ABSTRACT

Alzheimer's disease is a progressive neurodegenerative disorder in which early diagnosis and evaluation highly depend on Magnetic Resonance Imaging (MRI). Although deep learning models based on Convolutional Neural Network (CNN) are capable of achieving high classification accuracy, these CNN models are often considered as black boxes due to their lack of transparency in explaining clinical decision-making processes. This study aims to compare the performance of EfficientNet-B0 and ResNet50 architectures in classifying four stages of Alzheimer's disease and to evaluate model interpretability using the Gradient-weighted Class Activation Mapping (Grad-CAM) method. The dataset used consists of 6,400 MRI images divided into four classes: Non-Demented, Very Mild Demented, Mild Demented and Moderate Demented. The experiments were conducted under several scenarios, including the original dataset, data augmentation, class weighting, and a combination of data augmentation and class weighting. The results showed that EfficientNet-B0 demonstrates better stability and generalization capability, achieving a validation accuracy of 96.5% and a recall of 95% in the augmentation scenario. Meanwhile, ResNet50 achieved the highest accuracy of 97% on the original dataset but tended to suffer from overfitting and shortcut learning phenomena. Grad-CAM visualization indicates that EfficientNet-B0 consistently focuses on clinically relevant brain anatomical regions compared to ResNet50. Based on these findings, EfficientNet-B0 is considered superior in maintaining a balance between classification performance, generalization capability and clinical interpretability.

Keywords: Alzheimer's disease, CNN, EfficientNet-B0, ResNet50, Grad-CAM