ercim-news.ercim.eu Number 134 July 2023

ERCIM NEWS

Special theme:

eXplainable AI

Also in this issue

Research and Innovation:

Fibre-Optic Sensing for Road-Traffic Monitoring in Remote Areas

Actually, analysing neighbourhoods – rather than just attribute values – is the key advantage of BrightBox. This means that diagnostic attributes can reveal valuable characteristics of specific data instances, providing domain experts and data scientists with the needed information to create better prediction models. Focusing on neighbourhoods, the technology offers deeper insights and more meaningful guidance for model improvement.

Our goal is to further develop BrightBox in practical applications (see e.g. [2]). This approach can be highly beneficial in, for example, detecting errors in models submitted for online data science competitions, allowing their organisers and sponsors to gain valuable insights into the performance of each solution. Such insights can then suggest improvements to the models and facilitate their deployment in production-ready environments. In addition, it is possible to construct improved solutions from the existing ones. For instance, by using information about different degrees of risk (uncertainty) of decision-making by different classifiers of an ensemble, it is possible to modify the procedure for resolving conflicts in voting.

References:

[1] A. Janusz, et al., "BrightBox – a rough set based technology for diagnosing mistakes of machine learning models," Applied Soft Computing, vol. 141, pp. 110285, 2023. https://doi.org/10.1016/j.asoc.2023.110285

[2] A. Janusz and D. Ślęzak, "KnowledgePit meets BrightBox: a step toward insightful investigation of the results of data science competitions," in Proc. FedCSIS 2023 in ACSIS vol. 30, pp. 393-398. https://doi.org/10.15439/2022F309

Please contact:

Andżelika Zalewska-Küpçü,
QED Software, Poland
andzelika.zalewska-kupcu@qed.pl

Dominik Ślęzak,
University of Warsaw & QED Software, Poland
slezak@mimuw.edu.pl

An Explanation that LASTS: Understanding Any Time Series Classifier

by Francesco Spinnato (Scuola Normale Superiore and CNR-ISTI), Riccardo Guidotti (University of Pisa) and Anna Monreale (University of Pisa)

We present LASTS, an XAI framework that addresses the lack of explainability in black-box time series classifiers. LASTS utilises saliency maps, instance-based explanations and rule-based explanations to provide interpretable insights into the predictions made by these classifiers. LASTS aims to bridge the gap between accuracy and explainability, specifically in critical domains.

In recent years, the availability of high-dimensional time series data has led to the widespread usage of time series classifiers in various domains, including health care and finance. These classifiers play a crucial role in applications such as anomaly detection in stock markets and the automated diagnosis of heart diseases.

The existing landscape of time series classifiers encompasses a range of approaches. However, despite their effectiveness in achieving high classification accuracy, most of these classifiers suffer from a critical limitation: they are black-box models, offering little insight into their decision-making process [1].

The lack of explainability in black-box time series classifiers poses challenges, particularly in critical domains where human experts must understand the reasons behind the model's predictions. In applications such as clinical diagnosis, the interpretability of the models used by artificial intelligence (AI) systems becomes essential for building trust and facilitating reliable interaction between machines and human experts. Meaningful explanations can enhance the cognitive ability of domain experts, allowing them to make informed decisions and supporting AI accountability and responsibility in the decision-making process [L1].

We propose the LASTS (Local Agnostic Subsequence-based Time Series Explainer) framework to address the need for explainability in time series classification, providing interpretable explanations for any black-box predictor. By unveiling the logic behind the decisions made by these classifiers, LASTS enhances transparency and facilitates a deeper understanding of the classification process. The first version of LASTS was published in [2]. Since then, we have made significant advancements, introducing heterogeneous explanations and a novel salience map extraction to explain both univariate and multivariate time series. Compared to the previous version, these enhancements provide a more comprehensive, interpretable, and versatile approach to explaining black-box time series classifiers.

The input to the LASTS framework is a time series, X, and a black-box classifier, while the output is an explanation for the black-box's decision. The explanation comprises three parts: a

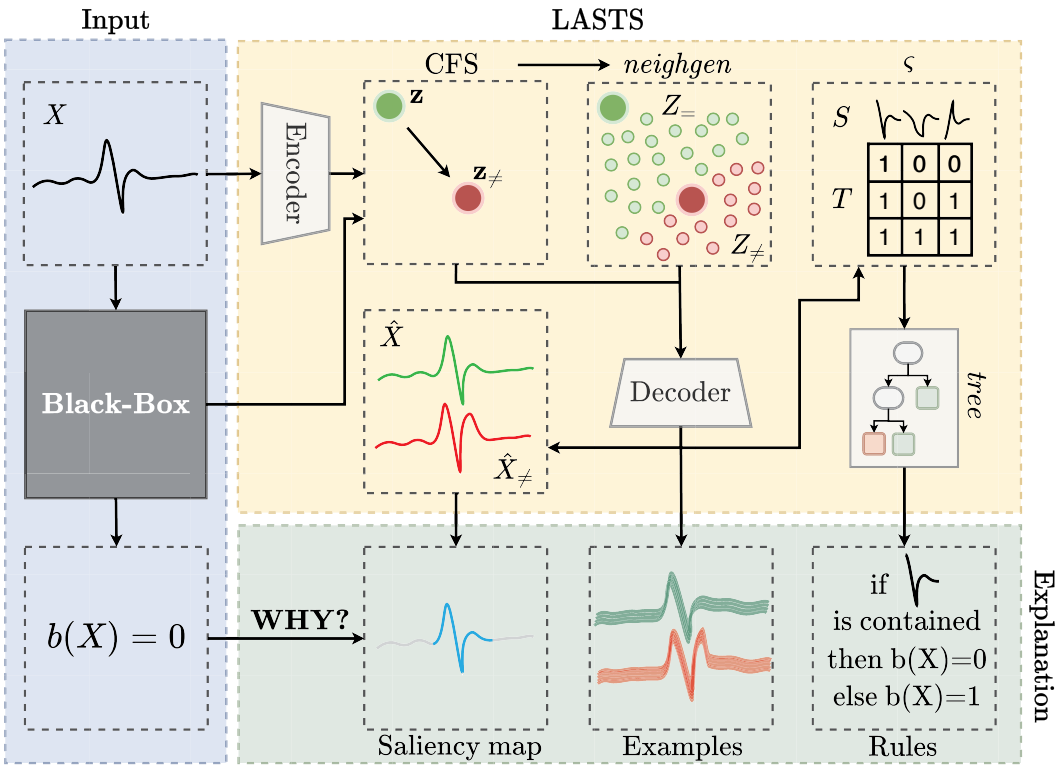

Figure 1: A comprehensive schema of LASTS.

saliency map, an instance-based explanation, and a rule-based explanation (a schema is presented in Figure 1). The saliency map highlights the most influential parts of the time series that contribute to the classifier's decision. This visualisation provides an immediate assessment of the critical timesteps, enabling users to better understand the driving factors behind the classification outcome. The instance-based explanation employs a set of exemplar and counterexemplar time series, offering concrete examples that align with or diverge from the black-box classifier's decision. These instances help identify common patterns and highlight the necessary modifications for obtaining different classification outcomes. Finally, the rule-based explanation utilises logical conditions based on interpretable time series subsequences, providing factual and counterfactual rules that reveal the reasons for the classification. The novel component of LASTS is its integration of multiple explanation components, which provide a comprehensive and interpretable set of explanations, that can be useful for different kinds of users.

From a technical standpoint, the framework leverages a trained variational autoencoder to encode and decode time series into a latent space. Once the input time series is encoded into its latent representation, LASTS finds the closest counterexemplar using a novel search algorithm, generating a synthetic neighbourhood around the black-box's decision boundary. The neighbourhood is then decoded, obtaining the black-box predictions for these synthetic instances. The saliency map is extracted based on the distance between X and the closest decoded counterexemplar, while other exemplar and counterexemplar instances are derived from the neighbourhood. Lastly, the neighbourhood is transformed into a set of subsequences, and a decision-tree surrogate is trained on the transformed data to extract the factual and counterfactual rules.

An example explanation for a real electrocardiogram from the ECG5000 dataset [L2] is presented in Figure 2. The black box classifies the instance being explained as a "normal" heartbeat. The explanation for this prediction highlights the main differ-

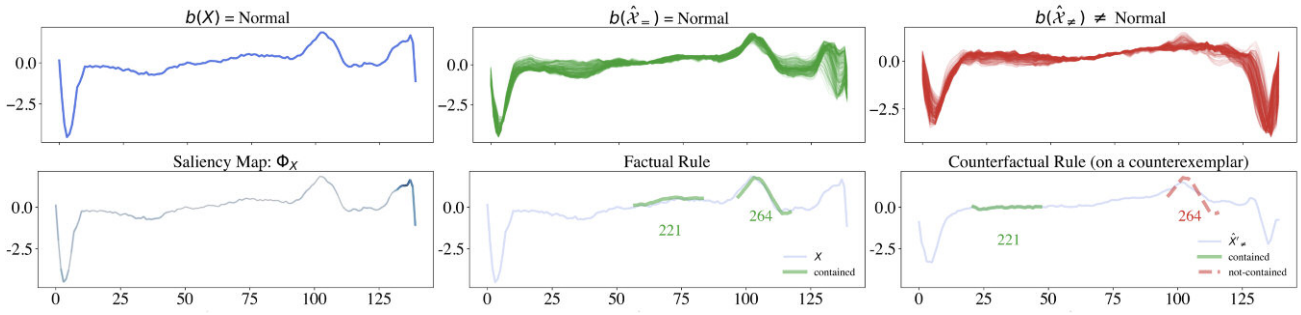

Figure 2: Explanation of a black-box prediction for a heartbeat of the ECG5000 dataset. From left to right: (top) instance to explain, exemplars, counterexemplars, (bottom) saliency map, factual rule and counterfactual rule (shown over a counterexemplar).

ence between normal and abnormal heartbeats: the lower and evident V-shape in the rightmost part of the time series. The saliency map confirms this observation by highlighting the last observations of the time series. The rules further illustrate the differences between classes, with the factual rule indicating the presence of a specific subsequence in normal instances and the counterfactual rule indicating its absence in abnormal instances. While the saliency map and rules may not cover the exact same areas, they provide complementary insights into the discriminative features of the time series.

Overall, LASTS represents a significant advancement in the field of time series explanation, with promising potential for future research and practical applications. In our future research, we plan to explore several directions to enhance LASTS. Firstly, we intend to evaluate the framework on longer and more complex real-world time series datasets, aiming to validate its performance in challenging scenarios. Additionally, we aim to extend LASTS to other types of sequential data, such as trajectories, text, and shopping transactions, in order to broaden its scope of applicability. Secondly, we will delve deeper into the relationship between the latent and subsequence spaces, conducting further investigations to gain a comprehensive understanding of their interactions. Finally, we intend to conduct human decision-making tasks guided by LASTS explanations, offering practical evaluation and valuable insights into the effectiveness of the explanations in real-world decision scenarios.

This article is coauthored with Mirco Nanni, Fosca Giannotti, and Dino Pedreschi (CNR-ISTI, Scuola Normale Superiore, Università di Pisa).

Links:
[L1] https://artificialintelligenceact.eu/
[L2] https://kwz.me/hxK

References:
[1] A. Theissler, et al., "Explainable AI for time series classification: a review, taxonomy and research directions," IEEE Access, 2022.
[2] R. Guidotti, et al., "Explaining any time series classifier," 2020 IEEE 2nd Int. Conf. on Cognitive Machine Intelligence (CogMI), IEEE, 2020.

Please contact:
Francesco Spinnato
Scuola Normale Superiore and CNR-ISTI, Pisa, Italy
francesco.spinnato@sns.it
Riccardo Guidotti
University of Pisa, Pisa, Italy
riccardo.guidotti@unipi.it

Explaining Ensemble Models for Lung Ultrasound Classification

by Antonio Bruno, Giacomo Ignesti and Massimo Martinelli (CNR-ISTI)

Correct classification is the main aspect in evaluating the quality of an artificial intelligence system, but what happens when you reach top accuracy and no method explains how it works? In our study, we aim at addressing the black-box problem using an ad-hoc built classifier for lung ultrasound images.

In the last few years, the novelties of artificial intelligence (AI) and computer vision (CV) significantly increased, allowing new algorithms to obtain meaningful information from digital images. Medicine is a field in which the use of this technology is experiencing fast growth. In 2020, in the USA alone, the production of 600 million medical images was reported, and this number seems to increase steadily. Robust and trustworthy algorithms need to be developed in a multidisciplinary collaboration.

During the SARS-CoV-2 pandemic, a fast and safe response became even more necessary. The use of point-of-care ultrasound (POCUS) to detect SARS-CoV-2 (viral) pneumonia and the bacterial infection emerged as one of the most peculiar emerging case studies, involving the use of on-site ultrasound examinations rather than a dedicated facility. As well as being faster, safer and less expensive, lung ultrasound (LUS) also appears to detect signs of lung diseases as well as or even better than other methods, such as X-ray and computed tomography (CT).

The employment of lung POCUS seemed an optimal solution for both quarantined and hospitalised subjects. CT and magnetic resonance imaging (MRI) are far more precise and reliable examinations, but both have downsides over mass screening. In our study, an efficient adaptive minimal ensembling model was developed to classify LUS using the largest publicly available dataset, the COVID-19 lung ultrasound dataset [L1], composed of 261 ultrasound videos and images from 216 different patients. The General Data Protection Regulation

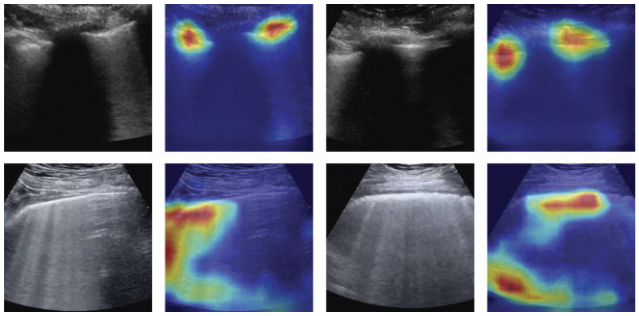

Figure 1: COVID-19 – Original and Grad-CAM-processed samples are shown for subjects with COVID-19; different images within different subjects show similar activation maps.