Enhancing Trust in Alzheimer's Disease Classification using Explainable Artificial Intelligence: Incorporating Local Post Hoc Explanations for a Glass-box Model

Abraham Varghese, PhD* Ben George, PhD* Vinu Sherimon, PhD* Huda Salim Al Shuaily, PhD**

ABSTRACT

Background: Alzheimer's disease (AD) leads to cognitive dysfunction among older people worldwide, making it nearly impossible for them to carry out their daily lives. Due to the inherent characteristics of Alzheimer's disease and its impact on the brain, timely intervention is crucial to delay its onset and mitigate its progression. Currently, the diagnosis of Alzheimer's disease often occurs at a stage where it is too late for effective prevention measures, allowing the disease to cause significant damage to the brain. The use of machine learning and deep learning models is critical for the classification of demented and non-demented cases, but most highly accurate models are non-linear and less transparent, not revealing the logic behind the predictions. Therefore, incorporating interpretability components into the models will make them more transparent and trustworthy. This study is aimed to develop appropriate diagnostic methods capable of assessing Mild Cognitive Impairment (MCI), the early stage of Alzheimer's disease that occurs before the irreversible loss of neurons.

Methods: Explainable artificial intelligence (XAI) refers to AI systems that can provide explanations for their decisions or predictions. In the context of AD classification, explainable AI systems aim to provide insights into the features or characteristics of the model used to make a prediction. This XAI provides a mechanism to understand and interpret the basis of a model's predictions which is more important for improving the trust in the system and its results. As such, a non-linear neural network is employed in this work to distinguish between demented and non-demented cases while local post hoc explanations are incorporated to make it a glass-box model using the XAI techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model Agnostic Explanations (LIME).

Results: The application of LIME provided valuable insights into the impact of various factors on predictions. Notably, factors such as CDR, Age, and ASF aligned with clinical knowledge and proved instrumental in predicting dementia cases. Conversely, features like nWBV, MMSE, and eTIV adversely affected the predictions, highlighting their significance in identifying non- demented cases. Similarly, exploring SHAP values yielded a comprehensive understanding of the decision-making process employed by the model in detecting Alzheimer's disease.

Conclusion: Through the utilization of explainable artificial intelligence (XAI) methods, this study endeavors to develop a dependable and transparent technique for early detection, monitoring, and personalized interventions in the realm of Alzheimer's disease.

Keywords: Machine learning, Alzheimer's disease, Interpretable Machine learning, LIME, SHAP, Explainable AI, Neural network

Bahrain Med Bull 2023; 45 (2): 1471 - 1478

* College of Computing and Information Sciences
University of Technology and Applied Sciences
Muscat, Sultanate of Oman.
E-mail: abraham.varghese@utas.edu.om

** Deputy Vice-Chancellor Office
University of Technology and Applied Sciences
Muscat, Sultanate of Oman.