WWU MÜNSTER
Institut für Medizinische Informatik
IMI
medizinische fakultät
Westfälische Wilhelms-Universität Münster

Titel:

Concept-based Interpretability in Electroencephalography Abnormality Detection via Machine Learning

Abstract:

Despite significant performance improvements in recent years, deep learning models are not yet widely used in clinical practice. One of the main reasons often cited for this is the "black box" nature of such systems. This refers to the lack of transparency and interpretability of complex neural networks, whose prediction process is based on an extremely large number of parameters and is not traceable for humans. These problems have led to an increased interest in concept-based interpretability approaches, where human-understandable, abstract concepts are used to explain model behavior. The method referred to as Testing with Concept Activation Vectors (TCAV) has previously been applied to the medical domain in the context of medical images and electronic health records. This study explores the use of TCAV in biosignals, specifically electroencephalography (EEG) data, with a state-of-the-art deep learning model. We show how new concepts can be defined and integrated into the testing. In the context of abnormality detection, we analyze whether the representative model is sensitive to specific pathological waveforms, brain regions or frequency bands. Based on the results, we discuss advantages, potential bias and limitations.