Evolutionary Intelligence (2024) 17:417–418
https://doi.org/10.1007/s12065-024-00906-x

EDITORIAL

Special issue on "Towards robust explainable and interpretable artificial intelligence"

Stefania Tomasiello[1,2] · Feng Feng[3] · Yichuan Zhao[4]

Published online: 17 January 2024

The complexity of modern artificial intelligence (AI) models has raised the question about their interpretability. The terms interpretability and explainability have been used interchangeably by researchers. These two terms sound very closely related, although they have to be meant differently. Interpretability is mostly related to the outcome of the cause-and-effect relationship given the inputs of a system. Explainability deals with the internal logic of a machine learning system. It aims to characterize model accuracy and transparency in AI-powered decision making. It is clear that there is a need for a proper mathematical formalism that is still missing. Hence, there is a trade-off between the performance of a machine learning model and its ability to produce explainable and interpretable predictions. The study of robust systems, which are also explainable and interpretable is still under way.

It is also worth mentioning that explainability and interpretability have become a requirement to comply with government regulations for sensitive applications, such as in finance, public health, and transportation. In fact, this issue has received attention from the European Parliament whose general data protection regulation recognizes the right to receive an explanation for algorithmic decisions. This also justifies the attention on this topic.

This special issue aims to collect some latest advances in the field, exploring the application of interpretable or explainable approaches in different areas. The special issue comprises fifteen papers, resulting from a meticulous selection, with an acceptance rate of less than 40%.

The contributions deal with interpretable and explainable approaches applied to problems in the medical and non-medical fields. Regarding the first group, there are:

- diabetic risk prognosis by means of tree ensemble machine learning models;
- multiple myeloma chemotherapy treatment recognition by using explainable AI tools jointly with the most popular black-box models;
- intravenous drug administration through a control system based on linear fractional-fuzzy differential equations;
- detection of activity patterns in electromyographic signals via decision trees.

The other contributions focus on.

- an interpretable classifier for spoken transcripts and written text;
- a template-based algorithm for automatic recognition of music chords;
- a fuzzy version of the resource description framework in the context of the semantic web;
- an attention-based named entity recognition for Chinese agricultural diseases and pests;
- a hybrid improved particle swarm optimization and differential evolution for a localization problem in wireless sensor networks;
- the use of a swin transformer for locating transmission line defects;

✉ Stefania Tomasiello
stomasiello@unisa.it; stefania.tomasiello@ut.ee

Feng Feng
fengf@xupt.edu.cn; fengnix@hotmail.com

Yichuan Zhao
yichuan@gsu.edu

[1] Department of Industrial Engineering, University of Salerno, Salerno, Italy

[2] Institute of Computer Science, University of Tartu, Tartu, Estonia

[3] Department of Applied Mathematics, Xi'an University of Posts and Telecommunications, Xi'an, China

[4] Department of Mathematics and Statistics, Georgia State University, Atlanta, USA

- a fused graph attention network for aspect-based sentiment analysis;
- generative adversarial networks for fingerprint image denoising and inpainting.

Most applications exploited existing concepts and tools for interpretability and explainability.

From a theoretical perspective, the main contributions deal with.

- conceptual interpretations of fuzzy differential equations;
- a post-hoc method for the interpretability of highly adaptive lassoes, termed highly adaptive regression trees;
- chaotic adaptive butterfly optimization algorithm and its visualization.

Acknowledgements We would like to thank all the authors for their contributions and reviewers for their timely and constructive comments. Stefania Tomasiello acknowledges support from the European Social Fund via the IT Academy programme and the Estonian Research Council, grant PRG1604.