Fraunhofer IPA

Fraunhofer Institute for Manufacturing Engineering and Automation IPA

Figure 1:
Use of a local explanation method to determine the image areas that contribute to the recognition of the human hand.

xAI – Explainable Machine Learning Methods

Market requirements

Machine learning methods are gaining in popularity in various fields of application like in manufacturing, medicine or the service sector. For example, machine vision or sensor data analysis can be used to detect rejects and remove them from the process at an early stage. Deep neural networks are also increasingly being tested for programming robots. However, there are many scenarios in which highly accurate predictions alone are less important than trust, acceptance, or compliance with regulations. In such cases, critical decisions must be accompanied by explanations so that users can understand the results or how the algorithm works in general. By providing explanations, not only is it possible to check that the models are functioning correctly, but also to investigate any discrepancies beween decisions made by humans and those made by algorithms. For example, answers can be found as to why a certain manipulated variable was output for a controller. Explanations can also offer considerable added value in fields where safety plays a critical role, such as autonomous driving or human-robot collaborations. Explanations for decisions made by a model help experts improve their understanding of the model and assess risks.

Our approach

There are basically two ways of making machine learning methods explainable. When it comes to explaining a model (globally), the focus is on understanding the model as a whole. The aim is to trace the internal decision paths of a black box model as best as possible. In contrast, local explanation methods are used to explain individual decisions (see Figure 1).

At Fraunhofer IPA, we can help you make your existing machine learning models (ML models) explainable by generating explanations - both locally and globally. Even in the course of new developments, Fraunhofer IPA enables you to consider explainability as an elementary component of the process and end product right from the start.

The current state of the art in the research field of explainable AI comprises a multitude of different methods. Since not every method is equally suitable for every application, choosing the best method is both a time-consuming and research-intensive task. For this reason, Fraunhofer IPA is also developing a software toolbox that reconciles existing explanatory methods. Proprietary algorithms and procedures developed at Fraunhofer IPA are also integrated. By using this universal toolbox, a comprehensive understanding of the model can be achieved through the rapid generation of explanations and the comparison of different techniques.

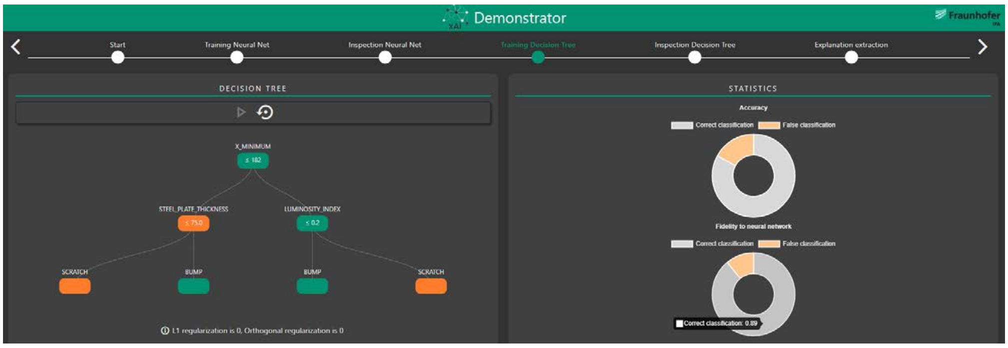

Figure 2:
Approximation of a neural network using a decision tree.

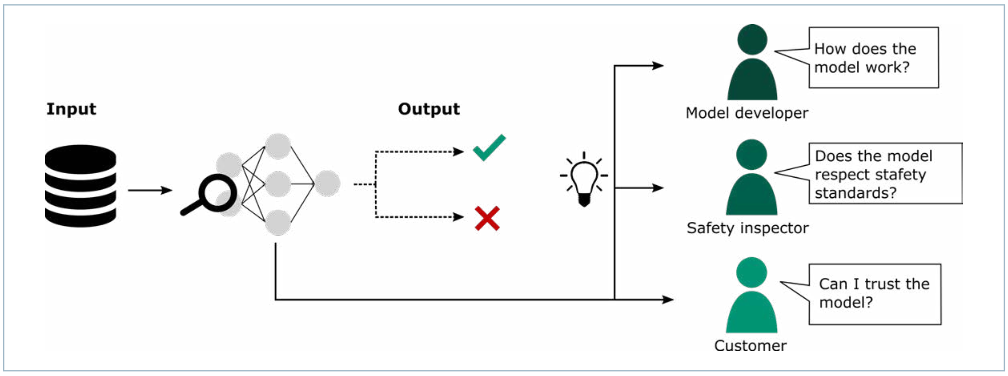

Figure 3:
Making machine learning methods explainable is important for different interest groups.

In addition to explaining the ML model, the data underlying the model also play a central role. The software toolbox provides methods that quantify the influence of the training data on the learned model. In this way, it is possible to determine which data are particularly valuable and in which areas there is insufficient data available.

Your advantages

This gives you as a user the following advantages:

Model validation
Check if your ML models work as expected. This aspect is particularly relevant for safety-critical applications, which, for example, have to be approved by an internal or external test center.

Model debugging
If you find out that the model is making wrong decisions, these can be examined in detail based on the characteristics on which the decision was made (see Figure 2). Debugging enables reasons why the model is malfunctioning to be identified and subsequently eliminated.

Data Value
Support a data-centric view that allows you to evaluate the quality and value of your data. You will gain insight into which data particularly influences the ML model. In addition, you will be shown which data is superfluous and which data still needs to be collected for a high quality ML model.

Acceptance and trust
Insufficient understanding of the decisions made by highly complex ML algorithms often makes people very reluctant to start using them. Explanations for decisions made by algorithms can strengthen the user's faith in the system and lead to a higher level of acceptance.

Gaining insight
By studying the learned ML model, general correlations can be understood, e.g., the relevance of particular input data and the way individual input data interact with each other.

Contact

Prof. Dr.-Ing. Marco Huber
Phone +49 711 970-1960
marco.huber@ipa.fraunhofer.de

M.Sc. Danilo Brajovic
Phone +49 711 970-3647
danilo.brajovic@ipa.fraunhofer.de

Fraunhofer-Institute for Manufacturing Engineering and Automation IPA
Nobelstrasse 12, 70569 Stuttgart
www.fraunhofer.de