Explaining Deep Learning Time Series Classification Models using a Decision Tree

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

[1] Technological University Dublin, *correspondence: D22125038@mytudublin.ie @ (Github) ephrem-eth.github.io

Summary

- We proposed a novel post-hoc XAI method for explaining deep learning time series classification models using a decision tree.
- We conducted preliminary experiements on three time series datasets.
- We objectively evaluated the generated explanation using metrics such as accuracy, fidelity, depth and number of nodes.

Motivation

- Deep learning models have demonstrated remarkable performance in time series classification tasks; however, they are often considered as black boxes [1].
- XAI methods for image and tabular data may not be suitable for time series data due to its temporal nature [1,2].
- Heatmaps, the primary explanation medium of XAI methods for time series data, may be challenging to interpret for general users unfamiliar with the underlying data [2].

How can the inference process of deep learning time series classification models be explained using a decision tree ?

Proposed Method

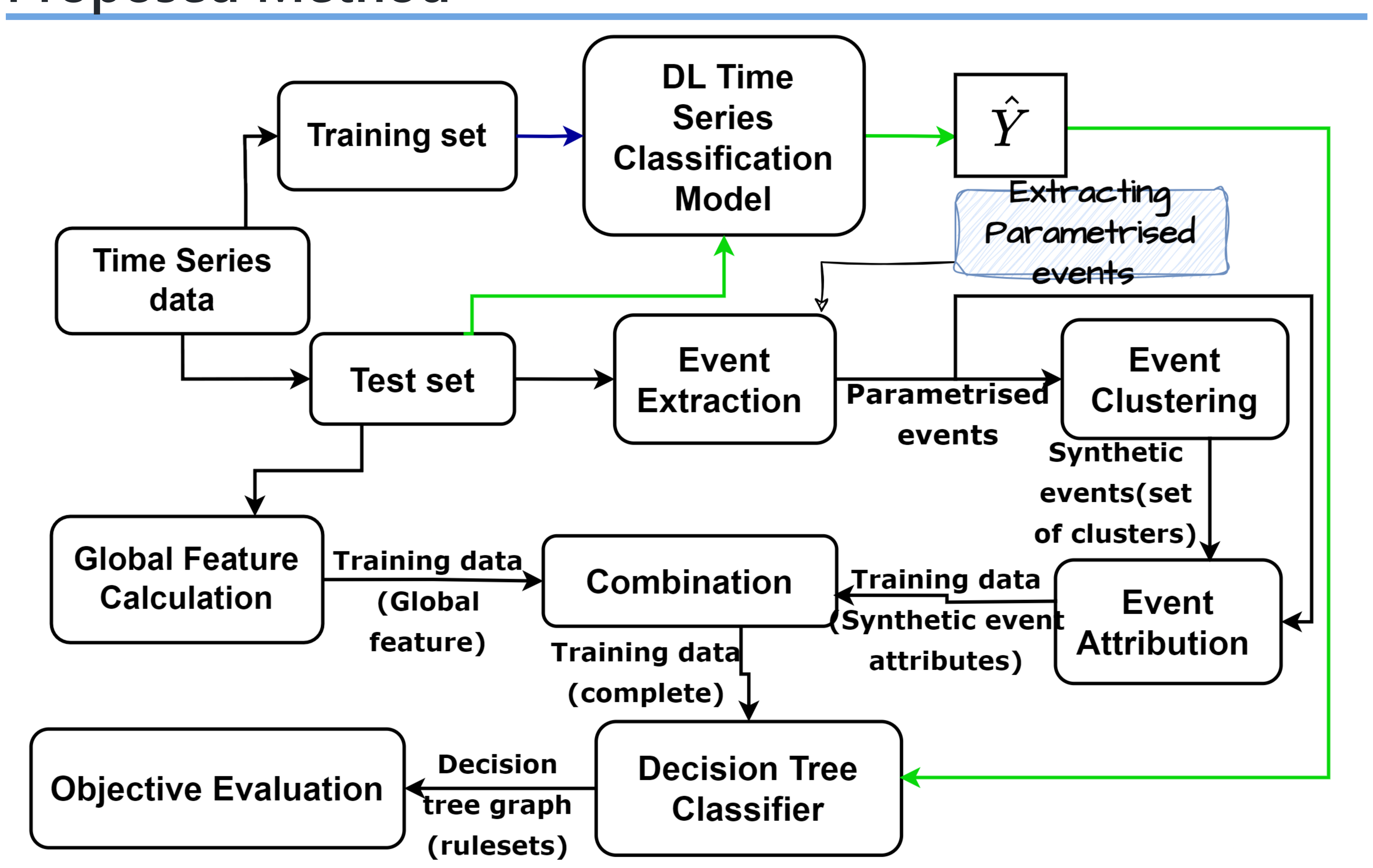

Fig 1A: The proposed XAI method for DL time series classification models

- This work proposed a new post-hoc XAI method to explain deep-learning time series classification models using a decision tree. The method consists of two phases:
  i. Train a Deep Learning-based time series classification model and evaluate its performance.
  ii. Generate synthetic training data from the evaluation set, using the model's predictions as the target variable to train the decision tree.
- Extracting parametrized event primitives (PEPs) from a time series helps to represent the temporal characteristics of events as parameters, which facilitates learning for interpretable models such as decision tree [3].

These PEPs include:

- Increasing and decreasing events:
  $PEP_{\mathrm{inc/dec}} = (\mathrm{start\_time, duration, avg\_gradient})$
- Local maximum and minimum events:
  $PEP_{\mathrm{max/min}} = (\mathrm{time, value})$

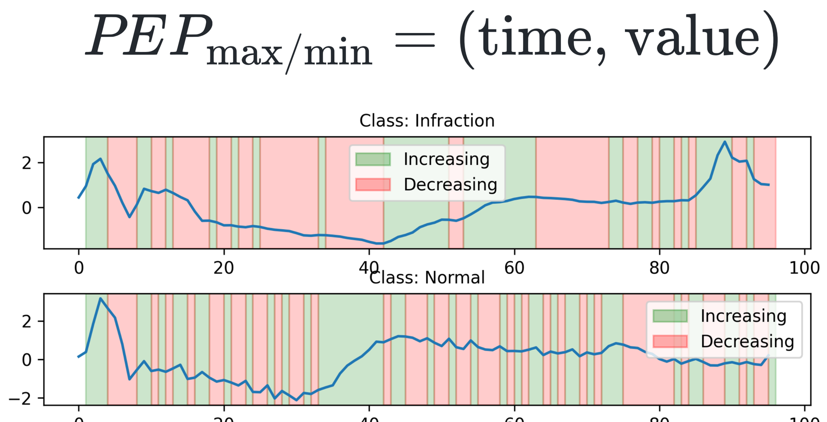

Fig 2A: Regions of the extracted increasing and decreasing events of a single time series per class from ECG dataset.

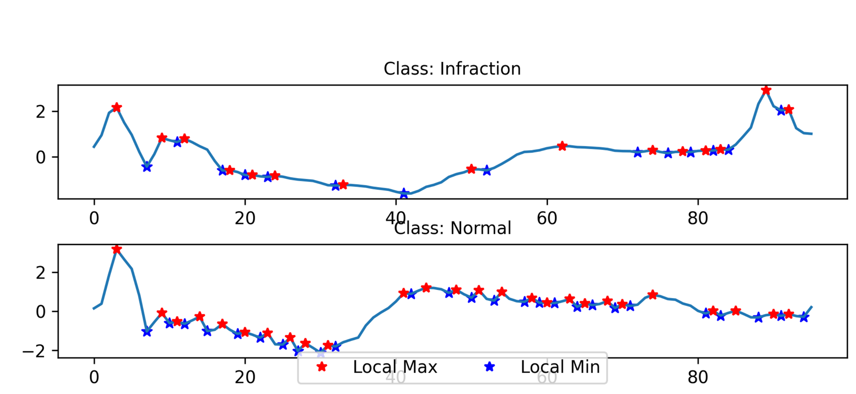

Fig 2B: The extracted local max and local min events of a single time series per class from ECG dataset.

Objective Evaluation

- Four objective metrics:
  - $Accuracy = \frac{\text{Number of correctly classified instances}}{\text{Total number of instances}}$
  - $Fidelity = \frac{\text{Number of instances with agreement}}{\text{Total number of instances}}$
  - $TreeDepth = D$
  - $NumberofNodes = N$

Result and Discussion

Table 1: Objective metrics results for decision tree-based explanations

Dataset	Accuracy	Fidelity	Depth	# Nodes
CBF	83.7	87.8	6	31
ECG	80.5	88.0	2	5
FordA	76.8	85.8	8	87

Table 2: LSTM model accuracy

Dataset	Test_Acc	Valid_Acc
CBF	98.0	96.9
ECG	80.0	76.0
FordA	91.5	89.5

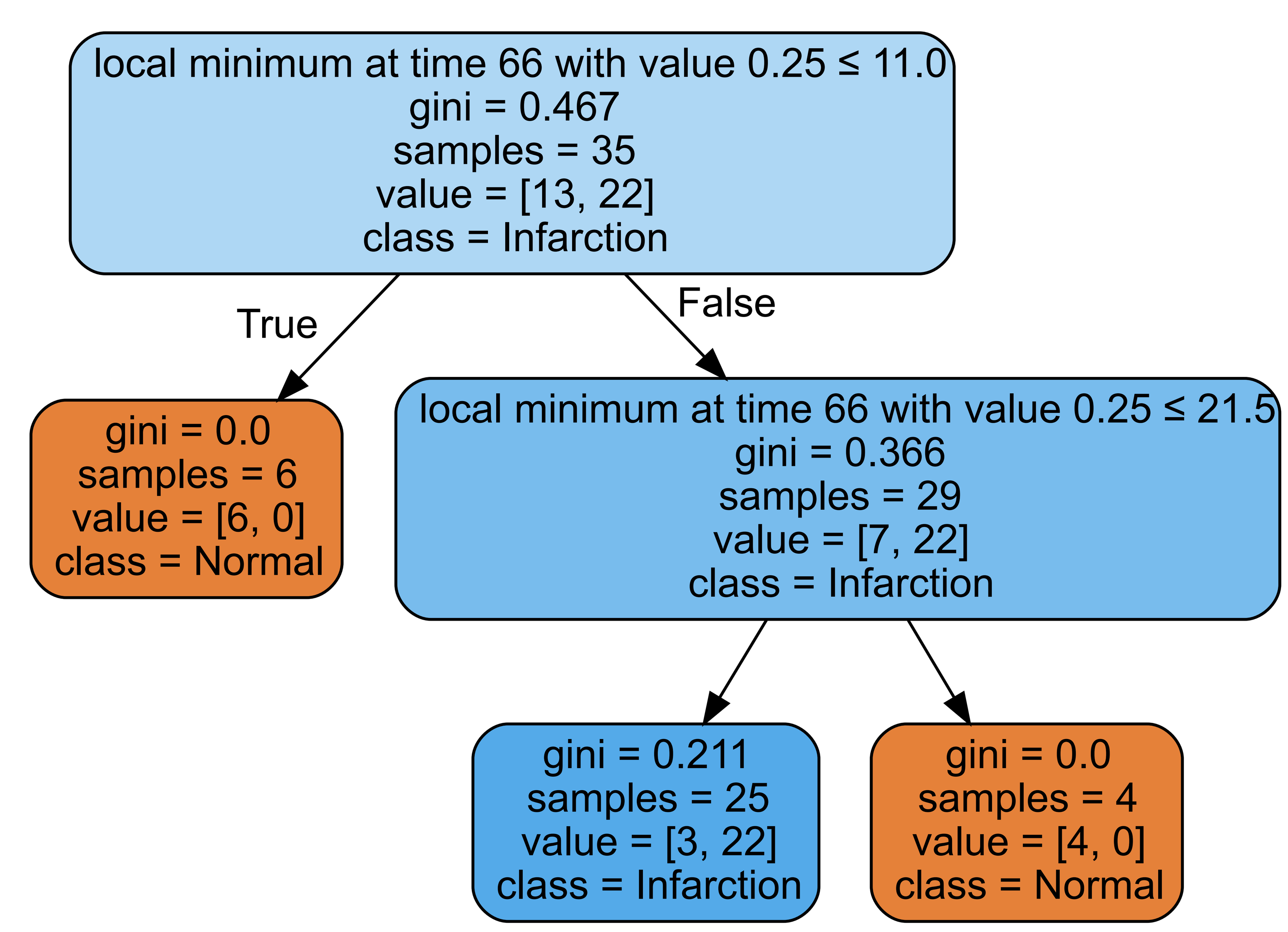

Fig 3A: A decision tree graph produced by the proposed method using ECG data.

- **List of Extracted Rules:**
  - Rule 1: Local minimum at time 66 with value 0.25 ≤ 11.0 ⇒ Normal
  - Rule 2: Local minimum at time 66 with value 0.25 > 11.0 and local minimum at time 66 with value 0.25 ≤ 20.5 ⇒ 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 ⇒ Infarction
- The extracted rules provide valuable insights into the impact of specific time steps and corresponding events on the model's predictions.

Conclusion and Future works

- Our proposed method demonstrates promising performance in terms of accuracy, fidelity, and interpretability on time series datasets.
- The decision tree-based explanations generated by our approach provide valuable insights into the factors influencing predictions.
- Future work aims to enhance the method's capability to handle more complex datasets while preserving interpretability.

Acknowledgements

The scholarship of TU Dublin is acknowledged for their generous support.

References

[1] Theissler et al. : Explainable ai for time series classification: A review, taxonomy and research directions. IEEE Access (2022)
[2] Jeyakumar et al.: How can i explain this to you?(2020)
[3] Kadous:Learning comprehensible descriptions of multivariate time series(1999)