Towards Interpretable Concept Learning over Time Series via Temporal Logic Semantics

Irene Ferfoglia[a,*]**, Simone Silvetti**[a]**, Gaia Saveri**[a]**, Laura Nenzi**[a] **and Luca Bortolussi**[a]

[a]Università degli Studi di Trieste

ORCID (Irene Ferfoglia): https://orcid.org/0000-0003-1585-6576, ORCID (Simone Silvetti): https://orcid.org/0000-0001-8048-9317, ORCID (Gaia Saveri): https://orcid.org/0009-0003-2948-7705, ORCID (Laura Nenzi): https://orcid.org/0000-0003-2263-9342, ORCID (Luca Bortolussi): https://orcid.org/0000-0001-8874-4001

Abstract. Time series classification is a task of paramount importance, as this kind of data often arises in safety-critical applications. However, it is typically tackled with black-box deep learning methods, making it hard for humans to understand the rationale behind their output. To take on this challenge, we propose a neuro-symbolic framework that unifies classification and explanation through direct embedding of trajectories into a space of Signal Temporal Logic (STL) concepts. By introducing a novel STL-inspired kernel that maps raw time series to their alignment with predefined STL formulae, our model jointly optimises for accuracy and interpretability, as each prediction is accompanied by the most relevant logical concepts that characterise it. This enables classification grounded in human-interpretable temporal patterns and produces both local and global symbolic explanations. Early results show competitive performance while offering high-quality logical justifications for model decisions.

1 Introduction

Time series data are pervasive across domains, from industrial monitoring to safety-critical applications. While machine learning models have demonstrated strong performance in analysing such data [19, 5], their black-box nature poses significant challenges to transparency and trust, especially in high-stakes contexts governed by ethical [8, 10] and legal [7, 4] concerns.

Explainable AI (XAI) techniques for time series often adapt methods originally developed for images or text [22, 18]. Prominent approaches include attribution-based methods [11, 9], which highlight influential subsequences in the input. However, these post-hoc techniques have been criticised for their potential to misrepresent the model's internal reasoning [23], and their explanations can be difficult to interpret due to the abstract nature of time series patterns.

Symbolic formalisms such as Signal Temporal Logic (STL) [17, 13] offer a compelling alternative for describing temporal behaviours in an interpretable and formal way. STL has been widely used in requirement mining and rule extraction [15, 21, 1, 14], but most prior work provides global descriptions of class behaviour, without explaining individual predictions. More recent neuro-symbolic models attempt to embed STL in decision tree structures or neural architectures [24, 12], yet they still lack the flexibility to generate instance-level explanations aligned with input semantics.

We address this gap through a concept-based approach [6, 25] that integrates STL formulae directly into the model as interpretable units. Unlike prior works that generate STL-based classifiers or tree-based rules [24, 16], our method computes both local and global explanations by embedding time series and STL concepts into a shared space using a robustness-inspired kernel [2]. This enables predictions to be interpreted directly in terms of temporal logic conditions, with relevance scores computed via a cross-attention mechanism. Importantly, our method is designed to handle both multivariate signals, by supporting STL formulae over multiple variables, and multiclass settings, through class-specific discriminability and attention mechanisms.

Our intended contributions are: (i) a novel time series classifier leveraging STL-based kernels for interpretable concept learning; (ii) local and global explanations via STL formulae, clarifying individual predictions and overall model behaviour; and (iii) empirical validation on real-world cyber-physical systems and scalability benchmarks.

2 Background

Concept-based models aim to make machine learning inherently interpretable by incorporating human-understandable concepts directly into the learning process [6, 25]. Unlike post-hoc explanation methods, these models use concepts, often representing domain knowledge, as part of the prediction pipeline.

Signal Temporal Logic (STL) is a linear-time temporal logic for specifying properties of continuous-time trajectories $\tau : I \to \mathcal{X}$ [13]. Its syntax combines Boolean and temporal operators:

$$\varphi := \top \mid \pi \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \varphi_1 \boldsymbol{U}_{[a,b]}\varphi_2$$

where predicates π are inequalities over signal values. Common derived operators include $\boldsymbol{F}_{[a,b]}$ (eventually), $\boldsymbol{G}_{[a,b]}$ (globally), and $\boldsymbol{U}_{[a,b]}$ (until). STL supports both Boolean satisfaction and quantitative semantics via robustness $\rho(\varphi, \tau, t)$, which measures how strongly a formula holds.

---

* Corresponding Author. Email: irene.ferfoglia@phd.units.it

3 Trajectory embedding kernel

A kernel for Signal Temporal Logic (STL) can be constructed using the quantitative robustness semantics, where each formula φ is viewed as a functional $\rho(\varphi, \cdot) : \mathcal{T} \to \mathbb{R}$ mapping trajectories to real values. By equipping the trajectory space $\mathcal{T}$ with a probability measure μ_0, a similarity kernel between two formulae φ and ψ is defined as:

$$k(\varphi, \psi) = \langle \rho(\varphi, \cdot), \rho(\psi, \cdot) \rangle = \int_{\tau \in \mathcal{T}} \rho(\varphi, \tau) \rho(\psi, \tau) \, d\mu_0(\tau)$$

This kernel captures behavioural similarity based on how formulae evaluate over likely trajectories, enabling a continuous embedding of symbolic logic into a Hilbert space. Extending this idea, we define a dual embedding for trajectories by introducing a functional $\rho_\tau : \mathcal{T} \to [-1, 1]$ given by:

$$\rho_\tau(\xi) = 2 \cdot \exp\left(-\frac{d_\tau(\xi)}{\varepsilon}\right) - 1$$

where $d_\tau(\xi) = \|\xi - \tau\|^2$ and $\varepsilon > 0$ controls locality. This yields a cross-kernel between a trajectory τ and a formula φ:

$$k(\tau, \varphi) = \int_{\xi \in \mathcal{T}} \rho_\tau(\xi) \rho(\varphi, \xi) \, d\mu_0(\xi)$$

allowing both formulae and trajectories to be embedded in a shared semantic space. This facilitates alignment between data and logic, forming the mathematical foundation for our interpretable concept-based classification framework.

4 The model

In our framework, concepts are defined as human-interpretable temporal patterns encoded as STL formulae. To construct a semantically rich yet interpretable concept set, we sample candidate formulae from a probabilistic grammar with constraints on the number of variables and syntactic complexity. Each formula is evaluated on a fixed dataset of time series signals using robustness semantics, resulting in robustness signature vectors that capture their behaviour across trajectories. Concept selection proceeds iteratively: new candidates are retained only if their signatures are sufficiently dissimilar (measured via cosine similarity) from those already selected. Simpler formulae are preferred in the presence of redundancy. Each formula may involve one or more signal dimensions, allowing the model to express and attend to complex relationships in multivariate inputs. The final set is generated once using a fixed time window, and can be rescaled for datasets of different lengths by adjusting temporal thresholds. This ensures that the selected formulae are both diverse and interpretable, capturing a wide range of temporal behaviours.

4.1 Architecture

Our architecture performs interpretable time series classification by integrating symbolic reasoning based on Signal Temporal Logic (STL) with neural representations of trajectories. It operates over a fixed set of human-interpretable STL formulae $\Phi_C = \{\varphi_1, \ldots, \varphi_n\}$, used as temporal logic concepts. Given an input time series $\boldsymbol{x}$, we compute its embedding $\mathcal{H}_{\boldsymbol{x}} \in \mathbb{R}^n$ by evaluating the robustness $\rho(\varphi_i, \boldsymbol{x})$ for each concept $\varphi_i \in \Phi_C$. This results in a fully interpretable vector where each dimension reflects the degree to which a known temporal property is satisfied. To assess how informative each concept is for classification, we compute class-specific discriminability scores. The model supports an arbitrary number of classes K, enabling it to scale naturally to multiclass classification tasks. For each class k and formula φ_i, we define a discriminability matrix $\mathcal{G}_{\boldsymbol{x}} \in \mathbb{R}^{n \times K}$ with entries

$$\mathcal{G}_i^k = \frac{\rho(\varphi_i, \boldsymbol{x}) - \mu_{k^*, i}}{\sigma_{k^*, i} + \varepsilon_G}$$

where $k^* \neq k$, and $\mu_{k^*, i}, \sigma_{k^*, i}$ are the mean and standard deviation of robustness values over class k^*, and ε_G is a small constant. This normalisation highlights whether a concept is atypical for other classes. A soft attention mechanism identifies the most relevant concepts via:

$$\alpha(\boldsymbol{x}) = \mathrm{softmax}\left(\mathcal{H}_{\boldsymbol{x}} / T_{attn}\right) \in \mathbb{R}^n$$

where T_{attn} is a temperature parameter. The weighted combination of attention and discriminability produces the modulated score matrix:

$$\boldsymbol{z}(\boldsymbol{x}) = \alpha(\boldsymbol{x}) \odot \mathcal{G}_{\boldsymbol{x}} \in \mathbb{R}^{n \times K}$$

which is flattened and passed through a linear layer to obtain the class prediction. This architecture ensures interpretability both through attention weights, indicating which concepts influenced the prediction, and through discriminability scores, explaining why those concepts support or oppose the predicted class.

4.2 Explanation extraction

Explanations are derived by combining three elements: attention weights $\alpha(\boldsymbol{x})$, discriminability scores $\mathcal{G}_{\boldsymbol{x}}$, and classifier weights $\boldsymbol{W}$. For a predicted class $\hat{y}$, the explanation scores are computed as $\boldsymbol{r} = \boldsymbol{W}\hat{y} \cdot \boldsymbol{z}(\boldsymbol{x})\hat{y}$, where $\boldsymbol{z}(\boldsymbol{x})$ reflects the relevance of each concept. The most influential STL formulae are selected either by choosing the top-γ highest absolute relevance scores or by applying a cumulative relevance threshold. These selected formulae $\mathcal{F}_{\boldsymbol{x}}$ are then simplified by adjusting thresholds, polarities, and structure, resulting in a refined set $\mathcal{F}'_{\boldsymbol{x}}$. The local explanation is a conjunction:

$$E_l(\boldsymbol{x}, \hat{y}) = \bigwedge_{\varphi'_i \in \mathcal{F}'_{\boldsymbol{x}}} \varphi'_i$$

which serves as a sufficient, human-readable condition for predicting $\hat{y}$. To extract global explanations for a class k, all local explanations of training samples with label k are aggregated into a set $\mathcal{F}_k$. A discriminative subset $\mathcal{F}'_k$ is then selected by solving a minimum-cost set cover problem, where cost reflects syntactic complexity. The resulting global explanation is a disjunction:

$$E_g(k) = \bigvee_{\varphi_i \in \mathcal{F}'_k} \varphi_i$$

capturing sufficient symbolic conditions for class membership.
To further enhance clarity and relevance, all formulae undergo three post-processing steps: (i) logical simplification (e.g. removing redundancies), (ii) threshold adjustment and polarity correction to better align with data, and (iii) data-aware pruning of consistently true/false subformulae. These steps improve the interpretability and discriminative power of both local and global explanations.

5 Planned experimental evaluation

To evaluate both the predictive performance and interpretability of our approach, we plan to conduct experiments on a wide selection of datasets from the UCR Time Series Archive [3]. This benchmark includes over 140 univariate and multivariate classification tasks spanning diverse domains such as physiological monitoring, sensor data, and motion capture, and offers a robust testbed for assessing generalisability.

Our method will be compared against a broad set of state-of-the-art time series classification models, including both high-performing black-box architectures and models explicitly designed

for interpretability. This comparison will enable us to assess trade-offs between accuracy and explainability across modelling paradigms. All models will be trained and evaluated using the standard train/test splits provided by the UCR repository, ensuring consistency with prior benchmarks [19, 20].

6 Conclusions

We have introduced a neuro-symbolic framework for time series classification that leverages Signal Temporal Logic (STL) as a foundation for interpretable concept learning. By embedding both trajectories and STL formulae into a shared semantic space, our approach enables faithful, logic-grounded explanations at both local and global levels. Each prediction is accompanied by temporally meaningful logical conditions, supporting model transparency and potential downstream use in high-stakes or regulated settings.

References

[1] G. Bombara, C.-I. Vasile, F. Penedo, H. Yasuoka, and C. Belta. A decision tree approach to data classification using signal temporal logic. In *Proceedings of the 19th International Conference on Hybrid Systems: Computation and Control*, HSCC '16, page 1–10, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450339551. doi: 10.1145/2883817.2883843. URL https://doi.org/10.1145/2883817.2883843.

[2] L. Bortolussi, G. M. Gallo, J. Kretínský, and L. Nenzi. Learning model checking and the kernel trick for signal temporal logic on stochastic processes. In *Tools and Algorithms for the Construction and Analysis of Systems - 28th International Conference, TACAS 2022,Proceedings, Part I*, 2022.

[3] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, and E. Keogh. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019. doi: 10.1109/JAS.2019.1911747.

[4] European Commission. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[5] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33:917 – 963, 2018.

[6] A. Ghorbani, J. Wexler, J. Zou, and B. Kim. Towards automatic concept-based explanations, 2019.

[7] D. Helbing. *Societal, Economic, Ethical and Legal Challenges of the Digital Revolution: From Big Data to Deep Learning, Artificial Intelligence, and Manipulative Technologies*, pages 47–72. Springer International Publishing, Cham, 2019. ISBN 978-3-319-90869-4. doi: 10.1007/978-3-319-90869-4_6.

[8] A. Ho. Deep ethical learning: Taking the interplay of human and artificial intelligence seriously. *The Hastings Center report*, 49:36–39, 02 2019. doi: 10.1002/hast.977.

[9] T.-Y. Hsieh, S. Wang, Y. Sun, and V. Honavar. Explainable multivariate time series classification: A deep neural network which learns to attend to important variables as well as informative time intervals, 2020. URL https://arxiv.org/abs/2011.11631.

[10] M. Iman, H. R. Arabnia, and R. M. Branchinst. Pathways to artificial general intelligence: A brief overview of developments and ethical issues via artificial intelligence, machine learning, deep learning, and data science. In H. R. Arabnia, K. Ferens, D. de la Fuente, E. B. Kozerenko, J. A. Olivas Varela, and F. G. Tinetti, editors, *Advances in Artificial Intelligence and Applied Cognitive Computing*, pages 73–87, Cham, 2021. Springer International Publishing. ISBN 978-3-030-70296-0.

[11] F. Karim, S. Majumdar, H. Darabi, and S. Chen. Lstm fully convolutional networks for time series classification. *IEEE Access*, 6:1662–1669, 2018. ISSN 2169-3536. doi: 10.1109/access.2017.2779939. URL http://dx.doi.org/10.1109/ACCESS.2017.2779939.

[12] D. Li, M. Cai, C.-I. Vasile, and R. Tron. Tlinet: Differentiable neural network temporal logic inference, 2024. URL https://arxiv.org/abs/2405.06670.

[13] O. Maler and D. Nickovic. Monitoring temporal properties of continuous signals. volume 3253, pages 152–166, 01 2004. ISBN 978-3-540-23167-7. doi: 10.1007/978-3-540-30206-3_12.

[14] S. Mohammadinejad, J. V. Deshmukh, A. G. Puranic, M. Vazquez-Chanlatte, and A. Donzé. Interpretable classification of time-series data using efficient enumerative techniques. In *HSCC '20: 23rd ACM International Conference on Hybrid Systems: Computation and Control*, pages 9:1–9:10. ACM, 2020.

[15] L. Nenzi, S. Silvetti, E. Bartocci, and L. Bortolussi. A robust genetic algorithm for learning temporal specifications from data. In *Quantitative Evaluation of Systems - 15th International Conference, Beijing, China, Proceedings*, volume 11024 of *Lecture Notes in Computer Science*, pages 323–338. Springer, 2018.

[16] G. Pagliarini, S. Scaboro, G. Serra, G. Sciavicco, and E. I. Stan. Neural-symbolic temporal decision trees for multivariate time series classification. In *Time*, 2024.

[17] A. Pnueli. The temporal logic of programs. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, pages 46–57, 1977.

[18] T. Rojat, R. Puget, D. Filliat, J. D. Ser, R. Gelin, and N. Díaz-Rodríguez. Explainable artificial intelligence (xai) on timeseries data: A survey, 2021.

[19] A. P. Ruiz, M. Flynn, J. Large, M. Middlehurst, and A. Bagnall. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 35:401 – 449, 2020.

[20] A. P. Ruiz, M. Flynn, J. Large, M. Middlehurst, and A. Bagnall. The great multivariate time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 35(2):401–449, Dec 2021. doi: 10.1007/s10618-020-00727-3.

[21] G. Saveri and L. Bortolussi. Retrieval-augmented mining of temporal logic specifications from data, 2024.

[22] A. Theissler, F. Spinnato, U. Schlegel, and R. Guidotti. Explainable ai for time series classification: A review, taxonomy and research directions. *IEEE Access*, 10:100700–100724, 2022.

[23] H. Turbé, M. Bjelogrlic, C. Lovis, and G. Mengaldo. Evaluation of post-hoc interpretability methods in time-series classification. *Nature Machine Intelligence*, 5:250–260, 2023.

[24] R. Yan, T. Ma, A. Fokoue, M. Chang, and A. A. Julius. Neuro-symbolic models for interpretable time series classification using temporal logic description. *2022 IEEE International Conference on Data Mining (ICDM)*, pages 618–627, 2022.

[25] C.-K. Yeh, B. Kim, S. Arik, C.-L. Li, T. Pfister, and P. Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20554–20565. Curran Associates, Inc., 2020.