Opening the Black-Box: Extracting Medical Reasoning from Machine Learning Predictions

Marius FERSIGAN[a,*] and Marius MĂRUŞTERI[b]

[a] Iuliu Haţieganu University of Medicine and Pharmacy, Louis Pasteur Str., no. 6, 400349 Cluj-Napoca, Romania
[b] University of Medicine, Pharmacy, Science and Technology of Târgu Mureş, Gheorghe Marinescu Str., no. 38, 540139 Târgu Mureş, Romania
E-mails: marius.fersigan@algunion.com; marius.marusteri@umftgm.ro

* Author to whom correspondence should be addressed

Abstract

Background: A transparent machine learning model enables the end-user to investigate the decision process from (clinical) input data to the final prediction of the model. Modern (Deep Learning) machine learning models are highly complex and opaque - lacking the transparency required by both the clinician and the patient. Not knowing why a model predicted a particular outcome in a clinical context would likely hinder the clinician's trust in the model and ultimately impact the patient's health outcome. *Aim*: To integrate the model training into a unique automated pipeline that will output the prediction along with the relevant explanation. *Materials and Methods*: The workflow was build using the Julia programming language and employing a wide range of machine learning packages from the MLJ universe. The interoperability with the explainer packages from the Python ecosystem was enabled using the PyCall.jl package. *Results*: An automated machine learning pipeline that a) exposes the (clinical) decision process inside various black-box models (convolutional neural networks, random forest, boosted models), b) enables the training of explainable models without compromising the performance metrics, c) validates the resulted method against the medical domain-knowledge. Our pipeline combines the known non-healthcare-specific explainers - LIME, SHAP, ShapML, and InterpretML into a global explainer tailored to healthcare-specific data. *Conclusions*: The performance metrics of machine learning models trained on healthcare data are not necessarily sacrificed on the altar of transparency and interpretability. Using model-agnostic and model-specific explainers, we can satisfy the clinician's need for a transparent decision and diagnosis process and the superior performance metrics that ensure the predictive power of the model.

Keywords: Explainable AI (Artificial Intelligence); Machine Learning; Interpretability; Explainability; Transparency