Interpreting Black-Box Time Series Classifiers using Parameterised Event Primitives

Ephrem T. Mekonnen[1,2,*], Luca Longo[1,2] and Pierpaolo Dondio[1]

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

Abstract

Amidst the remarkable performance of deep learning models in time series classification, there is a pressing demand for methods that unveil their prediction rationale. Existing feature importance techniques often neglect the temporal nature of time series data, focusing solely on segment importance. Addressing this gap, this paper introduces a local model-agnostic method akin to LIME, which generates neighbouring samples by randomly perturbing segments of the original instance. Subsequently, weights are computed for each neighbouring instance based on its distance from the original, elucidating its influence. Parameterised event primitives (PEPs) are then extracted from these perturbed samples, encompassing increasing and decreasing events and local maxima and minima points. These primitives are clustered to form prototypical events that capture the temporal essence of the data. Leveraging these events, computed weights, and black box predictions, a simple linear regression model is trained to provide local explanations. Preliminary experiments on real-world datasets showcase the method's efficacy in identifying salient subsequences and points and their importance scores, thereby enhancing comprehension of produced explanations.

Keywords

Explainable Artificial Intelligence, Model-Agnostic, Time Series, Post-hoc, Deep Learning,

1. Introduction

The ubiquity of sensors has facilitated the generation of extensive time series data across domains such as finance [1], healthcare [2, 3], human activity recognition [4], and environmental monitoring [5]. These data, crucial for informed decision making, require effective time series classification techniques. However, despite the success of deep learning models in various domains, including time series classification tasks, their lack of interpretability remains a significant challenge. Explainable Artificial Intelligence (XAI) has emerged to address this issue, aiming to provide transparent explanations for machine learning models. There are a multitude of XAI methods for image and tabular data; however, applying such methods to time series data presents unique challenges due to the temporal nature of the data and the requirement for domain knowledge [6, 7]. Locally Interpretable Model-Agnostic Explanation (LIME) has

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

*Corresponding author.
D22125038@mytudublin.ie (E. T. Mekonnen); luca.longo@tudublin.ie (L. Longo); pierpaolo.dondio@tudublin.ie (P. Dondio)
https://ephrem-eth.github.io/ (E. T. Mekonnen)
0000-0002-0877-7063 (E. T. Mekonnen); 0000-0002-2718-5426 (L. Longo); 0000-0001-7874-8762 (P. Dondio)

become a popular method for explaining black-box models [8]. However, its application to time series data is hindered by the difficulty of segmenting data while preserving temporal characteristics [9]. To address these challenges, we propose a novel Local Model Agnostic XAI method, akin to LIME, for interpreting black-box time series classifiers. Our approach does not require the segmentation of time series data. It provides detailed explanations of salient parts, including detecting trends such as increasing and decreasing local maxima and local minima. By enhancing the interpretability of black box time series classifiers, our method fosters a deeper understanding of model decisions and facilitates informed decision-making.

2. Related Works

Recent advancements in explainable artificial intelligence (XAI) have sparked significant interest in understanding black box models, particularly in time series classification. Although XAI research has focused predominantly on computer vision and natural language processing tasks, adapting these methods to time series analysis presents unique challenges due to the temporal nature of the data [6]. Schlegel et al. [7] explored various common XAI techniques, including saliency [10], LIME [8], SHAP [11] and LRP [12], to interpret deep learning-based time series classification models. Zhou et al. [13] have enriched the interpretability landscape by enhancing Class Activation Maps (CAM) and Grand-CAM with backpropagation. Simultaneously, the work described in [14] introduced TSViz, a saliency map-based methodology later integrated into TSXplain [15] to uncover the logic behind Deep Neural Networks (DNNs) in time series. These methodologies combine salient regions, instances, and statistical features, fostering natural language explanations. Furthermore, Vielhaben et al. [16] introduced DFT-LRP, a tailored variant of Layer-wise Relevance Propagation (LRP), specifically designed to address the complexities of time series analysis by incorporating a virtual inspection layer.

While many existing methods are model-specific and rely on internal model structures, there is a growing interest in model-agnostic explanations that identify key features without being tied to a particular model architecture. However, adapting feature importance-based explanations to time series data requires careful consideration of the temporal dimension. Among feature-importance methods, LIME stands out as a popular approach, but its direct application to time series data requires thoughtful preprocessing to ensure interpretability. Guileme et al. [17] and Neves et al. [18] adapted LIME for deep learning-based time series classification by using longer segments for perturbation. Still, these approaches are limited by fixed window sizes. To overcome this limitation, Silvio et al. [19] introduced the NNsegment, which identifies homogeneous regions in time series and employs various perturbation techniques for robust explanations. Furthermore, Schlegel et al. [20] expanded the LIME approach by employing six distinct segmentation methods, but the challenge remains to understand the significance of identified segments. Hence, we present a local model-agnostic Explainable Artificial Intelligence (XAI) approach, akin to LIME, tailored for elucidating deep learning time series classifiers. Our method effectively highlights crucial input data segments that significantly impact the black-box model's inferential process. Additionally, it provides insights into why these identified segments are important by providing information about their nature, such as whether they are increasing/decreasing events or local minima/maxima points on the time series input.

3. Method

This section introduces the proposed Local Model Agnostic (XAI) method tailored for time series classifiers. The steps involved in the approach are detailed in Figure 1.

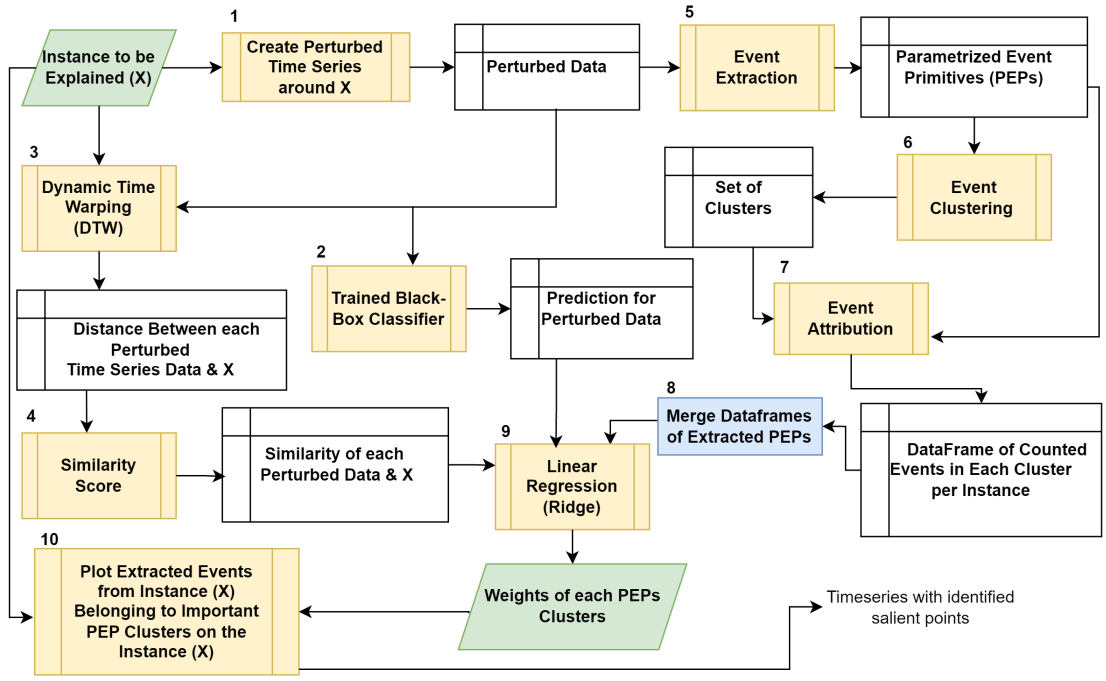


Figure 1: Step-by-step illustration of the proposed approach.

3.1. Generating Neighbourhood Samples

Our approach distinguishes itself from existing methods by avoiding fixed-interval segmentation for interpreting time series classifiers. Instead, we employ random perturbation of segments in the original time series, offering a flexible and tailored approach to generating perturbed data. These segments can be replaced by zero, the segment mean, or the total mean of the series. Importantly, perturbation is used solely to generate neighbourhood samples, and rather than employing segments as features for the linear regression model, as detailed in subsection 3.3, we utilise clusters of parameterised event primitives.

3.2. Distance Computation and Neighborhood Weighting

After generating neighbouring samples through perturbation, we calculate the distance (d) between the explained instance (X_i) and the neighbouring sample. In our scenario, we utilize dynamic time warping (DTW) as the distance metric, which is ideal for handling temporal data with varying speeds or time scales. Subsequently, we calculate the weight of each neighbouring

instance according to an exponential kernel, denoted as π_{X_i}, which assigns higher weights to instances similar to X_i. The exponential kernel is defined as: $\pi_{X_i} = e^{-\left(\frac{d^2}{\sigma^2}\right)}$.

Here, σ (sigma) represents the bandwidth parameter that controls the width of the kernel. It regulates how quickly the weight assigned to neighbouring instances decreases with increasing distance from the instance being explained. Lower values of σ indicate a narrower kernel, focusing more on closer neighbours, while higher values result in a broader influence, considering distant neighbours as well.

3.3. Transforming Perturbed Data via Parameterised Event Primitives (PEPs)

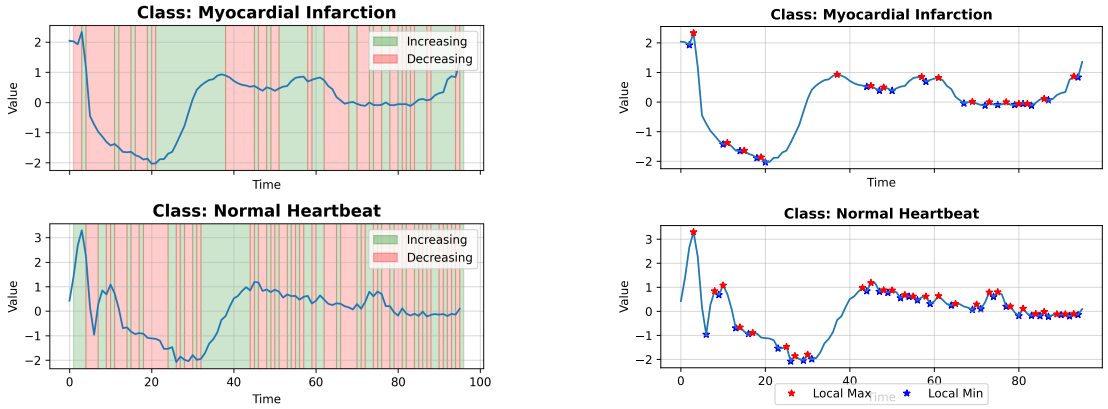
Parameterised Event Primitives (PEPs) are vital for capturing domain-specific events within the time series data. By extracting PEPs as shown in Figure2, we can effectively represent the temporal characteristics of events as parameters, thus facilitating the learning process for interpretable models such as linear regression and decision trees [21, 22]. These PEPs encompass various event types, including increasing and decreasing events, which capture parameters such as start time, duration, and average gradient value, and local maximum and minimum events, which capture time and corresponding value parameters. A structured three-step process was implemented to transform neighbouring samples in a manner conducive to training interpretable models. Initially, parameterized events were extracted from each time series sequence within the perturbed data. These events were encapsulated as tuples containing the relevant parameters. Subsequently, the parameterized events were flattened to enable the application of clustering algorithms, such as KMeans, resulting in the generation of distinct clusters. Determining the optimal number of clusters was facilitated by leveraging the silhouette method. Finally, event attribution was carried out, mapping the extracted events to their respective clusters. This process yielded a matrix wherein each cell represented the count of events associated with a specific cluster for a given instance. The event attribution matrices for each parameterized event primitive were combined to create a tabular dataset suitable for training interpretable models.

3.4. Training Linear Model

In our approach, we utilize transformed data, black box predictions of neighbouring samples, and their corresponding weights to train interpretable models similar to LIME. We employ ridge regression, a regularised linear model renowned for its interpretability and robustness. Ridge regression aims to minimize the following loss function:

$$\hat{\beta} = \text{argmin}_{\beta} \sum_{z \in \mathcal{Z}} \pi_{X_i}(z) \left(\hat{y}_z - (z) \cdot \beta\right)^2 + \lambda ||\beta||_2^2$$

Here, $\hat{\beta}$ represents the optimized coefficients obtained by minimizing the weighted sum of squared errors. $\pi_{X_i}(z)$ assigns weight to each neighbouring sample z, $\hat{y}_z$ is the probability score predicted by the classifier for the perturbed instance z, (z) is a perturbed instance and λ serves as the regularisation parameter, which governs the penalty imposed on the coefficients to prevent overfitting. In this case, the weights of the linear model learnt using a least-squares procedure denote the relative importance of each feature or each PEP cluster.



(a) Increasing and decreasing events.

(b) Local max and local min events.

Figure 2: Examples of events extracted from a single time series in the ECG200 dataset (a) increasing and decreasing events (b) local max and local min events .

After training the interpretable linear model, we identify the most significant features based on their importance scores. Here, the features correspond to clusters of Parameterised Event Primitives (PEPs) such as increasing cluster1, increasing cluster2, decreasing cluster1, and so on. We then visualise the extracted events of the instance to be explained, which belong to the top clusters, as shown in Figure 3.

4. Experimental Setup

Our preliminary experiment evaluated our method on two widely used univariate time series datasets: ECG200 and GunPoint from the UCR Archive [23], a renowned repository for time series classification tasks.

Our method provided local explanations for a black box model, the Fully Convolutional Network (FCN), built using the PyTorch-based tsai library [24]. The FCN was configured with default kernel sizes 7, 5, 3 and filter sizes 128, 256, 128 for its convolutional layers. The datasets were partitioned into training sets (70%), validation sets (15%), and test sets (15%) to facilitate robust evaluation. We used early stopping during training to avoid overfitting, with a patience parameter set to 15 and a minimum delta of 0.001. Additionally, to ensure accuracy and stability, the model was trained 100 times with randomised splits for training, validation, and testing. In particular, our method achieved an average accuracy of 85% and 86% on the ECG200 dataset for training and testing, respectively, and 98% for both the validation and testing sets of the GunPoint dataset.

5. Result and Discussion

In this section, we present the results of our experiments and discuss their implications. The method was deployed to offer local explanations for predictions generated by a deep learning-

based time series classifier, with fidelity metrics that evaluate the faithfulness of these explanations. We computed the local fidelity score across different replacement methods for the perturbation to generate neighbouring samples. From each dataset, 100 instances were randomly selected from the test set, and the resulting average fidelity score and standard deviation were calculated. Table 1 presents the fidelity scores obtained using the zero and mean replacement methods. In the ECG200 dataset, the fidelity scores were 0.76 and 0.67 for the zero and mean replacements, respectively. Similarly, in the GunPoint dataset, the fidelity scores were 0.64 and 0.44 for zero and mean replacements, respectively.

Table 1
Mean and standard deviation of explanation faithfulness across various perturbation replacement methods on ECG200 and GunPoint datasets.

Dataset	Zero (Std)	Mean (Std)
ECG200	0.76 (0.08)	0.67 (0.10)
GunPoint	0.64 (0.10)	0.44 (0.17)

These results indicate that our method demonstrates varying fidelity across different datasets and replacement methods. The higher fidelity scores obtained using zero replacement suggest that this method may better preserve the local interpretability of the model predictions compared to mean replacement. Furthermore, the observed standard deviations highlight the variability in the fidelity scores, indicating potential sensitivity to perturbation methods. This underscores the importance of careful consideration when selecting perturbation techniques to ensure reliable and consistent explanations.

Figure 3: The explanation generated by the method 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.

The explanation produced by our method, as depicted in Figure 3, not only highlights the significance of each part of the time series instance for the black box model's decision-making process but also provides the relevance score associated with each segment or point, along with the types of events, such as increasing, decreasing, local maximum, and local minimum. Overall, our results demonstrate the effectiveness of our method in providing local explanations for predictions of deep learning-based time series classifiers. However, further analysis and experiments are needed to fully understand the factors influencing fidelity and optimize our approach for broader applications.

6. Conclusion

Our XAI method, incorporating random perturbation and transformation using parameterized event primitives, shows promising results in enhancing interpretability for time series classifiers. While our current experiment has focused on two univariate time series datasets, future research will extend to other univariate and multivariate data to widen its applicability. Further exploration into diverse perturbation techniques and comparative analyses with existing methods will provide a comprehensive understanding of our approach's effectiveness. Overall, our method contributes to advancing explainable AI in time series classification, offering valuable insights into model predictions with ongoing efforts for refinement and expansion.

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. 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.

[5] 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.

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

[7] 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.

[8] 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.
[9] L. Longo, M. Brcic, et al., Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion (2024) 102301.
[10] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034 (2013).
[11] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).
[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] L. Zhou, C. Ma, X. Shi, D. Zhang, W. Li, L. Wu, Salience-cam: Visual explanations from convolutional neural networks via salience score, in: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021, pp. 1–8.
[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] 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.
[16] J. Vielhaben, S. Lapuschkin, G. Montavon, W. Samek, Explainable ai for time series via virtual inspection layers, arXiv preprint arXiv:2303.06365 (2023).
[17] M. Guillemé, V. Masson, L. Rozé, A. Termier, Agnostic local explanation for time series classification, in: 2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI), IEEE, 2019, pp. 432–439.
[18] I. Neves, D. Folgado, S. Santos, et al., Interpretable heartbeat classification using local model-agnostic explanations on ecgs, Computers in Biology and Medicine 133 (2021) 104393.
[19] T. Sivill, P. Flach, Limesegment: Meaningful, realistic time series explanations, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 3418–3433.
[20] 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.
[21] M. W. Kadous, Learning comprehensible descriptions of multivariate time series., in: ICML, volume 454, 1999, p. 463.
[22] E. T. Mekonnen, P. Dondio, L. Longo, Explaining deep learning time series classification models using a decision tree-based post-hoc xai method, CEUR Workshop Proceedings (2023).
[23] 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.
[24] 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.