Evaluating Explanation Methods of Multivariate Time Series Classification through Causal Lenses

Etienne Vareille, Adel Abbas, Michele Linardi, Vassilis Christophides

Evaluating Explanation Methods of Multivariate Time Series Classification through Causal Lenses

Etienne Vareille*, Adel Abbas[†], Michele Linardi[‡], Vassilis Christophides[§]
ETIS UMR-8051 Laboratory, CY Cergy Paris Université, ENSEA
Email: *etienne.vareille@ensea.fr, [†]adel.abbas@ensea.fr, [‡]michele.linardi@cyu.fr, [§]Vassilis.Christophides@ensea.fr

***Abstract*—Explainable machine learning techniques (XAI) aim to provide a solid descriptive approach to Deep Neural Networks (NN). In Multi-Variate Time Series (MTS) analysis, the most recurrent techniques use relevance attribution, where importance scores are assigned to each TS variable over time according to their importance in classification or forecasting. Despite their popularity, post-hoc explanation methods do not account for causal relationships between the model outcome and its predictors. In our work, we conduct a thorough empirical evaluation of model-agnostic and model-specific relevance attribution methods proposed for TCNN, LSTM, and Transformers classification models of MTS. The contribution of our empirical study is three-fold: (i) evaluate the capability of existing post-hoc methods to provide consistent explanations for high-dimensional MTS (ii) quantify how post-hoc explanations are related to sufficient explanations (i.e., the direct causes of the target TS variable) underlying the datasets, and (iii) rank the performance of surrogate models built over post-hoc and causal explanations w.r.t. the full MTS models. To the best of our knowledge, this is the first work that evaluates the reliability and effectiveness of existing xAI methods from a temporal causal model perspective.**

***Index Terms*—Multivariate Time Series, Explainability, Classification, Causal Explanations, Consistency, Benchmarking**

I. Introduction

Multivariate Time Series (MTS) are omnipresent in many science and engineering domains, including health-care, sustainable energy, geoscience, and high-performance computing. An MTS is composed of more than one time-dependent variables that may depend not only on its past values but also on other variables. Neural Network (NN) architectures (e.g., TCNN [1], LSTM [2] Transformer [3]) are today state-of-the-art solutions in order to implement MTS Classification [4]–[6].

Deep NN models are essentially uninterpretable black boxes, where one feeds an input and obtains an output without understanding the motivations behind that decision. We have recently witnessed a consistent effort to enhance NN models with *explanation* capabilities at various levels allowing users to track the time-dependent variables or the training timestamped samples that drive a NN model toward a certain decision. Specifically, eXplainable Artificial Intelligence (xAI) techniques such as dCAM [1], XCM [7], Dynamask [8], TimeSHAP [9] aim to reveal which subsets of MTS variables are mostly involved in deciding a particular class label.

Post-hoc explanation methods like feature attribution [10], [11] or saliency maps [12] can help indicate which features or pixels of input data are most 'relevant' to the output decision of a NN model being 'explained'. However, they are designed with the assumption that a single sample is self-explanatory and analysts can easily distinguish between two samples, which is not easy in complex MTS with thousands of time-dependent variables. Additionally, the intrinsic measures underlying xAI methods are purely associational and may expose potentially spurious and misleading correlations in input data used to train a NN model. Clearly, correlation-based explanations are not informative enough in order to produce *surrogate models* of reduced complexity (i.e., by projecting input data on the relevant features) or to *debug models* (e.g., by identifying data instances and/or model components that cause incorrect predictions).

These kinds of tasks call for more expressive forms of explanations, which can reveal causal-effect relationships between TS variables (input or target) along with adequate lag information. *Causal explanations* are usually distinguished between *sufficient explanations* and *counterfactual explanations* [13]. The former permits users to understand the conditions in which a particular action (e.g., selecting a subset of TS variables as predictors) will produce a desired model outcome. The latter instead identifies actions (e.g., changing the values of TS variables) that can alter an observed input, ensuring a change in a previously observed output.

In our work, we are experimentally evaluating the consistency and effectiveness of model-specific (i.e, dCAM [1], Dynamask [8], XCM [7]), and model-agnostic (i.e, Feature Ablation [14], [15], Feature Permutation [16], Feature Occclusion [16], Integrated Gradients [17], TimeSHAP [9], Gradient Shap [11]) post-hoc explanations of MTS models built using different NN architectures (i.e, TCNN [1], LSTM [2] and Transformer [3]). More precisely, the contribution of our empirical study is three-fold: (i) *evaluate the ability of existing post-hoc methods to provide consistent explanations* for high-dimensional MTS (ii) *quantify how post-hoc explanations are related to sufficient explanations* (i.e., the direct causes of the target TS variable) underlying the datasets, and (iii) *rank the performance of surrogate models built over post-hoc and causal explanations* w.r.t. the full MTS models.

None of the previous empirical studies ([6], [18]) compare both model-agnostic and model-specific post-hoc explanation methods for a variety of MTS classification models (LSTM, CNN, and Transformer) using as a cause-effect ground truth underlying the benchmark datasets as validated by the domain experts. More precisely, [6] focuses only on MTS agnostic xAI methods (not tailored for MTS classification models) evaluated using AUPR (Area under the precision-recall curve) given as ground truth feature relevance constructed independently of the underlying data generation processes (i.e., by adding or subtracting a fixed constant from the data). The Exathlon benchmark [18] compares two anomaly explanation methods (Macrobase, Xstream) with LIME explanations for three NN-based anomaly detectors (LSTM, Auto Encoder, and BiGAN). Exathlon provides ground truth only for the range-based anomalies occurring in Spark execution traces and explanation conciseness and consistency are evaluated independently of the true anomaly causes. As a matter of fact, these metrics report only the number of features used in the explanations and how often anomalies of the same type receive the same explanation. Although the need for experimentally evaluating sufficient explanations of target variables over time has been raised in previous works [8], [13], this is the first empirical study demonstrating that surrogate models built over causal relationships outperform not only models built over relevant features detected by xAI methods but also advanced NN models trained over the whole data feature space.

The rest of the paper is structured as follows: Section II introduces the core notions of post-hoc and sufficient explanations. Section III presents the main aspects of the analyzed xAI methods. Section IV describes our empirical evaluation setting and metrics. Section V details the results of our experimental evaluation and highlights the main insights. Finally, Section VI summarizes our findings and presents plans for future works.

II. Notation and Preliminaries

We denote a multivariate time series (MTS) as $X \in \mathbb{R}^{N \times T}$, where T is the number of time steps and N is the number of dimensions. Each MTS dimension is a univariate TS variable denoted as X^i, where $1 \leq i \leq N$. The length of a TS variable X^i is denoted as $|X^i|$. Since all univariate TS variables of an MTS have the same length, we use the notation $|X|$ to also denote the number of time steps T.

We define a sub-sequence of X^i of length ℓ, starting at time-step t, with $X^i_{t:t+\ell-1} \in \mathbb{R}^{N \times \ell}$, and the value at time t of the i^{th} TS variable as $X^i_t \in \mathbb{R}$. Given X, we define a multivariate sub-sequence (set of sub-sequences of all TS variables) of length ℓ and starting at time-step t as $X^{(1..N)}_{t:t+\ell-1} = \{X^1_{t:t+\ell-1}, .., X^N_{t:t+\ell-1}\}$. Note that a multivariate sub-sequence is a MTS itself.

To consider a scale-free representation of multivariate sub-sequences we apply Z-normalization [19], [20]. Hence, given a MTS X, each TS variable is always in the form $X^i = \{\frac{X^i_1 - \mu}{\sigma}, ..., \frac{X^i_{|X^i|} - \mu}{\sigma}\}$, where μ and σ are the mean and standard of X^i respectively.

We define MTS *classification* as a function $\mathcal{F}^c$ that given a multivariate sequence outputs the probability of the sequence to belong to a given class. Formally, we have: $\mathcal{F}^c(X) = [S_1(X), ..., S_C(X)] \in \mathbb{R}^{\mathbb{C}}$, where C is the total number of classes and $S_i(X)$ represents the probability of the sequence X to belong to the i^{th} class. Hence, $\sum_{i=1}^{C} S_i(X) = 1$.

We consider the explainability of an MTS classification model in terms of *relevance attribution*, which assigns a score to each TS variable at each time step. This score typically reports the feature's importance with respect to a class assignment. Fig. 1 illustrates an example of a Relevance Attribution Matrix (Fig. 1(b)) computed on a MTS X (Fig. 1(a)).

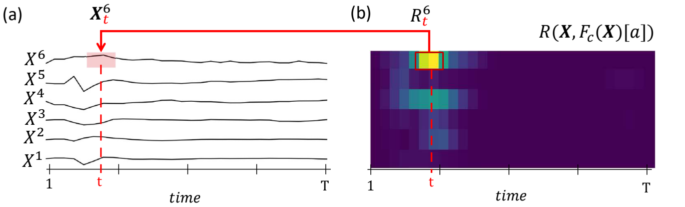


Fig. 1. *(a)* Synthetic MTS X, *(b)* Heatmap of a Relevance Attribution matrix computed on X. The highest value R^6_t is highlighted.

Definition 1 (Relevance Attribution Matrix [6], [16], [21]): Given an MTS X, for which a model $\mathcal{F}^c$ provides a probability to belong to the class a, the Relevance Attribution Matrix is defined as $R(X, \mathcal{F}^c(X)[a]) \in \mathbb{R}^{N \times \ell}$. Each $R_{i,j} \in R(X, \mathcal{F}^c(X)[a])$ reports the importance score of a TS value X^i_j for the model $\mathcal{F}^c$ in determining the prediction for the class a. We denote by $R^k(X, \mathcal{F}^c(X)[a])$ the set containing the $k \in \mathbb{N}$ highest values in $R(X, \mathcal{F}^c(X)[a])$.

In our work, we also consider *causal explanations* [22] in terms of cause-effect relationships among MTS variables. Causal relationships are typically stated through a temporal graphical model that we introduce hereafter.

Definition 2 (Window Causal Graph [23]): Given an MTS $X \in \mathbb{R}^{N \times T}$ a window causal graph is defined by $G^t = (V, E)$ at time t and maximal lag ℓ, where $V = \{X^i_{t-s} | 1 \leq i \leq N \wedge 0 \leq s \leq \ell\}$ is the set of vertices each one representing a feature value. A directed edge $(X^a_{t-\tau}, X^b_t) \in E$ exists if X^a causes the effect X^b at time t with lag $\tau \leq \ell$. Note that if $a = b \implies \tau > 0$ (self-causation).

A window causal graph is consistent throughout time when all the causal relationships remain constant (*Consistency Throughout Time* [23] i.e., Stationary), and always precede their effects (*Temporal Priority* [23]). In this case, we have that $G^t(V, E) = G^{t'}(V, E) \forall t, t'.1 \leq t, t' \leq |X|$ and thus we denote a single window causal graph by $G(V, E)$. In that case, we simply refer to the MTS causal graph. In fact, finding a causal explanation of a model outcome is equivalent to giving the actual causes of that outcome [24], [25]. Hence, causal explanations contain the direct causes of a target TS variable as defined in a causal graph.

Definition 3 (Causal Explanation [26]): Given a MTS X faithful to a causal graph $G(V, E)$ with maximal lag ℓ, the set of direct causes of an outcome X^i_t is defined as follows: $C^i_t(X) = \{(a, t') | t - \ell \leq t' < t \wedge (X^a_{t'}, X^i_t) \in E\}$.

Under the assumptions of (I) faithfulness (only the conditional independence relations true in the data are entailed by the Causal Markov condition applied to the graph) (II) causal sufficiency (no unobserved latent factor influences two observed vertices), (III) correct independence tests (a.k.a. Causal Markov condition [23]: nodes in a causal graph are conditional independent of its nondescendant, given their parents), (IV) temporal priority (causes occur before their effects), (V) minimality (the causal graph does not contain dependencies not present in the data), and (VI) absence of selection bias, the set of direct causes contains all the observed TS variables that are parents of a variable X_t^i in $G(V, E)$, and it is sufficient to predict a target variable X_t^i [26]. Formally, variable X_t^i is independent of $X_{t'}^a$ conditionally on $C_t^i(X)$ where $t' < t$.

III. Explanation Methods

In this work, we focus on *model-specific* and *model-agnostic* post-hoc explanations of TS classification models produced by *Perturbation-based* and *Gradient-based* methods.

A. Perturbation based methods

This family of xAI methods assesses feature relevance by perturbing their values (masking, removing, altering, permuting) and measuring the impact on the model outcome.

Dimension-wise Class Activation Map (dCAM) [1] is a model-specific method that adapts a Class Activation Mapping (CAM) [27] to MTS classification. CAM represents one of the earliest explainability techniques natively applied to image classification in Convolutional Neural Networks (CNN). CAM aims to compute the discriminating data region that induces the model to assign a particular class to a given instance.

We denote by $A \in \mathbb{R}^Z$ the activation map generated by an MTS $X \in \mathbb{R}^{N \times T}$ in the last convolutional filter of the network (before Global Average Pooling layer). Note that $A_{i,j}$ denotes the activation at time step i ($1 \leq i \leq T$) of the j^{th} kernel. Given w_c^j, the weight connecting the activation to the neuron of a class c to the kernel j, the CAM score for class c at time step i is computed as follows:

$$CAM_i^c = \sum_j w_c^j A_{i,j} \tag{1}$$

Each CAM score captures the activation conveying to a single class neuron at timestamp i. Note that such scores would not permit filtering out important features for predicting each class. As a matter of fact, state-of-the-art xAI methods for images (e.g., CAM) are not specifically tailored to MTS, where discovering the relevance of features over different time intervals becomes a crucial requirement. In [6] is empirically shown that such limitation is principally due to the conflation of time and feature representations in the NN layers.

To overcome this limitation, dCAM proposes a new CNN architecture in which the first convolutional filter applies to a TS cube containing all the features (a.k.a TS variables) permutations (of the input instance) along time. To compute relevance over time, dCAM does not require investigating every possible permutation but summarizes feature importance at each timestamp, by applying a heuristic over the permutations cube to obtain the final relevance attribution matrix.

Dynamask (DM) [8] relies on a perturbation operator to create a modified version of the model input in order to generate relevance scores. The operator is defined by considering neighboring values of each feature at different timestamps. By comparing the perturbed prediction to the original prediction, the produced errors can be backpropagated to adjust the matrix scores in the final output. Unlike DM and dCAM tailored to explain DL-based MTS classifiers, the following perturbation-based methods are model agnostic and aim to assess the impact of individual features in terms of model outcome difference, when the features are masked or permuted.

SHapley Additive exPlanations (SHAP) [11] adopt a fundamental concept from cooperative game theory, which consists of assigning importance value (a.k.a SHAP value) according to the effect of a given feature on the model prediction. SHAP values are regression coefficients whose computation requires retraining the model to explain each feature, and thus, it rapidly becomes a computationally expensive solution. Hence, many approximations have been proposed in this respect, for non-sequential and sequential models (a.k.a time series). **GradientSHAP [11] (GS)** is a hybrid approach (gradient-based + perturbation) that approximates SHAP values by adding white noise to each input sample multiple times. It works by selecting a random baseline and a random point along the path between the baseline and the input, computing the gradient of outputs w.r.t. the random points. The final SHAP values represent the expected values of gradients multiplied by the difference between the input and the baselines. **TimeSHAP [9] (TS)** works by creating a parallel linear model to the one to explain, training it with slightly perturbed data (over time). Since the space of different feature sets to perturb grows exponentially, TimeSHAP proposes a sampling strategy, that merges sets (a.k.a. coalitions) of semantically equivalent MTS dimensions and time-steps. The perturbation must change the output of the model, but at the same time, it must be large enough to identify the most important features for predictions. **Feature Ablation** (FA) [14], [15] performs the classification task by replacing each timestamp of an observed TS variable with a given baseline value, and then computing the difference from the outcome generated by the original instance. Such a difference quantifies the importance of the masked feature. **Feature Occlusion**(FO) [16] masks all attributes over a time interval (multivariate sub-sequence), obtaining thus an importance score per time window. The final attribution of each value becomes the mean value of all the window scores in which the value appears. When Z-Normalization is applied, it is common to consider a zero value and a zero-valued TS baseline in FA and FO, respectively. **Feature Permutation**(FP) [16] produces relevance scores by permuting each feature individually and drawing a new feature value randomly in batches.

B. Gradient-based Approaches

This family of xAI methods relies on gradient computation of neural network outputs to quantify feature relevance. Gradient CAM (Grad-CAM) [27] is a generalization of CAM aiming to output relevance scores based on *all the network layers*. Such scores are calculated by backpropagating the gradient of each class with respect to the activation generated by each observed TS variable at each timestamp.

XCM [7] was the first model-specific method adapting Grad-CAM to obtain feature relevance scores for MTS classification.

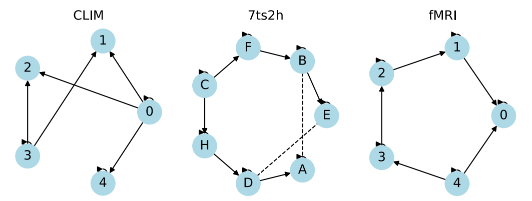


Fig. 2. Causal graphs underlying datasets of the three families: nodes are labeled by TS variables, directed edges represent direct causal relationships. Dashed lines indicate the presence of latent confounders (i.e., non-observable TS variables) that cause both linked observable TS variables).

Formally, the importance in terms of the neuron activation of a feature f for a given class y_c is denoted by:

$$\alpha_f^c = \frac{1}{N \times T} \sum_i \sum_j \frac{\delta y_c}{\delta A_{i,j}^f} \quad (2)$$

The coefficient α_f^c reports the global average of the (back-propagated) gradients of a class score y_c (output of the softmax layer) w.r.t. the activation map A^f obtained by the application of the first 2-dimensional convolution filter on a TS variable instance $X_t^i \in \mathbb{R}^{N \times T}$. To compute the *relevance attribution*, Grad-CAM involves a combination of the relevance coefficient α_f^c and all the feature maps. Hence, the final Grad-CAM score attribution for a class c is given according to the following formulae:

$$L_{2D}^c = ReLU(\sum_f \alpha_f^c A^f) \quad (3)$$

Note that here, a linear rectifier (ReLU) filter keeps only positive attributions, while L_{2D}^c reports a score for each observed TS variable at each timestamp.

Integrated Gradients (IG) is a model-agnostic method [17] that considers the integral of gradients of the model output with respect to the inputs along the straight line path (in $\mathbb{R}^{N \times T}$) from a baseline instance X' to the input X.

IV. Methodology and Experimental Framework

Testbed Datasets Our experimental evaluation considers families of datasets that dispose of casual graph validated by domain experts. Other available time series benchmarks like Exathlon [18] require to first run and evaluate causal discovery algorithms [23], [28]–[31] that is outside the scope of our work. Each dataset family is a collection of instances (MTS) with similar properties. For the sake of reproducibility, the complete framework containing the classification models and the explanation algorithms is available online [32]. We have integrated into our testbed the source code of the model agnostic methods provided by the library Captum[1], as well as, the original implementations of model-specific methods (dCAM, XCM, TimeShap and Dynamask) provided by their authors.

Table I summarizes the main characteristics of the datasets of the three families included in our testbed. Each dataset family contains several MTS instances (number reported in the previous table) of diverse length and dimensionality (i.e., TS variables) committing to causal graphs (see Fig. 2) of varying structure (w.r.t. node number and in-degree) and temporal dependencies (i.e., linear vs nonlinear).

TABLE I

Datasets characteristics: fMRI and CLIM instances commit to different causal graphs (in-degrees are aggregated).

Dataset	Instances	Variables	Timestamps	Avg in-degree	Max lag
7ts2h [23]	10	7	4000	1.8	1
fMRI [31]	17	5	200 to 5000	2.0 to 2.6	1
CLIM [29]	200	5	250	2.0 to 4.4	2

Instances of the `CLIM` family [29] encode linear temporal relationships estimated from a Vector Auto-Regressive model fit on real-world data and processed by domain experts of climate modeling. Instances of the `fMRI` family [33] simulate fMRI-BOLD levels in a brain network using a domain-specific model and includes nonlinear relations as it is generated from a linear ODE with constant coefficients. Finally, instances of `7ts2h` [34] are fully synthetically generated, using a structural model that includes sines, cosines, or absolute values.

A. Datasets

`7ts2h` and `fMRI` exhibit different distributions of MTS instances, as the generative stochastic process relies on different weights and coefficients from one instance to another. On the other hand, their causal graphs differ by two edges in the worst case. `7ts2h` is the only family where the data-generating process contains hidden confounders (i.e., non-observed TS variables), which are part of true causes in the causal graphs.

The causal graph (with lag values information) is available for each instance in all families except for `CLIM`. For this latter, we obtain causal information running SLARAC [35] algorithm, as suggested by Runge et al. in their benchmark [29]. In the obtained causal model, we observe a significant variance in the number of edges across the instances. Moreover, we consider that all the cause-effect pairs have a lag, either one or two, as indicated in the context of the `causality4climate` competition [29].

Problem Settings In our study, we consider the problem of predicting the value of a target TS variable at time $t{+}1$, namely X_{t+1}^i, given the MTS subsequence of the preceding window (of length ℓ) denoted by $X_{(t-\ell+1):t}^{1,..,N}$ (a.k.a predictor). Since the evaluated explanation methods apply to (deep learning) classification models, we compute binary labels from continuous target values. In this respect, we rely on the *KBinsDiscretizer* available in Scikit-learn. Hence, the probability of X_{t+1}^i to belong to one class or another is given by $\mathcal{F}^c(X_{(t-\ell+1):t}^{1,..,N})$. The total number of timestamps for which we predict the value of the target TS variable is $|X| - \ell$.

NN Architectures for Classification To implement MTS classification we rely on state-of-the-art Neural Network (NN) architectures such as Temporal Convolutional Neural Network (TCNN)(dCAM [1], XCM [7]), Long-short Term Memory Network (LSTM) [2] and Transformer [3] Neural Networks.

Unlike the standard deployment of TCNN and LSTM architectures [6], we have made adjustments to the Transformers

[1] https://captum.ai/api/

TABLE II
AVERAGE AUCROC AND NUMBER OF MODELS WITH AUCROC ≥ 0.7 PER NN ARCHITECTURE AND DATASET FAMILY.

	7ts2h	CLIM	fMRI
LSTM	0.916 (51)	0.784 (116)	0.817 (45)
XCM	0.901 (43)	0.801 (49)	0.826 (19)
DCAM	0.867 (38)	0.786 (59)	0.793 (26)
Transformer	0.903 (51)	0.792 (23)	0.906 (16)

NNs architecture commonly adopted in the literature [36]. In particular, we use only a transformer encoder replacing ReLU with GELU activations (in the Position-Wise Feed-Forward Layer). To obtain the final outcome, a single feature vector (per batch) resulting from max-pooling the temporal features is fed to a two-layer multi-layer perceptron, which uses GELU activation prior to producing the classification results (in the softmax layer). Our implementation relies on [37] that learns positional encoding instead of the default sine-cosine scheme in order to obtain a better classification performance overall.

Model training We train a model of each DL architecture type on the prediction task, for each target variable and each MTS instance (in total 1155 models). To obtain meaningful insights from explanation methods, for each NN architecture we consider models built with MTS instances that have an AUCROC above 0.7. The final number of retained models per dataset family and DL architecture is reported in Table II, for a total of 537 models. We perform MTS classification over windows (i.e., sub-sequences) of eight observations. Such window length permits to cover maximal lags between causes and effects in the causal graph, along with several previous timestamps whose influence might be correlated to the target TS variable. We use a 70:30 splitting ratio for training/validation and test respectively, adopting Forward Chaining Cross Validation [38] (FCCV). During training, we oversampled MTS instances to balance classes. We rely on a stratified K-Fold data split to tune NN hyperparameters using the `optuna` tool. During this process, we apply K-Fold to a subset of instances (up to three, depending on the dataset size and time complexity limit) and uniquely on training data.

B. Explanation evaluation metrics

xAI methods are usually evaluated along two axes: (i) to what extent the generated explanations meet user expectations (*Plausibility*); (ii) how accurately they reflect the predictive model to explain (*Fidelity*). In this paper, we focus on the latter criterion, and specifically, we consider the following metrics: (a) *temporal consistency* of explanations produced in different windows; (b) the *relevance of explaining features to the underlying causal graph*; (c) the *predictive performance of the surrogate models trained only on the explaining features*.

1) Temporal Explanation Consistency: In explainability, the notion of *consistency* typically measures to what extent two samples of the same class are explained in the same manner. In the case of the datasets we use, where each MTS value originates from a unique generation process, we expect that explanations are locally consistent over time. Hence, we measure the temporal consistency of an xAI method as the degree of randomness exhibited by explanations across different windows. Specifically, we adopt a definition of consistency similar to the concordance metric introduced in the Exathlon benchmark [18].

More precisely, given an MTS $X \in \mathbb{R}^{N \times T}$ and a classification model $\mathcal{F}^c$, we compute consistency over the k highest scores of relevance attribution matrices resulting from the application of a classification model $\mathcal{F}^c$ on each sub-sequence in X of length ℓ. Recall that the number of sub-sequences of length ℓ in X is given by the number of target TS variables we predict in X, namely $|X| - \ell$. Hence, we define:

$$\text{Consistency} = -\sum_{i=1}^{N}\sum_{j=1}^{\ell} p_k(i,j)\log_2 p_k(i,j) \tag{4}$$

$$p_k(i,j) = \frac{\sum_{t=1}^{|X|-\ell} \mathbb{1}[R_{i,j} \in R^k(X^{1,..,N}_{t,(t+\ell-1)}, \mathcal{F}^c(X^{1,..,N}_{t,(t+\ell-1)})[a])]}{k(|X|-\ell)} \tag{5}$$

In equation (4), $p_k(i,j)$ corresponds to the relevance scores frequency of all sub-sequences at position i, j in the top k scores of each relevance attribution. Each relevance attribution matrix is computed with respect to a class a, where each sub-sequence belongs to. Consistency is thus the entropy of this distribution. In equation (5) $\mathbb{1}[]$ is the indicator function. The sum of all frequencies $\sum_{i=1}^{N}\sum_{j=1}^{\ell} p_k(i,j)$ is equal to 1. Consistency is bounded between a fixed theoretical maximum $\log_2(N \times \ell)$ and a theoretical minimum $\log_2 k$. As Consistency measures the entropy of explanations across time, the lower the value the more consistent the set of k-length explanations we obtain across time.

2) Explanation precision w.r.t. a Causal Graph: To measure whether highly scored relevance features correspond to true causes of the target variable, we introduce Precision and Recall for a fixed explanation length k, using as ground truth the causal graph of each dataset.

Given a MTS $X \in \mathbb{R}^{N \times T}$ and a classification model F_c applied on a number $|X| - \ell$ of sliding windows of X to predict the class of values in a target X^i, we define:

$$P = \sum_{t=1}^{|X|-\ell} \frac{|\{R_{a,b} \in R^k(X^{1,..,N}_{t,(t+\ell-1)}|(a,b) \in C^i_{t+\ell}(X)\}|}{k(|X|-\ell)} \tag{6}$$

$$R = \sum_{t=1}^{|X|-\ell} \frac{|\{(a,b) \in C^i_{t+\ell}(X)|R_{a,b} \in R^k(X^{1,..,N}_{t,(t+\ell-1)}\}|}{|C^i_{t+\ell}(X)|(|X|-\ell)} \tag{7}$$

P measures on average how many relevant features are direct causes of the target attributes in X. Symmetrically, R reports the average number of direct causes that are relevant according to the explanation provided in the top k scores of each relevance attribution matrix.

3) Predictive Performance of Explanations: The predictive performance of explanations is often evaluated using the Insertion Deletion method [39] that measures the impact of top k salient features on the classification performance (e.g., AUC), either by masking those or by masking all other features. While useful for comparing different xAI methods, this method does not suit binary causal relationships with a target MTS variable. While it is possible to define the causal strength of directed feature relationships [40], such information is not available in our ground truth.

In our study, we are interested in measuring the predictive performance of the final classification outcome of the known causes. In this respect, we rely on the remove-and-retrain method (ROAR) [41], which consists of re-training a model with reduced data dimensionality and testing its performance. In a nutshell, for each MTS and model we build and compare two different models: I) the Explanation-Only model (`EO model`), which is is trained with only the k most frequently salient features for p_k, and II) the Cause-Only model (`CO model`) trained with only the features that are direct causes of the target variable.

To compare the performance of these two kinds of surrogate models, we measure the Area Under the Receiver Operating Characteristics Curve (ROC AUC) while performing FCCV. It is worth mentioning that we can not exploit the Insertion Deletion method [39], measuring the AUC of model accuracy variation, when k changes. Since no order exists on the causes' importance in the causal graph ground truth, k must remain fixed to the number of true causes ($k = |C_t^i(X)|$).

V. Experimental Results

In this Section, we report the performance results of the explanation methods presented in Section III according to the metrics we introduced in Section IV. We finally highlight the main conclusions drawn from our experiments.

We evaluate each metric on our 30% test data split. As relevance maps have to be computed w.r.t. a given class, we restrict our analysis to values of the target variable belonging to the positive class of each MTS instance. The number of relevance attribution maps over which metrics are computed varies across MTS instances, for a minimum of 17 in fMRI and CLIM and 47 for 7ts2h.

A. Temporal Consistency of xAI Methods

The series of experiments reported in this section aims to answer the following questions: (i) Is there an xAI method that is more consistent than others across windows? (ii) Is there a NN architecture that is favored by an xAI method when looking for temporally consistent explanations? (iii) How does the length k of explanations affect their consistency, with respect to a random and theoretical best baseline?

Fig. 3 illustrates the temporal consistency of different xAI methods per NN architecture and dataset family, when varying the length k of explanations. The minimal (or optimal) consistency metric corresponding to a deterministic explanation is plotted in dashed font as a reference. Note that given a MTS $X \in \mathbb{R}^{N \times T}$, the largest value for k is thus given by $N \times T$.

The Feature Occlusion (FO) method is evaluated only for k that are multiples of the number of TS variables. Note that the relevance attribution of all TS variables in a given timestamp is identical since the occlusion patch covers all features in a timestamp. In this sense, the FO method estimates temporal importance rather than feature importance. Hence, we consider only k values that correspond to selecting all features in the most salient timestamps.

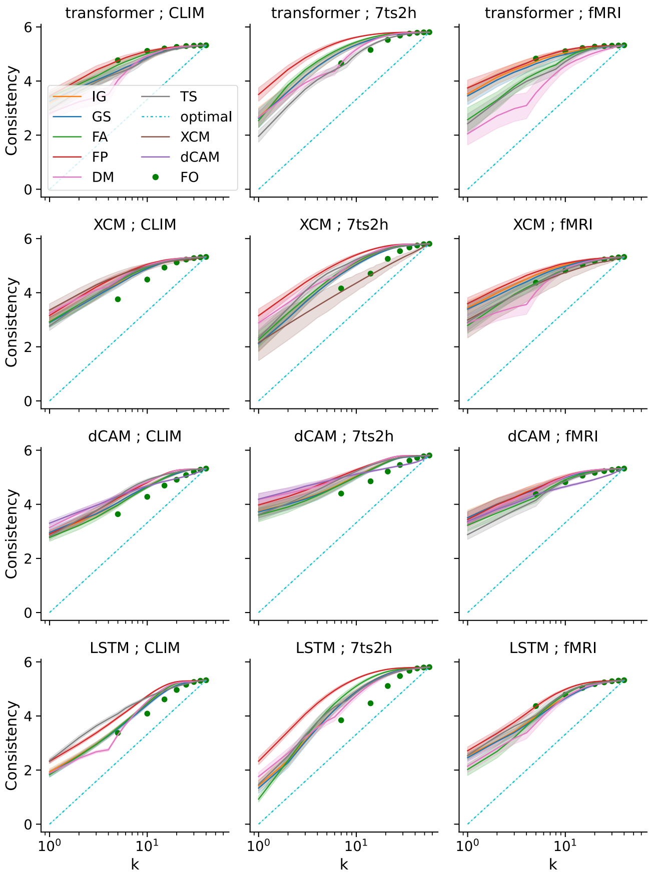


Fig. 3. Mean temporal consistency when varying k per dataset family and NN architecture. Lines represent average consistency across all target TS variables in a dataset and areas the 95% confidence interval of predictions.

We first focus on the temporal consistency of xAI methods (i). As we can observe in Fig. 3 with the exception of dCAM and XCM, explanation methods exhibit similar trends for the same NN architecture and dataset family. The two gradient methods IG and GS have nearly the same consistency, which we can attribute to the high similarity between the two methods. FA is similar but with better consistency at $k < 5$, showing that the method's higher saliency features tend to vary less.

The lower consistency metric value that DM shows around $k = 4, 5, 6$ is a consequence of the sparsity loss term optimized by the method. It forces an adjustable number of relevance coefficients to zero and the rest to 1 (and selects by itself this adjustable parameter). There is thus a reduced amount of randomness in its top k features as k gets close to this threshold. Method TS performs close to or better than FA, except for LSTMs. As for FO, it is more consistent than other methods due to ranking timestamps and not individual features, which decreases the entropy of the associated distribution. As expected, methods with coalition constraints (FO, DM, TS) are generally more consistent than univariate alternatives, except FA.

The randomness of the FP method could be attributed to the batch computation. To permute features, FP explains at the same time a batch of windows. The values of the permuted features come from the same batch. Since our batch size is 32, the distribution of permuted features varies in practice across batches.

The dCAM method shows a distinct behavior, especially at high k, due to the internals of its architecture. We observe that CAM coefficients regularly have low variance in the last time steps of the window. This side effect is likely due to the input padding, which consists of adding zero coefficients to the input of each convolutional layer. It derives that the feature relevance in the last timestamps of each window is consistently low.

As for the XCM method in 7ts2h, we observe that for about 60% of the models, the saliency maps produced consist of zero-valued coefficients. XCM essentially runs two parallel stacks of convolution layers with temporal (1D) and spatiotemporal (2D) kernels respectively. To build explanations, XCM only uses the spatiotemporal stack. We note that the temporal pipeline has wider kernels and three to four times more trainable weights. Obtaining zero-valued explanations denotes that the model has learned to use only this part. For the other 40% of the models, we observe a consistency at around four for $k = 1$. In fMRI and CLIM, the produced saliency maps are different from zero for more than 90% of the models.

Then, we investigate the effect of different NN architectures (ii). Our first observation is that up to $k = 10$, explanations of LSTM models are systematically more consistent than their counterparts computed on TCNN and Transformer (except FA on Transformer). For $k = 1$ in particular, for each xAI method and dataset family, the consistency metric on LSTM is at least one point below the consistency metric on the other types of models. The consistency at low k (LSTM, XCM, Transformer, DCAM) is ordered by the number of trainable parameters in our NN architectures. More complex models might distinguish more finely between different inputs, leading to a higher variation of explanations. A more extensive study of the impact of the model size on the temporal consistency of explanations is left as future work.

We finally study the consistency of explanations for different k values (iii). Our general observation is that the more features in the explanation, the less consistent the explanation is, across all NN architectures and xAI methods. For LSTM especially, the explanation methods get farther away from the minimal consistency as k increases. This indicates that the top few features are less randomly distributed than the rest ones, especially when it comes to the last 15 features, where consistency becomes close to a fully random explanation. We observe the ability of the xAI methods to report somewhat consistent most salient features while having very random small coefficients for others. Another noteworthy result is that for the LSTM model on 7ts2h with FA, explanations of length one have about the same entropy as a uniform distribution of support two. This means that the explanation methods are mostly concentrating on one or two features.

Summary: xAI methods produce more consistent explanations of LSTMs, which could be due to their small size. On the same NN architecture and dataset family, coalition-based method tend to be more consistent than univariate alternatives, with FA as an exception for low explanation sizes. dCAM and XCM have side effects leading to distinct behaviors.Finally, we observe that explanations quickly become random as we include features of lower relevance.

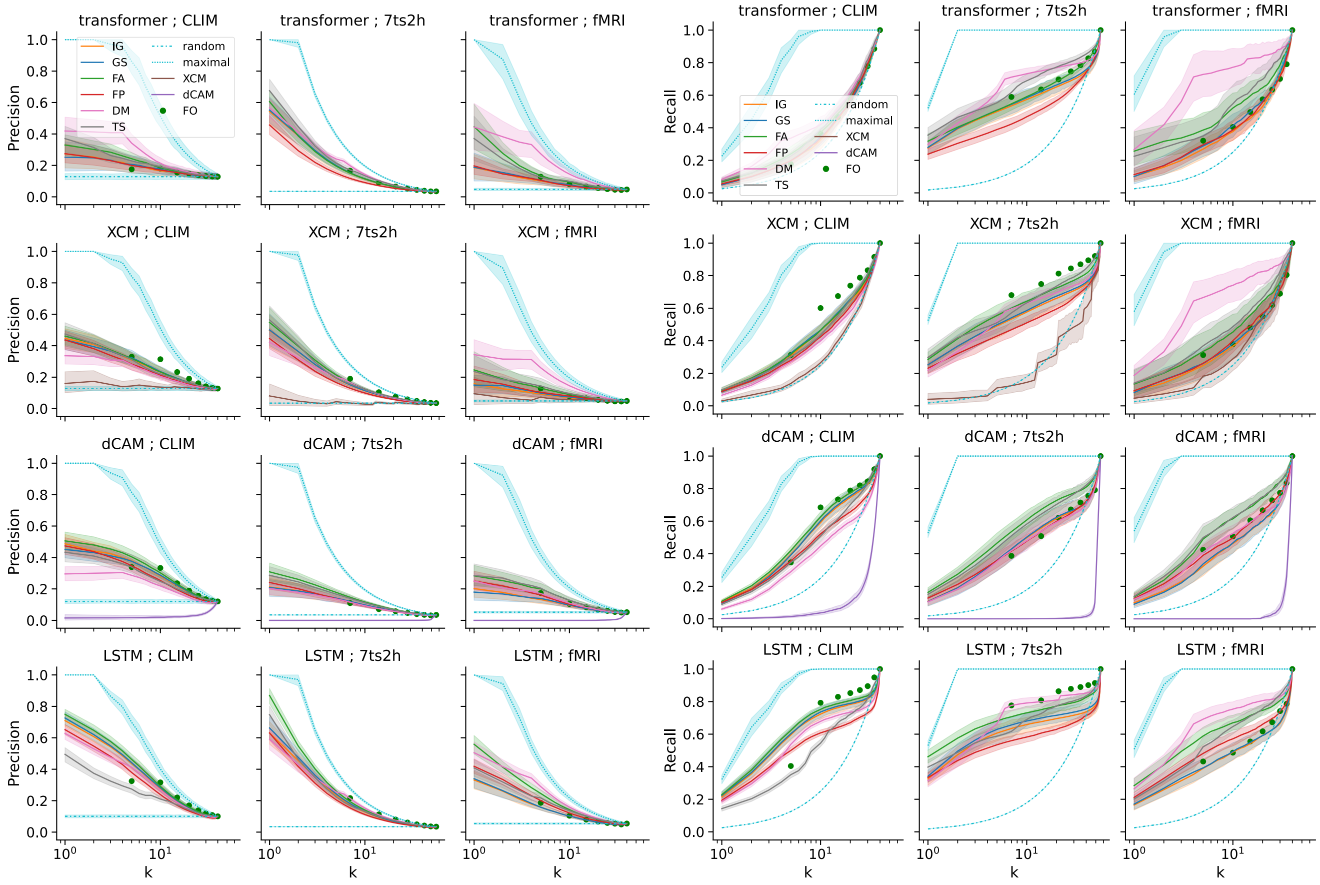
B. Causal Relevance of xAI Methods

The series of experiments reported in this section aims to identify the extent to which xAI methods are able to discover the true causes of the predicted variable in the causal graph. Similarly to the previous section, we are interested in the comparative performance i) of the different xAI methods, ii) of the different NN architectures, iii) and how performance metrics are affected as k grows, with respect to a random and theoretical best baseline.

We measure for each xAI method the precision and recall per explanation length k using as ground truth the respective causal graph. We first compute these metrics for individual MTS instances and predicted variables and average them over the time t. Then, we average across MTS instances and targets to plot in Fig. 4 a single curve per dataset family and NN architecture. We add two baselines: max and random, corresponding respectively to the maximal precision or recall that can be obtained, and the expectation of these metrics if the explanation was a uniformly randomly chosen feature set.

On TCNNs and LSTM, the xAI methods that achieves the best precision and recall are FA and TS on most datasets with close performances. DM seems well suited to the fMRI dataset. The sparsity constraint around $k = 4, 5, 6$ is effective in selecting the best features. Gradient-based methods IG and GS obtain near identical results. FP exhibits a performance either above or below IG and GS depending on the NN architecture and dataset family. dCAM explanations are extremely inaccurate: similarily to consistency, dCAM explanations miss the ground truth causes as they are usually situated within the last few timestamps of the saliency map that have systematically low coefficients. On the opposite, FO has a high precision and recall starting from the second marker on CLIM, since FO often identifies the two timestamps in which the ground truth causes are located. XCM produces close-to-random explanations. As seen in the previous section, XCM explains the less relevant spatiotemporal part of the network. The problem of 0-valued saliency maps is a particular case of this partial model explanation. Finally, we remark that in general, all xAI methods exhibit a similar standard deviation of the precision across all datasets and model combinations (differing by less than 0.1 to 0.05), decreasing as k increases. On top of it, we remark no clear ordering on all (dataset, model) pairs, except that GS and IG are nearly identical and have the highest variance on 7ts2h and CLIM.

Considering model differences (ii), we see that LSTM dominates over the others on all methods and dataset families.



(a) Precision

(b) Recall

Fig. 4. Mean Precision (a) and Recall (b) as a function of k per NN architecture and family of datasets. Lines represent metrics at a given k averaged over all predicted features and instances and areas the 95% confidence intervals of predictions.

At $k = 1$, each xAI method dominates its counterparts in other models by at least 0.1. This is a surprising result as the AUCROC of the models used for explanation are not higher in LSTM compared to XCM and Transformer (see Table II). Nevertheless, we observe that the LSTM models obtain an AUCROC higher than 0.7 on more MTS instances. We conjecture that even if two architectures have similar performance, the architecture that can be applied reliably on many MTS instances will focus more on the true causes. The recurrent structure of the LSTM permits it to outperform other models since it is capable of better learning cause and effect with a short lag (1 or 2).

Regarding the effect of the explanation length (iii), we observe that for $k = 1$ LSTM paired with FA achieves a close to 0.9 precision for the most salient feature in 7ts2h, and above 0.75 for CLIM. The recall of explanations methods excluding dCAM and XCM shows that up to $k = 8$, the k-th feature is identifying true causes better than a random guessing. Afterward, added features bring little information until the recall reaches the random baseline. It becomes clear that only the top few features include the true causes of our target variable, while the rest of the explanation contains noise.

Finally, we observe a lower starting precision on fMRI than CLIM or 7ts2h. Table II reports a lower number of successful models for fMRI than for CLIM and 7ts2h. We reach a similar conclusion as in the analysis of per model type behavior: the ability of a model to achieve acceptable performances on a diverse set of MTS of similar dynamics is linked to model reliance on true causes.

Summary: Overall, methods other than dCAM and XCM succeed in recovering at least one of the true causes. Still, none of them are able to recover all causes, as features outside of the top 10 bring little information. FA and DM have a small lead on other xAI methods depending on the model and the dataset. Finally, models that perform well on more MTS instances, as LSTM does, have explanations closer to the causal graph.

C. *Predictive Performance of xAI Methods*

The last part of our analysis concerns the predictive performance of surrogate models trained only over the explaining features. We seek to know if the top salient features are sufficient predictors for training surrogate models and compare their performance to a baseline model where only the direct causes from ground truth are included as input. Thus, we focus on i) whether the features that are most frequently included in explanations can be used to predict the target TS variable as the original model, ii) how the performance of these surrogate

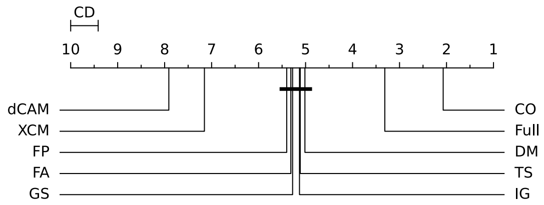


Fig. 5. Critical difference (CD) diagram of the AUCROC of the EO, CO, and Full models (Nemenyi test CD=0.518). The scale denotes the average ranking of each method, and bold lines denote non significant differences.

models compares to models exclusively built over the direct causes set of the target TS variable.

Both questions i) and ii) relate to the relative performance of the explanation-only models with FA, IG, GS, FP, DM, XCM, dCAM, TS (EO models), along with cause-only models (CO models) and Full (original) models. We build a critical difference diagram using the AutoRank library [42] to test if there are any significant differences between the 10 models [43]. We rely on the non-parametric Friedman test [43] as an omnibus test, and reject the null hypothesis that all methods come from the same distribution (with a significance level of 5%). Next, we use the post-hoc Nemenyi test, in order to compare the methods in pairs. Two methods are significantly different when their average ranks differ more than a critical distance (CD) of 0.585 (at a significance level of 0.05). We conduct our test on the 536 data points that are formed by successfully trained models (Table II). Fig. 5 depicts the statistically significant models in a critical difference diagram.

We observe that i) EO models exhibit on average a poor predictive performance compared to Full models, and ii) that CO models outperform on average not only the EO models but also the Full models. xAI methods definitively fail to identify a minimal set of features necessary to the prediction task, unlike causal predictors leading on average to a higher model quality.

Fig. 6 finally depicts how often EO models exhibit a worse performance than the CO models. Only in a few MTS instances, we observed that EO models are better than their CO counterparts. We speculate that this difference is attributed to two factors. The first is the independence relations not entailed by the causal graph (i.e. the faithfulness of the causal graph to the distribution it explains). Specifically, it is possible that some TS variables in our true causes have low causal strength. The second factor is specific to the CLIM dataset, where the causal ground truth is observed to be slightly inaccurate for some particular MTS instances.

Summary: EO models clearly underperform compared to the Full and CO models. This demonstrates not only the predictive power of direct causes w.r.t. the salient features but also w.r.t. to the full models trained on the original set of features. As a matter of fact, full models exploit correlations between all available features to gain predictive performance that are unlikely to generalize in different environments, especially in high-dimensional settings.

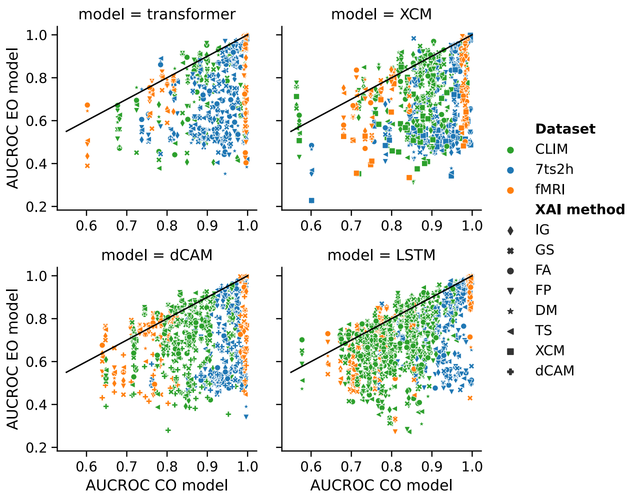


Fig. 6. AUCROC of explanation-only restricted (EO) models compared to the causal-only restricted (CO) models, for each predicted variable and xAI method. Points above the diagonal represent datasets where the EO outperforms the CO model.

VI. Conclusion

The main conclusion drawn from our experiments is that relevance attribution methods do not produce consistent explanations across time and seldom discover the true causes of the predicted variables. As expected, the intrinsic measures underlying post-hoc methods are purely associational and usually expose potentially spurious and misleading correlations in input data used to train a NN model. For this reason, surrogate models build over salient features systematically underperform w.r.t. MTS classifiers build over the full feature space. On the contrary, sparse NN models build over causal explanations generalize much better than the full models.

We let as future work the study of how the size/complexity of the models influences the consistency of the xAI methods. We also remark on the need of adding more stochastically generated dataset families, ideally with controllable noise and diverse dynamics. This would allow experimenting on a link between the discovery of true causes and the robustness of the model type (to which extent the same NN architecture can learn on different MTS instances of shared dynamics). Dataset families with a large number of covariates could be added to study if NN can learn sparse causal information despite the curse of dimensionality, and if sparsity constraints could make the networks focus on causal variables. Another direction of improvement would be enlarging the causal evaluation of xAI methods to counterfactual explanations [44], to forecasting and multi-class models interpretability methods [45].

References

[1] P. Boniol, M. Meftah, E. Remy, and T. Palpanas, "dcam: Dimension-wise class activation map for explaining multivariate data series classification," in *SIGMOD*, 2022.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] M. A. Nemer, J. Azar, J. Demerjian, A. Makhoul, and J. Bourgeois, “A review of research on industrial time series classification for machinery based on deep learning,” in *4th IEEE MENACOMM*, 2022, pp. 89–94.

[5] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. Muller, “Deep learning for time series classification: a review,” *Data Min. Knowl. Discov.*, vol. 33, no. 4, pp. 917–963, 2019.

[6] A. A. Ismail, M. Gunady, H. Corrada Bravo, and S. Feizi, “Benchmarking deep learning interpretability in time series predictions,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6441–6452.

[7] K. Fauvel, T. Lin, V. Masson, É. Fromont, and A. Termier, “Xcm: An explainable convolutional neural network for multivariate time series classification,” *Mathematics*, vol. 9, p. 3137, Dec 2021.

[8] J. Crabbé and M. van der Schaar, “Explaining time series predictions with dynamic masks,” in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds.

[9] J. Bento, P. Saleiro, A. F. Cruz, M. A. T. Figueiredo, and P. Bizarro, “Timeshap: Explaining recurrent models through sequence perturbations,” in *KDD ’21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2021*, 2021.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, “”why should i trust you?”: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144.

[11] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.

[12] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” in *ICLR (Workshop Poster)*, 2014.

[13] S. Beckers, “Causal explanations and XAI,” in *First Conference on Causal Learning and Reasoning*, 2022.

[14] W. Freeborough and T. van Zyl, “Investigating explainability methods in recurrent neural network architectures for financial time series data,” *Applied Sciences*, vol. 12, no. 3, 2022.

[15] H. Suresh, N. Hunt, A. E. W. Johnson, L. A. Celi, P. Szolovits, and M. Ghassemi, “Clinical intervention prediction and understanding using deep networks,” *CoRR*, vol. abs/1705.08498, 2017.

[16] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *13th European Conference on Computer Vision ECCV-Part I*, vol. 8689, 2014, pp. 818–833.

[17] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *Proceedings of the 34th International Conference on Machine Learning, ICML*, D. Precup and Y. W. Teh, Eds., 2017.

[18] V. Jacob, F. Song, A. Stiegler, B. Rad, Y. Diao, and N. Tatbul, “Exathlon: A benchmark for explainable anomaly detection over time series,” *Proc. VLDB Endow.*, vol. 14, no. 11, p. 2613–2626, oct 2021.

[19] T. R. et al., “Searching and mining trillions of time series subsequences under dynamic time warping,” in *SIGKDD*, 2012.

[20] E. J. Keogh and S. Kasetty, “On the need for time series data mining benchmarks: A survey and empirical demonstration,” *Data Min. Knowl. Discov.*, 2003.

[21] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, “Towards better understanding of gradient-based attribution methods for deep neural networks,” in *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[22] J. Pearl, “The seven tools of causal inference, with reflections on machine learning,” *Commun. ACM*, vol. 62, no. 3, p. 54–60, feb 2019.

[23] C. K. Assaad, E. Devijver, and É. Gaussier, “Survey and evaluation of causal discovery methods for time series,” *J. Artif. Intell. Res.*, vol. 73, pp. 767–819, 2022.

[24] J. Woodward, *Making Things Happen: A Theory of Causal Explanation*. Oxford Univ. Press, 2003.

[25] J. Y. Halpern and J. Pearl, “Causes and explanations: A structural-model approach. part i: Causes,” *British Journal for the Philosophy of Science*, vol. 56, no. 4, pp. 843–887, 2005.

[26] J. Runge, “Causal network reconstruction from time series: From theoretical assumptions to practical estimation,” *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 7, 2018.

[27] 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.*, vol. 128, no. 2, pp. 336–359, 2020.

[28] M. Nauta, D. Bucur, and C. Seifert, “Causal discovery with attention-based convolutional neural networks,” *Mach. Learn. Knowl. Extr.*, 2019.

[29] J. Runge, X.-A. Tibau, M. Bruhns, J. Muñoz Marí, and G. Camps-Valls, “The causality for climate competition,” in *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, vol. 123, 2020, pp. 110–120.

[30] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich, “Network modelling methods for fmri,” *NeuroImage*, vol. 54, no. 2, 2011.

[31] Y. Huang and S. Kleinberg, “Fast and accurate causal inference from time series data.” in *FLAIRS Conference*, 2015, pp. 49–54.

[32] E. Vareille, A. Abbas, M. Linardi, and V. Christophides, “https://github.com/mlinardiCYU/Evaluating_xAI_Time_Series_Causal_Lens.git,” 2023.

[33] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich, “Network modelling methods for fmri,” *Neuroimage*, vol. 54, no. 2, pp. 875–891, 2011.

[34] K. Assaad, E. Devijver, and E. Gaussier, “7ts2h structure,” 2020. [Online]. Available: https://doi.org/10.7910/DVN/UC7JME

[35] S. Weichwald, M. E. Jakobsen, P. B. Mogensen, L. Petersen, N. Thams, and G. Varando, “Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values,” in *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, 2020, pp. 27–36.

[36] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, “A transformer-based framework for multivariate time series representation learning,” in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 2114–2124.

[37] P. Lippe, “UvA Deep Learning Tutorials,” 2022.

[38] C. Bergmeir and M. Benítez, “On the use of cross-validation for time series predictor evaluation,” *Inf. Sci.*, vol. 191, 2012.

[39] V. Petsiuk, A. Das, and K. Saenko, “Rise: Randomized input sampling for explanation of black-box models,” in *British Machine Vision Conference*, 2018.

[40] B. Fitelson and C. Hitchcock, “60029 Probabilistic measures of causal strength,” in *Causality in the Sciences*. Oxford Univ. Press, 03 2011.

[41] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, “A benchmark for interpretability methods in deep neural networks.” in *NeurIPS*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 9734–9745.

[42] S. Herbold, “Autorank: A python package for automated ranking of classifiers,” *Journal of Open Source Software*, vol. 5, no. 48, p. 2173, 2020.

[43] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” *Journal of Machine Learning Research*, vol. 7, no. 1, pp. 1–30, 2006.

[44] E. Ates, B. Aksar, V. J. Leung, and A. K. Coskun, “Counterfactual explanations for multivariate time series,” in *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*, 2021, pp. 1–8.

[45] O. Ozyegen, I. Ilic, and M. Cevik, “Evaluation of interpretability methods for multivariate time series forecasting,” *Applied Intelligence*, vol. 52, no. 5, pp. 4727–4743, 2022.