NeuroPredict: Deep Learning and LLM-Based Framework for SSRI Response in Major Depressive Disorder

Md Fahimul Kabir Chowdhury[1] MdFahimulKabirChowdhury@my.unt.edu
Jayed Mohammad Barek[1] JayedMohammadBarek@my.unt.edu
Md Ariful Hasan[1] MdArifulHasan@my.unt.edu
Haili Wang[1] Haili.Wang@unt.edu
[1] *University of North Texas*

Editors: Under Review for MIDL 2026

Abstract

Major Depressive Disorder (MDD) is a leading cause of disability worldwide, yet selective serotonin reuptake inhibitors (SSRIs) yield highly variable outcomes across patients. To address the need for individualized treatment prediction, we introduce NeuroPredict, a non-invasive EEG-based framework that leverages rich time–frequency representations of pre-treatment signals. EEG data were rigorously preprocessed with artifact removal and transformed using Wigner-Ville distribution (WVD), continuous wavelet transform (CWT), and constant-Q transform (CQT), then classified with a tailored convolutional neural network (CNN). We systematically benchmarked NeuroPredict against state-of-the-art pre-trained models (Xception, DenseNet201, MobileNetV2) and EEG-specific baselines (EEGNet, SleepEEGNet, DeepConvNet). Our proposed CNN consistently outperformed these approaches, achieving a peak accuracy of 98.28% and ROC of 98.20%, alongside strong precision, recall, and F1 score. Beyond predictive gains, we also highlight how large language model (LLM)-based support can enhance interpretability and streamline the analytic pipeline. These findings establish NeuroPredict as a scalable and clinically viable tool for precision prediction of SSRI treatment response, advancing the integration of data-driven methods into personalized psychiatry. Code available at: https://github.com/fahimulkabir/NeuroPredict

Keywords: SSRI, Major Depressive Disorder, CNN, Mental Depression, Machine Learning.

1. Introduction

MDD is a debilitating illness that disrupts emotional well-being, cognitive processes, and everyday functioning. A central challenge in its management is the considerable variability in treatment response across individuals. Although SSRIs remain a standard first-line therapy, only a subset of patients achieve full remission, with many requiring multiple treatment attempts before identifying an effective option (Zhou et al., 2021). This trial-and-error process is time-consuming, burdensome, and places significant emotional strain on both patients and their families.

In response to these challenges, recent research has increasingly focused on predicting antidepressant outcomes at baseline or during the early stages of therapy using machine learning approaches. Several studies have leveraged clinical and symptom profiles to train

predictive models of SSRI remission, demonstrating that baseline characteristics and early symptom changes can help identify patients more likely to benefit (Zhou et al., 2021; Li et al., 2025; Korani et al., 2025). Parallel efforts have examined neurophysiological signals, with EEG-based models incorporating features such as spectral power, connectivity, and entropy to capture mechanisms associated with treatment response (Li et al., 2025). A systematic review further emphasizes that models built on clinical, EEG, and other biological markers outperform traditional methods, while also underscoring the need for larger, more clinically applicable frameworks (Yadav et al., 2025).

In (Zhou et al., 2021), a cohort of 400 patients with MDD was analyzed to predict 8-week SSRI outcomes using baseline sociodemographic, clinical, psychological, and neurocognitive variables. After feature selection with LASSO, with the support vector machine(SVM) achieving the best performance (74.5% accuracy, AUC 0.734). In (Li et al., 2025), an EEG-based machine learning framework was developed using data from 27 patients with depression and validated on an independent cohort of 5. A SVM classifier achieving 96.8% accuracy on 12-second EEG windows. In (Yadav et al., 2025), a systematic review of 30 studies examined machine learning approaches for predicting antidepressant response. Across diverse datasets, SVM consistently achieved reliable performance (AUC 0.65-0.74) using clinical and symptom features, while EEG-based models reported accuracies up to 88%.

In (Albizu et al., 2024), using data from 16 participants in the ELECT clinical trial, individualized neuroimaging-derived computational models were applied. The framework achieved over 90% accuracy in classifying treatment responders (AUC 0.90, F1 0.92). Precision dosing further increased the predicted likelihood of response to nearly 100%. In (Lin et al., 2025), a multi-polygenic score framework was evaluated in two Taiwanese cohorts (N=177 and N=245) to predict SSRI treatment outcomes. The ensemble model achieved modest accuracy (AUC 0.631), but incorporating early symptom improvement substantially increased predictive performance (AUC 0.859). In (Byun et al., 2025), A study of 147 participants (MDD: 41, PD: 47, HC: 59) evaluated machine learning models for stress detection. Random forest and multilayer perceptron classifiers achieved moderate accuracies (0.67–0.73), which improved substantially with longitudinal scaling (up to 0.94 for MDD and 0.96 for HCs).

In (Wang et al., 2025), EEG functional connectivity features were analyzed in 30 untreated depression patients across multiple resting and video-watching conditions to predict SSRI treatment response. Using feature selection and SVM classification, accuracies exceeded 95% in cross-validation and reached over 85% on an independent test set. In (Moncy et al., 2024), a double-blinded, placebo-controlled trial of home-based tDCS, baseline EEG recordings from 21 participants were analyzed using portable four-electrode devices. Among multiple deep learning models, 1D-CNN achieved 85.7% accuracy, with 71.4% specificity and 92.8% sensitivity. In (Shahabi and Shalbaf, 2025), a Vision Transformer architecture was applied to pre-treatment EEG signals from 30 patients with MDD to predict SSRI response. The ViT model, comprising six encoder layers with multi-head attention, achieved 96% training accuracy, 92% validation accuracy, and 91% test accuracy, demonstrating strong generalizability for distinguishing responders from non-responders.

Building on this foundation, our contributions focus on predicting how patients will respond to SSRI using time-frequency image representations converted from the EEG signals.

The main contributions of this work are as follows:

- We introduce NeuroPredict, a non-invasive EEG-based framework for forecasting SSRI treatment response in MDD, leveraging rich time–frequency representations rather than relying solely on clinical or imaging variables.
- We investigate multiple time–frequency transformations (WVD, CWT, CQT) applied to rigorously preprocessed EEG signals, supported by modules for acquisition, artifact removal, and representation.
- To validate the proposed CNN, we perform comparisons against state-of-the-art pre-trained networks (Xception, DenseNet201, MobileNetV2) and EEG-specific baselines (SleepEEGNet, EEGNet, DeepConvNet), demonstrating clear performance advantages.
- We present a comprehensive evaluation of NeuroPredict, reporting accuracy, F1 score, precision, recall, and ROC metrics for SSRI outcome prediction, and further discuss how LLM-based support can enhance interpretability and pipeline organization.

2. Methodology

All experiments were conducted on a high-performance computing platform equipped with two NVIDIA H100 NVL GPUs, each offering 95,830 MB of dedicated memory and running CUDA 12.2. This computational infrastructure provided the capacity necessary for efficient model training, fine-tuning, and evaluation throughout the study. A schematic representation of the proposed framework is shown in Fig. 1.

2.1. EEG Data Acquisition

In this work, we analyze pre-treatment resting-state EEG recordings from 30 individuals diagnosed with MDD. All recordings were obtained under eyes-closed (EC) conditions (Mumtaz et al., 2017). The study protocol was reviewed and approved by the Human Ethics Committee of Hospital University Sains Malaysia, Kelantan. Clinical diagnoses of MDD adhered to the criteria specified in the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV). Of the participants, 12 were classified as treatment responders, defined by a reduction of at least 50% in their Beck Depression Inventory (BDI) scores following antidepressant therapy relative to baseline. EEG signals were recorded over a 5-minute interval using a 19-channel montage arranged according to the international 10–20 system and referenced to linked ears. To reduce artifacts, power-line contamination at 50 Hz was removed using a notch filter, and the data were band-pass filtered between 0.5 and 70 Hz. Signals were digitized at 256 Hz. The EEG signals were separated into periods of 15 seconds. Each 300-second resting-state EEG was segmented into non-overlapping 15-s windows, yielding 244 R and 344 NR segments. No overlapping windows were used. Table 1 summarizes the clinical characteristics of the participants.

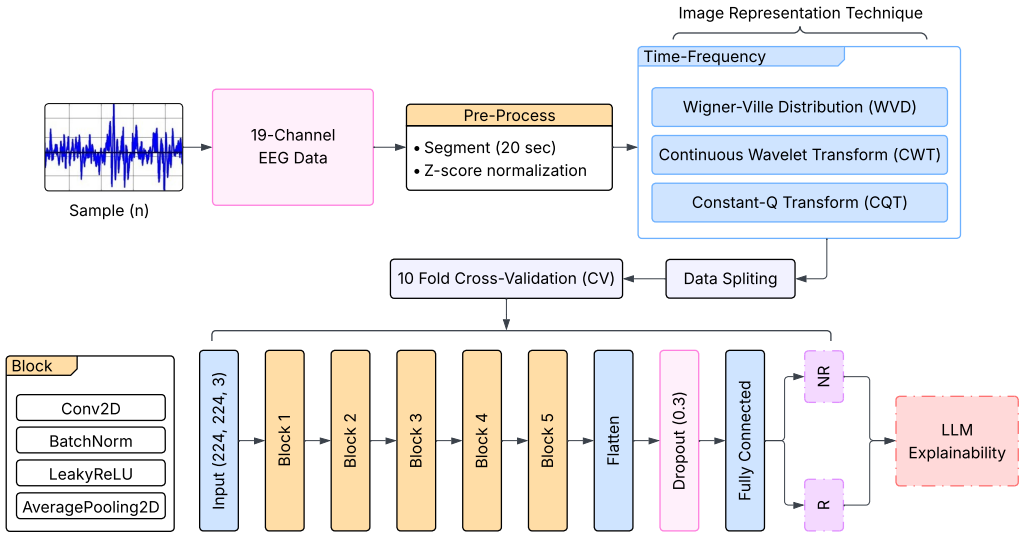


Figure 1: A summary of the EEG data preprocessing workflow and the transformation of signals into image representations using time-frequency techniques, which are then utilized as inputs for the proposed custom CNN model and LLM explanation.

Table 1: Demographic and clinical overview of MDD participants grouped by treatment response.

Characteristic	**Responder (R)**	**Non-Responder (NR)**	**Total**
Age (years)	40.7 ± 13.0	41.1 ± 12.5	40.3 ± 12.9
Gender (Male/Female)	8 / 8	9 / 9	17 / 17
BDI-II (Pre-Treatment)	18.4 ± 7.4	22.8 ± 12.5	20.6 ± 8.6
BDI-II (Post-Treatment)	9.1 ± 6.3	22.1 ± 3.3	15.6 ± 4.5

2.2. Time-frequency methods

CNN's are utilized for EEG classification by transforming the EEG signals into the image domain through three distinct time-frequency analysis methods. These techniques convert the signals into time-frequency representations, which are then used as inputs for the CNN shown in Fig. 2. The CNN models were trained successfully and exhibited enhanced learning capabilities when time-frequency images were incorporated, leading to improved performance in EEG classification tasks. The study employs three specific time-frequency analysis methods: Constant-Q Transform (CQT), Continuous Wavelet Transform (CWT), and Wigner-Ville Distribution (WVD).

Constant-Q Transform (CQT)

The CQT is a time-frequency analysis technique designed to produce frequency bins with a constant ratio between frequency and resolution. Unlike traditional Fourier transforms, the

CQT offers a logarithmic frequency scale, making it especially suited for signals with varying frequencies. For EEG signals, CQT converts the one-dimensional time-domain signal $x(t)$ into a two-dimensional time-frequency representation, where the time axis corresponds to the signal's temporal evolution and the frequency axis corresponds to the varying frequency bins. The CQT of a signal can be computed as:

$$\text{CQT}(t, f) = \int_{-\infty}^{\infty} x(t)\, g(t, f)\, dt$$

where $g(t, f)$ is the analysis window with a frequency-dependent resolution. The constant-Q property ensures that the time resolution is inversely proportional to the frequency, making it particularly effective for detecting both high and low-frequency components in EEG signals. This property is beneficial for CNNs because it allows the network to learn features across multiple scales of frequency, particularly in distinguishing between subtle patterns that may correspond to different mental states or neurological conditions.

Continuous Wavelet Transform (CWT)

The CWT is another time-frequency analysis method that offers a multi-resolution representation of the signal. The CWT transforms a one-dimensional EEG signal into a two-dimensional time-frequency image by convolving the signal with a set of wavelet functions at different scales and positions. Mathematically, the CWT is defined as:

$$\text{CWT}(t, \tau, s) = \int_{-\infty}^{\infty} x(t')\, \psi^* \left(\frac{t' - t}{s} \right) dt'$$

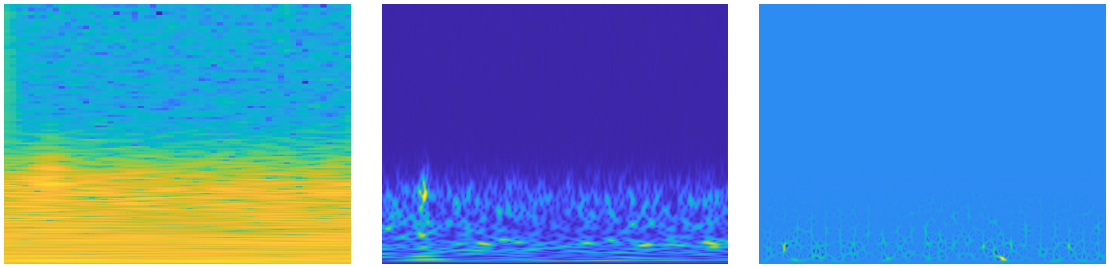
where ψ is the wavelet function, s is the scale parameter, and t is the time index. The CWT provides a detailed time-frequency representation with the flexibility to adjust the scale (or frequency) of the wavelet, making it effective for analyzing non-stationary signals like EEG. For CNN-based classification, this technique is advantageous because it allows the model to capture both high-frequency transient events and low-frequency oscillations.

Wigner-Ville Distribution (WVD)

The Wigner-Ville Distribution (WVD) is a time-frequency method that provides a high-resolution representation of a signal by calculating its instantaneous frequency content. It is defined as the Fourier transform of the autocorrelation function of the signal and is given by:

$$\text{WVD}(t, f) = \int_{-\infty}^{\infty} x(t + \tau/2) x^*(t - \tau/2) e^{-i2\pi f \tau}\, d\tau$$

where $x(t)$ is the EEG signal and f is the frequency. The WVD provides a precise time-frequency representation with sharp frequency localization, making it well-suited for capturing instantaneous spectral changes in EEG signals. For CNN applications, the high time-frequency resolution of the WVD allows the model to effectively distinguish between rapidly changing frequency components, enhancing the network's ability to identify transient patterns in EEG signals.



CQT CWT WVD

Figure 2: Examples of time-frequency (CQT, CWT, WVD) representations generated from the EEG data.

2.3. Data Splitting and Cross-Validation Setup

The dataset was split using a 10-fold cross-validation approach to ensure a robust evaluation of the model's performance. The original dataset was divided into 10 subsets, each used once as the test set while the remaining 9 subsets were used for training. This splitting process was automated using the `KFold` function from `sklearn.model_selection`, which randomly shuffles the data to ensure that each fold is representative of the overall dataset. The resulting data was organized into separate directories for training and testing, maintaining class labels for each image. This approach helps mitigate overfitting by allowing the model to train and test on different subsets of the data, providing a more reliable estimate of its generalization ability.

2.4. Proposed CNN Architecture

The proposed CNN architecture is designed for classifying the responder and non-responder to SSRI therapy task. The model takes an input image of size $224 \times 224 \times 3$ and processes it through a series of convolutional layers with increasing filter sizes. The first convolutional layer uses 32 filters with a kernel size of 5×5, followed by batch normalization, LeakyReLU activation, and average pooling with a pool size of 2×2. This process is repeated with increasing filter sizes: 64, 128, 256, and 512, respectively, at each subsequent layer. After each convolutional block, batch normalization ensures stable training, while LeakyReLU activation allows the model to learn non-linear relationships. Average pooling is applied after each convolutional block to reduce the spatial dimensions and retain the most relevant features.

The final convolutional block is followed by a flattening layer, which converts the 2D feature maps into a 1D vector. A dropout rate of 0.3 is applied to the flattened output to reduce overfitting. The model ends with a fully connected layer with a single output neuron, using a sigmoid activation function for binary classification. The model is compiled using the AdamW optimizer with a learning rate of 1×10^{-4}, binary cross-entropy as the loss function, and accuracy as the evaluation metric.

2.5. Transfer Learning (TL) Architecture

To provide a fair comparison with the proposed CNN, several state-of-the-art pre-trained models were implemented, including DenseNet201, MobileNetV2, and Xception. All models were initialized using weights pre-trained on ImageNet and configured with a fixed input shape of $224 \times 224 \times 3$. Each model followed a consistent architectural design to ensure comparability: a global average pooling layer, followed by a fully connected dense layer with 1024 units, a LeakyReLU activation function, a dropout layer with a rate of 0.5, and a softmax output layer for four-class classification. By maintaining a uniform classification head across all pre-trained networks, we were able to isolate the impact of the feature extractors themselves, thus minimizing any variations due to differences in the architectures of the models.

2.6. EEG-Specific Baseline Models

Well-established EEG based deep learning architectures commonly used in EEG analysis: EEGNet, DeepConvNet, and SleepEEGNet. EEGNet (Lawhern et al., 2018) utilizes depth-wise and separable convolutions to efficiently capture both temporal and spatial structures, offering a low computational cost while being well-suited for compact EEG representations. DeepConvNet (Schirrmeister et al., 2017) employs a deeper sequence of convolutional layers, augmented with max-pooling and dropout, to learn robust features from multi-channel EEG data. SleepEEGNet (Mousavi et al., 2019), originally designed for sleep stage classification, integrates temporal convolutions with hierarchical feature extraction modules to capture complex patterns in the EEG signal.

For a fair comparison, all three models were adapted to process our time-frequency EEG inputs of size $(224 \times 224 \times 3)$ and were trained using the same 10-fold cross-validation protocol to ensure consistency across experiments.

2.7. Model evaluation and validation

Although numerous metrics exist to find out the performance of classification models, accuracy remains a core and commonly adopted criterion. In this work, we focus primarily on accuracy as the principal evaluation metric.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

where TP is indicated as true positives, TN is indicated true negatives, FP is false positives, false negative is indicated as FN.

We also utilize the confusion matrix, which provides a tabulated summary of correct and incorrect predictions across both classes. From this, we derive additional key metrics:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F-score} = \frac{2}{3}\sum_{c=1}^{3}\frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{4}$$

Additionally, we incorporate the Receiver Operating Characteristic (ROC) curve as a visual and quantitative tool to evaluate model performance. The Area Under the Curve (AUC) serves as a critical measure of class separability. A perfect model yields an AUC of 1.0, indicating flawless discrimination between responder and non-responder classes.

3. Results and Discussion

In our first experiment, we evaluated different time–frequency image-based representations to determine which modality best supports our proposed model. As shown in Table 2, the CWT consistently outperformed the other approaches across all metrics. Specifically, CWT achieved the highest precision (99% NR, 98% R), recall (99% NR, 98% R), and F1-scores (99% NR, 98% R), resulting in an overall accuracy of 98.28% and ROC of 98.20. In comparison, the WVD obtained strong but lower performance (96.39% accuracy, 96.22 ROC), while the CQT lagged behind with 95.81% accuracy and 95.52 ROC. These results clearly demonstrate that CWT provides the most discriminative and robust time-frequency representation, making it the optimal choice for subsequent experiments. Fig. 3 shows the confusion matrix of MobileNetv2, SleepEEGNet and Proposed CNN. These results highlight the advantage of CWT-based time-frequency representations, which provide stable, low-noise structure.

Table 2: Performance comparison of time-frequency image-based representations using the proposed model. CWT achieves the highest overall accuracy. Shorthand: Prec. = Precision; Rec. = Recall; F1 = F1-score; Acc. = Accuracy; ROC = Receiver Operating Characteristic.

Method	**Prec. NR**	**Prec. R**	**Rec. NR**	**Rec. R**	**F1 NR**	**F1 R**	**Acc. (%)**	**ROC**
CQT	96	95	97	94	97	95	95.81	95.52
WVD	97	96	97	95	97	95	96.39	96.22
CWT	**99**	**98**	**99**	**98**	**99**	**98**	**98.28**	**98.20**

In the second experiment, we compared the performance of the best-performing time frequency representation (CWT) against a range of pretrained and EEG-specific models. As shown in Table 3, the proposed CNN consistently outperformed all baselines across every metric. It achieved near-perfect precision, recall, and F1 scores for both NR and R classes, culminating in an overall accuracy of 98.28% and a ROC score of 98.20. Among the pretrained models, DenseNet201 showed competitive results (92.97% accuracy), while EEGNet and DeepConvNet led the EEG-specific group with accuracies of 94.95% and 96.44%, respectively. These findings highlight the effectiveness of tailored time-frequency features combined with a dedicated CNN architecture, surpassing both general-purpose and domain-specific alternatives. Fig. 4 shows the ROC curve of DenseNet201, EEGNet, and Proposed CNN.

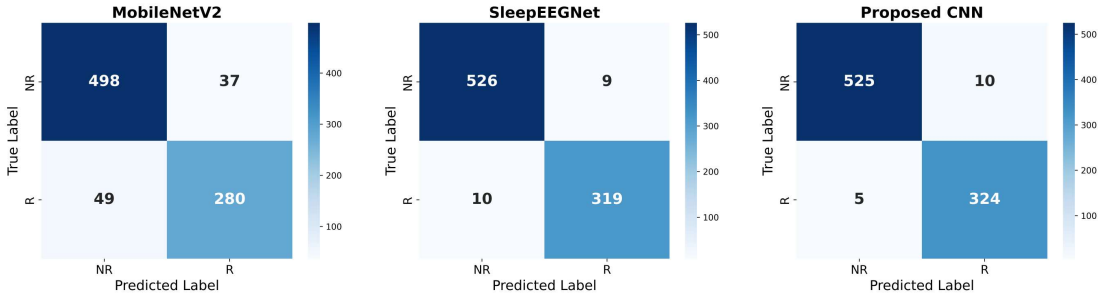


Figure 3: Confusion matrix of two benchmark models and proposed CNN

Table 3: Performance comparison of the best time-frequency representation (CWT) with pretrained and EEG-specific models.

Model	Prec. NR	Prec. R	Rec. NR	Rec. R	F1 NR	F1 R	Acc. (%)	ROC
Xception	86	87	93	78	89	82	86.77	85.30
MobileNetV2	90	88	92	85	91	87	89.50	88.74
DenseNet201	93	93	95	89	94	91	92.97	92.37
EEGNet	95	94	96	93	96	94	94.95	94.60
SleepEEGNet	96	96	97	94	97	95	95.82	95.44
DeepConvNet	96	96	98	95	97	95	96.44	96.11
Proposed CNN	**99**	**98**	**99**	**98**	**99**	**98**	**98.28**	**98.20**

To enhance interpretability of our model's predictions, we integrated a large language model (LLM)-based explanation module into our pipeline. After classifying patients as Responders or Non-responders to SSRI therapy using EEG-derived time-frequency images, we employed a clinician-facing explanation system powered by Meta-Llama-3-8B-Instruct via the Hugging Face Inference API. This module generates concise, non-prescriptive summaries contextualizing the model's output, including confidence scores and optional clinical metadata. The explanation prompt is carefully structured to avoid direct medical advice, instead offering general insights into EEG-based prediction patterns and emphasizing that final decisions must be made by qualified clinicians shown in Fig. 5.

This study presents NeuroPredict, a non-invasive EEG-based framework for predicting SSRI treatment response in patients with Major Depressive Disorder. Among the evaluated modalities, CWT yielded the most discriminative representations, enabling our proposed CNN to achieve a peak accuracy of 98.28%, outperforming both pretrained vision models and EEG-specific baselines. Furthermore, the inclusion of LLM-based support modules enhances interpretability and streamlines the analytic pipeline, offering a practical bridge between technical outputs and clinical decision-making. Notably, while our results demonstrate strong generalizability, the sample size remains modest. Future work should explore cross-site validation, larger cohorts, and multimodal fusion with genetic, behavioral, or imaging data to further improve robustness and clinical utility.

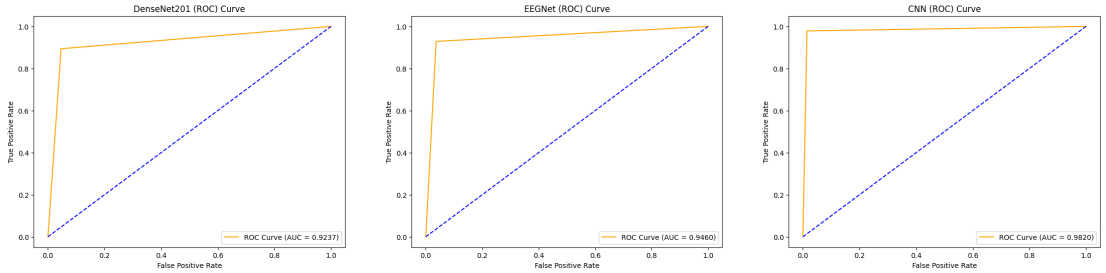


Figure 4: ROC curve of two benchmark models and proposed CNN

LLM Explanation Module: Prompt–Response Example

System instruction:
You are an assistant helping clinicians interpret predictions from an EEG-based research model for SSRI response in major depressive disorder. Be clear, concise, cautious, and avoid any prescriptive medical advice.

User prompt:
MODEL OUTPUT:
– Predicted class: `Responder`
– Model confidence (probability of "Responder"): `0.87` (87.0%)
TASK (3–6 sentences):
1) Restate the prediction in clear, non-technical language.
2) Briefly explain EEG-based prediction patterns.
3) Include a disclaimer about clinical decision-making.

LLM response:
The model suggests that this patient is likely to respond to SSRI therapy, with a confidence score of 0.87. This prediction reflects EEG patterns resembling those observed in individuals who improved during treatment. Features such as connectivity and time–frequency dynamics may contribute to this classification. However, this is a research-based estimate and should not be interpreted as a guarantee of benefit. Final decisions must be made by qualified clinicians using comprehensive clinical evaluation.

Figure 5: Illustration of the LLM explanation workflow, showing how model outputs are translated into clinician-friendly language with appropriate disclaimers.

4. Conclusion

We introduced NeuroPredict, a deep learning framework that utilizes time–frequency EEG representations to forecast SSRI treatment outcomes in MDD. Our proposed CNN achieved state-of-the-art performance, with 98.28% accuracy, demonstrating the feasibility of EEG-based precision psychiatry. By combining rigorous preprocessing, optimized signal encoding, and comparative benchmarking, NeuroPredict offers a scalable and interpretable solution for individualized treatment planning. These findings support the broader integration of neurophysiology-aware models into clinical workflows and highlight the potential of EEG as a low-cost, portable biomarker for antidepressant response prediction.

References

Alejandro Albizu, Aprinda Indahlastari, Paulo Suen, Ziqian Huang, Jori L Waner, Skylar E Stolte, Ruogu Fang, Andre R Brunoni, and Adam J Woods. Machine learning-optimized non-invasive brain stimulation and treatment response classification for major depression. *Bioelectronic Medicine*, 10(1):25, 2024.

Sangwon Byun, Ah Young Kim, Min-Sup Shin, Hong Jin Jeon, and Chul-Hyun Cho. Automated classification of stress and relaxation responses in major depressive disorder, panic disorder, and healthy participants via heart rate variability. *Frontiers in Psychiatry*, 15: 1500310, 2025.

Natalia Jaworska, Sara De la Salle, Mohamed-Hamza Ibrahim, Pierre Blier, and Verner Knott. Leveraging machine learning approaches for predicting antidepressant treatment response using electroencephalography (eeg) and clinical data. *Frontiers in psychiatry*, 9:768, 2019.

Ahmad Khodayari-Rostamabad, James P Reilly, Gary M Hasey, Hubert De Bruin, and Duncan J MacCrimmon. A machine learning approach using eeg data to predict response to ssri treatment for major depressive disorder. *Clinical Neurophysiology*, 124(10):1975–1985, 2013.

Wael Korani, Md Fahimul Kabir Chowdhury, Sadam AlQadi, Priyan Malarvizhi Kumar, Reza Rostami, and Reza Kazemi. Predicting the outcome of rtms depression therapy using eeg signals and cnn. In *Recent Trends in Image Processing and Pattern Recognition*, 2025.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Gang Li, Boyi Huang, Yuling Wang, Bin Zhou, Fo Hu, and Linbing Wang. Neurophysiological mechanisms and predictive modeling of ssri treatment response in depression disorder based on multidimensional eeg features. *Journal of Affective Disorders*, page 120424, 2025.

Shu-Chin Lin, Chiu-Ping Fang, Chia-Lin Hsu, An-Nie Chung, Tzu-Ting Chen, Kai-Hsiang Hsu, Chueh-Chun Yeh, Jingyi Zheng, ChaoYu Liu, Chi-Shin Wu, et al. Integrating multipolygenic scores for enhanced prediction of antidepressant treatment outcomes in an east asian population. *Neuropsychopharmacology*, pages 1–8, 2025.

Shota Minami, Masaki Kato, Shunichiro Ikeda, Masafumi Yoshimura, Satsuki Ueda, Yosuke Koshikawa, Yoshiteru Takekita, Toshihiko Kinoshita, and Keiichiro Nishida. Association between the rostral anterior cingulate cortex and anterior insula in the salience network on response to antidepressants in major depressive disorder as revealed by isolated effective coherence. *Neuropsychobiology*, 81(6):475–483, 2022.

Jijomon Chettuthara Moncy, Yong Fan, and Cynthia HY Fu. Prediction of treatment outcome to transcranial direct current stimulation in major depression based on deep

learning of eeg data. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 1123–1128. IEEE, 2024.

Sajad Mousavi, Fatemeh Afghah, and U Rajendra Acharya. Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PloS one*, 14 (5):e0216456, 2019.

Wajid Mumtaz, Likun Xia, Mohd Azhar Mohd Yasin, Syed Saad Azhar Ali, and Aamir Saeed Malik. A wavelet-based technique to predict treatment outcome for major depressive disorder. *PloS one*, 12(2):e0171409, 2017.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

Mohsen Sadat Shahabi and Ahmad Shalbaf. Deep vision transformer model for the prediction of antidepressants outcome in major depressive disorder using raw eeg signals. *Frontiers in Biomedical Technologies*, 12, 2025.

Mohsen Sadat Shahabi, Ahmad Shalbaf, and Arash Maghsoudi. Prediction of drug response in major depressive disorder using ensemble of transfer learning with convolutional neural network based on eeg. *Biocybernetics and Biomedical Engineering*, 41(3):946–959, 2021.

Fanglan Wang, Zifan You, Tingkai Zhang, Kai Xu, Liangliang Wang, Jingqi He, and Jinsong Tang. Predicting the treatment response of patients with major depressive disorder to selective serotonin reuptake inhibitors using machine learning techniques and eeg functional connectivity features. *Depression and Anxiety*, 2025(1):9340993, 2025.

Nishant Yadav, Anamika Gulati, Varun Gulati, and Prashant Yadav. Evaluating the impact of machine learning models on adult major depressive disorder using conventional treatment strategies: a systematic review approach. *Discover Public Health*, 22(1):410, 2025.

Shuzhe Zhou, Qinhong Ma, Yiwei Lou, Xiaozhen Lv, Hongjun Tian, Jing Wei, Kerang Zhang, Gang Zhu, Qiaoling Chen, Tianmei Si, et al. Machine learning to predict clinical remission in depressed patients after acute phase selective serotonin reuptake inhibitor treatment. *Journal of Affective Disorders*, 287:372–379, 2021.

Appendix A. Training Hyperparameter Configuration

To ensure a fair and consistent evaluation, all model categories were trained under identical hyperparameter settings shown in Table 4. The input dimensions, batch size, optimizer, activation function, and cross-validation strategy were standardized across EEG-specific networks, pretrained vision models, and the proposed CNN. Minor variations in dropout rates reflect the differing regularization needs of each architecture, but the overall framework was kept uniform. This alignment guarantees that performance differences arise from the models themselves and the representation of EEG signals, rather than from unequal training conditions.

Table 4: Training hyperparameters across EEG-specific models, pretrained architectures, and the proposed CNN.

Hyperparameter	EEG Models	Pretrained	Proposed
Input shape	(224,224,3)	(224,224,3)	(224,224,3)
Batch size	32	32	32
Dropout rate	0.2 / 0.3	0.5	0.3
Learning rate	1e-4	1e-4	1e-4
Optimizer	Adam	Adam	Adam
Activation	LeakyReLU	LeakyReLU	LeakyReLU
Cross-validation	10-fold	10-fold	10-fold

Appendix B. Confusion Matrix Comparisons Across Models

The appendix includes confusion matrices for several baseline models: Xception, DenseNet201, EEGNet, and DeepConvNet, trained under identical hyperparameter settings shwon in Fig. 6. These visualizations provide a clearer view of each model's classification strengths and weaknesses, particularly in distinguishing responders from non-responders. While all architectures achieved competitive performance, the confusion matrices highlight that misclassifications were more frequent in the pretrained vision models compared to EEG-specific networks. By contrast, our proposed CNN demonstrated the most balanced and accurate predictions across both classes, reinforcing the quantitative results reported in the main text.

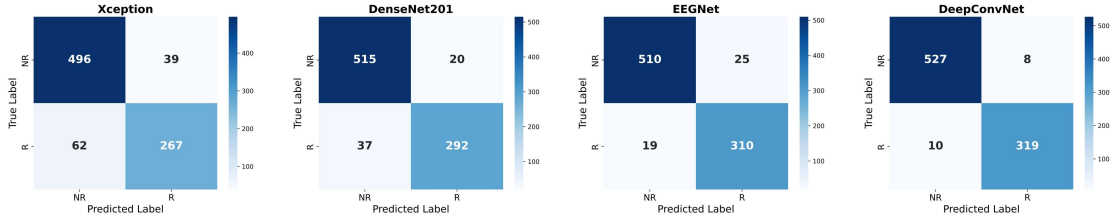


Figure 6: Confusion matrix of other benchmark deep learning & EEG based models

Appendix C. Additional Comparative Results

Table 5 summarizes prior EEG-based approaches for predicting antidepressant treatment response, highlighting differences in methodology, feature domains, and reported performance. Earlier studies employed techniques such as logistic regression on time–frequency features, nonlinear EEG descriptors with discriminant classifiers, or demographic and source-localized features combined with random forests. While these methods achieved moderate accuracy, sensitivity, or specificity, their performance was often limited by small sample sizes or reliance on handcrafted features. More recent work using CWT-based image representations and transfer learning ensembles reported improved results, approaching accuracies above 96%. In contrast, our proposed CNN, trained on CWT, CQT, and WVD representations, achieved the highest overall performance with 98.28% accuracy, 98% sensitivity, and 99%

specificity. This comparison underscores the advantage of deep learning applied to rich time-frequency EEG representations, demonstrating clear improvements over both traditional statistical approaches and pretrained vision models.

Table 5: Comparative analysis of EEG-based SSRI response prediction studies. Abbreviations: SEN = Sensitivity; SPE = Specificity

Reference	Methodology	Domain	Sample Size	ACC (%)	SEN (%)	SPE (%)
(Khodayari-Rostamabad et al., 2013)	Nonlinear EEG features, Fisher ratio, MFA classifier	Time-frequency	11 R / 11 NR	87.40	94.90	80.90
(Mumtaz et al., 2017)	Time–frequency analysis, ROC, logistic regression	Time-frequency	16 R / 18 NR	91.60	90.00	90.00
(Jaworska et al., 2019)	Demographics, EEG, source-localized PCA + RF	Time domain	27 R / 24 NR	88.00	77.00	99.00
(Shahabi et al., 2021)	CWT-based images, ensemble of pretrained models	Time-frequency	12 R / 18 NR	96.55	96.01	96.95
(Minami et al., 2022)	LORETA, coherence, functional connectivity, ROC	Time domain	12 R / 18 NR	–	82.00	86.00
Ours	CWT, CQT, WVD based deep learning	Time-frequency	12 R / 18 NR	**98.28**	**98.00**	**99.00**

Appendix D. LLM Explanation Module Examples

Appendix D presents illustrative outputs from the LLM explanation module for both responder and non-responder cases shown in Fig. 7. These examples demonstrate how raw model predictions are translated into concise, clinician-friendly language with appropriate disclaimers. The responder case highlights high-confidence predictions linked to EEG patterns resembling treatment success, while the non-responder case shows how altered spectral and connectivity features inform a cautious estimate of limited benefit. Together, these examples underscore the module's role in bridging technical outputs with interpretability, ensuring that probabilistic results are communicated responsibly for clinical contexts.

Appendix E. Extended Discussion and Future Directions

While we have included a focused discussion and outlook in the main paper, Appendix E provides extended reflections on the implications of this work. Beyond demonstrating strong predictive performance, our study highlights the importance of integrating explainability modules to bridge technical outputs with clinical interpretation. Future work may explore larger, multi-site EEG datasets to improve generalizability, as well as multimodal approaches that combine EEG with clinical or genetic information. Another promising direction is refining the LLM explanation module, exploring different models to adapt dynamically to diverse clinical contexts, ensuring that probabilistic outputs are communicated

LLM Explanation Module: Responder Example

System instruction:
You are an assistant helping clinicians interpret predictions from an EEG-based research model for SSRI response in major depressive disorder. Be clear, concise, cautious, and avoid any prescriptive medical advice.

User prompt:
MODEL OUTPUT:
– Predicted class: `Responder`
– Model confidence: `0.93` (93.0%)

LLM response:
The model indicates that this patient is very likely to respond positively to SSRI treatment, with a confidence score of 0.93. This estimate is based on EEG activity patterns that closely resemble those observed in individuals who benefited from therapy. Spectral and connectivity features in the time–frequency domain contributed to this classification. While the confidence is high, this remains a probabilistic assessment derived from research data. Clinical decisions must always be made by qualified professionals using a full evaluation of the patient's condition.

LLM Explanation Module: Non-Responder Example

System instruction:
You are an assistant helping clinicians interpret predictions from an EEG-based research model for SSRI response in major depressive disorder. Be clear, concise, cautious, and avoid any prescriptive medical advice.

User prompt:
MODEL OUTPUT:
– Predicted class: `Non-Responder`
– Model confidence: `0.88` (88.0%)

LLM response:
The model suggests that this patient is unlikely to benefit from SSRI treatment, with a confidence score of 0.88. This estimate is based on EEG activity patterns that resemble those observed in individuals who did not show improvement during therapy. Features such as reduced spectral coherence and altered time–frequency dynamics may have contributed to this classification. It is important to note that this is a probabilistic, research-based output and not a definitive clinical judgment. Treatment decisions must always be made by qualified clinicians using a comprehensive evaluation of the patient's condition.

Figure 7: Side-by-side illustration of the LLM explanation workflow for responder (left) and non-responder (right) cases.

responsibly. These extensions underscore the potential of our framework to evolve into a more comprehensive tool for precision psychiatry.