Towards Explainable Artificial Intelligence for Centrifugal Compressor Operating Conditions Classification

Przemysław Kucharski
Lodz University of Technology
Institute of Applied Computer Science
POLAND

pkuchars@iis.p.lodz.pl

Bartosz Kowalewski
Lodz University of Technology
Institute of Applied Computer Science
POLAND

Mateusz Stajuda
University of Edinburgh
Institute for Infrastructure and Environment
UNITED KINGDOM

mateusz.stajuda@ed.ac.uk

Grzegorz Liśkiewicz
Lodz University of Technology
Institute of Turbomachinery
POLAND

grzegorz.liskiewicz@p.lodz.pl

Keywords: Distributed Engine Control, Artificial Intelligence, Explainability, Aerodynamic Instabilities, Centrifugal Compressors.

ONEPAGER

The aim of the study is to present the possible path towards creating an explainable artificial intelligence model for classification of a centrifugal compressor operating conditions. The operation of a convolution network classifier, being a base for explainable system was presented with use of a pressure signal from different locations inside the machine. It was trained with a use of quasi-dynamic measurements, being a series of separate measurements for a given operating conditions. The expert knowledge, which can be implemented into the Artificial Intelligence model to increase its explainability comes from the non-linear signal analysis methods that could be implemented to find characteristic features that can be tied to specific operating conditions. The resulting explainable model should be able to provide a number of new, potentially better features for detection of operating conditions or demonstrate which of the currently applied features are the most valuable for identification of operating conditions.

AI and machine learning methods can be effective in detection and classification of dangerous phenomena in turbomachines. Preliminary studies performed by the authors show that with simple fully-connected neural networks accuracy of over 90% can be achieved. However, even with simple deep networks, an attempt to explain its behaviour can be very difficult. The main problem is that the network is, in theory, trained to generalize the problem that is introduced to it with the use of available data – which can lead to unexpected

behaviour if data contains patterns that the user is not aware of. Furthermore, in some cases, the phenomena which the model was created to detect are underrepresented in measurement data – e.g. because triggering some events in real life scenarios may lead to serious damage.

Another interesting aspect of a way towards explainable AI models in machine disfunction monitoring is that, with explicit introduction of known patterns, model based on the available data can be trained to detect more subtle patterns, that could not be extracted with a usual approach in supervised learning. These known patterns can come directly from expert knowledge or can arise from non-linear signal analysis methods, which when used properly can serve as experts. Such methods, used in the field of compressor instabilities detection are Continuous Wavelet Transform, Singular Spectrum Analysis or Empirical Mode Decomposition. These methods can also serve as a stand-alone detectors, but they are often inferior to black boxes - complicated deep networks that are applied in the industry to solve this class of problems.

There are numerous ways to deal with black-box models, one of which is feature analysis to find those that have major contributions to solving the problem. The shortcomings of such approaches include the fact that only explicitly defined features can be analyzed, which limits the preprocessing stage to known methods. The approach explored in this study is quite different, since it tries to constraint the reasoning of a machine learning model, fed with raw measurement data, so that the knowledge built in the model can be easily interpreted and modified. This allows to introduce explicit knowledge during learning.

The main assumption made in our work is that the model can be restricted to minimize the spread of information across the learned features, so that the mutual information for any subset of features is as small as possible. This promotes modularity and interpretability of said features. Although learning takes longer and can be under some circumstances less accurate that traditional models. Initial research shows that the classification results can be transferred back to observable phenomena in the measurement data.

It is expected that extracting knowledge from models trained under this constraints may lead to finding less obvious features and, more importantly, the knowledge base can be transferred to systems with different properties, where model with knowledge introduced during initialization stage has to be adjusted, not trained from scratch.