

TSInterpret: A unified framework for time series interpretability

Jacqueline Höllig

HOELLIG@FZI.DE

*Information Process Engineering
FZI Forschungszentrum Informatik
76131 Karlsruhe, Germany*

Cedric Kulbach

KULBACH@FZI.DE

*Information Process Engineering
FZI Forschungszentrum Informatik
76131 Karlsruhe, Germany*

Steffen Thoma

THOMA@FZI.DE

*Information Process Engineering
FZI Forschungszentrum Informatik
76131 Karlsruhe, Germany*

Abstract

With the increasing application of deep learning algorithms to time series classification, especially in high-stake scenarios, the relevance of interpreting those algorithms becomes key. Although research in time series interpretability has grown, accessibility for practitioners is still an obstacle. Interpretability approaches and their visualizations are diverse in use without a unified api or framework. To close this gap, we introduce TSInterpret¹, an easily extensible open-source Python library for interpreting predictions of time series classifiers that combines existing interpretation approaches into one unified framework. The library features (i) state-of-the-art interpretability algorithms, exposes a (ii) unified API enabling users to work with explanations in a consistent way, and provides (iii) suitable visualizations for each explanation.

Keywords: Time Series; Interpretability; Feature Attribution; Counterfactual Explanation

1. Introduction

Although time series data are omnipresent in industry and daily life, deep learning methods have only been applied to time series data in the last decade. Before, time series classification (TSC) primarily focused on classifications based on discriminatory features (e.g., learning based on the whole time series, intervals, or shapelets (Bagnall et al. (2017))). Compared to deep methods, they do not have the "black-box"-character as a significant drawback. However, with the increasing accuracy of deep learning models and their ability to cope with vast amounts of data, their application on time series classification increased, leading to a need for interpretability, especially in high-risk settings. Various interpretability methods for time series classification are available (Rojat et al. (2021)). However, the usage of those methods is not yet standardized: The proposed methods often lack a) open code (e.g., Siddiqui et al. (2021)), b) an easy-to-use interface (e.g., Ismail et al. (2020)), or

1. <https://github.com/JHoelli/TSInterpret>

c) visualization (e.g., Guilleme et al. (2019)), making the application of those methods inconvenient and thereby hindering the usage of deep learning methods on safety-critical scenarios.

Although unifying frameworks (e.g., tf-explain (Meudec (2021)), Alibi (Klaise et al. (2021)), or captum (Kokhlikyan et al. (2020))) have already been developed for various data types (tabular, image, text) and algorithms (e.g., GradCam (Selvaraju et al. (2020)), Counterfactuals (Mothilal et al. (2020)), SHAP (Lundberg and Lee (2017))), an interpretability framework for time-series data is still missing. Due to the different structures and properties of time-ordered data, most approaches and frameworks are not directly applicable to time series (Ismail et al. (2020)). Further, time series are not intuitive for human understanding (Siddiqui et al. (2019)) and need additional visualizations. Therefore, we propose TSInterpret, a framework implementing algorithms for time series classification. In this work, we provide

- a review of existing interpretability libraries,
- a unified framework for time series interpretability,
- and unified visualizations for the implemented algorithms.

2. Need for Interpretability in Time Series Classification

Temporal data is ubiquitous and encountered in many real-world applications ranging from electronic health records (Rajkomar et al. (2018)) to cyber security (Susto et al. (2018)). Although almost omnipresent, time series classification has been considered one of the most challenging problems in data mining for the last two decades (Yang and Wu (2006); Esling and Agon (2012)). With the rising data availability and accessibility (e.g., provided by the UCR / UEA archive (Bagnall et al. (2017); Dau et al. (2019))), hundreds of time series classification algorithms have been proposed. Although deep learning methods have been successful in the field of Computer Vision (CV) and Natural Language Processing (NLP) for almost a decade, application on time series has only occurred in the past few years (e.g., Fawaz et al. (2019); Rajkomar et al. (2018); Susto et al. (2018); Ruiz et al. (2021)). Deep learning models have been shown to achieve state-of-art results on time series classification (e.g., Fawaz et al. (2019)). However, those methods are black-boxes due to their complexity which limits their application to high-stake scenarios (e.g., in medicine or autonomous driving), where user trust and understandability of the decision process are crucial.

Although much work has been done on interpretability in CV and NLP, most developed approaches are not directly applicable to time series data. The time component impedes the usage of existing methods (Ismail et al. (2020)). Thus, increasing effort is put into adapting existing methods to time series (e.g., LEFTIST based on SHAP / Lime (Guilleme et al. (2019)), Temporal Saliency Rescaling for Saliency Methods (Ismail et al. (2020)), Counterfactuals (Ates et al. (2021); Delaney et al. (2021))), and developing new methods specifically for time series interpretability (e.g., TSInsight based on autoencoders (Siddiqui et al. (2021)), TSViz for interpreting CNN (Siddiqui et al. (2019))). For a survey of time series interpretability, please refer to Rojat et al. (2021).

Compared to images or textual data, humans cannot intuitively and instinctively understand the underlying information contained in time series data. Therefore, time series

data, both uni- and multivariate, have an unintuitive nature, lacking an understanding at first sight (Siddiqui et al. (2019)). Hence, providing suitable visualizations of time series interpretability becomes crucial.

For instance, health data tasks like cardiac disease detection from electrocardiogram data (ECG5000, Dau et al. (2019)) or seizure prediction from the data of a tri-axial accelerometer (Epilepsy, Bagnall et al. (2018)) are examples for the need of interpretability within time series data. Those classification tasks are in a sensitive field regarding people’s lives. Therefore, decisions need to be carefully taken based on solid evidence. Wrong medications or unidentified diseases can have long-lasting effects on a patient’s health. To allow physicians to make such data-driven decisions with the help of machine learning, interpretations of black-box models are crucial. It is not only relevant if a patient has epilepsy or a cardiac disease but why or which indications are there for such. Further, implementing a non-interpretable machine learning model in medicine raises legal and ethical issues (Amann et al. (2020)).

3. Related Work

Name	Backend	Data				Task		Scope				Output		
		Tabular	Text	Image	Time Series	Classification	Regression	Model Agnostic	Model Specific	Local	Global	FA	IB	RB
AIX360 (Arya et al. (2019))	PYT,TF, BB	X	X	X		X	X	X	X	X	X	X	X	X
ALIBI (Klaise et al. (2021))	BB	X	X	X		X	X	X	X	X	X	X	X	
Anchor (Ribeiro et al. (2018))	BB	X	X					X		X			X	
Captum (Kokhlikyan et al. (2020))	PYT	X	X	X		X	X	X	X	X		X		
carla (Pawelczyk et al. (2021))	PYT, TF, SK	X				X	X	X		X			X	
DALEX (Baniecki et al. (2021))	SK,TF	X				X	X	X		X	X	X		
DeepExplain (Ancona et al. (2018))	TF		X	X		X	X			X	X	X		
Dice (Mothilal et al. (2020))	PYT, TF, BB	X				X	X	X		X			X	
ELI5 (eli5 org (2017))	BB	X	X	X		X		X		X		X		
H2O (Hall et al. (2019))	-	X			X	X	X	X	X	X	X	X		
iNNvestigate (Alber et al. (2019))	TF			X		X			X	X	X	X		
InterpretML (Nori et al. (2019))	BB, WB	X				X	X	X	X	X	X	X		
Lucid (tensorflow (2018))	TF			X		X	X		X	X	X	X		
OmniXAI (Yang et al. (2022))	PYT, TF, SK	X	X	X	X	X	X	X	X	X	X	X	X	
pytorch-cnn-visualizations (Ozbulak (2019))	PYT			X		X			X	X		X		
SHAP (Lundberg and Lee (2017))	PYT, TF, BB	X	X	X		X	X	X	X	X	X	X		
Skater (Oracle (2018))	BB	X	X	X		X	X			X	X	X	X	
Tf-Explain (Meudec (2021))	TF			X		X			X	X		X		
TorchRay (Fong et al. (2019))	PYT			X		X			X			X		
What-if (Wexler et al. (2019))	TF	X	X	X		X	X	X					X	
wildboar (Samsten (2020))	-				X	X	X		X	X			X	

Table 1: Overview over recent explanation libraries and their time series capabilities. BB: Black-Box, TF: Tensorflow, PYT: PyTorch, SK: scikit-learn

Numerous open-source tools for machine learning evolved in the last few years. In the following, we focus on libraries implementing various post-hoc interpretability methods in-

stallable via PyPI. For the summary in Table 2, only active libraries, i.e., libraries with some development activity in the past 12 months on GitHub, are included. The comparison is conducted according to the scope of included explanations (Model Agnostic vs. Model-specific, Global vs. Local) and library properties (supported model libraries, data types, and tasks). Model Agnostic methods apply to any classification algorithms, whereas model-specific algorithms only work on a subset of classification algorithms (e.g., Cam on Convolutional Neural Networks). Global interpretability methods interpret the model’s overall decision-making process, while local interpretability focuses on the interpretation of a single instance and the prediction of that instance². Further, the return type of the explanation methods are taken into account. Feature Attribution methods (FA) return a per-feature attribution score based on the feature’s contribution to the model’s output (e.g., GradCam (Selvaraju et al. (2020)) or SHAP (Lundberg and Lee (2017))), instance-based methods (IB) calculate a subset of relevant features that must be present to retain or remove a change in the prediction of a given model (e.g., counterfactuals (Mothilal et al. (2020)) or anchors (Ribeiro et al. (2018))), while rule-based methods (RB) derive rules.

Except for the libraries wildboar (Samsten (2020)), OmniXAI (Yang et al. (2022)), and H2O (Hall et al. (2019)), no library provides interpretability methods on time series. However, the support of OmniXAI (Yang et al. (2022)) is restricted to anomaly detection, univariate time series and includes only non-time-series-specific interpretability methods. Wildboar (Samsten (2020)) focuses on temporal machine learning (classification, regression, and anomaly detection) and only provides a limited number of (counterfactual) interpretability methods as additional features. A unified framework for time series interpretability like captum or ALIBI enabling easy-to-use interpretability is still missing.

4. Library Design

The API Design is orientated on the scikit design paradigms consistency, sensible defaults, composition, nonproflarity of classes, and inspection (Buitinck et al. (2013)). Listing 1

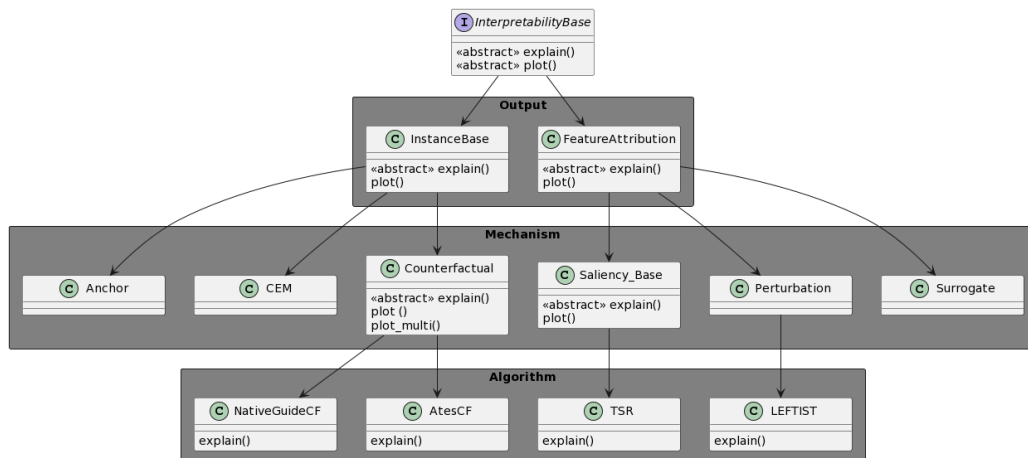


Figure 1: Structure of TSInterpret.

2. More information on the taxonomy are available in Rojat et al. (2021).

shows the workflow of the library in a coding sample. Given a trained machine learning model and an instance to be classified: First, the desired interpretability method is imported (line 1) and instantiated (line 2), followed by explaining the instance (line 3), and finally, the generation of the plot (line 4).

Code Listing 1: API Usage Example.

```
from TSInterpret.InterpretabilityModels.Saliency.TSR import TSR

int_mod = TSR(model, item.shape[-1], item.shape[-2])
exp = int_mod.explain(item, labels = label, TSR = True)
int_mod.plot(item, exp)
```

Consistency All implemented objects share a consistent interface. Every interpretability method inherits from the interface `InterpretabilityBase` to ensure that all methods contain a method `explain` and a `plot` function. The `plot` function is implemented on the level below based on the output structure provided by the interpretability algorithm to provide a unified visualization experience (e.g., in the case of Feature Attribution, the plot function visualizes a heatmap on the original sample). If necessary, those plots are refined by the Mechanism layer. This is necessary to ensure suitable representation as the default visualization can sometimes be misinterpreted (e.g., the heatmap used in the plot function of `InterpretabilityBase` allows positive and negative values, while `TSR` is scaled to $[0, 1]$). Using the same color pattern for both scales would lead to a high risk of misinterpreting results while comparing `TSR` with `LEFTIST`). The `explain` function is implemented on the method level.

Sensible Defaults `TSInterpret` provides reasonable defaults for most parameters by providing the default parameterizations for each method from the designated papers. Those parameters can easily be changed if needed by providing alternative values during model instantiation.

Composition Many interpretability methods for time series classification are based on already existing methods for tabular, image, or text data (e.g., Ismail et al. (2020) is based amongst others on Lundberg and Lee (2017)). Whenever feasible, existing implementation of such algorithms are used (e.g., SHAP (Lundberg and Lee (2017)), captum (Kokhlikyan et al. (2020)), or tf-explain (Meudec (2021))).

Nonproflarity of classes

`TSInterpret` implements the interpretability algorithms as custom classes. Datasets, instances, and results are represented as NumPy arrays, Lists, or Tuples, instead of classes. For instance, the counterfactual method returns a Tuple of the counterfactual time series and the label (list, int). Hyperparameters are regular strings or numbers.

Inspection `TSInterpret` stores and exposes the parameter of the interpretability algorithms as public attributes. In some methods, parameters have a significant impact on the obtained results. Making those parameters publicly available through attributes facilitates experimenting with hyperparameters.

5. Algorithm Overview

As depicted in section 3, our framework provides support for various time series interpretability types, resulting in interpretations and visualization for a variety of use cases. The current version of the library includes the interpretability algorithms listed in Table 2. Their implementation in TSInterpret is based on code provided by the authors of the algorithms, which got adapted to the framework and extended to ensure the availability on multiple model backends (e.g., PyTorch or TensorFlow).

Method	Model	Explanations	Type	Dataset
NUN-CF (Delaney et al. (2021))	TF, PYT, SK	IB	uni	y
CoMTE (Ates et al. (2021))	TF, PYT, SK	IB	multi	y
LEFTIST (Guilleme et al. (2019))	TF, PYT, SK	FA	uni	y
TSR (Ismail et al. (2020))	TF, PYT	FA	multi	n

Table 2: Interpretability Methods implemented in TSInterpret. BB: Black-Box, TF: TensorFlow, PYT: PyTorch, SK: scikit-learn

NUN-CF ³ Delaney et al. (2021) proposed using the K-nearest neighbors from the dataset belonging to a different class as native guide to generate counterfactuals. They propose three options for transforming the original time series with this native guide: the plain native guide, the native guide with bary centering, and transformation based on the native guide and class activation mapping.

CoMTE ⁴ Ates et al. (2021) proposed CoMTE as a perturbation-based approach for multivariate time series counterfactuals. The goal is to exchange the smallest possible number of features with reference data by applying random-restart hill climbing to obtain a different classification.

LEFTIST ⁵ Agnostic Local Explanation for Time Series Classification by Guilleme et al. (2019) adapted LIME for time series classification and proposed to use prefixed shapelets as the interpretable components. Each shapelet is a non-overlapping subsection of the original time series with a prefixed length. The feature importance is provided on a shapelet basis.

TSR ⁶ Temporal Saliency Rescaling (Ismail et al. (2020)) calculates the importance of each timestep, followed by the feature importance based on different Saliency Methods, both back-propagation based and perturbation based. We refer the reader to our code documentation for a complete list of implemented methods. The implementation in TSInterpret is based on tf-explain (Meudec (2021)), SHAP (Lundberg and Lee (2017)) and captum (Kokhlikyan et al. (2020)).

3. https://github.com/e-delaney/Instance-Based_CFE_TSC

4. <https://github.com/peaclab/CoMTE>

5. https://www.dropbox.com/s/y1xq5bhpf0irg2h/code_LEFTIST.zip?dl=0

6. <https://github.com/ayaabdel salam91/TS-Interpretability-Benchmark>

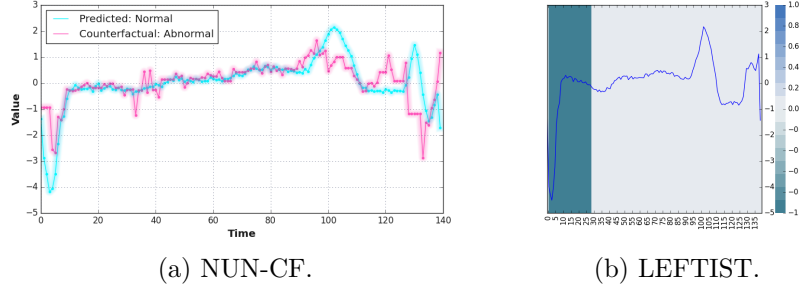


Figure 2: Interpretations obtained with TSInterpret the univariate ECG5000 dataset.

Returning to the use case described in section 2, we train a 1D-conv ResNet on ECG5000 and Epilepsy. Figure 2 and Figure 3 show interpretations obtained with TSInterpret for the 1D-conv ResNet on the first item of the testset for the univariate ECG5000 dataset (fig. 2) and the multivariate Epileptical dataset (fig. 3). With the help of the interpretations shown in Figure 2 a physician can gather evidence on the functionality of the ResNet classification on ECG 5000. The feature attribution method LEFTIST (fig. 2b) identifies sections with positive/ negative influence on the current time series prediction. In this case, the first section of the time series has the most considerable impact on the current prediction being a normal electrocardiogram. If a physician wants to know why the electrocardiogram shows a normal instance and not an abnormal beat, conclusions can be drawn from the counterfactual approach NUN-CF. Compared to the original time series (in blue fig. 2a) classified as normal, the classification is changed if the drop at the beginning is lower and the drop at the end is replaced with a rise (pink line in fig. 2a). Figure 3 shows interpretations for the epilepsy dataset. Suppose a physician wants to know how the original instance is predicted as walking (blue) instead of running. In such a case, the practitioner applies CoMTE to generate a counterfactual in the class direction of running, resulting in a different acceleration of the x-axis in the first 200 time steps (see fig. 3a). If the physician is only interested in seeing the most important features TSR (fig. 3b) can be applied. With the help of interpretability, a physician can gather evidence for the predicted classification and gain trust in the system (Amann et al. (2020)). In general, the visualization for the

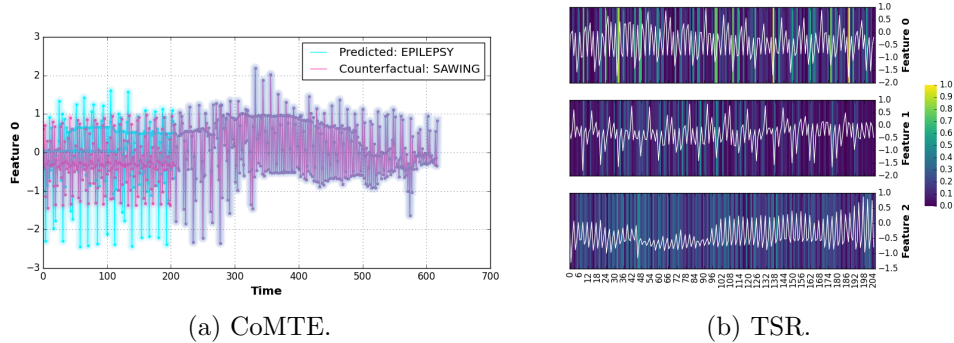


Figure 3: Interpretations obtained with TSInterpret the multivariate Epilepsy dataset.

counterfactual approaches CoMTE and NUN-CF have a similar style, although NUN-CF visualizes univariate and CoMTE multivariate data. Note that CoMTE only visualizes the changed features. For the feature attribution methods LEFTIST and TSR, the visualizations have similar styles but a different color map. This is necessary as LEFTIST returns a positive and negative influence (range $[-1, 1]$), while TSR returns normalized time slices and feature importances (range $[0, 1]$).

The easy-to-use interface of TSInterpret allows the out-of-box usage of interpretability models on time series data for different use cases, time series types, and classification models. Further, the framework can be easily extended by inheriting from either `InterpretabilityBase` or one of the more customized classes (e.g., `Feature Attribution` or `InstanceBased`). All classes below the Output layer also come with a default plot function to match the visualizations obtained by the already implemented algorithms.

6. Outlook

TSInterpret provides a cross-backend unified API for the interpretability of time series classifications enabling interpretation generation with just three lines of code. In the first phase of the development, TSInterpret provides four interpretability algorithms for both uni- and multivariate time series classification and functions for visualizing the interpretation results. However, the Framework is easily extensible by inheritance and provides a simple API design. In the future, we will include support for additional interpretability algorithms, time series prediction, as well as metrics to progress into an easy-to-use benchmarking tool for time series interpretability. In order to focus the research and application of time series prediction models not only on performance but also on understanding the model in a practical environment, a further development step is to integrate TSInterpret into the landscape of existing frameworks and thus establish it as an easily usable standard.

Acknowledgments

This work was carried out with the support of the German Federal Ministry of Education and Research (BMBF) within the project "MetaLearn" (Grant 02P20A013).

References

- M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans. iNNvestigate Neural Networks! *Journal of Machine Learning Research*, 20(93):1–8, 2019.
- J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*, 20(1):310, Dec. 2020. ISSN 1472-6947. doi: 10.1186/s12911-020-01332-6.
- M. Ancona, E. Ceolini, C. Öztireli, and M. Gross. Towards better understanding of gradient-based attribution methods for Deep Neural Networks, Mar. 2018.
- V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghaven-

- dra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques, Sept. 2019.
- E. Ates, B. Aksar, V. J. Leung, and A. K. Coskun. Counterfactual Explanations for Machine Learning on Multivariate Time Series Data. *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*, pages 1–8, May 2021. doi: 10.1109/ICAPAI49758.2021.9462056. arXiv: 2008.10781.
- A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Min Knowl Disc*, 31(3):606–660, May 2017. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-016-0483-9.
- A. Bagnall, H. A. Dau, J. Lines, M. Flynn, J. Large, A. Bostrom, P. Southam, and E. Keogh. The UEA multivariate time series classification archive, 2018. *arXiv:1811.00075*, Oct. 2018.
- H. Baniecki, W. Kretowicz, P. Piatyszek, J. Wisniewski, and P. Biecek. dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python. *Journal of Machine Learning Research*, 22(214):1–7, 2021.
- L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. Vanderplas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: experiences from the scikit-learn project, Sept. 2013. Number: arXiv:1309.0238 arXiv:1309.0238 [cs].
- 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. *arXiv:1810.07758*, Sept. 2019.
- E. Delaney, D. Greene, and M. T. Keane. Instance-Based Counterfactual Explanations for Time Series Classification. In A. A. Sánchez-Ruiz and M. W. Floyd, editors, *Case-Based Reasoning Research and Development*, volume 12877, pages 32–47. Springer International Publishing, Cham, 2021. ISBN 978-3-030-86956-4 978-3-030-86957-1. doi: 10.1007/978-3-030-86957-1_3. Series Title: Lecture Notes in Computer Science.
- eli5 org. ELI5, 2017. URL <https://github.com/eli5-org/eli5>.
- P. Esling and C. Agon. Time-series data mining. *ACM Comput. Surv.*, 45(1):1–34, Nov. 2012. ISSN 0360-0300, 1557-7341. doi: 10.1145/2379776.2379788.
- H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. Deep learning for time series classification: a review. *Data Min Knowl Disc*, 33(4):917–963, July 2019. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-019-00619-1. arXiv: 1809.04356.
- R. Fong, M. Patrick, and A. Vedaldi. Understanding Deep Networks via Extremal Perturbations and Smooth Masks, Oct. 2019.

- M. Guilleme, V. Masson, L. Roze, and A. Termier. Agnostic Local Explanation for Time Series Classification. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 432–439, Portland, OR, USA, Nov. 2019. IEEE. ISBN 978-1-72813-798-8. doi: 10.1109/ICTAI.2019.00067.
- P. Hall, N. Gill, M. Kurka, W. Phan, and A. Bartz. Machine Learning Interpretability with H2O Driverless AI. *H2O. ai*, page 40, 2019.
- A. A. Ismail, M. Gunady, H. C. Bravo, and S. Feizi. Benchmarking Deep Learning Interpretability in Time Series Predictions. *arXiv:2010.13924*, Oct. 2020.
- J. Klaise, A. V. Looveren, G. Vacanti, and A. Coca. Alibi Explain: Algorithms for Explaining Machine Learning Models. *Journal of Machine Learning Research*, 22(181):1–7, 2021.
- N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch, Sept. 2020.
- S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.
- R. Meudec. tf-explain, Feb. 2021. URL <https://github.com/sicara/tf-explain>. doi: 10.5281/ZENODO.5711704.
- R. K. Mothilal, A. Sharma, and C. Tan. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, Jan. 2020. doi: 10.1145/3351095.3372850. arXiv: 1905.07697.
- H. Nori, S. Jenkins, P. Koch, and R. Caruana. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223*, 2019.
- Oracle. Skater, 2018. URL <https://github.com/oracle/Skater>.
- U. Ozbulak. PyTorch CNN Visualizations, 2019. URL <https://github.com/utkuozbulak/pytorch-cnn-visualizations>. Publication Title: GitHub repository.
- M. Pawelczyk, S. Bielawski, J. v. d. Heuvel, T. Richter, and G. Kasneci. CARLA: A Python Library to Benchmark Algorithmic Recourse and Counterfactual Explanation Algorithms. *arXiv:2108.00783*, Aug. 2021.
- A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboun, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. D. Howell, C. Cui, G. S. Corrado, and J. Dean. Scalable and accurate deep learning with

- electronic health records. *npj Digital Med*, 1(1):18, Dec. 2018. ISSN 2398-6352. doi: 10.1038/s41746-018-0029-1.
- M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-Precision Model-Agnostic Explanations. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.
- T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, and N. Díaz-Rodríguez. Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey. *arXiv:2104.00950*, 2021.
- 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 Min Knowl Disc*, 35(2):401–449, Mar. 2021. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-020-00727-3.
- I. Samsten. isaksamsten/wildboar: wildboar, Nov. 2020. URL <https://github.com/isaksamsten/wildboar>. doi: 10.5281/ZENODO.4264063.
- R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int J Comput Vis*, 128(2):336–359, Feb. 2020. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-019-01228-7. arXiv: 1610.02391.
- S. A. Siddiqui, D. Mercier, M. Munir, A. Dengel, and S. Ahmed. TSViz: Demystification of Deep Learning Models for Time-Series Analysis. *IEEE Access*, 7:67027–67040, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2912823. arXiv: 1802.02952.
- S. A. Siddiqui, D. Mercier, A. Dengel, and S. Ahmed. TSInsight: A Local-Global Attribution Framework for Interpretability in Time Series Data. *Sensors*, 21(21):7373, Nov. 2021. ISSN 1424-8220. doi: 10.3390/s21217373.
- G. A. Susto, A. Cenedese, and M. Terzi. Time-Series Classification Methods: Review and Applications to Power Systems Data. In *Big Data Application in Power Systems*, pages 179–220. Elsevier, 2018. ISBN 978-0-12-811968-6. doi: 10.1016/B978-0-12-811968-6.00009-7.
- tensorflow. Lucid, 2018. URL <https://github.com/tensorflow/lucid>.
- J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viegas, and J. Wilson. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Trans. Visual. Comput. Graphics*, pages 1–1, 2019. ISSN 1077-2626, 1941-0506, 2160-9306. doi: 10.1109/TVCG.2019.2934619.
- Q. Yang and X. Wu. 10 Challenging Problems In Data Mining Research. *Int. J. Info. Tech. Dec. Mak.*, 05(04):597–604, Dec. 2006. ISSN 0219-6220, 1793-6845. doi: 10.1142/S0219622006002258.
- W. Yang, H. Le, S. Savarese, and S. C. H. Hoi. OmniXAI: A Library for Explainable AI, June 2022. arXiv:2206.01612.