Comparison of White Box and Black Box Models in the Context of Artificial Intelligence for ARDS Classification

(Master thesis)

LISA HARTUNG

Motivation

The motivation for this work lies in the current research focus on artificial intelligence (AI) models for the classification of patients with Acute Respiratory Distress Syndrome (ARDS). The condition is still under-diagnosed or diagnosed too late, and crucial time for the early initiation of targeted therapy could be shortened through by early prediction. As AI is increasingly integrated into medical diagnostics and therapeutic decision making, there is an urgent need to elucidate the available white and black box models. In the literature, white box models are recognized as machine learning algorithms that are easy to understand and transparent, while black box models earn their label due to their mathematical complexity, which is challenging to explain and comprehend. The transitions between the two types of models are sometimes fluid.

State of the art

In the literature, there is ongoing debate about whether white box models are inherently more transparent than black box models or if the distinctions are more nuanced. The question of whether there is a performance tradeoff between explainability and model effectiveness is also a recurring theme. Several white and black box models have been applied to classify ARDS in time-series data, leveraging different available datasets. These models vary in their architecture, training data, and overall complexity. In efforts to enhance interpretability and transparency, explainable AI (xAI) methods have been applied to these models. A systematic comparison, particularly with regard to performance and explainability, has not yet been conducted in this context.

Objective

The main objective of this research is to conduct a comprehensive analysis and comparison of White-Box and Black-Box models for the classification of ARDS in time-series data. Additionally, extra AI procedures will be incorporated to augment the versatility of the models. This could, for example, involve utilizing linear or logistic regression as a representative of white box models, as well as employing XGBoost as a representative of black box models, leveraging the data made available by the department. The results of the investigation may provide recommendations regarding the selection of suitable models, with performance, explainability, and applicability in medical diagnostics being essential criteria.

Procedure

This study commences with a comprehensive literature review on ARDS classification and the integration of artificial intelligence. White-Box and Black-Box models are then identified for subsequent analysis. Possible xAI methods are identified and applied to the selected AI models. The study concludes with an outlook on potential advancements and future research directions in the realm of ARDS classification using AI. The results and findings are then summarized in a written report.

Betreuer: Simon Fonck
21.11.2023

Informatik 11
Embedded Software
RWTH AACHEN UNIVERSITY