Channel Selection Guided by Layer-wise Relevance Propagation for CNN-Based EEG Classification of Major Depressive Disorder

Woo-Seok Ahn[1–3], Seung-Hwan Lee[3–5], and Han-Jeong Hwang[1,2*]
[1] Department of Electronics and Information Engineering, Korea University, Sejong, Republic of Korea
[2] Artificial Intelligence Smart Convergence Technology, Korea University, Sejong, Republic of Korea
[3] Clinical Emotion and Cognition Research Laboratory, Department of Psychiatry, Inje University, Republic of Korea
[4] Psychiatry Department, Ilsan Paik Hospital, Inje University, Goyang, Republic of Korea
[5] Bwave Inc, Goyang, Republic of Korea
E-mail: harry4746@korea.ac.kr, *hwanghj@korea.ac.kr, Tel: +82-44-860-1762
E-mail: lshpss@hanmail.net, Tel: +82-31-811-8430

Abstract— **This study presents a deep learning-based computer-aided diagnostic (CAD) system incorporating explainable artificial intelligence (XAI) techniques to support the accurate diagnosis of major depressive disorder (MDD). Resting-state electroencephalography (EEG) data from 40 drug-naïve male MDD patients and 41 male healthy controls were analyzed using a shallow convolutional neural network (Shallow ConvNet) combined with layer-wise relevance propagation (LRP). The proposed system achieved a classification accuracy of 99% without relying on hand-crafted feature engineering. The topographical map of relevance scores revealed higher relevance in the prefrontal, central, and occipital regions for MDD patients, and predominantly in the occipital regions for healthy controls. In addition, the LRP-based channel selection method enabled a substantial reduction in the number of EEG channels—from 62 to 10—while maintaining over 90% classification accuracy. These findings demonstrate the potential of XAI-based CAD systems not only to enhance diagnostic performance but also to provide novel insights into the neurophysiological mechanisms underlying MDD.**

I. Introduction

Major depressive disorder (MDD) is a common psychiatric condition that affects about one in ten adults. Without early diagnosis and treatment, MDD can progress to severe outcomes, including suicide [1]. However, when diagnosed early, more than 54% of patients recover well, showing the importance of early treatment [2]. Currently, psychiatrists have diagnosed MDD through subjective interviews and questionnaires based on patient self-report. These methods rely heavily on the patient's honesty and the clinician's judgment, lacking objective physiological indicators. As a result, diagnostic errors remain common, delaying appropriate treatment and lowering recovery rates [3].

To overcome these limitations, many studies now focus on developing computer-aided diagnostic (CAD) systems. Resting-state electroencephalography (EEG) has gained attention because it shows clear neurophysiological differences between MDD patients and healthy controls (HCs) [4]. EEG-based diagnostics offer more objective and reliable markers compared to traditional methods. With advances in artificial intelligence (AI), deep learning-based CAD systems have further improved diagnostic accuracy for MDD.

However, despite their strong performances, these systems have critical challenges. Deep-learning models often lack explainability due to the black-box nature of their architecture, which limits clinical use [5]. In addition, it is complex to use multi-channel EEG data. This challenge leads to a decrease in practical usability. As a result, it has become a key goal to improve model explainability and minimize the number of required channels.

The present study has aimed to develop an interpretable CAD system to classify drug-naïve male MDD patients and HCs using resting-state EEG data. We applied explainable artificial intelligence (XAI) techniques, specifically layer-wise relevance propagation (LRP), to enhance model explainability [6]. We also have used relevance scores from LRP to design a channel selection approach that reduces the number of EEG channels without compromising diagnostic performance, improving the system's practical utility. Ultimately, the proposed CAD system seeks to enable early and accurate identification of MDD, enhancing its potential clinical utility.

II. Methods

A. Data acquisition

This study included 40 drug-naïve male patients diagnosed with MDD and 41 age- and sex-matched HCs. The diagnosis of MDD was confirmed by board-certified psychiatrists based on

979-8-3315-7206-8/25/$31.00

the Diagnostic and Statistical Manual of Mental Disorders, 5th Editon (DSM-5) criteria. The severity of depression and anxiety symptoms was assessed using the Hamilton Depression Rating Scale (HAM-D) and the Hamilton Anxiety Rating Scale (HAM-A), respectively. Exclusion criteria included neurological disorders, substance abuse, developmental delays, history of head injury with loss of consciousness, prior electroconvulsive therapy, or psychotic symptoms lasting over 24 hours. All participants provided written informed consent, and the study protocol was approved by the Institutional Review Board (IRB) of Inje University Ilsan Paik Hospital, in accordance with the Declaration of Helsinki.

Table 1 presents the demographic characteristics of drug-naïve male MDD patients and healthy controls, including comparisons of age and education level, which were statistically assessed using independent t-tests.

Table 1. Demographic characteristics of drug-naïve male MDD and HCs.

	MDD	HC	*p*-value
Case (N)	40	41	
Age (years)	33.88±12.24	34.15±11.37	0.92
Education (years)	14.83±1.06	15.27±0.95	0.05
HAM-D	24.58±5.01		
HAM-A	29.30±6.67		

B. EEG data analysis

Eyes-closed resting-state EEG data were recorded for 3 – 5 min at a sampling rate of 1000 Hz using a NeuroScan SynAmps2 system (Compumedics, USA) with 64 Ag/AgCl electrodes mounted on a QuickCap according to the extended international 10 – 20 system. Data from 62 electrodes, excluding the reference electrodes (M1 and M2), were analyzed after bandpass filtering between 1 and 55 Hz. Independent component analysis (ICA) was applied to remove artifacts such as eye blinks, electrocardiography (ECG), and electromyography (EMG). The data were then re-referenced using the common average reference (CAR). The cleaned EEG data were down-sampled to 200 Hz, segmented into approximately 3-min intervals, and processed using MATLAB R2022a (MathWorks, USA).

C. Deep learning strategy

A shallow convolutional neural network (Shallow ConvNet), informed by the filter bank common spatial patterns (FBCSP) framework, was implemented to extract spatio-temporal frequency features through two convolutional layers, a mean pooling layer, and a rectified linear unit (ReLU) activation function. We trained the model with a batch size of 2 for 200 epochs, using a learning rate of 0.001 and a dropout rate of 0.5. Model performance was rigorously evaluated using leave-one-out cross-validation (LOOCV), with one subject was held out for testing while the remaining participants were split into training and validation sets at a 4:1 ratio.

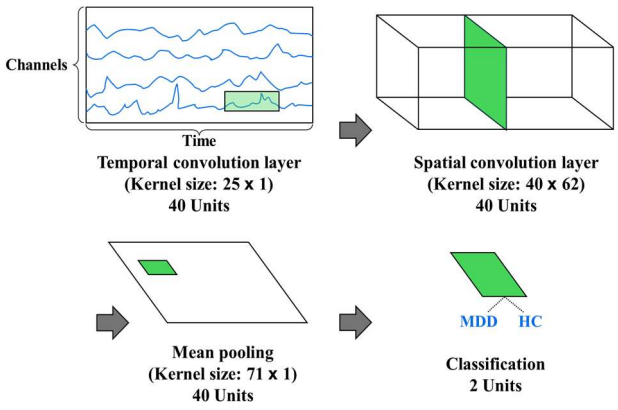

Fig. 1 Schematic of the Shallow ConvNet architecture for EEG classification, highlighting key convolutional, pooling, and classification layers for temporal and spatial feature extraction.

D. XAI-based channel selection approach

We applied layer-wise relevance propagation (LRP), an XAI method, to improve the diagnostic model's interpretability and utility. We computed channel-wise relevance scores from validation data, averaged them over time, and normalized them within each class.

The LRP relevance score indicates how much each input contributes positively to the model's decision, meaning that high positive scores suggest channels that well explain MDD features. Negative relevance scores are known to hinder the model's decision, so we explored whether these channels explain HCs better. Therefore, channels were ranked using three strategies: original value, absolute value, and zero-clipped value strategies.

Original value strategy is sorting its original relevance scores in descending order. To better understand the meaning of negative scores, we applied the absolute value strategy, allowing channels with large negative scores to be included in selection. In absolute value strategy, we first computed the absolute values of the relevance scores and then ranked the channels in descending order based on these absolute values for channel selection. We also applied the zero-clipped value strategy to further investigate the role of negative scores. For zero-clipped value strategy, we set negative relevance scores to zero and then ranked the channels in descending order for selection.

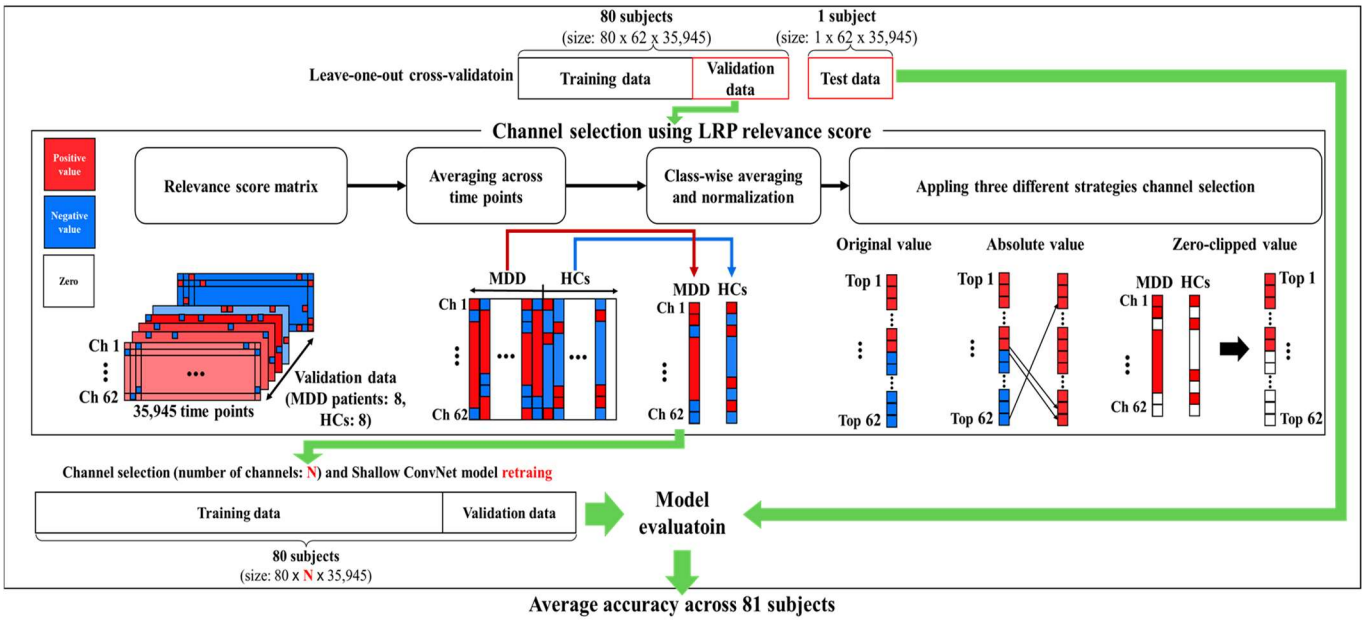

Fig. 2 Scheme of the relevance score-based channel selection method using three sorting strategies: original relevance, absolute relevance, and zero-clipped relevance.

Based on these, we selected top-ranked subsets and evaluated classification performance with fewer channels. We compared 10 and 5 channels to test practicality and assess robustness under limited input. Fig. 2 illustrates the channel selection process based on LRP relevance scores. This approach not only identified the minimal set of channels needed to maintain reliable performance but also demonstrated the model's potential for clinical translation.

III. Results

Table 2 presents the classification performance of the proposed CAD system in distinguishing MDD patients from HCs using varying numbers of EEG channels. When all 62 channels were utilized, the system achieved a classification accuracy of 99%. With a reduced set of ten channels, classification performance remained consistently above 90% across the three-relevance score sorting strategies: 96% for the zero-clipped strategy, 92% for the absolute value strategy, and 91% for the original value strategy. However, when further reduced to five channels, a notable decline in accuracy was observed, particularly for the original value strategy (74%), while the absolute value and zero-clipped strategies yielded slightly higher accuracies of 75% and 77%, respectively.

Table 2. Deep learning performance across LRP-based channel selection strategies. The best performance is shown in bold.

Selected channels (N)	Strategy	Accuracy
62		99%
10	Original value	91%
	Absolute value	92%
	Zero-clipped value	**96%**
5	Original value	74%
	Absolute value	75%
	Zero-clipped value	77%

Fig. 3 illustrates the topographical distributions of the EEG channels selected using relevance score-based sorting under each of the three strategies: original value, absolute value, and zero-clipped value. Across all sorting methods, the most frequently selected channels were primarily located in the central and occipital regions, highlighting the potential neurophysiological significance of these areas in distinguishing individuals with MDD from HCs.

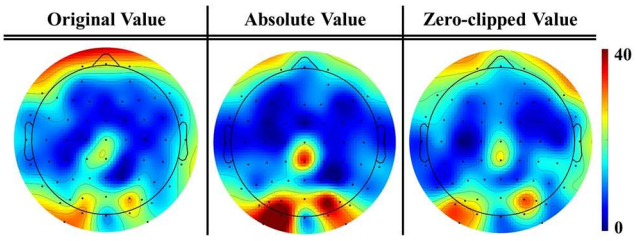

Fig. 3 Topographical maps showing the most frequently selected EEG channels under each LRP-based sorting strategy: original relevance (left), absolute relevance (middle), and zero-clipped relevance (right).

IV. Conclusion

The LRP-based channel selection approach effectively identified key neurophysiological features, maintaining high

diagnostic performance even when the number of EEG channels was reduced from 62 to 10–an approximate 84% reduction. Incorporating neurophysiological characteristics identified in the male MDD population, in conjunction with general MDD-related features, further improved classification performance.

Notably, alterations in neural activity were observed across several brain regions. In the occipital lobes, which are responsible for visual processing and sensory integration, altered neural activity may reflect impairments in visual and sensory functions among patients with MDD [7]. In the central region, typically associated with sensorimotor integration, deviations in neural activity potentially indicate disruptions in sensory-motor coordination [8]. Furthermore, in the prefrontal cortex, a region critically involved in emotional regulation and executive function, the observed alterations were directly linked to deficits in emotional control [9].

Future research should prioritize expanding the datasets and developing integrated diagnostic systems applicable to both male and female patient populations. Furthermore, advancing channel selection methods will not only improve the CAD system's practical utility and clinical applicability but also enhance its potential to facilitate earlier and more precise diagnosis of MDD.

V. Acknowledgment

This research was supported by the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00455484) and the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2025-RS-2023-00258971) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

[1] World Health Organization, World Health Statistics 2025: Monitoring Health for the SDGs, Sustainable Development Goals. Geneva: WHO Press, 2025, ISBN 978-92-4-011049-6.

[2] A. Fernández, A. Pinto-Meza, et al., "Is major depression adequately diagnosed and treated by general practitioners? Results from an epidemiological study," Gen. Hosp. Psychiatry, vol. 32, no. 2, pp. 201–209, 2010.

[3] M. Guha, "Diagnostic and Statistical Manual of Mental Disorders: DSM-5 (5th edition)," Ref. Rev., vol. 28, no. 3, pp. 36–37, 2014.

[4] C. Greco, O. Matarazzo, et al., "Discriminative power of EEG-based biomarkers in major depressive disorder: A systematic review," IEEE Access, vol. 9, pp. 112850–112870, 2021.

[5] E. Şahin, N. N. Arslan, and D. Özdemir, "Unlocking the black box: An in-depth review on interpretability, explainability, and reliability in deep learning," Neural Comput. Appl., vol. 37, pp. 859–965, 2025.

[6] S. Bach, A. Binder, et al., "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," PLoS One, vol. 10, no. 7, p. e0130140, 2015.

[7] J. Yang, J. Li, S. Zhao, Y. Zhang, B. Li, and X. Liu, "Fusion of eyes-open and eyes-closed electroencephalography in resting state for classification of major depressive disorder," Biomed. Signal Process. Control, vol. 100, part B, p. 106964, 2025.

[8] X. Liu, H. Zhang, Y. Cui, T. Zhao, B. Wang, X. Xie, S. Liang, S. Sha, Y. Yan, X. Zhao, and L. Zhang, "EEG-based major depressive disorder recognition by neural oscillation and asymmetry," Front. Neurosci., vol. 18, p. 1362111, 2024.

[9] Xie, X. M., Sha, S., Cai, H., Liu, X., Jiang, I., Zhang, L., & Wang, G. (2024). Resting-State Alpha Activity in the Frontal and Occipital Lobes and Assessment of Cognitive Impairment in Depression Patients. *Psychology Research and Behavior Management*, *17*, 2995–3003.