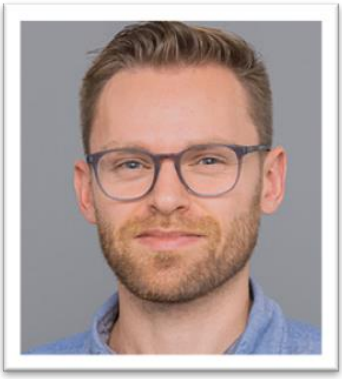
Development of a Framework for post-hoc Explainability of Fault Detection in ICU-Data

(Master Thesis)



JULIAN TREIBER

Motivation

For the survival of patients in Intensive Care Units (ICUs) fast and reliable detection of changing conditions, like the acute respiratory distress syndrome (ARDS), is critical. As part of the SMITH project, the chair is researching algorithmic solutions to detect complications of ICU patients and detect faults in ICU data using data-driven Machine Learning (ML) approaches. Modern ML approaches, like Neural Networks, are black-boxes from a user perspective. Especially in the medical field, transparency and explainability are of key importance to facilitate trust in the decision-making process. Explainable AI (XAI) is a research field with the aim of adding transparency to ML methods.

The chair is developing the Novelty Detection Analysis System (NDAS), which is used for finding novelty data or outliers in ICU data using ML approaches. The system can benefit from adding human readable explanations, since it is intended to be used by medical professionals.

State of the Art

There are two levels of explanations for black-box models: Global explanation approaches aim at explaining the inner workings of a model, whereas local explanations interpret a specific output of the model. Examples of this are feature-based explanations like LIME or SHAP, where LIME calculates the importance of input features on a global or local scale with the help of simpler “proxy” models, and SHAP infers the marginal contributions of each input feature for a given a specific input and output pair, with extensions for global explanations. Another approach is to use example-based explanations, such as “counterfactual explanations” or “prototyping”. Counterfactual explanations try to find the slightest change for a given input, which causes it to be classified differently, whereas prototyping produces representative examples for the general behavior of the model.

Objective

The thesis’ objective is to enhance NDAS with XAI methods. Options to view global explanations, giving a better understanding of the chosen algorithm, will be added. Additionally, local explanations will be added to the classification outputs. Automatic evaluation of the explanation methods will be implemented, and visualization of the explanations will be made available. The integration of explainability into NDAS will be modular, to allow additional algorithms in the future.

Planned Procedure

As a first step, a literature research into model-agnostic explainability will be conducted. Suitable global and local methods for the explanation of the outlier detection will be explored and then integrated as an extension of the NDAS framework. Afterwards, an evaluation on the transparency and usability of the implemented explainability techniques will be conducted.

Betreuer: Alexander Kruschewsky
20.04.2023

Informatik 11
Embedded Software
RWTH AACHEN UNIVERSITY