Quantitative analysis methods for post-hoc interpretability in neural networks for biosignal analysis

Nicolai Spicher, University Medical Center Göttingen (UMG)

ABSTRACT

The increasing availability of large-scale datasets containing biosignals, such as electrocardiography (ECG) or electroencephalography (EEG) led to a surge of machine learning methods. Especially end-to-end deep learning pipelines showed remarkable results in many clinical tasks, often outperforming human experts. However, a pitfall lies in the fact that these are assumed to be "black box" models often based on agnostic features. While they bear the theoretical potential to aid clinicians in diagnostics or treatment decisions, clinicians need to be able to comprehend the reasoning behind the algorithm, as a ”Clever Hans” prediction, based on spurious or artifactual correlations, might lead to wrong decisions and adverse consequences for patients. Diverse post-hoc interpretability methods have been developed in recent years but they are used in most cases only qualitatively on single signals of individual patients and give only anecdotal evidence.

In this talk, I will present the activities of the UMG biosignal processing research group towards extracting quantitative information of these interpretability methods to explain what deep neural networks for biosignal classification actually "learned". Examples stem from the field of cardiac disease classification and sleep studies.

Website: https://medizininformatik.umg.eu/en/about-us/scientific-research-groups/biosignal-processing/