Software Frameworks for Explainable AI in Time Series Classification: A Systematic Review

Louis Peter[1,2][0009−0001−7478−1888], Nils Gumpfer[1,2][0000−0001−8644−9885], Jana Fischer[1,2][0009−0005−0445−0847], Christin Seifert[2,3][0000−0002−6776−3868], and Jennifer Hannig[1,2][0000−0002−2789−5540] (✉)

[1] Technische Hochschule Mittelhessen - University of Applied Science, Friedberg, Germany
[2] Hessian Center for AI (hessian.AI), Darmstadt, Germany
{louis.peter,nils.gumpfer,jana.fischer,jennifer.hannig}@kite.thm.de
[3] Marburg University, Marburg, Germany christin.seifert@uni-marburg.de

Abstract. Time series arise in a wide range of application domains and are analyzed using machine learning in decision-critical settings. Time series classification (TSC) is one of the most widely studied and relevant tasks. In this context, ensuring the transparency and trustworthiness of TSC models has become an important requirement, motivating the use of explainable artificial intelligence (XAI) methods. Despite growing interest, research on XAI for TSC remains fragmented, and a systematic understanding of the available software frameworks for explanation generation, their evaluation practices, and practical limitations is still lacking. Prior work largely focused on individual explanation methods, while cross-framework consistency, time-series-specific evaluation, and reproducibility have received little attention. In this survey, we analyze existing software frameworks for explanation generation and evaluation in TSC. We compare them along multiple dimensions, including supported XAI methods, evaluation metrics, usability, benchmarking support, and reproducibility, providing the first time-series-specific survey of frameworks with implementation comparisons and an analysis of frequency-domain support. We identify six frameworks that explicitly support time series and reveal common limitations: only one method supports frequency-domain explanations despite their relevance; only two evaluation metrics have been developed specifically for time series; and identical XAI methods can yield substantially different explanations across frameworks. Based on these findings, we discuss open challenges and outline directions for future research, highlighting the need for unified, time-series-specific XAI frameworks that enable faithful, reproducible, and time-series-aware explanations.

Keywords: Explainable Artificial Intelligence · Time Series Classification · XAI Evaluation · XAI Software Frameworks

1 Introduction

Time series data are a central part of many real-world applications in healthcare, industrial monitoring, finance, and audio processing. The increasing use of deep learning models for time series classification (TSC) has led to substantial gains in predictive performance, but these improvements often come at the cost of interpretability and transparency. In high-risk domains, understanding and justifying automated decisions is critical, both to foster user trust and to comply with regulatory requirements such as the European Union's Artificial Intelligence Act (AI Act), in particular Article 14, which requires human oversight to ensure that users can interpret and effectively oversee AI outputs [13]. These regulatory developments further highlight the necessity of reliable and faithful explanation methods for TSC.

Explainable artificial intelligence (XAI) has consequently emerged as an important research area. However, most existing XAI research has been developed and evaluated primarily in the context of image data [28]. As a result, many explanation methods are applied to time series without being specifically adapted or systematically evaluated [30], leaving their faithfulness and practical usefulness for time series data unclear.

Explanations for time series are less intuitive than for other data modalities, increasing the risk of misleading interpretations [29]. While image classification tasks usually allow even non-experts to visually distinguish between classes such as cats and dogs, many real-world time series lack a clear notion of interpretable components. The raw time series signal often appears noisy and does not exhibit visually interpretable structure, even for domain experts. A representative example is an audio signal of a spoken digit, which is not visually interpretable in the time domain (see the left part of Fig. 4). In such cases, explanations based solely on the time domain are often insufficient, and alternative representations in the frequency and time-frequency domains are usually considered. This highlights the importance of incorporating these representations to generate meaningful explanations. In contrast, the electrocardiogram (ECG) is a multivariate time series for which meaningful and clinically interpretable patterns are well defined and visible in the time domain. Pathology-related patterns can be reliably interpreted by domain experts, particularly cardiologists, and often arise from complex cross-channel dependencies, reflecting spatial and temporal interactions of cardiac activity (see Fig. 1). Consequently, explaining multivariate time series requires XAI methods that explicitly account for cross-channel dependencies, which can be synchronous (i.e., dependencies between channels occurring at the same time step) or asynchronous.

Overall, time series pose unique challenges for explainability. Meaningful patterns often emerge from complex temporal dynamics and cross-channel dependencies, including interactions across time, frequency, or time-frequency domains. As a result, XAI for TSC differs fundamentally from image-based explanations and motivates the need for the systematic comparison, evaluation, and development of XAI methods tailored to the unique properties of time series.

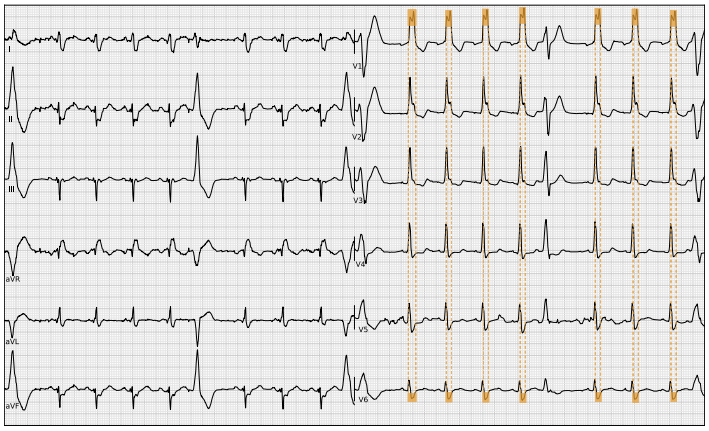

Fig. 1: The electrocardiogram (ECG) signal is a multivariate time series in which each channel corresponds to a distinct measurement location (lead). All channels reflect the same underlying cardiac process with strong temporal and synchronous cross-channel dependencies. An example with right bundle branch block (RBBB) shows characteristic patterns in leads V1 and V6, highlighting the need for multichannel analysis.

Prior surveys have identified key challenges of XAI for time series data. Theissler et al. [36] and Rojat et al. [29] reviewed existing XAI methods and emphasized the need for systematic evaluation but focused primarily on methods rather than the frameworks[4] used to implement, evaluate, and benchmark them in practice. Although multiple frameworks for evaluation exist, the survey by Le et al. [26] showed that, as of 2022, only a single framework explicitly supported time series data. At the same time, Theissler et al. [36] called for unified frameworks to enable comparative and reproducible evaluation of XAI methods for time series. It remains unclear how far this call has been addressed. In particular, there is no consolidated understanding of how recent XAI frameworks support TSC, how they differ in explanation generation and evaluation, or how they align with emerging regulatory requirements.

Addressing this gap motivates the present survey, which provides the first systematic review and comparative analysis of recent XAI frameworks for TSC. In contrast to prior surveys that focus on XAI methods [36,29] or analyze frameworks across different data types [26], we provide the first framework-centric analysis specific to XAI for TSC. We analyze how existing frameworks support the generation, evaluation, and benchmarking of explanations for TSC and assess their usability and practical limitations.

RQ1: Which XAI frameworks currently support TSC, and how do they differ in usability and practical applicability? (Sec. 3.1)

[4] Throughout, we use "framework" to refer to a *software* framework, i.e., a library or tool for generating and/or evaluating explanations.

RQ2: How comprehensively do these frameworks support explanation methods and evaluation metrics, and to what extent do they support univariate and multivariate time series? (Sec. 3.1)
RQ3: To what extent do current frameworks rely on reusing generic XAI methods and evaluation metrics, and how well do they support alternative signal representations? (Sec. 3.2-3.3)
RQ4: How do existing frameworks support benchmarking of XAI methods for TSC, particularly with respect to dataset support and available ground-truth information? (Sec. 3.4)
RQ5: To what extent do differences in framework design and implementation affect the explanations and evaluation results produced by ostensibly identical XAI methods? (Sec. 3.5)

By addressing these questions, this survey provides a consolidated view of the current landscape of XAI frameworks for TSC, highlights systematic strengths and shortcomings, and identifies open challenges for robust, comparable, and time-series-aware XAI frameworks.

2 Methodology

We performed a systematic review of available XAI frameworks for TSC. In this work, an XAI framework denotes any software – such as library or tool – that facilitates either the comparative generation of explanations and/or the evaluation of XAI methods. The review process is illustrated in Fig. 2. We searched GitHub, which is the dominant platform for hosting open-source projects, using the following queries, resulting in 750 entries[5]: `"explainable-ai time-series"`, `"explainable-ai evaluation"`, `"xai time-series"`, `"xai evaluation"`, and `"explanation time-series"`.

After removing duplicates, 589 repositories remained. To exclude unmaintained or incomplete repositories and ensure community relevance, we filtered repositories with fewer than four stars[6] and fewer than four forks[7] resulting in 115 repositories. We then manually assessed the remaining repositories and excluded those that 1.) did not claim the support of time series in the corresponding publication, README file, or documentation, *or* 2.) contained less than two XAI methods and less than two evaluation metrics, *or* 3.) were not installable as a python package, *or* 4.) had no associated publication or an insufficient README file[8].

We identified six candidate repositories. We further examined the frameworks dependencies to identify additional frameworks, resembling a backward search

[5] The search was conducted on Dec 5th, 2025.
[6] GitHub stars represent the number of users who have marked a repository as a favorite.
[7] GitHub forks represent user copies of a repository to work on.
[8] A README was regarded insufficient if it provided neither a tutorial nor documentation for usage.

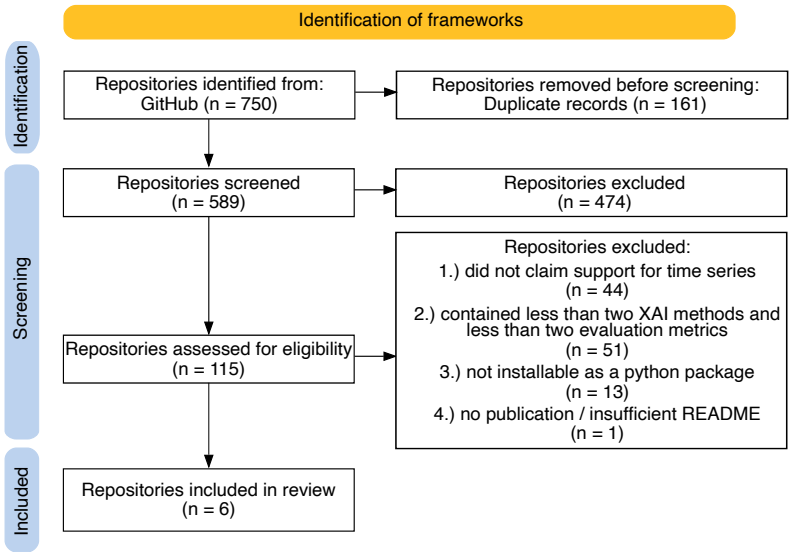

Fig. 2: Flowchart of selection and review process.

for frameworks; however, all identified dependencies met at least one exclusion criterion. As a result, the final survey included six XAI frameworks. For each framework, we conducted a systematic review of the source code repositories, associated publications, documentation, supported datasets, and – where available – the original publications describing the implemented XAI methods and evaluation metrics. The analysis was structured around the following dimensions:

Frameworks. For each framework, we record the number of XAI methods and evaluation metrics, the supported machine-learning backends, whether it is available at a package index, and the corresponding GitHub repository metadata (last update, stars, and forks). To assess general usability, we adopted the usability scores proposed by Le et al. [26], which evaluate the usability of frameworks along three dimensions: *active maintenance, interaction with community*, and *documentation*. Each dimension is scored on a scale from 0 to 5 based on predefined criteria.

XAI Methods. We categorized the XAI methods into counterfactual, gradient-based, and perturbation-based approaches; methods that did not fit these categories were grouped under *other*. We examined whether methods explicitly claim to support time series. For time-series-specific methods, we analyzed their applicability to univariate or multivariate time series and the domains in which explanations are generated (time, frequency, or time-frequency). For methods implemented in multiple frameworks, we compared the resulting explanations to assess cross-framework reproducibility.

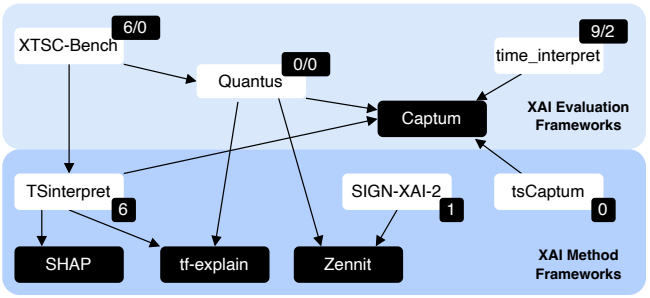

Fig. 3: Dependency ecosystem of XAI frameworks (arrows indicate dependencies). White nodes support time series; black nodes are core dependencies without time series support. For XAI method frameworks, x indicates the number of time-series-specific XAI methods. For XAI evaluation frameworks, X/Y indicates the number of time-series-specific XAI methods (X) and evaluation metrics (Y).

Evaluation Metrics. We investigated whether evaluation metrics are compatible with time series (as explicitly stated) or specific to time series (i.e., originally developed for time series in the corresponding publication). For metrics specific to time series, we analyzed the domains in which explanations can be evaluated (time, frequency, or time-frequency). We categorized metrics into perturbation-based and ground-truth-based metrics; metrics that did not fit these categories were grouped under *other*. For metrics implemented in multiple frameworks, we compared the resulting scores to assess reproducibility across implementations.

Benchmarking Capabilities. We assessed whether frameworks support benchmark datasets for comparing and evaluating XAI methods. Datasets were considered if they can be accessed from the installed packages. We further examined whether the datasets provide ground-truth information (known or expected reference explanations) and whether they contain uni- or multivariate time series.

3 Results

We present our findings along the five questions introduced above. We first characterize the six frameworks and their usability (Sec. 3.1), then analyze XAI methods (Sec. 3.2), evaluation metrics (Sec. 3.3), and benchmarking capabilities (Sec. 3.4), and finally assess cross-framework reproducibility (Sec. 3.5).

3.1 Frameworks

Of the six identified frameworks, three support both explanation generation and evaluation by providing XAI evaluation metrics in addition to XAI methods (time_interpret [12], Quantus [19], and XTSC-Bench [21]), while the remaining three focus exclusively on XAI methods (TSinterpret [20], tsCaptum [32], and

Table 1: Overview of the frameworks with a reference to the papers proposing frameworks or to the software release (URLs are given in the reference entries). last update on GitHub; : number of stars on GitHub; : number of forks on GitHub; : availability local (L) or as installable package (P); </>: supported machine-learning backend (TensorFlow (T), PyTorch (P), and Scikit-learn (S)); : Usability scores in order of active maintenance, interaction with community, and documentation (3 scores denoted as X–X–X, each in the range 0–5 from lowest to best); #: number of supported XAI methods (implemented methods and, in parenthesis, wrapped or reused implementations of other frameworks); : number of methods specific to univariate time series; : number of methods specific to multivariate time series; : number of methods generating explanations in the frequency domain; : number of supported XAI metrics (implemented metrics and, in parenthesis, wrapped or reused implementations of other frameworks); : number of metrics specific to time series.

							XAI Methods				Metrics	
XAI Method Frameworks							#					
TSInterpret [20]	11/2025	144	18	P	T, P, S	5-4-5	6 + (10)	2	4	0	0	0
tsCaptum [32]	11/2024	10	1	P	P	3-2-4	(5)	0	0	0	0	0
SIGN-XAI-2 [18]	12/2025	6	0	P	P	5-5-5	2 + (9)	1	0	1	0	0
XAI Evaluation Frameworks												
Quantus [19]	07/2025	631	83	P	T, P	5-4-5	(26)	0	0	0	36	0
time_interpret [12]	09/2025	71	10	P	P	5-4-5	18	0	9	0	17	2
XTSC-Bench [21]	10/2023	4	0	L	T, P, S	2-2-5	(16)	(2)	(4)	0	(11)	0

SIGN-XAI-2 [18]). Table 1 summarizes the frameworks and their main characteristics. All analyzed frameworks support PyTorch. TSInterpret has the broadest backend support, with native compatibility with PyTorch, scikit-learn, and TensorFlow; XTSC-Bench inherits these backends as it builds on TSInterpret. Usability scores (active maintenance, interaction with the community, and documentation) are consistently high, with mean values of (4.3, 3.6, 4.6) for XAI method frameworks and (4, 3.3, 5) for XAI evaluation frameworks. Quantus, which has the highest numbers of GitHub stars and forks, supports the largest set of XAI methods (26) by reusing or wrapping implementations from Captum [24], Zennit [1], and tf-explain [27]. However, none of these XAI methods are specific for time series (see Fig. 3). Quantus also provides the largest collection of evaluation metrics (36). The largest set of time-series-specific XAI methods (9, all of which support multivariate time series) and metrics (2) is provided by time_interpret. SIGN-XAI-2 extends Zennit, while tsCaptum wraps Captum and applies chunking to reduce computational complexity, despite not providing time-series-specific methods.

3.2 XAI Methods

The frameworks contain 51 XAI methods, including 16 perturbation-based, 25 gradient-based, 4 counterfactual, and 6 other methods (see Table A.1 in the

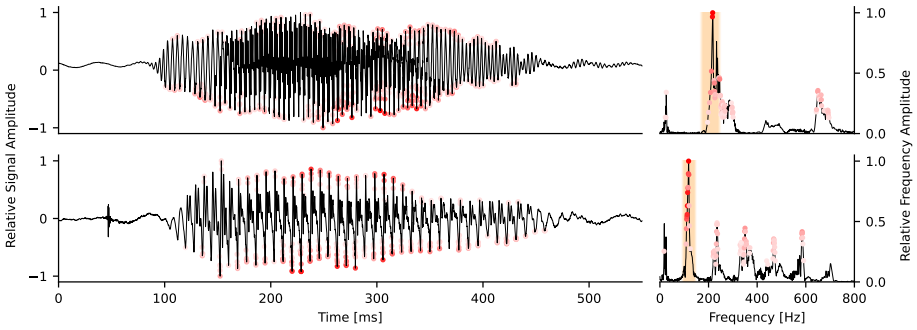

Fig. 4: Spoken digit nine from the AudioMNIST data by a female (upper) and a male (lower) speaker. DFT-LRP explanations (SIGN-XAI-2 framework) are shown in the time (left) and frequency domain (right). Red dots indicate features (time steps or frequencies) with positive relevance scores; higher intensity corresponds to higher relevance. The expected ground truth in the frequency domain is highlighted in orange.

Appendix). Only 16 of these methods explicitly state that they were developed for time series, although the remaining methods may still be applicable. The majority of these 16 methods target multivariate time series, while only three are restricted to univariate time series (DFT-LRP [37], LEFTIST [16], and NUN-CF [10]). The multivariate methods are in principle also applicable to univariate signals (with two exceptions), whereas the three univariate-specific methods are not designed for the multivariate case. Two XAI methods can be used to adapt existing XAI methods to time series. Temporal Saliency Rescaling calculates two relevance scores [22]: one for all time points and one for all features by masking them individually; the final explanation is the product of the two. Time Forward Tunnel (TFT) calculates the relevance for each time point using only the time points preceding it, disallowing the usage of future time points [12]. This can be used to force XAI methods to respect the temporal aspect of time series data. These methods can be options to apply XAI methods of other domains in TSC tasks while considering temporal structure.

For a time series as input, only a single method can generate explanations in the frequency and time-frequency domains. DFT-LRP combines Fourier transforms with Layer-wise Relevance Propagation (LRP) [2] to generate explanations for the time, frequency, and time-frequency domains and is developed for univariate data [37]. The applied Fourier transforms can also be combined with other methods than LRP. In many TSC tasks, class-relevant patterns are present in multiple domains (time, frequency, and time-frequency). Fig. 4 illustrates an example of such a TSC task: the digit nine spoken by a female (Fig. 4 upper part) and by a male speaker (Fig. 4 lower part) from the AudioMNIST data set [6]. The classification task is to distinguish between female and male speakers. We trained the model from [37] using samples of the digit nine, with a balanced set of female and male speakers (same preprocessing as in [37], 1,200 records,

66.7/16.7/16.7% train/validation/test split, 100% test accuracy). We classified a single instance from a male and a female recording and generated explanations for the resulting predictions using DFT-LRP within the SIGN-XAI-2 framework. In the time domain (Fig. 4 left side), the explanations are scattered across the signal and concentrated on positive and negative peaks. In contrast, the frequency domain (Fig. 4 right side) clearly reveals the characteristic difference between a female and a male pronunciation of the digit nine. For adult speakers, the fundamental frequency typically ranges from 165 to 255 Hz for female voices, whereas male voices usually exhibit fundamental frequencies between 90 to 155 Hz [5]. The frequency regions corresponding to these ranges are highlighted in Fig. 4 and serve as ground-truth information. For both samples, the highest relevance scores are concentrated within their ground-truth regions.

3.3 Evaluation Metrics

The frameworks support a total of 52 evaluation metrics: 27 perturbation-based, 15 ground-truth-based, and 10 other (see Table A.2 in the Appendix). There is a small overlap in supported metrics, as only one metric is supported by all three XAI evaluation frameworks. XTSC-Bench builds entirely on Quantus for the implementation of its metrics, therefore, the metrics of XTSC-Bench are a subset of the metrics available in Quantus. Out of the 52 evaluation metrics, 44 metrics are compatible with time series, six are not compatible (restricted to images) and only two are specifically developed for time series. The two metrics specific to time series are Mask Information and Mask Entropy [7], they work on subsequences of time series measuring how well a predicted relevance fits to a ground-truth mask. None of the identified metrics are specifically developed for evaluating explanations within the frequency or time-frequency domain.

3.4 Benchmarking Capabilities

Out of the six analyzed frameworks, three support datasets (XTSC-Bench, TSInterpret, and time_interpret) for the comparative evaluation of explanations. The datasets available in each framework are listed in Table 2. XTSC-Bench offers six synthetic datasets proposed by Ismail et al. [22], which contain ground-truth information and are variable in length and number of channels. Additionally, it supports the University of California Riverside (UCR) and University of East Anglia (UEA) time series archives [9,4], which are well established in AI research for TSC and comprise 128 and 30 datasets, respectively. Five datasets are supported by time_interpret, including three synthetic datasets and two real-world clinical datasets (of which one is not openly accessible). The synthetic datasets are variable in length and number of channels and provide ground-truth information. The clinical datasets contain patient laboratory data and biomarkers but lack ground-truth information. The synthetic datasets of both XTSC-Bench and time_interpret enable benchmarking of XAI methods for TSC due to their ground-truth information. TSInterpret also supports the UCR and UEA time series archives.

Table 2: Datasets of frameworks. **Frameworks (F)**: Frameworks supporting the dataset; **Channels and Length (L)**: Univariate (📈) or multivariate (🔀), synthetic datasets have variable (v) lengths and channel number; **Synthetic (S)**: ✓ indicates dataset is synthetic, mixed (m), and ✗ otherwise; **Ground Truth (GT)**: ✓ indicates dataset contains GT information and ✗ otherwise; *no open access; †Archives consist of multiple datasets; here we only indicate their variability at the collection level.

Dataset	Frameworks	Channels 📈	Channels 🔀	L	S	GT
Arma [7]	time_interpret	✓	✓	v	✓	✓
BioBank [33]	time_interpret	*	*	*	*	*
Hawkes [3]	time_interpret	✓	✓	v	✓	✓
Hidden Markow Model [7]	time_interpret	✓	✓	v	✓	✓
Mimic III [23]	time_interpret	✗	31	48	✗	✗
Gaussian, Harmonic, Pseudo Periodic, Autoregressive (AR), Continuous AR, Narma [22]	time_interpret	✓	✓	v	✓	✓
UCR Archive† [9]	TSInterpret, XTSC-Bench	✓	✗	v	m	✗
UEA Archive† [4]	TSInterpret, XTSC-Bench	✗	✓	v	m	✗

3.5 Reproducibility

We compared explanations and evaluation scores produced by XAI methods and evaluation metrics with different implementations across frameworks. Five XAI methods (Saliency, Occlusion, Feature Ablation, Deconvolution, and Integrated Gradients, others were excluded due to different parameter usage or different backend support) were compared on a convolutional neural network (CNN) with the architecture described in Gumpfer et al. [17], trained for 20 epochs on right bundle branch block (RBBB) records and an equal amount of healthy ECG records of the PTB-XL dataset [38] (full-length 12-lead ECGs, no preprocessing, 3,316 records, 80/10/10% train/validation/test split, 96.98% test accuracy). We explained the model predictions on the test set using Integrated Gradients (IG) [34] as it is implemented in TSInterpret via Captum and in SIGN-XAI-2 via Zennit. IG attributes relevance by integrating gradients along a straight-line path between a baseline and the input. To ensure comparability, we used identical parameters in both frameworks (zero baseline and 50 interpolation steps). One example of resulting explanations is shown in Fig. 5, illustrating one channel (lead V1) of the ECG example from Fig. 1. The relevance distributions differ noticeably and only the Zennit-based implementation (Fig. 5, right) explains the expected "M"-shaped pattern characteristic of RBBB [35]. To verify the results, we analyzed the whole test set of 443 instances: Captum- and Zennit-based IG yielded a mean Spearman rank correlation of -0.14 (median -0.17, std 0.11; 0 of 443 instances had $\rho > 0.5$) and a top-5% Jaccard overlap of 0.26 (± 0.09). These results confirm that the divergence in Fig. 5 reflects systematic disagreement rather than an instance-specific or thresholding

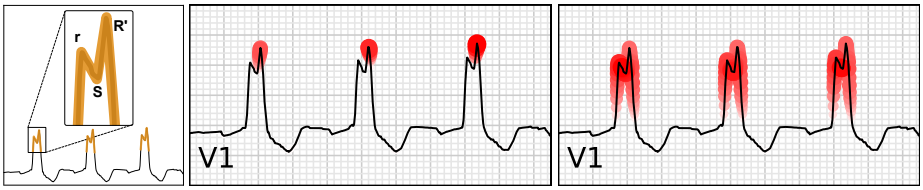

Fig. 5: Electrocardiogram (ECG) example with right bundle branch block (RBBB). Left: Illustration of characteristic "M"-shaped pattern (rSR'). Middle and right: Explanations generated using Integrated Gradients as implemented in TSInterpret (middle, based on Captum) and SIGN-XAI-2 (right, based on Zennit) for the same ECG instance, illustrating differences in relevance distributions across implementations. Relevance scores are indicated by red dots; higher intensity indicates higher relevance.

artifact. The other investigated methods provided identical explanations across frameworks, when configured with matching parameters. We also compared evaluation metrics across frameworks, but only Area under the Receiver Operating Characteristic Curve (ROC-AUC) [14] is implemented by two (Quantus and time_interpret). Unlike the explanations, the metric produced consistent values across both implementations.

4 Open Challenges & Further Research

Compared to 2022, when only one framework supported time series data [26], we now identified five additional frameworks that claim to support time series, indicating increased research activity. This development suggests that earlier calls for time-series-aware XAI frameworks [36] have been acknowledged by the community. However, our analysis shows that explicit support for time-series-specific properties remains limited.

Method Applicability. Although many supported XAI methods are technically applicable to TSC, only a small fraction has been explicitly developed for time series. Two frameworks, Quantus and tsCaptum, do not include any XAI methods that are specifically designed for time series, but instead rely exclusively on reusing or wrapping methods developed for other data types. The remaining frameworks also partially rely on generic explanation methods. However, the suitability of the underlying methods for time series is often not explicitly validated. Further research should investigate to which extent these methods account for time-series-specific properties like temporal and cross-channel dependencies.

Cross-Channel Dependencies. While most supported XAI methods can highlight relevant regions in multiple channels, the dependencies of these relevant parts are not represented and would require an additional dimension of explanation. This issue also manifests in the common use of heatmap-based relevance visualizations, which are unable to highlight class-relevant synchronous

and asynchronous cross-channel dependencies. PAX-TS [25], a perturbation-based method, addresses this for forecasting tasks, but at higher computational cost. More generally, gradient-based approaches are efficient for high-dimensional inputs yet cannot explain cross-channel dependencies, whereas perturbation-based methods can, at increased cost, suggesting potential benefits of hybrid approaches for future research.

Frequency and Time-Frequency Domains. Another major limitation is the insufficient support for frequency and time-frequency domain explanations. In many TSC tasks, including audio, biomedical, and sensor-based applications, time-domain explanations alone are often insufficient, as class-relevant patterns are present in multiple domains (time, frequency, and time-frequency). DFT-LRP is currently the only XAI method of the analyzed frameworks that generates explanations in multiple domains for models using time-domain inputs. However, Fourier transformations, used by DFT-LRP, can also be combined with other XAI methods to extend their explanations to the frequency and time-frequency domains. Currently, among the analyzed frameworks only SIGN-XAI-2 implements a frequency-aware method, which remain rare in the literature as well.

Evaluation Metrics. Similar limitations exist for evaluation metrics. Although many metrics implemented in current frameworks are technically applicable to time series, it is often unclear to which extent they account for time-series-specific properties. An assumption that some metrics make is the independence of feature relevance, which is violated by temporal and cross-channel dependencies in time series. Only two of the 52 analyzed metrics were developed specifically for time series, in contrast to the literature, which offers time-series-specific metrics such as the Attribution Stability Indicator [31], Swap Time Points, and Mean Time Points [30]. Representations of time series in the frequency domain are not considered by any of the evaluation metrics. The application of existing perturbation-based metrics to evaluate explanations in the frequency domain is problematic, because small perturbations can impact the entire signal in the time domain. Future research should investigate to which extent supported metrics account for time-series-specific properties like temporal and cross-channel dependencies as well as frequency and time-frequency evaluations. We encourage researchers to develop dedicated metrics for these properties.

Benchmarking and Ground-Truth Availability. Only half of the frameworks support benchmarking datasets and only two include ground-truth information. All datasets with ground truth are based on synthetic or predefined data distributions. It remains unclear to what extent they represent real time series. Extending existing real-world datasets with ground-truth annotations would allow for more reliable benchmarking and transfer to real applications.

Reproducibility and Consistency. Our reproducibility analysis shows that explanation results may depend on the choice of the XAI framework. We ob-

served substantial differences across frameworks for explanations generated by IG. For one framework, the class-relevant pattern was not visible in the resulting explanations. Prior work by Le et al. [26] has shown that evaluation metrics can exhibit substantial variability across implementations. In contrast, we did not observe discrepancies for evaluation metrics in our experiments. However, this observation is likely limited due to the small overlap of evaluation metrics across frameworks, which restricts the scope of reproducibility analysis for metrics. We emphasize that researchers need to take care when comparing or reproducing published results (without knowledge of the used framework), this highlights current limitations of XAI research in generating reliable explanations.

Transferability. Although this survey focuses on TSC, many of the identified limitations may also affect other time series tasks, such as forecasting and anomaly detection, which are even less studied within current XAI research for time series despite their practical relevance.

Limitations. The systematic search underlying this review was conducted entirely on GitHub, which allowed us to specifically filter for software contributions. We acknowledge that this procedure might miss research software published on other platforms. Furthermore, our cross-framework reproducibility analysis is limited in scope: only five explanation methods and a single evaluation metric were implemented in more than one framework. Our results therefore demonstrate that explanations can depend on the framework implementation, but do not establish how prevalent this is across methods and metrics.

5 Conclusion

We presented the first systematic survey of XAI frameworks for TSC and identified substantial limitations in the current landscape. These include limited support for time-series-specific XAI methods and evaluation metrics, minimal incorporation of alternative signal representations such as frequency and time-frequency domains for XAI methods, a lack of ground-truth annotations for real-world benchmark datasets, and a lack of methods for explaining cross-channel dependencies in multivariate time series. We argue that researchers should exercise caution when applying methods or evaluation metrics for TSC from existing frameworks, as different implementations can yield substantially different results. We provide Tables 1 and 2 as references to select frameworks and datasets for TSC tasks (Tables A.1 and A.2 in the Appendix can serve as additional resources for selecting time-series specific methods and metrics).

Publicly funded, community-driven efforts in other disciplines, such as the platform Galaxy [15] in bioinformatics, highlight the potential of shared infrastructure for reproducible and collaborative research. A standardized, community-driven initiative would greatly benefit XAI research by enabling faithful, reproducible, and time-series-aware explanations as well as robust, systematic comparison and evaluation of XAI methods.

Acknowledgments. This work was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) through ExperTeam4KI (grant no. 16IS24063). We gratefully acknowledge support from the hessian.AI Service Center (funded by the BMFTR, grant no. 16IS22091) and the hessian.AI Innovation Lab (funded by the Hessian Ministry for Digital Strategy and Innovation, grant no. S-DIW04/0013/003). This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project ID 536124560.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Anders, C.J., Neumann, D., Samek, W., Müller, K.R., Lapuschkin, S.: Software for dataset-wide xai: From local explanations to global insights with zennit, corelay, and virelay. PLOS One **21**(1), 1–38 (2026). https://doi.org/10.1371/journal.pone.0336683, https://github.com/chr5tphr/zennit
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS One **10**(7), 1–46 (2015). https://doi.org/10.1371/journal.pone.0130140
3. Bacry, E., Bompaire, M., Deegan, P., Gaïffas, S., Poulsen, S.V.: tick: a python library for statistical learning, with an emphasis on hawkes processes and time-dependent models. Journal of Machine Learning Research **18**(214), 1–5 (2018)
4. Bagnall, A., Dau, H.A., Lines, J., Flynn, M., Large, J., Bostrom, A., Southam, P., Keogh, E.: The UEA multivariate time series classification archive, 2018. Computing Research Repository **abs/1811.00075** (2018). https://doi.org/10.48550/arXiv.1811.00075
5. Baken, R.J., Orlikoff, R.F.: Clinical measurement of speech and voice. Singular Thomson Learning, 2. edn. (2000)
6. Becker, S., Vielhaben, J., Ackermann, M., Müller, K.R., Lapuschkin, S., Samek, W.: AudioMNIST: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. Journal of the Franklin Institute **361**(1), 418–428 (2024). https://doi.org/10.1016/j.jfranklin.2023.11.038
7. Crabbé, J., Van Der Schaar, M.: Explaining time series predictions with dynamic masks. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning (ICML). vol. 139, pp. 2166–2177. Proceedings of Machine Learning Research, online (2021)
8. Dasgupta, S., Frost, N., Moshkovitz, M.: Framework for evaluating faithfulness of local explanations. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning (ICML). vol. 162, pp. 4794–4815. Proceedings of Machine Learning Research, Baltimore, Maryland, USA (2022)
9. Dau, H.A., Bagnall, A., Kamgar, K., Yeh, C.C.M., Zhu, Y., Gharghabi, S., Ratanamahatana, C.A., Keogh, E.: The UCR time series archive. IEEE/CAA Journal of Automatica Sinica **6**(6), 1293–1305 (2019). https://doi.org/10.1109/JAS.2019.1911747

10. Delaney, E., Greene, D., Keane, M.T.: Instance-based counterfactual explanations for time series classification. In: Sánchez-Ruiz, A.A., Floyd, M.W. (eds.) Case-Based Reasoning Research and Development. pp. 32–47. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-86957-1_3
11. DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., Wallace, B.C.: ERASER: A benchmark to evaluate rationalized NLP models. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 4443–4458. Association for Computational Linguistics, online (2020). https://doi.org/10.18653/V1/2020.ACL-MAIN.408
12. Enguehard, J.: Time Interpret: A unified model interpretability library for time series. Computing Research Repository **abs/2306.02968** (2023). https://doi.org/10.48550/ARXIV.2306.02968, https://github.com/josephenguehard/time_interpret
13. European Union: Regulation (EU) 2024/1689 of the European Parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) (2024)
14. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters **27**(8), 861–874 (2006). https://doi.org/10.1016/j.patrec.2005.10.010
15. Galaxy Community, T.: The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. Nucleic Acids Research **52**(W1), W83–W94 (2024). https://doi.org/10.1093/nar/gkae410
16. Guillemé, M., Masson, V., Rozé, L., Termier, A.: Agnostic local explanation for time series classification. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). pp. 432–439. Institute of Electrical and Electronics Engineers, Portland, Oregon, USA (2019). https://doi.org/10.1109/ICTAI.2019.00067
17. Gumpfer, N., Dinov, B., Sossalla, S., Guckert, M., Hannig, J.: Towards trustworthy ai in cardiology: A comparative analysis of explainable ai methods for electrocardiogram interpretation. In: Finkelstein, J., Moskovitch, R., Parimbelli, E. (eds.) Artificial Intelligence in Medicine. pp. 350–361. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-66535-6_36
18. Gumpfer, N., Prim, J., Keller, T., Seeger, B., Guckert, M., Hannig, J.: SIGNed explanations: Unveiling relevant features by reducing bias. Information Fusion **99**, 101883 (2023). https://doi.org/10.1016/j.inffus.2023.101883, https://github.com/TimeXAIgroup/signxai2
19. Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., Höhne, M.M.: Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. Journal of Machine Learning Research **24**, 34:1–34:11 (2023), https://github.com/understandable-machine-intelligence-lab/Quantus
20. Höllig, J., Kulbach, C., Thoma, S.: TSInterpret: A unified framework for time series interpretability (2022). https://doi.org/10.48550/arXiv.2208.05280, https://github.com/fzi-forschungszentrum-informatik/TSInterpret
21. Höllig, J., Thoma, S., Grimm, F.: XTSC-Bench: Quantitative benchmarking for explainers on time series classification. In: 2023 International Conference on Machine Learning and Applications (ICMLA). pp. 1126–1131. Institute of Electrical and Electronics Engineers, Jacksonville, Florida, USA (2023). https://doi.org/10.1109/ICMLA58977.2023.00168, https://github.com/JHoelli/XTSC-Bench

22. Ismail, A.A., Gunady, M., Corrada Bravo, H., Feizi, S.: Benchmarking deep learning interpretability in time series predictions. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 6441–6452. Curran Associates, Inc., online (2020)
23. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific Data **3**(1), 160035 (2016). https://doi.org/10.1038/sdata.2016.35
24. Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., Reblitz-Richardson, O.: Captum: A unified and generic model interpretability library for PyTorch. Computing Research Repository **abs/2009.07896** (2020). https://doi.org/10.48550/arXiv.2009.07896, https://github.com/meta-pytorch/captum
25. Kreuzer, T., Zdravkovic, J., Papapetrou, P.: PAX-TS: model-agnostic multigranular explanations for time series forecasting via localized perturbations. Computing Research Repository **abs/2508.18982** (2025). https://doi.org/10.48550/ARXIV.2508.18982
26. Le, P.Q., Nauta, M., Nguyen, V.B., Pathak, S., Schlötterer, J., Seifert, C.: Benchmarking eXplainable AI - A survey on available toolkits and open challenges. In: Elkind, E. (ed.) Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI). pp. 6665–6673. International Joint Conferences on Artificial Intelligence Organization, Macau (2023). https://doi.org/10.24963/ijcai.2023/747
27. Meudec, R.: tf-explain (2021). https://doi.org/10.5281/zenodo.5711704, Software. https://github.com/sicara/tf-explain
28. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. ACM Computing Surveys **55**(13s), 295:1–295:42 (2023). https://doi.org/10.1145/3583558
29. Rojat, T., Puget, R., Filliat, D., Ser, J.D., Gelin, R., Rodríguez, N.D.: Explainable artificial intelligence (XAI) on timeseries data: A survey. Computing Research Repository **abs/2104.00950** (2021). https://doi.org/10.48550/arXiv.2104.00950
30. Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., Keim, D.A.: Towards a rigorous evaluation of xai methods on time series. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 4197–4201. Institute of Electrical and Electronics Engineers, Seoul, South Korea (2019). https://doi.org/10.1109/ICCVW.2019.00516
31. Schlegel, U., Keim, D.A.: Introducing the attribution stability indicator: a measure for time series XAI attributions. Computing Research Repository **abs/2310.04178** (2023). https://doi.org/10.48550/ARXIV.2310.04178
32. Serramazza, D.I., Nguyen, T.L., Ifrim, G.: A short tutorial for multivariate time series explanation using tsCaptum. Software Impacts **22**, 100723 (2024). https://doi.org/10.1016/j.simpa.2024.100723, https://github.com/mlgig/tscaptum
33. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R.: UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLOS Medicine **12**(3), 1–10 (2015). https://doi.org/10.1371/journal.pmed.1001779

34. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML). vol. 70, p. 3319–3328. Proceedings of Machine Learning Research, Sydney, NSW, Australia (2017)
35. Surawicz, B., Childers, R., Deal, B.J., Gettes, L.S.: AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram. Circulation **119**(10), e235–e240 (2009). https://doi.org/10.1161/CIRCULATIONAHA.108.191095
36. Theissler, A., Spinnato, F., Schlegel, U., Guidotti, R.: Explainable AI for time series classification: A review, taxonomy and research directions. IEEE Access **10**, 100700–100724 (2022). https://doi.org/10.1109/ACCESS.2022.3207765
37. Vielhaben, J., Lapuschkin, S., Montavon, G., Samek, W.: Explainable AI for time series via Virtual Inspection Layers. Pattern Recognition **150**, 110309 (2024). https://doi.org/10.1016/j.patcog.2024.110309
38. Wagner, P., Strodthoff, N., Bousseljot, R.D., Kreiseler, D., Lunze, F.I., Samek, W., Schaeffter, T.: PTB-XL, a large publicly available electrocardiography dataset. Scientific Data **7**(1), 154 (2020). https://doi.org/10.1038/s41597-020-0495-6

Appendix A

Table A.1: XAI methods of frameworks. **Compatibility**: ✓ indicates method is specific for time series, otherwise marked as – (not explicitly stated); **Dimensionality**: supports uni- (U), multivariate (M) time series or both (B); **Frequency**: ✓ generates explanations for the frequency and time-frequency domain; **Frameworks**: ✓ indicates that the framework implements the method, (✓) if the framework wrapped or reused implementations of other frameworks, and ✗ otherwise. Abbr.: Counterfactuals (CF)

						Frameworks				
	Methods	**Compatibility**	**Dimensionality**	**Frequency**	**Quantus**	**TSInterpret**	**time_interpret**	**tsCaptum**	**XTSC-Bench**	**SIGN-XAI-2**
Gradient-based	BetaSmooth	-	-	✗	✗	✗	✗	✗	✗	(✓)
	Conductance	-	-	✗	(✓)	✗	✗	✗	✗	✗
	Deconvolution	-	-	✗	(✓)	✗	✗	✗	✗	(✓)
	Deep Learning Important Features (DeepLIFT)	-	-	✗	(✓)	(✓)	✗	✗	(✓)	✗
	DeepLIFT Shapley Additive Explanations (SHAP)	-	-	✗	(✓)	(✓)	✗	✗	(✓)	✗
	DFT-LRP	✓	U	✓	✗	✗	✗	✗	✗	✓
	Discretized Integrated Gradients	-	-	✗	✗	✗	✓	✗	✗	✗
	Excitation Backpropagation	-	-	✗	✗	✗	✗	✗	✗	(✓)
	GeodesicIntegratedGradients	-	-	✗	✗	✗	✓	✗	✗	✗
	Gradient-weighted Class Activation Mapping (Grad-CAM)	-	-	✗	(✓)	(✓)	✗	✗	(✓)	✗
	Gradient × Input	-	-	✗	(✓)	✗	✗	✗	✗	✗
	GradientSHAP	-	-	✗	(✓)	(✓)	✗	✗	(✓)	✗
	Guided Backpropagation	-	-	✗	(✓)	✗	✗	✗	✗	(✓)
	Guided Grad-CAM	-	-	✗	(✓)	✗	✗	✗	✗	✗
	Integrated Gradients	-	-	✗	(✓)	(✓)	✗	✗	(✓)	(✓)
	InternalInfluence	-	-	✗	(✓)	✗	✗	✗	✗	✗
	Layer Activation	-	-	✗	(✓)	✗	✗	✗	✗	✗
	Layer-wise Relevance Propagation (LRP)	-	-	✗	(✓)	✗	✗	✗	✗	(✓)
	Saliency (Vanilla Gradients)	-	-	✗	(✓)	(✓)	✗	✗	(✓)	(✓)
	Sequential Integrated Gradients	-	-	✗	✗	✗	✓	✗	✗	✗
	SIGN	-	-	✗	✗	✗	✗	✗	✗	✓
	SmoothGrad	-	-	✗	(✓)	(✓)	✗	✗	(✓)	(✓)
	Testing with Concept Activation Vectors (TCAV)	-	-	✗	(✓)	✗	✗	✗	✗	✗

Continued on next page

Table A.1 – *continued from previous page*

	Methods	Compatibility	Dimensionality	Frequency	Frameworks: Quantus	TSInterpret	time_interpret	tsCaptum	XTSC-Bench	SIGN-XAI-2
	Temporal Integrated Gradients	✓	B	✗	✗	✗	✓	✗	✗	✗
	Tracing Gradient Descent with Checkpoints (TracInCP)	-	-	✗	(✓)	✗	✗	✗	✗	✗
Perturbation-based	Augmented Occlusion	✓	B	✗	✗	✗	✓	✗	✗	✗
	BayesKernelSHAP	-	-	✗	✗	✗	✓	✗	✗	✗
	BayesLIME	-	-	✗	✗	✗	✓	✗	✗	✗
	DynaMask	✓	B	✗	✗	✗	✓	✗	✗	✗
	Extremal Mask	✓	B	✗	✗	✗	✓	✗	✗	✗
	Feature Ablation	-	-	✗	(✓)	(✓)	✓	(✓)	(✓)	✗
	Feature Permutation	-	-	✗	(✓)	✗	✗	(✓)	✗	✗
	KernelSHAP	-	-	✗	(✓)	✗	✗	(✓)	✗	✗
	LEFTIST	✓	U	✗	✗	✓	✗	✗	(✓)	✗
	Local Interpretable Model-agnostic Explanations (LIME)	-	-	✗	(✓)	✗	✗	(✓)	✗	✗
	Local Outlier Factor (LOF)-LIME	-	-	✗	✗	✗	✓	✗	✗	✗
	LOF-KernelSHAP	-	-	✗	✗	✗	✓	✗	✗	✗
	Occlusion	-	-	✗	(✓)	(✓)	✓	✗	(✓)	(✓)
	Shapley Value Sampling	-	-	✗	(✓)	(✓)	✗	(✓)	(✓)	✗
	Temporal Augmented Occlusion	✓	B	✗	✗	✗	✓	✗	✗	✗
	Temporal Occlusion	✓	B	✗	✗	✗	✓	✗	✗	✗
CF	COMTE	✓	M	✗	✗	✓	✗	✗	(✓)	✗
	NUN-CF	✓	U	✗	✗	✓	✗	✗	(✓)	✗
	Shapelet explainer for time series (SETS)	✓	B	✗	✗	✓	✗	✗	(✓)	✗
	TSEvo	✓	B	✗	✗	✓	✗	✗	(✓)	✗
Other	Activations Visualization	-	-	✗	✗	✗	✗	✗	✗	✗
	Feature Importance in Time (FIT)	✓	B	✗	✗	✗	✓	✗	✗	✗
	Layer/Neuron Activations	-	-	✗	✗	✗	✗	✗	✗	✗
	Reverse Time Attention model (RETAIN)	✓	B	✗	✗	✗	✓	✗	✗	✗
	Time Forward Tunnel	✓	B	✗	✗	✗	✓	✗	✗	✗
	Temporal Saliency Rescaling	✓	M	✗	✗	✓	✗	✗	(✓)	✗

Table A.2: Evaluation metrics of frameworks. Legend: **Compatibility**: ✓ indicates metric is specific for time series, C if the metric is compatible with time series, and ✗ otherwise; **Quantus, time_interpret, XTSC-Bench**: ✓ indicates that the framework implements the metrics, (✓) if the framework wrapped or reused implementations of other frameworks, and ✗ otherwise.

	Metrics	Compatibility	Quantus	time_interpret	XTSC-Bench
Perturbation-based	Accuracy	C	✗	✓	✗
	Avg-Sensitivity	C	✓	✗	(✓)
	Comprehensiveness	C	✗	✓	✗
	Continuity	✗	✓	✗	✗
	Consistency	C	✓	✗	✗
	Cross_entropy	C	✗	✓	✗
	Faithfulness Correlation	C	✓	✗	(✓)
	Faithfulness Estimate	C	✓	✗	(✓)
	Infidelity	✗	✓	✗	✗
	Iterative Removal of Features (IROF)	✗	✓	✗	✗
	Lipschitz_max	C	✗	✓	✗
	Local Lipschitz Estimate	C	✓	✗	✗
	Log_odds	C	✗	✓	✗
	Max-Sensitivity	C	✓	✗	(✓)
	Mean absolute error (MAE)	C	✗	✓	✗
	Mean squared error (MSE)	C	✗	✓	✗
	Monotonicity (Arya)	C	✓	✗	(✓)
	Monotonicity (Nguyen)	C	✓	✗	(✓)
	Pixel Flipping	C	✓	✗	✗
	Region Perturbation	C	✓	✗	✗
	Relative Input Stability	C	✓	✗	✗
	Relative Output Stability	C	✓	✗	✗
	Relative Representation Stability	C	✓	✗	✗
	Remove and Debias (ROAD)	✗	✓	✗	✗
	Selectivity	✗	✓	✗	✗
	SensitivityN	C	✓	✗	✗
	Sufficiency[1] by DeYoung et al. [11]	C	✗	✓	✗

Continued on next page

Table A.2 – *continued from previous page*

	Metrics	Compatibility	Quantus	time_interpret	XTSC-Bench
Ground-Truth-based	Attribution Localisation	C	✓	✗	✗
	Area under precision curve (AUP)	C	✗	✓	✗
	Area under precision-recall curve (AUPRC)	C	✗	✓	✗
	Area under recall curve (AUR)	C	✗	✓	✗
	Area under the Receiver Operating Characteristic Curve (ROC-AUC)	C	✓	✓	(✓)
	Focus	✗	✓	✗	✗
	MAE	C	✗	✓	✗
	Mask entropy	✓	✗	✓	✗
	Mask information	✓	✗	✓	✗
	MSE	C	✗	✓	✗
	Pointing Game	C	✓	✗	(✓)
	Relevance Mass Accuracy	C	✓	✗	(✓)
	Relevance Rank Accuracy	C	✓	✗	(✓)
	Root mean squared error (RMSE)	C	✗	✓	✗
	Top-K Intersection	C	✓	✗	✗
Other	Complexity	C	✓	✗	(✓)
	Effective Complexity	C	✓	✗	✗
	Efficient MPRT	C	✓	✗	✗
	Input Invariance	C	✓	✗	✗
	Model Parameter Randomization Test (MPRT)	C	✓	✗	✗
	Non-Sensitivity	C	✓	✗	✗
	Random Logit Test	C	✓	✗	✗
	Smooth MPRT	C	✓	✗	✗
	Sparseness	C	✓	✗	✗
	Sufficiency[1] by Dasgupta et al. [8]	C	✓	✗	✗

[1] Two metrics share the name *Sufficiency* and are differentiated by the reference to their original publications.