpose because it records brain electrical activity non-invasively with high temporal resolution and comparatively modest acquisition cost.[5] Its clinical use also provides a practical basis for exploring quantitative EEG markers in neuropsychiatric assessment, although EEG findings should be interpreted as decision-support information rather than a stand-alone diagnostic substitute.[6] Recent diffusion-based models and systematic reviews show that MDD-related EEG classification is an active field, while also emphasising substantial methodological heterogeneity across preprocessing, feature representation, dataset composition, and validation design.[7-9]

A central challenge is that multichannel EEG contains interacting temporal, spatial, spectral, and connectivity information. Deep neural networks can reduce dependence on manually engineered features by learning hierarchical representations, but reported performance varies with how these dimensions are represented and how data are partitioned. Effective-connectivity approaches, spectrogram-based CNNs, and machine-learning pipelines using multiple EEG features have all produced promising MDD-HC discrimination, illustrating that useful information can be represented in several complementary forms.[10-12] These differences also complicate direct comparison across studies because high accuracy can reflect not only model quality but also epoch definition, sample size, preprocessing, and the degree of subject independence in evaluation.

Compact architectures are particularly relevant when the aim is efficient analysis of short EEG windows or eventual implementation in resource-constrained settings. CNN-Transformer models have shown that convolutional feature extraction and self-attention can be combined for depression detection with reduced channel configurations, while lightweight CNN-Transformer and temporal-spatial-frequency models seek to improve efficiency or generalisation.[13-15] Recent work with a compact 1D-CNN has likewise highlighted the importance of subject-wise evaluation when estimating performance on previously unseen participants.[16] These findings motivate direct comparison of a lightweight convolutional baseline with a small attention-based architecture under an identical input representation and training framework.

Interpretability is another requirement for clinically oriented EEG classification. A high discrimination score alone does not reveal which channels or spatial regions a trained network depends on, and model-specific explanations should not be equated with established physiological biomarkers. Recent explainable and quantitative EEG studies have reported contributions from frontal, temporal, occipital, parietal, motor, spectral, and connectivity-related patterns in depression, supporting a distributed rather than single-channel view of MDD-related information.[17-22] Comparing channel-importance profiles across two architecturally distinct models may therefore help identify whether learned spatial dependencies are model-specific or reproducible across classifiers.

Against this background, the present study evaluated two compact deep-learning models - a Lightweight 1D Convolutional Neural Network (Light 1D-CNN) and a CNN-MiniTransformer - using 19-channel, eyes-closed resting-state EEG segmented into non-overlapping 1-s windows. The study had three objectives: (1) to compare the two architectures under the same preprocessing and training framework; (2) to assess whether their performance is retained on a completely held-out independent cohort; and (3) to use channel-ablation analysis to compare global, MDD-specific, HC-specific, and regional EEG importance patterns. The analysis was designed to pair predictive evaluation with a transparent assessment of channel-level model dependence while explicitly distinguishing segment-level internal performance from generalisation to held-out recordings.

Materials and Methods

Study Population and EEG Data Acquisition

There is no need for ethics committee approval. This retrospective study included 112 resting-state EEG recordings obtained from the clinical EEG archive of NPİSTANBUL Brain Hospital, Istanbul, Türkiye, comprising 56 individuals with Major Depressive Disorder (MDD) and 56 healthy controls (HC). The development cohort included 92 recordings (46 MDD and 46 HC) and was used for model development and internal evaluation. A separate cohort of 20 recordings (10 MDD and 10 HC) was reserved for independent external validation and was not used during model training, hyperparameter optimisation, or internal evaluation.

The MDD group included patients with a clinical diagnosis of MDD established by the treating psychiatrist according to DSM-5 criteria. Healthy controls had no documented psychiatric or neurological disorder. Eligible records required a documented MDD diagnosis or HC status, an adequate resting-state EEG recording, and complete coverage of the required EEG channels. Recordings with incomplete channel coverage, insufficient duration, or substantial technical artefacts were excluded.

EEG recordings had originally been acquired as part of routine clinical assessment, predominantly using the same EEG system and standardised acquisition protocol. Recordings were obtained in a quiet, dimly lit room under controlled environmental conditions (approximately 21–26°C and 30–60% relative humidity). Participants were seated comfortably in an approximately 120° reclined chair and instructed to remain relaxed, minimise movement and excessive blinking, and remain awake.

Resting-state EEG was recorded for approximately 3 minutes under eyes-closed conditions. Nineteen scalp channels (Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, Fz, Cz, and Pz) were acquired using active electrodes mounted on an EEG cap (Wuhan Greentek Pty. Ltd., Wuhan, China) according to the International 10-20 system. Electrode impedance was maintained below 5 kΩ, and the ground electrode was positioned midway between Fpz and Fz.

EEG Preprocessing and Segmentation

EEG recordings were imported from EDF files using MNE-Python. Each recording was checked for the 19 required EEG channels, and recordings with incomplete channel coverage were excluded. The retained channels were reordered according to the predefined channel configuration to ensure consistent spatial organization across all model inputs.

The sampling frequency was standardised to 125 Hz. Recordings already sampled at 125 Hz were retained unchanged, while recordings acquired at other sampling frequencies were resampled to 125 Hz. No additional digital filtering was applied within the deep learning pipeline because the EDF files used in the analysis had already been treated as preprocessed clinical EEG recordings.

Each continuous EEG recording was divided into non-overlapping 1-s segments. At 125 Hz, each segment contained 125 temporal samples across 19 channels, resulting in an input representation of 125 × 19. Only complete segments were retained; residual samples shorter than 1 s were discarded, and no zero-padding was applied.

Following segmentation, signal-quality control was performed by calculating the standard deviation of each segment across its temporal and channel dimensions. Flat or near-zero segments with a standard deviation $\leq 1 \times 10^{-8}$ were excluded. The remaining 1-s segments were used as the input units for both deep learning models. The binary classification task was defined as HC (class 0) versus MDD (class 1), and model performance was evaluated at the segment level.

Dataset Preparation, Splitting, and Normalisation

Following preprocessing and quality control, valid HC and MDD segments were balanced by random undersampling of the larger class to match the number of segments in the smaller class. A fixed random seed of 42 was used for reproducibility.

The balanced dataset was randomly shuffled and divided using stratified segment-level splitting. First, 10% of the segments were reserved as the internal test set. The remaining 90% were divided into training and validation sets, with 15% assigned to validation, resulting in an approximate overall allocation of 76.5% training, 13.5% validation, and 10% internal testing. Stratification preserved the HC–MDD class distribution across the subsets.

EEG amplitudes were standardised channel-wise using a StandardScaler fitted exclusively on the training data. The training tensor was reshaped so that all temporal samples contributed to the estimation of the mean and standard deviation for each of the 19 channels. The fitted scaler was then applied without refitting to the validation and internal test sets.

The same preprocessing pipeline and training-derived scaler were applied to the independent external validation recordings. The scaler was used without refitting, and no external-validation data contributed to model training, hyperparameter optimisation, model selection, or normalisation-parameter estimation.

Deep Learning Models

Two deep learning architectures were developed and evaluated using the same standardised 1-s EEG segments. The first was a computationally efficient one-dimensional convolutional neural network (Light 1D-CNN), designed as a compact baseline for temporal EEG feature extraction. The second was a hybrid CNN-MiniTransformer architecture that combined convolutional feature extraction with self-attention to capture both local and longer-range temporal dependencies. Both models received identical input tensors of 125 temporal samples × 19 EEG channels and produced a single sigmoid probability corresponding to the MDD class.

Lightweight 1D-CNN Architecture

The first architecture was designed as a computationally efficient one-dimensional convolutional neural network (Light 1D-CNN) capable of learning local temporal patterns from short EEG segments without the computational complexity associated with recurrent or attention-based architectures. The network received a standardised EEG segment of 125 temporal samples × 19 channels as input.

The convolutional feature extractor consisted of three successive one-dimensional convolutional layers. The first layer applied 32 filters with a kernel size of 7 and the same padding, followed by a rectified linear unit (ReLU) activation and max pooling with a pool size of 2. The second convolutional layer increased the representation to 64 filters with a kernel size of 5, again followed by ReLU activation and max pooling. A third convolutional layer used 96 filters with a kernel size of 3 and ReLU activation, allowing progressively higher-level temporal features to be extracted from the EEG signal.

Following convolutional feature extraction, Global Average Pooling was applied across the temporal dimension to transform the resulting feature maps into a fixed-length representation. This representation was passed through a fully connected layer containing 64 ReLU-activated units, followed by dropout with a rate of 0.20 to reduce overfitting. The final classification layer consisted of a single sigmoid unit that generated the estimated probability of the MDD class.

CNN-MiniTransformer Architecture

The second architecture combined convolutional feature extraction with a lightweight Transformer encoder. The model was designed to retain the ability of convolutional layers to identify local temporal EEG patterns while introducing self-attention to model dependencies across the compressed temporal representation. The CNN-MiniTransformer received the same standardised 125 × 19 EEG input used by the Light 1D-CNN, enabling direct comparison between the two architectures.

The convolutional front end consisted of two sequential blocks. The first block applied a one-dimensional convolution with 32 filters, a kernel size of 7, same padding, and no convolutional bias, followed by batch normalisation, Gaussian Error Linear Unit (GELU) activation, and max pooling with a pool size of 2. The second block increased the representation to 64 filters using a kernel size of 5 and was similarly followed by batch normalisation, GELU activation, and max pooling. Following the two pooling operations, the original 125-sample temporal sequence was reduced to approximately 31 temporal tokens, with each token represented by a 64-dimensional feature vector.

Learnable positional embeddings were added to the convolutionally derived token sequence to retain information regarding temporal position before the self-attention operation. The resulting sequence was processed by a single pre-normalised Transformer encoder block. The multi-head self-attention module employed two attention heads with a model dimension of 64 and an attention dropout rate of 0.10. A residual connection was used to add the attention output to the input representation.

The subsequent feed-forward subnetwork expanded the representation to a hidden dimension of 128 using GELU activation and a dropout of 0.10 before projecting it back to 64 dimensions. A second residual connection combined the feed-forward output with its input. Layer normalisation was applied before the self-attention and feed-forward operations, and an additional normalisation stage was applied after the Transformer encoder.

The final encoded temporal sequence was summarised using Global Average Pooling. The resulting representation was passed through a fully connected layer containing 64 GELU-activated units, followed by dropout with a rate of 0.25. As in the Light 1D-CNN, the final classification layer consisted of a single sigmoid unit that generated the probability of the MDD class.

Model Training and Optimisation

Both deep learning models were trained using the Adam optimizer and binary cross-entropy as the loss function. A batch size of 128 was used for both architectures, and training was allowed to continue for a maximum of 50 epochs. Because the two architectures differed in complexity and optimisation behaviour, the Light 1D-CNN was initialized with a learning rate of 1×10^{-3}, whereas the CNN-MiniTransformer was trained with a lower initial learning rate of 5×10^{-4}.

Model selection was based on validation area under the receiver operating characteristic curve (AUROC). During training, the model checkpoint achieving the highest validation AUROC was saved and subsequently used for internal test and external validation analyses. This ensured that the internal test set was not used for checkpoint selection.

Adaptive learning-rate reduction and early stopping were incorporated to improve optimisation and limit unnecessary training. ReduceLROnPlateau monitored validation loss and reduced the learning rate by a factor of 0.5 when no improvement was observed for four consecutive epochs, with the minimum learning rate restricted to 1×10^{-6}. Early stopping also monitored validation loss, using a patience of 10 epochs and restoring the best available weights when training was terminated.

The same overall training and checkpoint-selection framework was applied to both architectures, allowing their subsequent performance to be compared under a consistent experimental protocol.

Performance Evaluation Metrics

Model performance was evaluated at the segment level using accuracy, balanced accuracy, precision, sensitivity (recall), specificity, F1-score, AUROC, and confusion matrices. MDD was defined as the positive class. AUROC was calculated from the continuous sigmoid probabilities, whereas a threshold of 0.50 was used for threshold-dependent metrics and confusion matrices.

The same metrics and classification threshold were applied to both deep learning models for internal testing and independent external validation, allowing direct comparison of their classification performance.

Independent External Validation

Independent external validation was performed using 20 completely held-out EEG recordings, comprising 10 MDD and 10 HC recordings. These data were not used during training, validation, internal testing, hyperparameter optimisation, checkpoint selection, or normalisation-parameter estimation.

The external recordings were processed using the same preprocessing pipeline described for the development cohort. The StandardScaler fitted exclusively on the training data was applied without refitting. The best saved checkpoints of the Light 1D-CNN and CNN-MiniTransformer were then applied directly to the external segments without retraining, fine-tuning, or adaptation.

External performance was evaluated at the segment level using the same metrics and classification threshold applied to the internal test set.

Explainable Artificial Intelligence Analysis

A perturbation-based channel-ablation analysis was performed to investigate the spatial EEG information contributing to the predictions of the Light 1D-CNN and CNN-MiniTransformer. The complete internal test set was used for both models, allowing direct comparison of their channel-importance profiles.

Baseline predictions were first obtained from the unmodified EEG segments. Because both models used a single sigmoid output representing the probability of MDD, true-class probability was defined as p(MDD) for MDD segments and $1 - p$(MDD) for HC segments. Each of the 19 EEG channels was then ablated individually by replacing all temporal samples of that channel with zero while leaving the remaining channels unchanged. Since the inputs had been standardised using training-derived statistics, zero approximately represented the training-set mean of each channel.

Channel importance was defined as the decrease in mean true-class probability after ablation. Larger decreases therefore indicated greater model dependence on the corresponding channel. Negative changes were truncated to zero, and the resulting positive importance values were normalised to sum to one. Importance was calculated globally using all test segments and separately for MDD and HC segments to obtain class-specific profiles.

For regional analysis, electrodes were grouped into frontal (Fp1, Fp2, F3, F4, F7, F8, Fz), central (C3, C4, Cz), parietal (P3, P4, Pz), occipital (O1, O2), and temporal (T3, T4, T5, T6) regions. Regional importance was calculated as the mean channel importance within each region to account for differences in the number of electrodes and was subsequently normalised across the five regions.

Finally, the channel- and region-level importance profiles of the two models were compared to examine whether the architectures relied on similar spatial EEG patterns.

Results

Internal Test Performance

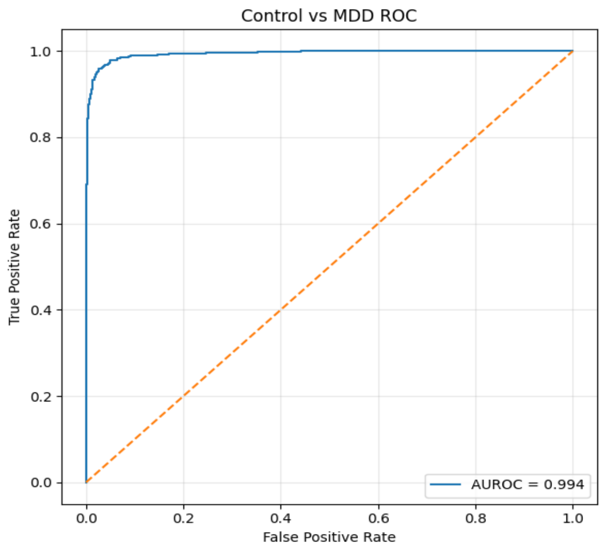
Both deep learning architectures achieved high discrimination performance on the internal segment-level test set, which consisted of 1,455 EEG segments, including 728 HC and 727 MDD segments.

The Light 1D-CNN achieved an accuracy of 96.43%, balanced accuracy of 96.43%, precision of 97.34%, sensitivity of 95.46%, specificity of 97.39%, F1-score of 96.39%, and AUROC of 0.9941. The confusion matrix showed that 709 of 728 HC segments and 694 of 727 MDD segments were correctly classified, corresponding to 19 HC segments misclassified as MDD and 33 MDD segments misclassified as HC. During model development, the selected Light 1D-CNN checkpoint was obtained at epoch 21, with a validation AUROC of 0.9914 and validation accuracy of 96.69%.

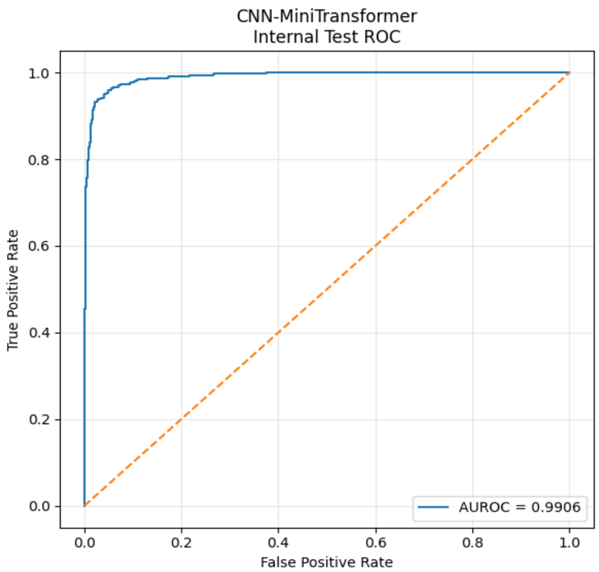
The CNN-MiniTransformer also demonstrated strong internal discrimination, achieving an accuracy of 95.26%, balanced accuracy of 95.26%, precision of 94.22%, sensitivity of 96.42%, specificity of 94.09%, F1-score of 95.31%, and AUROC of 0.9906. Of the 728 HC segments, 685 were correctly classified, and 43 were misclassified as MDD, whereas 701 of the 727

MDD segments were correctly identified and 26 were misclassified as HC. The selected CNN-MiniTransformer checkpoint was obtained at epoch 6, with a validation accuracy of 96.49% and validation AUROC of 0.9899.

Overall, both architectures achieved internal test accuracies above 95% and AUROC values above 0.99. The Light 1D-CNN yielded higher overall accuracy, balanced accuracy, precision, specificity, F1-score, and AUROC, whereas the CNN-MiniTransformer achieved higher sensitivity for detecting MDD segments. The ROC curves of both models on the internal test set are presented in Figure 1, showing strong discrimination between MDD and healthy-control EEG segments. The corresponding confusion matrices are shown in Figure 2, providing a detailed view of the classification performance and error distribution for both models.

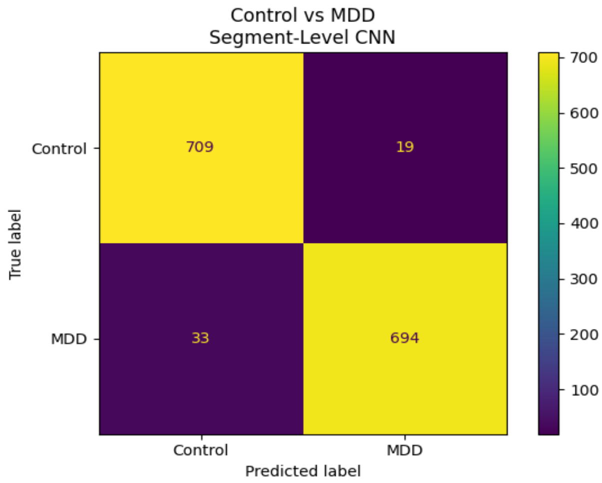


(a)

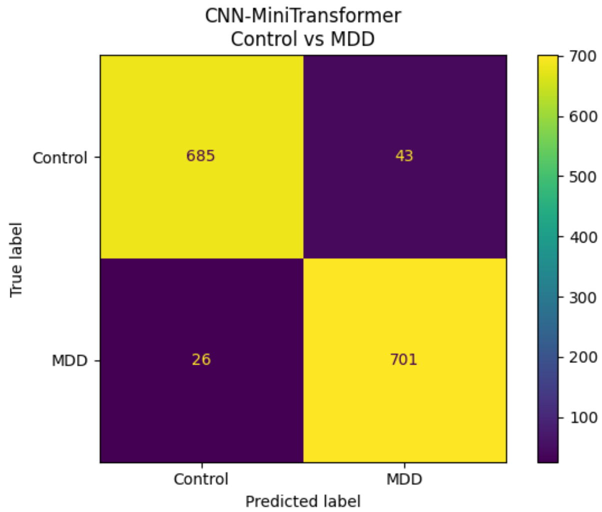


(b)

Figure 1. Internal test receiver operating characteristic (ROC) curves for the two deep learning models: (A) Light 1D-CNN and (B) CNN-MiniTransformer.



(a)



(b)

Figure 2. Internal test confusion matrices for the two deep learning models: (A) Light 1D-CNN and (B) CNN-MiniTransformer.

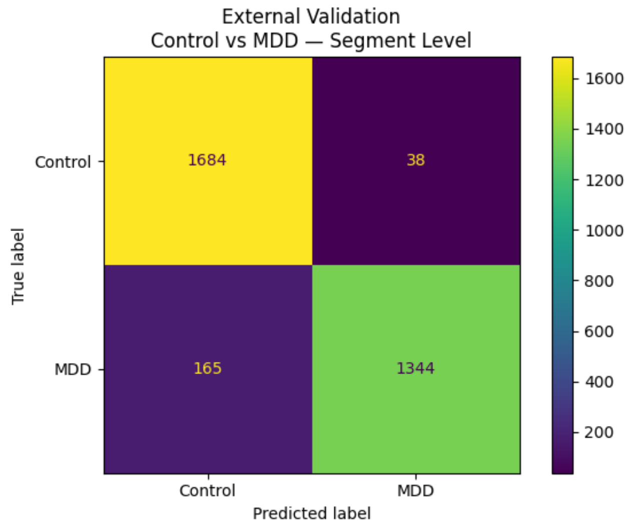
Independent External Validation Performance

Both trained models maintained high classification performance when applied without retraining or adaptation to the independent external validation cohort consisting of 10 HC and 10 MDD recordings. Segment-level external evaluation comprised 3,231 EEG segments, including 1,722 HC and 1,509 MDD segments.

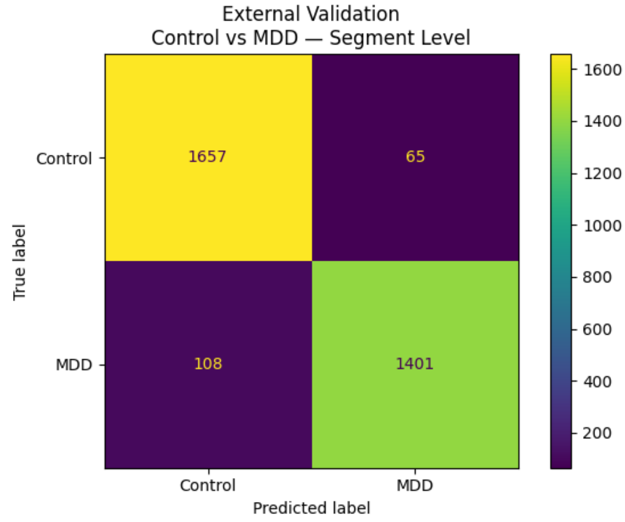
The Light 1D-CNN achieved an external validation accuracy of 93.72%, balanced accuracy of 93.43%, precision of 97.25%, sensitivity of 89.07%, specificity of 97.79%, F1-score of 92.98%, and AUROC of 0.9785. The model correctly classified 1,684 of 1,722 HC segments and 1,344 of 1,509 MDD segments. A total of 38 HC segments were classified as MDD, whereas 165 MDD segments were classified as HC.

The CNN-MiniTransformer achieved an external validation accuracy of 94.65%, balanced accuracy of 94.53%, precision of 95.57%, sensitivity of 92.84%, specificity of 96.23%, F1-score of 94.18%, and AUROC of 0.9872. The corresponding confusion matrix showed correct classification of 1,657 of 1,722 HC segments and 1,401 of 1,509 MDD segments, with 65 HC segments misclassified as MDD and 108 MDD segments misclassified as HC.

Both architectures therefore retained strong discrimination on the independent external dataset, with AUROC values remaining above 0.97. In contrast to the internal test results, the CNN-MiniTransformer achieved higher external accuracy, balanced accuracy, sensitivity, F1-score, and AUROC, whereas the Light 1D-CNN retained higher precision and specificity. The confusion matrices for the independent external validation cohort are presented in Figure 3, showing the classification patterns of the Light 1D-CNN and CNN-MiniTransformer on previously unseen EEG recordings.



(a)



(b)

Figure 3. External validation confusion matrices for the two deep learning models: (A) Light 1D-CNN and (B) CNN-MiniTransformer.

Comparative Performance of the Deep Learning Models

Table 1 summarizes the segment-level performance of both architectures on the internal test and independent external validation datasets.

Table 1. Segment-level performance on internal and external validation datasets

Metric	Light CNN Internal	MTI*	Light CNN External	MTE*
Accuracy	0.9643	0.9526	0.9372	0.9465
Balanced Accuracy	0.9643	0.9526	0.9343	0.9453
Precision (MDD)	0.9734	0.9422	0.9725	0.9557
Sensitivity (MDD)	0.9546	0.9642	0.8907	0.9284
Specificity	0.9739	0.9409	0.9779	0.9623
F1 Score	0.9639	0.9531	0.9298	0.9418
AUROC	0.9941	0.9906	0.9785	0.9872

*** MTI: MiniTransformer Internal, MTE: MiniTransformer External**

The Light 1D-CNN showed the strongest overall performance on the internal test set, with an accuracy of 96.43% and an AUROC of 0.9941. On independent external validation, however, the CNN-MiniTransformer achieved the highest overall performance, reaching 94.65% accuracy and an

AUROC of 0.9872.

From the internal test set to the external validation set, Light 1D-CNN accuracy decreased from 96.43% to 93.72%, corresponding to an absolute reduction of 2.71 percentage points, while AUROC decreased from 0.9941 to 0.9785. The CNN-MiniTransformer showed a smaller change in accuracy, from 95.26% internally to 94.65% externally, corresponding to an absolute reduction of only 0.61 percentage points, while AUROC changed from 0.9906 to 0.9872.

The two models therefore exhibited slightly different performance profiles. The Light 1D-CNN provided the highest internal discrimination and the highest external specificity, whereas the CNN-MiniTransformer showed more stable performance between the internal and external datasets and achieved higher external sensitivity, F1-score, accuracy, balanced accuracy, and AUROC.

Explainable Artificial Intelligence Results

Channel-ablation analysis was performed on all 1,455 internal test segments, comprising 728 HC and 727 MDD segments. The analysis was conducted independently for the Light 1D-CNN and CNN-MiniTransformer using the same standardised test dataset. Global, MDD-specific, and HC-specific channel importance profiles were examined, followed by regional aggregation and cross-model comparison.

Global Channel Importance

The Light 1D-CNN demonstrated a concentrated global channel-importance profile. P4 was the highest-ranked channel, with a normalised importance of 0.2087, followed by Fz (0.1734) and F4 (0.1363). P3 (0.0747) and F3 (0.0630) completed the five highest-ranked channels. Additional contributions were observed for T6 (0.0575), O2 (0.0488), Fp2 (0.0461), Pz (0.0417), and O1 (0.0255).

The CNN-MiniTransformer produced a similar global channel ranking. P4 was again the highest-ranked channel (0.1816), followed by Fz (0.1370) and F4 (0.0934). F3 (0.0774), O1 (0.0726), P3 (0.0641), F7 (0.0592), T6 (0.0502), Pz (0.0444), and O2 (0.0370) constituted the remainder of the ten highest-ranked channels.

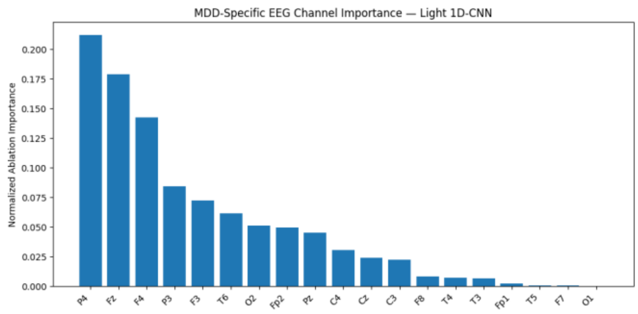
Thus, despite their architectural differences, both models independently identified P4, Fz, and F4 as the three most influential channels in the global analysis. P3, F3, T6, and Pz were also represented among the ten highest-ranked channels in both architectures.

Class-Specific Channel Importance

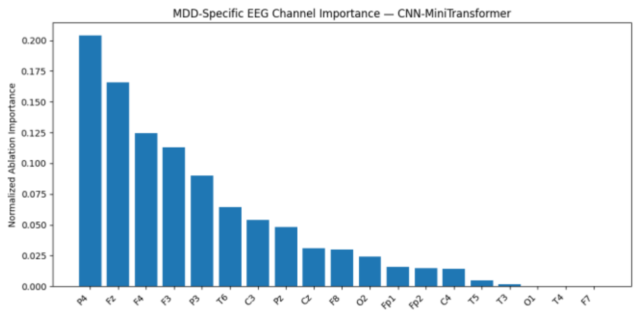
For the Light 1D-CNN, the mean correct-class probability for MDD segments before ablation was 0.9492. The MDD-specific channel ranking closely resembled the global profile. P4 showed the highest normalised importance (0.2121), followed by Fz (0.1789), F4 (0.1422), P3 (0.0843), and F3 (0.0724). The remaining top-ranked channels were T6 (0.0613), O2 (0.0510), Fp2 (0.0495), Pz (0.0454), and C4 (0.0307).

The MDD-specific analysis of the CNN-MiniTransformer yielded a mean correct-class probability of 0.9548. P4 was again the highest-ranked channel (0.2040), followed by Fz (0.1659), F4 (0.1245), F3 (0.1128), and P3 (0.0902). T6 (0.0642), C3 (0.0542), Pz (0.0481), Cz (0.0312), and F8 (0.0300) completed the top ten.

The MDD-specific results therefore showed substantial agreement between the two architectures. In both models, P4, Fz, and F4 occupied the first three positions, while P3, F3, T6, and Pz were also consistently represented among the most influential channels.



(a)



(b)

Figure 4. MDD-specific EEG channel importance obtained using channel-ablation analysis for the two deep learning models: (A) Light 1D-CNN and (B) CNN-MiniTransformer.

The HC-specific profiles differed substantially from the MDD-specific patterns. For the Light 1D-CNN, the mean correct HC probability before ablation was

0.9602. O1 showed the highest positive normalised importance (0.4994), followed by T5 (0.2851), F7 (0.1187), T4 (0.0544), Fp1 (0.0246), and F8 (0.0178).

A comparable pattern was observed for the CNN-MiniTransformer, for which the mean correct HC probability was 0.9298. O1 was again the highest-ranked channel (0.4004), followed by F7 (0.2534), T4 (0.1603), T5 (0.1110), T3 (0.0422), and O2 (0.0324). Thus, both architectures showed a distinct HC-specific importance profile characterised primarily by occipital and temporal electrodes.

Because negative ablation effects were truncated to zero before normalisation, the sparse HC-specific importance distributions represent positive model dependence under the predefined ablation procedure and should not be interpreted as absolute physiological effect sizes.

Regional EEG Importance

Regional analysis further demonstrated similarities between the two architectures. For the Light 1D-CNN, global regional importance was highest in the parietal region (0.4281), followed by frontal (0.2502), occipital (0.1467), temporal (0.0958), and central (0.0792) regions. In the MDD-specific analysis, the parietal region remained dominant (0.4576), followed by frontal (0.2605), central (0.1033), occipital (0.1024), and temporal (0.0762) regions.

The CNN-MiniTransformer exhibited a comparable regional distribution. Global importance was highest for the parietal region (0.3726), followed by frontal (0.2356), occipital (0.2112), temporal (0.1227), and central (0.0578) regions. In the MDD-specific analysis, parietal importance increased to 0.4692, followed by frontal (0.2724), central (0.1364), temporal (0.0724), and occipital (0.0496).

Accordingly, the parietal region ranked first, and the frontal region ranked second in both the global and MDD-specific analyses for both deep learning architectures.

Table 2. Regional channel-ablation importance for the Light 1D-CNN and CNN-MiniTransformer.

Region	Light CNN Global	Light CNN MDD	MTG*	MT* MDD
Parietal	0.4281	0.4576	0.3726	0.4692
Frontal	0.2502	0.2605	0.2356	0.2724
Occipital	0.1467	0.1024	0.2112	0.0496
Temporal	0.0958	0.0762	0.1227	0.0724
Central	0.0792	0.1033	0.0578	0.1364

*** MTG: Mini Transformer Global, MT: MiniTransformer**

Cross-Model Consistency of XAI Findings

A notable finding of the explainability analysis was the convergence of the two architectures on a common channel-importance pattern. P4 was the highest-ranked channel in both the global and MDD-specific analyses for both models. Furthermore, Fz and F4 occupied the second and third positions, respectively, in both MDD-specific rankings. P3, F3, T6, and Pz also appeared among the ten most influential MDD-related channels in both architectures.

The regional results showed a similarly consistent pattern. The parietal region was the highest-ranked region in the global and MDD-specific analyses of both models, while the frontal region consistently ranked second. The CNN-MiniTransformer additionally assigned relatively greater MDD-specific importance to the central region than the Light 1D-CNN.

Taken together, the channel- and region-level analyses demonstrated that two architecturally distinct classifiers converged on a common spatial importance pattern characterised by prominent parietal and frontal contributions to MDD classification. In contrast, correct HC classification showed a different positive-importance distribution, with both architectures assigning their highest HC-specific importance to O1 and additional contributions predominantly from temporal electrodes.

Discussion

The present study showed that short resting-state EEG segments contain useful information for distinguishing MDD from healthy controls. Both models achieved high classification performance, but their results differed slightly between internal and external evaluation. The Light 1D-CNN achieved the best internal performance, with an accuracy of 96.43%, an

F1-score of 96.39%, and an AUROC of 0.9941. The CNN-MiniTransformer achieved an accuracy of 95.26%, an F1-score of 95.31%, and an AUROC of 0.9906. In the independent external validation cohort, however, the CNN-MiniTransformer performed slightly better, reaching 94.65% accuracy and an AUROC of 0.9872, compared with 93.72% and 0.9785 for the Light 1D-CNN.

These findings are consistent with previous studies showing that deep learning can distinguish MDD from healthy controls using EEG. High classification performance has been reported using different EEG representations and architectures. Anik et al. reported 99.60% accuracy using an extended 1D-CNN with gamma-band EEG.[4] Saeedi et al. achieved 99.24% using effective-connectivity representations and a 1D-CNN-LSTM.[10] Loh et al. reported 99.58% using STFT-derived spectrograms and CNN classification.[11] More recent approaches have continued to explore temporal-spatial-frequency representations and compact convolutional frameworks for MDD detection.[15,16] However, direct comparison between studies should be made cautiously because EEG representation, epoch duration, sample size, preprocessing, and validation procedures differ considerably.

An interesting finding in the present study was the different behaviour of the two architectures on the independent validation data. The accuracy of the Light 1D-CNN decreased by 2.71 percentage points between internal and independent evaluation, whereas the CNN-MiniTransformer showed a smaller decrease of 0.61 percentage points. A similar pattern was observed for AUROC. This may be related to differences in how the two architectures process EEG information. CNNs mainly capture local signal patterns, while the self-attention component of the CNN-MiniTransformer can additionally model relationships across the learned temporal representation. Previous studies have also reported promising results with attention-based approaches. Xia et al. combined CNN-based learning with self-attention for EEG-based MDD classification.[3] Yu et al. reported strong performance with a CNN-Transformer using a reduced EEG channel configuration, and Liu et al. further explored a lightweight CNN-Transformer architecture.[13,14] The slightly stronger independent performance of the CNN-MiniTransformer in our study may therefore reflect more stable representations on unseen recordings, although the held-out cohort was relatively small and this interpretation requires confirmation in larger datasets.

The explainability analysis provided another important finding. Despite their different architectures, both models showed a similar MDD-specific channel-importance pattern. P4 was the most influential channel in both models, followed by Fz and F4. P3, F3, T6, and Pz were also consistently represented among the most influential channels. At the regional level, the parietal region showed the highest MDD-specific importance in both models, followed by the frontal region. Parietal importance was 0.4576 for the Light 1D-CNN and 0.4692 for the CNN-MiniTransformer, while frontal importance was 0.2605 and 0.2724, respectively. The agreement between the two different architectures suggests that these spatial patterns were not unique to a single model.

Previous explainable EEG studies have also reported that depression-related information is not distributed equally across the scalp. Naeim and Atadokht identified frontal EEG channels as important for classification using Grad-CAM and SHAP.[17] Other recent quantitative EEG studies have reported informative frontal and temporal connectivity, occipital-frontal patterns, and broader frontal, temporal, parietal, and motor network contributions.[18-20] Temporal-spatial-spectral EEG representation learning has also been explored in depression-related analysis in an emotion-EEG setting.[21] Although these studies used different EEG representations and explanation methods, together they support the broader view that depression-related EEG information may involve distributed spatial networks rather than a single universal scalp location.

The strong contribution of P4 in both models is particularly interesting, but it should not be interpreted as evidence that P4 alone represents a specific biomarker of MDD. Channel ablation measures how strongly a trained model depends on information from a given EEG channel rather than directly identifying its neurophysiological source. Our findings therefore indicate that both models relied strongly on information available through parietal and frontal channels. Spectral analysis may provide additional insight into this pattern. Kozulin et al. identified delta- and alpha-band measures as informative components in depression-related EEG analysis.[22] Combining channel ablation with frequency-band analysis in future work could help determine which oscillatory components contribute to the importance of P4, Fz, F4, and other channels.

The HC-specific analysis showed a different pattern. O1 was the most influential channel for correct HC classification in both architectures, while other important channels were mainly located in occipital and temporal regions. In contrast, MDD-specific predictions depended more strongly on parietal and frontal channels. This suggests that the models used partly different spatial information when identifying MDD and HC segments rather than relying on the same channels for both classes.

Several aspects of the study support the reliability of the comparison. Both architectures were trained and evaluated using the same preprocessing pipeline and EEG representation, the external recordings were kept separate from model development, and the same channel-ablation procedure was applied to both models. Nevertheless, the external validation cohort contained only 20 recordings, and larger and more diverse cohorts are needed to further examine generalizability. The present analysis was also based on one-second EEG segments, and future studies could examine recording-level or participant-level aggregation. In addition, channel ablation cannot determine which frequency bands or temporal patterns are responsible for the observed channel importance. Combining channel-level analysis with spectral, connectivity, SHAP, or gradient-based approaches could provide a more detailed neurophysiological interpretation. Because the internal training, validation, and test partitions were created at the segment level, segments originating from the same recording could occur in different internal subsets; the internal metrics should therefore be interpreted as segment-level estimates rather than subject-independent performance. The completely held-out 20-recording cohort provides the more informative assessment of generalisation in the present design.

Overall, both the Light 1D-CNN and CNN-MiniTransformer showed strong performance for MDD-HC classification. The Light 1D-CNN achieved the highest internal performance, while the CNN-MiniTransformer performed slightly better on the independent external data. Importantly, both architectures identified a consistent MDD-related spatial pattern centered on P4, Fz, F4, and the broader parietal and frontal regions. These findings support further investigation of compact and interpretable deep learning models for EEG-based MDD classification, particularly in larger and multi-center datasets.

Conclusion

This study investigated two compact deep learning architectures for distinguishing MDD from healthy controls using short resting-state EEG segments. Both the Light 1D-CNN and CNN-MiniTransformer achieved high classification performance on the internal test set and maintained strong performance on the independent external validation dataset. The Light 1D-CNN achieved the highest internal performance, while the CNN-MiniTransformer showed slightly better external performance and a smaller decrease between internal and external evaluation.

The explainability analysis provided additional insight into the EEG information used by the models. Despite their different architectures, both models showed a similar spatial importance pattern. P4 was consistently the most influential channel for MDD classification, with Fz and F4 also showing strong contributions. At the regional level, both architectures identified the parietal region as the most influential, followed by the frontal region. This agreement suggests that the two models relied on partly consistent spatial EEG information when distinguishing MDD from healthy-control segments.

Overall, the findings support the potential of compact and interpretable deep learning approaches for EEG-based MDD classification. The strong performance observed in the independent external validation cohort, together with the consistent spatial patterns identified by two different architectures, provides a useful basis for further EEG-based MDD research. Future studies with larger and multi-center cohorts could further evaluate the generalizability of these findings. Combining channel-level explainability with frequency-band and connectivity analyses may also help clarify the neurophysiological characteristics underlying the identified parietal and frontal EEG patterns.

Patient informed consent

Patient informed consent was not required.

Ethics committee approval

There is no need for ethics committee approval.

Financial support and sponsorship

No funding was received.

Conflict of interest

There is no conflict of interest to declare.

Author contributions subject and rate

Caglar Uyulan (100%): Designed the study; developed the methodology and software; performed the analyses and interpretation; prepared the visualisations; and wrote, reviewed, and approved the manuscript.

References

1. Duan L, Duan H, Qiao Y, Sha S, Qi S, Zhang X, Huang J, Huang X, Wang C. Machine learning approaches for MDD detection and emotion decoding using EEG signals. Front Hum Neurosci. 2020;14:284. doi:10.3389/fnhum.2020.00284.
2. Wang Y, Zhao S, Jiang H, Li S, Li T, Pan G. M-MDD: A multi-task deep learning framework for major depressive disorder diagnosis using EEG. Neurocomputing. 2025;636:130008. doi:10.1016/j.neucom.2025.130008.
3. Xia M, Zhang Y, Wu Y, Wang X. An end-to-end deep learning model for EEG-based major depressive disorder classification. IEEE Access. 2023;11:41337-41347. doi:10.1109/ACCESS.2023.3270426.
4. Anik IA, Kamal AHM, Kabir MA, Uddin S, Moni MA. A robust deep-learning model to detect major depressive disorder utilizing EEG signals. IEEE Trans Artif Intell. 2024;5(10):4938-4947. doi:10.1109/TAI.2024.3394792.
5. Biasiucci A, Franceschiello B, Murray MM. Electroencephalography. Curr Biol. 2019;29(3):R80-R85. doi:10.1016/j.cub.2018.11.052.
6. Jaseja H. Electroencephalography in the diagnosis and management of treatment-resistant depression with comorbid epilepsy: A novel strategy. Gen Psychiatry. 2023;36(2):e100868. doi:10.1136/gpsych-2022-100868.
7. Wang Y, et al. DiffMDD: A diffusion-based deep learning framework for MDD diagnosis using EEG. IEEE Trans Neural Syst Rehabil Eng. 2024;32:728-738. doi:10.1109/TNSRE.2024.3360465.
8. Safayari A, Bolhasani H. Depression diagnosis by deep learning using EEG signals: A systematic review. Med Nov Technol Devices. 2021;12:100102. doi:10.1016/j.medntd.2021.100102.
9. Rehman H, Khan DM, Amanullah H, Kamran L, Rehman OU, Siddiqui ST, Masroor K. Advances in EEG-based detection of major depressive disorder using shallow and deep learning techniques: A systematic review. Comput Biol Med. 2025;192:110154. doi:10.1016/j.compbiomed.2025.110154.
10. Saeedi A, Saeedi M, Maghsoudi A, Shalbaf A. Major depressive disorder diagnosis based on effective connectivity in EEG signals: A convolutional neural network and long short-term memory approach. Cogn Neurodyn. 2021;15(2):239-252. doi:10.1007/s11571-020-09619-0.
11. Loh HW, Ooi CP, Aydemir E, Tuncer T, Dogan S, Acharya UR. Decision support system for major depression detection using spectrogram and convolutional neural network with EEG signals. Expert Syst. 2022;39(3):e12773.
12. Ksibi A, Zakariah M, Menzli LJ, Saidani O, Almuqren L, Hanafieh RAM. Electroencephalography-based depression detection using multiple machine learning techniques. Diagnostics. 2023;13(10):1779.
13. Yu D, Li P, Su F, Wang K, Duan S, Yuan D. Automatic detection of depression using a CNN-Transformer model based on low-channel EEG data. Turk J Electr Eng Comput Sci. 2025;33(5):594-612. doi:10.55730/1300-0632.4146.
14. Liu X, Zhou L, Zhang W, Shao Z, Liu X, Chen C. LightDepressionNet: A lightweight depression diagnostic system based on CNN-Transformer. In: 2025 15th International Conference on Biomedical Engineering and Technology (ICBET); 2025. p. 6-10. IEEE. doi:10.1109/ICBET66094.2025.11470277.
15. Gan W, Zhao R, Ma Y, Ning X. TSF-MDD: A deep learning approach for electroencephalography-based diagnosis of major depressive disorder with temporal-spatial-frequency feature fusion. Bioengineering. 2025;12(2):95. doi:10.3390/bioengineering12020095.
16. Balakrishnan S, Murugesan RK, Ashraf H, Lim EL, Faisal A. EEGMetaNet: An enhanced Major Depressive Disorder detection framework using electroencephalogram and one-dimensional convolutional neural network. Appl Sci. 2026;16(15):7564. doi:10.3390/app16157564.
17. Naeim M, Atadokht A. Explainable temporal deep learning for EEG-based depression detection using resting-state brain dynamics. Int J Methods Psychiatr Res. 2026;35(2):e70088.
18. Liu S, Cui Y, Xu Y, Zhou B, Li G, Luo X. From EEG signals to quantitative assessment: Predicting depression severity using a novel deep learning framework. Sci Rep. 2026.
19. Rahmati M, Jalaeianbanayan A, Vahedi J, et al. Novel EEG-based diagnostic framework for Major Depressive Disorder using microstate and entropy features. Cogn Neurodyn. 2025;19:116. doi:10.1007/s11571-025-10305-2.
20. Zheng L, Fu J, Sun Z, Han Y, Shao T, Wang Q, Lin Z. Multi-scale EEG analysis identifies neural circuit signatures of iTBS responsiveness in major depressive disorder. NeuroImage. 2026:121697.
21. Peng D, Zheng WL, Lu BL. Enhancing depression detection from emotion EEG with temporal-spatial-spectral representation learning. In: 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society

(EMBC); 2025. p. 1-5. IEEE. doi:10.1109/EMBC58623.2025.11253710.

22. Kozulin I, Merkulova E, Savostyanov V, Shi H, Wang X, Bocharov A, Savostyanov A. Predicting major depressive disorder using neural networks from spectral measures of EEG data. Bioengineering. 2025;12(11):1251.