

Explaining Deep Learning Time Series Classification Models using a Decision Tree-Based Post-Hoc XAI Method

Ephrem T. Mekonnen^{1,2,*}, Pierpaolo Dondio² and Luca Longo^{1,2}

¹*Artificial Intelligence and Cognitive Load Research Lab*

²*School of Computer Science, Technological University Dublin*

Abstract

This preliminary study proposes a new post hoc method to explain deep learning-based time series classification models using a decision tree. Our approach generates a decision tree graph or rulesets as an explanation, improving interpretability compared to saliency map-based methods. The method involves two phases: training and evaluating the deep learning-based time series classification model and extracting prototypical events from the evaluation set to train the decision tree classifier. We conducted experiments on artificial and real datasets, evaluating the explanations based on accuracy, fidelity, number of nodes, and depth. Our preliminary findings suggest that our post-hoc method improves the interpretability and trust of complex time series classification models.

Keywords

Explainable Artificial Intelligence, Deep Learning, Time Series Classification, Decision Tree

1. Introduction

Time series classification is crucial in domains like finance [1], healthcare [2, 3], human activity recognition [4, 5], and environment monitoring [6]. Deep learning models have shown remarkable performance in this task. However, they are often seen as "black boxes" due to their complexity, limiting interpretability. Explainable Artificial Intelligence (XAI) [7, 8, 9] aims to address this issue by developing techniques that provide understandable and transparent explanations for deep learning models. To explain deep learning-based time series classification models, XAI methods like Local Interpretable Model-agnostic Explanations (LIME)[10], saliency maps [11], and Layer-wise Relevance Propagation (LRP)[12] have been adapted [13]. However, these methods struggle to generate easily understandable explanations for time series data [14], often catering more to developers[15]. Additionally, the temporal nature of time series data poses challenges for feature importance-based approaches. This paper proposes a novel post-hoc XAI method to explain deep learning-based time series classification models using


Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

*Corresponding author.

✉ D22125038@mytudublin.ie (E. T. Mekonnen); pierpaolo.dondio@tudublin.ie (P. Dondio); luca.longo@tudublin.ie (L. Longo)

🆔 0009-0009-3035-3441 (E. T. Mekonnen); 0000-0001-7874-8762 (P. Dondio); 0000-0002-2718-5426 (L. Longo)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

a decision tree. Decision trees are inherently interpretable and visually accessible [16]. Our approach aims to generate a decision tree graph or rules that are comprehensible to non-experts, enhancing understanding of the model’s predictions.

2. Related work

Explainable Artificial Intelligence (XAI) has gained significant attention in the machine learning field as a means to address the lack of transparency and interpretability in complex models. Two prominent approaches in XAI are attribution methods and attention-based methods. Attribution methods, such as LIME [10] and LRP[12], have been widely adopted in computer vision to identify salient parts of an input that contribute to model predictions. However, applying these techniques to explain time series data poses challenges due to the inherent non-intelligible nature of such data [13]. Siddique et al. propose TSViz in [17] to explain Convolutional Neural Networks (CNNs) using a saliency map, and the authors in [18] exploit TSViz to design TSXplain to explain the decisions of Deep Neural Networks (DNNs) in time series. TSXplain identifies the most salient regions responsible for a model’s prediction and the most important time series through TSViz [17]. These regions and instances are then combined with different statistical features used to generate natural language explanations.

Although attributions are used to attribute a relevance score to each input value of a model, generating explanations for time series using only attribution and their relevance scores is challenging due to the non-intelligible nature of time series [19]. Heatmaps, the primary explanation medium for attributions, are often promising for domain experts, but ineffective for general users, as relevance scores are difficult to interpret without additional underlying data knowledge [14].

Attention mechanisms, often employed in transformer networks, have demonstrated remarkable performance in language-related tasks [20]. As attention is embedded in the network architecture, most attention-based approaches are ante-hoc explanation methods. Karim et al. [21] combine Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) submodules to create a specialized time series classification model. The authors also propose a variant incorporating an attention mechanism, allowing us to explain the decision process of the LSTM model. Similar to attributions, attentions can highlight relevant parts of the input. However, attention only works if specific components are implemented into the model’s architecture. Like attributions, attentions are often visualised as heatmaps and are somewhat challenging to interpret in many cases [14]. To overcome the limitations of existing methods, our work proposes a novel approach for explaining deep learning-based time series classification models using a decision tree graph. Decision trees offer intuitive and structured explanations by representing the underlying logic of an ML model as rulesets[22, 23].

3. Proposed method

Our method consists of two phases: training a Deep Learning-based time series classification model, and evaluating its performance, generating synthetic training data from the evaluation set, using its predictions as the target variable.

Dataset preparation - Three publicly available datasets, namely cylinder-bell-funnel (CBF), ECG200, and FordA, were used for this study. These datasets were downloaded from the UCR time series classification archive [24]. The cylinder-bell-funnel dataset aims to classify time series into three classes: cylinder, bell, or funnel [25]. The ECG200 dataset consists of time series representing electrical activity during heartbeats, with two classes: normal heartbeat and myocardial infarction. The FordA dataset comprises engine noise time series data collected during typical operating conditions, specifically to classify the presence or absence of symptoms in the engine.

Model training - An LSTM model was constructed using the PyTorch-based tsai library [26], with a specific emphasis on explainability. The model was trained and evaluated on the following datasets: CBF, ECG200, and FordA. The test results showed an accuracy of 98.0% for CBF, 76.0% for ECG200, and 89.5% for FordA. This evaluation took into account the availability of other state-of-the-art models for these specific datasets.

Synthetic training data preparation - At this stage, the evaluation set was processed in two ways: Firstly, global feature calculation is applied to extract overall characteristics from the time series data, including global maxima, global minima, channel means, and stream duration. This helps provide a holistic understanding of the data. Additionally, parameterized event primitives (PEPs) are employed to capture specific events expected in the domain. Extracting PEPs from a time series helps to represent the temporal characteristics of events as parameters, which facilitates learning for interpretable models such as decision trees [27]. These PEPs include increasing and decreasing events, which capture start time, duration, and average gradient value parameters, as well as local maximum and minimum events, which capture time and corresponding value parameters. To prepare the synthetic training data, a three-step process was followed. In the first step, parameterized events were extracted from each time series sequence of the evaluation set. The events were represented as tuples containing relevant parameters. In the second step, the parameterized events were flattened to apply clustering algorithms, such as KMeans, and generate clusters. The optimal number of clusters was determined using the silhouette method. In the third step, event attribution was performed, mapping the extracted events to the clusters. This resulted in a matrix where each cell denoted the number of events belonging to a specific cluster for a particular instance. The event attribution matrices were combined with global features and the trained model's predictions on the evaluation set, instead of using ground truth. This created a complete synthetic training dataset.

Applying interpretable model - Following the generation of synthetic training data, the decision tree classifier was applied as the next step. The synthetic training data was divided into training and testing sets, with 70% of the data allocated for training and 30% for testing.

Objective evaluation - To ensure an objective and quantitative assessment of the interpretability of the proposed method, four metrics were selected: accuracy, fidelity, depth, and number of nodes. The evaluation process was conducted without any human intervention to maintain objectivity. Accuracy measures the proportion of correct predictions made by the model, while fidelity evaluates the consistency between the model's decisions. The depth and number of nodes indicate the complexity of the decision tree.

4. Results and Discussion

The proposed method was applied to three time series datasets (CBF, ECG200, and FordA), and its performance was evaluated based on accuracy, fidelity, number of nodes, and depth. The evaluation results are summarized in Table 1. The results demonstrated that the proposed method achieved notable accuracy and fidelity scores across all three datasets. The results show that the proposed method effectively explains predictions in deep learning-based time series classification models. Regarding the complexity of the generated decision tree graphs, the number of nodes and depth remained relatively low for all three datasets. This suggests that the proposed method can generate interpretable explanations using relatively simple decision trees, facilitating domain experts comprehension. The list of extracted rules below demonstrates

Table 1

Objective metrics results for decision tree-based explanations

Dataset	Accuracy (%)	Fidelity (%)	Nodes	Depth
CBF	83.7	87.8	31	6
ECG	80.0	88.0	5	2
FordA	76.8	85.8	87	8

the preliminary findings of our experiment using the ECG200 dataset. Each rule highlights the importance of particular time steps along with the corresponding events occurring at those steps, significantly impacting the model’s prediction. Additionally, if domain experts provide definitions for the conditional part of the rules, we can generate human-readable explanations for better comprehension.

Rule 1: Local minimum at time 66 with value $0.25 \leq 11.0 \Rightarrow$ Normal

Rule 2: Local minimum at time 66 with value $0.25 > 11.0$ and local minimum at time 66 with value $0.25 \leq 20.5 \Rightarrow$ Infarction

Rule 3: Local minimum at time 66 with value $0.25 > 11.0$ and local minimum at time 66 with value $0.25 > 20.5 \Rightarrow$ Infarction

5. Conclusion

In conclusion, the proposed method shows promising performance in terms of accuracy, fidelity, and interpretability in the selected time series datasets. The generated decision tree-based explanations provide valuable insights into the underlying factors influencing the model’s predictions. Future work could focus on enhancing the method’s capability to handle more complex datasets while preserving its interpretability.

References

- [1] X. Zhang, X. Liang, A. Zhiyuli, S. Zhang, R. Xu, B. Wu, At-lstm: An attention-based lstm model for financial time series prediction, in: *IOP Conference Series: Materials Science and Engineering*, volume 569, IOP Publishing, 2019, p. 052037.
- [2] P. Liu, X. Sun, Y. Han, Z. He, W. Zhang, C. Wu, Arrhythmia classification of lstm autoencoder based on time series anomaly detection, *Biomedical Signal Processing and Control* 71 (2022) 103228.
- [3] N. Strodthoff, P. Wagner, T. Schaeffter, W. Samek, Deep learning for ecg analysis: Benchmarks and insights from ptb-xl, *IEEE Journal of Biomedical and Health Informatics* 25 (2020) 1519–1528.
- [4] S. Mekruksavanich, A. Jitpattanakul, Lstm networks using smartphone data for sensor-based human activity recognition in smart homes, *Sensors* 21 (2021) 1636.
- [5] S. Joshi, E. Abdelfattah, Deep neural networks for time series classification in human activity recognition, in: *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, IEEE, 2021, pp. 0559–0566.
- [6] T. Shu, J. Chen, V. K. Bhargava, C. W. de Silva, An energy-efficient dual prediction scheme using lms filter and lstm in wireless sensor networks for environment monitoring, *IEEE Internet of Things Journal* 6 (2019) 6736–6747.
- [7] A. Theissler, F. Spinnato, U. Schlegel, R. Guidotti, Explainable ai for time series classification: A review, taxonomy and research directions, *IEEE Access* (2022).
- [8] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, A. Holzinger, Explainable artificial intelligence: Concepts, applications, research challenges and visions, in: *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020*, Dublin, Ireland, August 25–28, 2020, Proceedings, Springer, 2020, pp. 1–16.
- [9] F. Di Martino, F. Delmastro, Explainable ai for clinical and remote health applications: a survey on tabular and time series data, *Artificial Intelligence Review* (2022) 1–55.
- [10] M. T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you? explaining the predictions of any classifier, in: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.
- [11] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, *arXiv preprint arXiv:1312.6034* (2013).
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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).

- [16] G. Vilone, L. Longo, A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods, *Frontiers in artificial intelligence* 4 (2021) 717899.
- [17] 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.
- [18] M. Munir, S. A. Siddiqui, F. Küsters, D. Mercier, A. Dengel, S. Ahmed, Tsxplain: Demystification of dnn decisions for time-series using natural language and statistical features, in: *Artificial Neural Networks and Machine Learning–ICANN 2019: Workshop and Special Sessions: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings 28*, Springer, 2019, pp. 426–439.
- [19] U. Schlegel, D. A. Keim, Time series model attribution visualizations as explanations, in: *2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TRES)*, IEEE, 2021, pp. 27–31.
- [20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [21] F. Karim, S. Majumdar, H. Darabi, S. Chen, Lstm fully convolutional networks for time series classification, *IEEE access* 6 (2017) 1662–1669.
- [22] G. Vilone, L. Longo, Classification of explainable artificial intelligence methods through their output formats, *Machine Learning and Knowledge Extraction* 3 (2021) 615–661.
- [23] G. Vilone, L. Longo, A novel human-centred evaluation approach and an argument-based method for explainable artificial intelligence, in: *Artificial Intelligence Applications and Innovations: 18th IFIP WG 12.5 International Conference, AIAI 2022, Hersonissos, Crete, Greece, June 17–20, 2022, Proceedings, Part I*, Springer, 2022, pp. 447–460.
- [24] 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.
- [25] N. Saito, *Local feature extraction and its applications using a library of bases*, Yale University, 1994.
- [26] I. Oguiza, tsai - a state-of-the-art deep learning library for time series and sequential data, Github, 2022. URL: <https://github.com/timeseriesAI/tsai>.
- [27] M. W. Kadous, Learning comprehensible descriptions of multivariate time series., in: *ICML*, volume 454, 1999, p. 463.