Explaining Time Series Classifiers Through Post-Hoc XAI Methods Capturing Temporal Dependencies

Ephrem T. Mekonnen[1,2]

[1]School of Computer Science, Technological University Dublin, Ireland
[2]Artificial Intelligence and Cognitive Load Research Lab, Technological University Dublin, Ireland

Abstract

Time series classification is essential in domains such as healthcare and finance, where accurate predictions can have significant real-world consequences. However, in many high-stakes applications, understanding why a model makes a certain decision is just as important as the prediction itself. While deep learning models excel at capturing complex temporal patterns, their black-box nature limits transparency, making it difficult to trust and interpret their decisions. Although eXplainable AI (XAI) methods have advanced considerably for image and tabular data, applying them to time series remains challenging due to the intricate temporal dependencies and high dimensionality of the data. Post-hoc model-agnostic XAI techniques offer a promising solution by providing explanations without altering the underlying model. This research focuses on developing novel post-hoc model-agnostic XAI methods specifically for time series classifiers. By elucidating prediction processes while preserving temporal structures, these methods seek to enhance interpretability and trust, thereby facilitating informed decision-making in high-stakes applications.

Keywords

Explainable Artificial Intelligence, Time series, Model-agnostic, Post-hoc, Deep learning, XAI

1. Introduction

Time series data are crucial in many real-world applications, from healthcare and finance to environmental monitoring. With the growing availability of such data, machine learning models, particularly deep learning approaches, have demonstrated impressive performance in classifying complex temporal patterns. However, these models often function as "black boxes," making it difficult to understand their decision-making processes. In high-stakes scenarios, where interpretability and accountability are essential, this lack of transparency poses significant challenges. Explainable Artificial Intelligence (XAI) has emerged as a line of research aimed at addressing these issues by developing methods that provide insight into model predictions [1, 2].

Although XAI has seen significant advances, most methods have been developed for image and tabular data, which do not share the same characteristics as time series data. The high dimensionality and strong temporal dependencies of time series pose unique challenges that many existing XAI techniques struggle to handle. Popular approaches such as Local Interpretable Model-agnostic Explanations (LIME) [3], Saliency Maps [4], and Layer-wise Relevance Propagation (LRP) [5] have been adapted from computer vision tasks [6]. However, these methods often rely on visual heatmaps, which can be difficult to interpret and are mainly designed for developers rather than end users [7, 8]. Furthermore, feature attribution techniques such as LIME and SHAP tend to ignore temporal dependencies, treating each time step or segment as an independent feature [6, 2].

This research aims to develop novel post-hoc, model-agnostic XAI methods specifically tailored for deep learning-based time series classifiers. By addressing the unique challenges of interpretability in time series classification, this study seeks to fill a critical gap in the XAI landscape.

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey
D22125038@mytudublin.ie (E. T. Mekonnen)
https://ephrem-eth.github.io/ (E. T. Mekonnen)
0009-0009-3035-3441 (E. T. Mekonnen)

2. Related Work

Research in Explainable AI (XAI) for time series classification has evolved alongside broader developments in XAI. Nonetheless, the majority of existing techniques have primarily been designed for static data types, such as images and tabular data, which do not encapsulate the temporal dependencies inherent in time series data. Consequently, adapting these techniques to effectively address the unique challenges presented by temporal dependencies remains a non-trivial task [2, 7].

XAI methods can be categorised based on several criteria, one of which distinguishes between model-specific and model-agnostic approaches. Model-specific methods leverage the internal mechanisms of black-box models, whereas model-agnostic methods are applicable across various model architectures. Notably, many of these methods serve as post-hoc explanations, offering interpretability without requiring changes to the underlying model and are applied after model training.

2.1. Model-Specific Methods

A range of model-specific methods has been employed to elucidate black-box models trained on time series data. For instance, Class Activation Mapping (CAM), introduced by Zhou et al. [9], has been adapted for explaining convolutional neural network (CNN)-based models in time series classification. CAM identifies class-relevant regions by projecting weighted activations from the final convolutional layer onto a feature map. However, its implementation necessitates a specific CNN architecture featuring Global Average Pooling (GAP), thus limiting its broader applicability. In contrast, Grad-CAM [10] extends CAM by utilising gradient information from the last convolutional layer to identify critical regions and generate saliency maps, making it more adaptable to various CNN architectures without strict architectural constraints.

Schlegel et al. [6] have also investigated several standard XAI techniques, including Layer-wise Relevance Propagation (LRP) [5], to interpret deep learning-based time series classification models. Specifically, the work presented in [11] introduced DFT-LRP, a tailored variant of LRP designed to address the complexities associated with time series analysis by incorporating a virtual inspection layer. Despite these advancements, many of these explanations tend to be developer-centric and rely on latent layer activations, whereas end users often require higher-level, abstract explanations that enhance overall interpretability [12]. Most XAI methods focus on local explanations, but some also provide global insights for time series classifiers. For example, [13] extends CAM to all training samples within a class, generating an average CAM to highlight key discriminative features. Similarly, TsViz [14] identifies important input regions and evaluates filter importance for a given prediction. It also derives global insights by clustering filters with similar activation patterns, as they likely capture the same underlying concepts.

2.2. Model-Agnostic Methods

Conversely, model-agnostic methods, such as LIME [3] and SHAP [15], offer greater applicability across different model types and are particularly relevant in a post-hoc context. LIME approximates complex model predictions by creating a locally interpretable model, such as a linear classifier, around the instance to be explained. This process begins with the generation of a dataset of perturbed samples near the target instance, using the original model's predictions for these samples. A linear model is subsequently trained on this dataset, with samples weighted according to their proximity to the target instance, yielding feature importance scores as regression coefficients.

SHAP, rooted in cooperative game theory, explains individual predictions by attributing feature contributions through Shapley values. Although originally developed for image and tabular data, Schlegel et al. [6] adapted SHAP methods for interpreting time series classifiers. However, this adaptation often overlooks the temporal dependencies present in time series data, as each time step is treated as an independent feature. To mitigate these limitations, Guilleme et al. [16] and Neves et al. [17] have tailored LIME for deep learning-based time series classification by utilising longer segments

for perturbation. However, these approaches are constrained by fixed window sizes, which can limit their effectiveness.

Additionally, Sivill et al. [18] introduced the NNsegment method, which identifies homogeneous regions within time series data and employs diverse perturbation techniques to yield more robust explanations. Further, Schlegel et al. [12] expanded LIME by incorporating six segmentation methods; however, understanding the significance of the identified segments continues to present challenges.

2.3. Challenges in Visualisation and Interpretation

Moreover, many attribution-based and attention-based XAI methods [19] predominantly utilise heatmaps to visualise feature attributions. While such visual representations can be advantageous for domain experts, they often pose significant interpretability challenges for general users [8]. This underscores the necessity for more intuitive and user-friendly explanation techniques that extend beyond traditional heatmaps, facilitating clearer reasoning behind model predictions.

3. Research Questions and Objectives

The overall goal of this research is to develop novel model-agnostic post-hoc explainable artificial intelligence (XAI) methods specifically designed for time series classifiers by leveraging Parameterised Event Primitives (PEPs). PEPs provide a structured approach to defining and extracting events in a time series. An event is a specific pattern or behaviour expected to occur within the domain. Extracting PEPs from time series data enables the representation of temporal characteristics as parameters, facilitating the training of interpretable models such as decision trees [20]. These events can be characterised using PEPs, which include increasing and decreasing trends defined by parameters such as start time, duration, and average gradient, as well as local maxima and minima, which are characterised by the time they occur and their corresponding values. This parameterisation offers an intuitive and meaningful representation of temporal structures, enhancing the interpretability of time series models.

Furthermore, PEPs offer a significant advantage by eliminating the need for explicit segmentation of time series data. Segmentation can be complex and subjective, often overlooking critical patterns that exist between segments [12]. Conversely, treating each time step as an independent feature obscures the essential temporal dynamics that characterise time series, potentially leading to misinterpretations of the underlying patterns. This study is structured around the following key question:

How can model-agnostic XAI methods be designed to provide interpretable, faithful explanations while capturing temporal dynamics in the data for black-box time-series classifiers?

- To develop a global model-agnostic XAI method that approximates the inference process of deep learning-based time-series classifiers using decision trees, generating interpretable rule-based explanations while preserving temporal dependencies.
- To design a local post-hoc XAI method that generates instance-specific explanations for predictions made by time series classifiers, effectively capturing the temporal dynamics of the data.
- To identify and establish a set of evaluation metrics specifically tailored to evaluate the faithfulness and interpretability of the explanations generated by the proposed XAI methods for time-series classifiers.
- To investigate and develop methodologies for constructing global explanations for time series classifiers by aggregating local explanations, ensuring that the process maintains interpretability and adequately captures the temporal dependencies present in the data.

4. Research Plan and Methodology

To achieve these objectives, this research uses publicly available time series datasets from the UCR archive [21] with minimal preprocessing. The study involves training a time series classifier and

evaluating its performance before applying the proposed XAI methods. The research is structured into three main work packages, of which the first two have been done thus far, while the third is planned for the future. An overview of the research plan is provided in Figure 1.

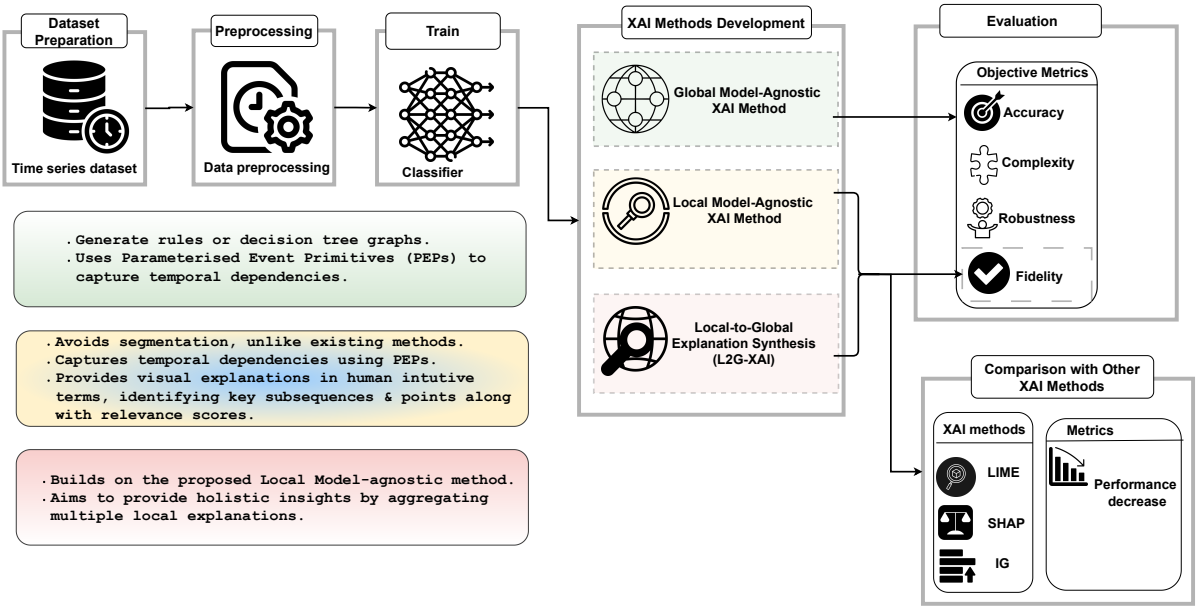


Figure 1: Overview of the proposed research plan.

The first work package focuses on developing a global model-agnostic XAI method that provides explanations in the form of decision rules or decision tree graphs. This process begins with initial data preprocessing to train and evaluate the time series classifier. Next, the test set is transformed using Parameterised Event Primitives (PEPs) to train the decision tree. This involves extracting PEPs from the test set and clustering them using an appropriate algorithm. The frequency of events belonging to each cluster is then counted to create a dataframe, where columns represent clusters and rows represent instances. The dataframes for various Parameterised Event Primitives (PEPs), including increasing and decreasing trends, as well as local maxima and minima, are subsequently merged. Finally, the decision tree is trained and evaluated using the transformed test set and the model's predictions. The method was rigorously evaluated using accuracy, fidelity, complexity, and robustness.

The second work package, currently under review, focuses on developing LOMATCE (Local Model-Agnostic Time-Series Classification Explanation), a local model-agnostic explainable artificial intelligence (XAI) method that generates instance-specific explanations for time series classifiers without requiring segmentation. LOMATCE begins by creating neighbouring instances around the instance to be explained and computes weights for these instances based on their distance from the original instance, elucidating their influence on the model's prediction. The neighbouring samples are transformed using Parameterised Event Primitives (PEPs), following a sequence of steps similar to the global model-agnostic method. Using the transformed data, the method employs the weights of the neighbouring samples along with predictions from the trained black-box classifier to train a linear regression model. From this model, the top cluster is identified based on the coefficients derived from the linear regression model. Finally, the events extracted from the original instance that belong to the identified top clusters are highlighted, along with their corresponding importance scores, thereby providing a clear and interpretable explanation of the original instance's prediction.

The third work package will explore the aggregation of local explanations generated by LOMATCE to construct global explanations, referred to as Local to Global eXplanation (L2GX). Drawing inspiration from Submodular-Pick LIME (SP-LIME) [3], L2GX first generates local explanations for each instance using LOMATCE, identifying the most important PEP clusters along with their corresponding importance scores. Similar clusters are merged to enhance coherence and reduce redundancy. An

instance-cluster matrix is then constructed, where rows represent instances, columns denote clusters, and each cell captures the importance score of a given cluster for the respective instance. This matrix serves as the foundation for computing global importance scores. Using a predefined budget B, L2GX employs submodular optimization techniques to select the top B instances that maximise coverage of clusters with positive importance scores, ensuring diverse contributions of information. Finally, PEPs from the selected instances are extracted, retaining only those belonging to the identified clusters as global explanations. This method will be evaluated based on faithfulness using fidelity metrics and compared to other XAI methods using performance drop metrics.

To the best of my knowledge, the proposed global model-agnostic XAI method represents the first attempt to generate global explanations for time series classifiers in the form of rules or decision tree graphs. As illustrated in the Figure 1, this study does not include a comparative analysis with existing methods but rather focuses on the development and evaluation of novel techniques tailored for time series classification.

5. Results and Contributions to Date

To date, I have been focusing on the first and second work packages, specifically the development of global and local model-agnostic XAI methods tailored for time series classifiers. One of the contributions is a global model-agnostic XAI method designed to generate explanations in the form of decision rules or decision graphs. This approach enhances the interpretability of model predictions by clearly identifying the time steps that significantly influence outcomes. Initial findings were showcased as late-breaking work at the XAI-2023 [22], followed by a full paper published in Frontiers in Artificial Intelligence [23]. The effectiveness of this method was evaluated using various objective metrics, including accuracy, fidelity, depth, number of nodes, and robustness, demonstrating that the decision tree graph effectively highlights crucial time steps, thereby facilitating a better understanding of the model's predictions (see Figure 2).

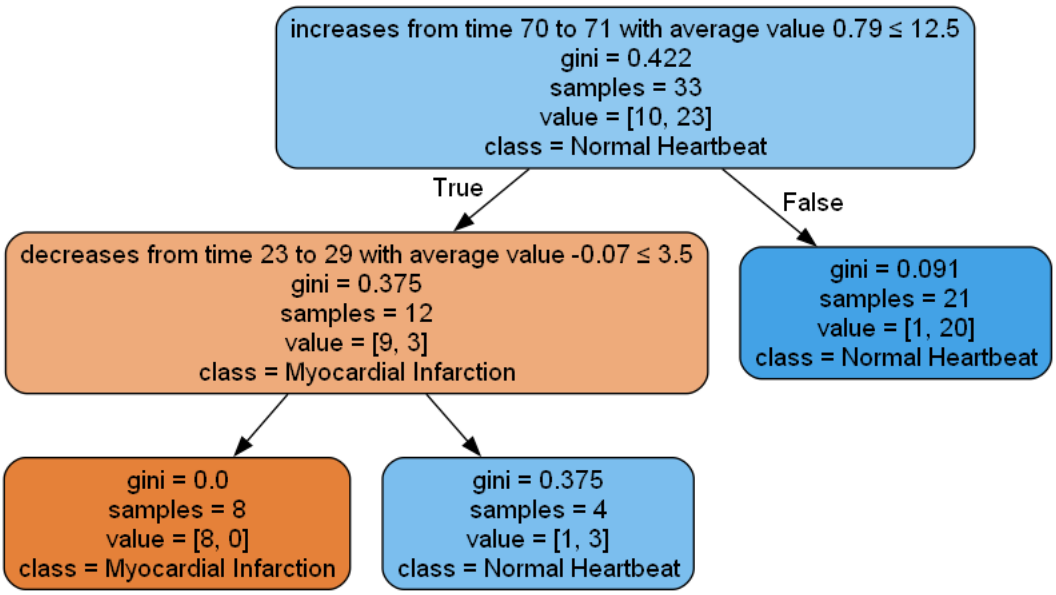


Figure 2: Visualization of decision tree graph produced by the proposed Global-model-agnostic method applied to ECG data for the FCN model .

Additionally, a local model-agnostic XAI method known as LOMATCE has been introduced, with preliminary results presented at XAI-2024 [24]. A full paper is currently under review at the IEEE Transactions on Artificial Intelligence (IEEE-TAI). LOMATCE provides instance-specific explanations by leveraging Parameterised Event Primitives (PEPs) to capture temporal dependencies and train a simple linear surrogate model. This method effectively captures essential patterns, such as increasing

and decreasing trends, as well as local maxima and minima, thereby offering valuable insights into the model's decision-making process (illustrated in Figure 3). The explanations generated by LOMATCE were compared with those produced by established methods, including Integrated Gradients (IG), LIME, and SHAP. For visual comparison, Figures 4 and 5 present two approaches to applying LIME and SHAP for time series data: (1) the segment-based approach, which involves dividing the time series into frames and assigning relevance to each segment. However, this method may overlook important relationships between segments and can yield different explanations depending on the selected segment width, thereby complicating the determination of optimal segmentation; and (2) the approach that treats each time step as a separate feature, which can obscure critical temporal dynamics. An evaluation of LOMATCE across various perturbation strategies, which are used to generate neighbouring samples around the instance to be explained, indicated that the choice of perturbation method plays a crucial role in ensuring the faithfulness of the explanations. Comparative analysis shows that LOMATCE performs competitively across diverse datasets, occasionally outperforming LIME and SHAP in terms of both interpretability and accuracy.

These ongoing efforts represent significant strides toward improving the transparency and interpretability of time series classifiers while effectively addressing the challenge of capturing temporal dependencies.

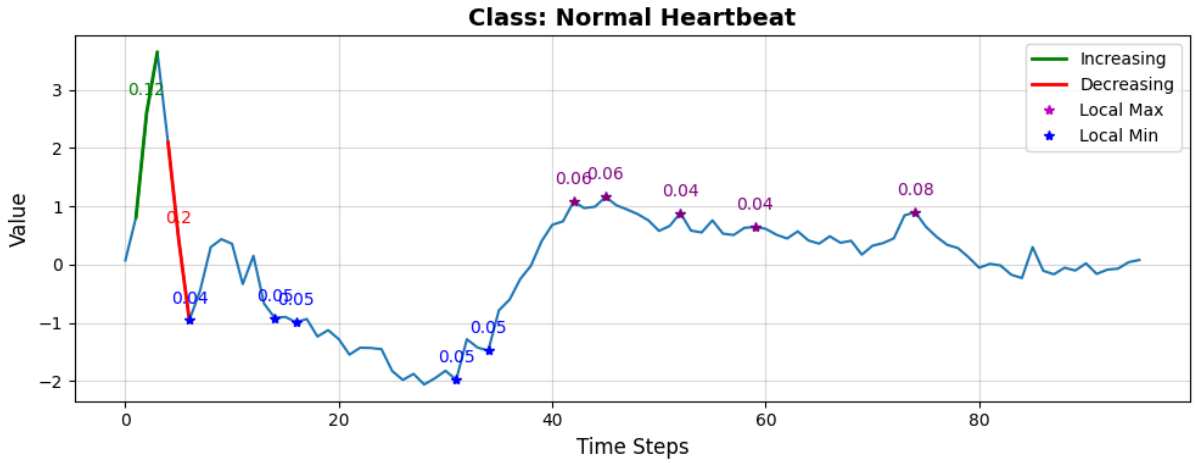
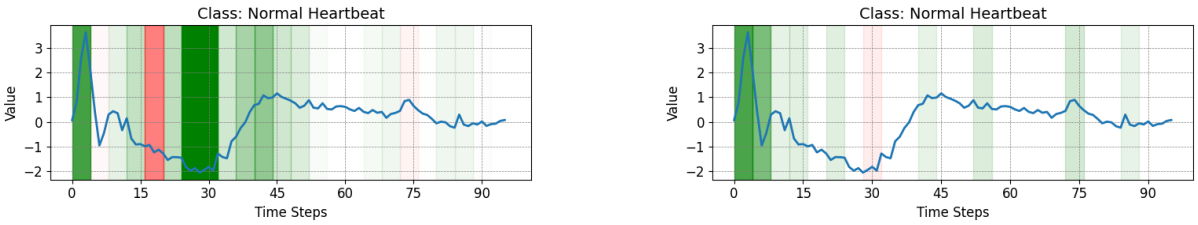


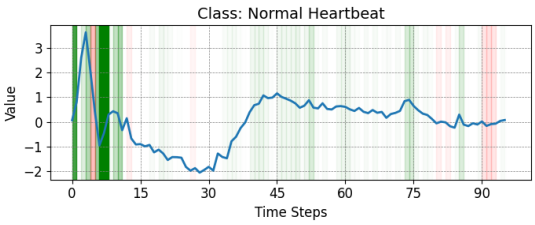
Figure 3: The explanation generated by the LOMATCE highlights segment significance, relevance scores, and event types (e.g., increasing, decreasing, local maximum, local minimum) in the time series data for the black box model.



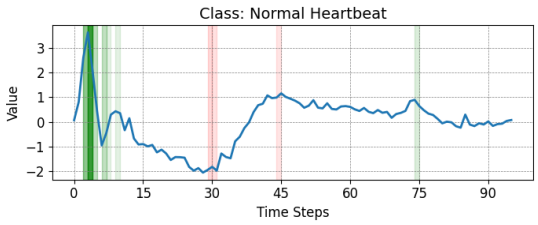
(a) SHAP explanation (segment-based).

(b) LIME explanation (segment-based).

Figure 4: Explanations generated by (a) SHAP (b) LIME for ECG200 instance of normal heartbeat class.



(a) SHAP explanation (point-based).



(b) LIME explanation (point-based).

Figure 5: Explanations generated by (a) SHAP (b) LIME for ECG200 instance of normal heartbeat class.

6. Next Research Step and Expected Final Contribution

The next phase of this research involves implementing the Local-to-Global eXplanation(L2GX) method, which aggregates local explanations to generate coherent global insights while preserving interpretability. Furthermore, the proposed methods will be rigorously validated across diverse univariate time series datasets to assess their robustness and extended to multivariate time series, addressing the added complexity of interdependent temporal variables and their dynamic relationships.

The expected final contribution of this research is the development of novel model-agnostic explainable AI (XAI) methods tailored for time series classifiers. A key component of this work is LOMATCE, a local model-agnostic XAI method that generates faithful, instance-specific explanations while capturing temporal dependencies without requiring predefined segmentation. Building on this, a second global method constructs global explanations by aggregating local explanations from selected instances, ensuring that the broader model behaviour is captured while preserving temporal dynamics. Additionally, a separate global rule-based XAI method approximates the decision-making process of black-box models using decision trees, providing interpretable explanations in the form of decision rules. These contributions aim to enhance the transparency, reliability, and interpretability of time series classifiers, facilitating their adoption in real-world applications where explainability is critical.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] L. Longo, et al., Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion 106 (2024) 102301. doi:https://doi.org/10.1016/j.inffus.2024.102301.

[2] A. Theissler, F. Spinnato, U. Schlegel, R. Guidotti, Explainable ai for time series classification: A review, taxonomy and research directions, IEEE Access (2022).

[3] M. T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you? explaining the predictions of any classifier, in: Proc. of the 22nd ACM SIGKDD Int. Conf. on knowledge discovery and data mining, 2016, pp. 1135–1144.

[4] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034 (2013).

[5] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (2015) e0130140.

[6] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, D. A. Keim, Towards a rigorous evaluation of xai methods on time series, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE, 2019, pp. 4197–4201.

[7] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, N. Díaz-Rodríguez, Explainable artificial intelligence (xai) on timeseries data: A survey, arXiv preprint arXiv:2104.00950 (2021).
[8] J. V. Jeyakumar, J. Noor, Y.-H. Cheng, L. Garcia, M. Srivastava, How can i explain this to you? an empirical study of deep neural network explanation methods, Advances in Neural Information Processing Systems 33 (2020) 4211–4222.
[9] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proc. of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.
[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, International journal of computer vision 128 (2020) 336–359.
[11] J. Vielhaben, S. Lapuschkin, G. Montavon, W. Samek, Explainable ai for time series via virtual inspection layers, arXiv preprint arXiv:2303.06365 (2023).
[12] U. Schlegel, D. L. Vo, D. A. Keim, D. Seebacher, Ts-mule: Local interpretable model-agnostic explanations for time series forecast models, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2021, pp. 5–14.
[13] F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. T. P. Hartono, S. Ramasamy, B. L. DeCost, S. I. Tian, G. Romano, et al., Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks, npj Computational Materials 5 (2019) 60.
[14] S. A. Siddiqui, D. Mercier, M. Munir, A. Dengel, S. Ahmed, Tsviz: Demystification of deep learning models for time-series analysis, IEEE Access 7 (2019) 67027–67040.
[15] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).
[16] M. Guillemé, V. Masson, L. Rozé, A. Termier, Agnostic local explanation for time series classification, in: 2019 IEEE 31st Int. Conf. on tools with artificial intelligence (ICTAI), IEEE, 2019, pp. 432–439.
[17] I. Neves, D. Folgado, S. Santos, M. Barandas, A. Campagner, L. Ronzio, F. Cabitza, H. Gamboa, Interpretable heartbeat classification using local model-agnostic explanations on ecgs, Computers in Biology and Medicine 133 (2021) 104393.
[18] T. Sivill, P. Flach, Limesegment: Meaningful, realistic time series explanations, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 3418–3433.
[19] F. Karim, S. Majumdar, H. Darabi, S. Chen, Lstm fully convolutional networks for time series classification, IEEE access 6 (2017) 1662–1669.
[20] M. W. Kadous, Learning comprehensible descriptions of multivariate time series., in: ICML, volume 454, 1999, p. 463.
[21] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, E. Keogh, The ucr time series archive, IEEE/CAA Journal of Automatica Sinica 6 (2019) 1293–1305.
[22] E. Mekonnen, P. Dondio, L. Longo, Explaining deep learning time series classification models using a decision tree-based post-hoc xai method, in: xAI-2023 Late-breaking Work, Demos and Doctoral Consortium Joint Proceedings, CEUR-WS.org, 2023.
[23] E. T. Mekonnen, P. Dondio, L. Longo, A global model-agnostic rule-based xai method based on parameterised event primitives for time series classifiers, Frontiers in Artificial Intelligence 7 (2024) 1381921.
[24] E. T. Mekonnen, L. Longo, P. Dondio, Interpreting black-box time series classifiers using parameterised event primitives, xAI-2024 Late-breaking Work, Demos & Doctoral Consortium Joint Proceedings (2024).