CausalConceptTS: Causal Attributions for Time Series Classification using High Fidelity Diffusion Models

Anonymous authors
Paper under double-blind review

Abstract

Despite the excelling performance of machine learning models, understanding their decisions remains a long-standing goal. While commonly used attribution methods from explainable AI attempt to address this issue, they typically rely on associational rather than causal relationships. In this study, within the context of time series classification, we introduce a novel model-agnostic framework to assess the causal effect of concepts, i.e., predefined segments within a time series, on specific classification outcomes. To achieve this, we leverage state-of-the-art diffusion-based generative models to estimate counterfactual outcomes. Our approach compares these causal attributions with closely related associational attributions, both theoretically and empirically. We demonstrate the insights gained by our approach for a diverse set of qualitatively different time series classification tasks. Although causal and associational attributions might often share some similarities, in all cases they differ in important details, underscoring the risks associated with drawing causal conclusions from associational data alone. We believe that the proposed approach is widely applicable also in other domains to shed some light on the limits of associational attributions.

1 Introduction

Machine learning has achieved remarkable success across diverse fields, thanks to the development of powerful hardware and the collection of large datasets. Time series data, widely present in domains such as natural sciences, medicine, and life sciences (Wang et al., 2023; Esteva et al., 2019; Miotto et al., 2018; Shen et al., 2017; Bepler & Berger, 2021) serve as invaluable resources for modeling temporal patterns and dependencies, particularly in widely considered classification settings (Rajkomar et al., 2018; Wang et al., 2019). However, complex models such as deep learning models often sacrifice interpretability for performance, a trade-off that can be critical in downstream tasks (Somani et al., 2021; Roy et al., 2019).

Need for explainability A lack of insights into the model's decision making process often represents a significant hurdle when it comes to the the deployment of deep learning models in particular in safety-critical domains. This led to the emergence of the subfield of explainable artificial intelligence (XAI), see (Lundberg & Lee, 2017; Montavon et al., 2018; Covert et al., 2021) for reviews. Existing literature on XAI for time series classifiers has explored various methods (Crabbé & Van Der Schaar, 2021; Raykar et al., 2023; Zhao et al., 2023; Rojat et al., 2021; Ismail et al., 2020). However, the majority of the proposed methods rely on associations whereas ultimately one is rather interested in uncovering causal effects. Moreover, a clear understanding of the precise differences between these two kinds of attributions, both on a theoretical level as well as on an empirical level, is lacking.

Need for causal insights Counterfactual inference is a type of causal reasoning that involves estimating the effect of a particular intervention or treatment on an outcome by comparing it to what would have happened if a certain intervention or treatment had been applied. In medical applications, counterfactual inference has been used to estimate the effect of a treatment on a patient's health outcome (Gillies, 2018). As nicely laid out in (Goyal et al., 2019), causal attributions provide a clear advantage in the case of correlated features. The hypothetical scenario where the classifier bases its decision only on one of two correlated features cannot

be resolved with associational attributions. Therefore, associational attributions possibly fail to capture the actual model behavior.

Main contributions In this paper, we introduce *Causal Concept Time Series Explainer (CausalConceptTS)*, a novel model-agnostic, causal attribution method, which was specifically designed to enhance the interpretability of time series classification tasks by leveraging causal concepts. More specifically, our main contributions can be described as follows: (1) We formalize the difference between causal and associational attributions for predefined segments within time series data (2) We demonstrate how counterfactual outcomes, required for causal attributions, can be estimated using state-of-the-art diffusion models. (3) We conduct a comparative analysis of causal and associational attributions for a diverse set of time series classification tasks, highlighting the necessity to overcome purely associational attributions for more reliable model insights.

2 Related work

Classification The taxonomy of traditional machine learning algorithms for time series classification is extensive, encompassing various approaches such as distance-based methods (Rakthanmanon & Keogh, 2013), feature-based techniques (Fulcher & Jones, 2017), interval-based models (Deng et al., 2013), shapelet-based algorithms (Hills et al., 2014), and dictionary-based methods (Schäfer, 2015). In addition to these traditional methods, numerous deep-learning techniques have been proposed for time series classification. These leverage different backbone architectures, including Convolutional Neural Networks (CNNs) (Ismail Fawaz et al., 2020), Recurrent Neural Networks (RNNs) (Karim et al., 2017), self-attention mechanisms (Rußwurm & Körner, 2020), and most recently state space models (Gu et al., 2022). This work, this work, rely on the latter architecture, but stress that the proposed method is applicable to any model, including non-deep-learning models.

Deep generative models The generation of synthetic time series data with deep learning has been implemented in various contexts such as conditional generation (Alcaraz & Strodthoff, 2023b), class imbalance (Hssayeni, 2022), anomaly detection (Bashar & Nayak, 2020), imputation (Tashiro et al., 2021; Alcaraz & Strodthoff, 2023a), or explainability (Goyal et al., 2019). While early backbone architectures involve VAEs and GANs, diffusion models have recently emerged as powerful alternative (Tashiro et al., 2021; Alcaraz & Strodthoff, 2023a). We therefore also rely on diffusion models as generative model to estimate high-fidelity counterfactual input sequences, while, again, wishing to add that the proposed method does not rely on a specific choice for the generative model.

Counterfactuals for time series data Several approaches have been explored for utilizing counterfactuals to handle time series data. Ates et al. (2021) experimented with multivariate settings for individual treatment effects, but their approach involves random sampling from appropriate training set samples, leading to discontinuous counterfactual samples. Delaney et al. (2021) proposed an instance-based framework that intervenes in samples until they belong to a different class of interest, however, the intervention areas are limited to neural network findings extracted via class activation mappings. Li et al. (2022) utilized motif discovery for identifying intervention areas, which represents a rather limited scenario due to its focus on precisely recurring patterns. Wang et al. (2021) introduced a framework for generating counterfactuals from the latent space of neural networks, capable of learning both low and high-level concepts, however, it is only applicable to univariate time series data. We are to the best of our knowledge the first to use high-fidelity diffusion models to estimate counterfactual time series inputs.

Attribution methods for time series classification Explainable methods for time series range across diverse downstream tasks as classification (Crabbé & Van Der Schaar, 2021), and forecasting (Raykar et al., 2023). For recent reviews we refer to see (Zhao et al., 2023) for post-hoc methods, emphasizing backpropagation, perturbation, and approximation methods and (Rojat et al., 2021) for ante-hoc methods. As already briefly mentioned above, the existing attribution methods focus almost exclusively on associational effects as opposed to the proposed approach, which aims to infer causal effects. In this respect, the most closely related prior work is the work of Goyal et al. (2019). They use a variational autoencoder to infer counterfactuals, albeit in the context of image classification models. They mostly rely on manually defined attributes as concepts, whereas our concepts are defined in combination with specific subsets of the input.

3 CausalConceptTS: Causal Concept Time Series Explainer

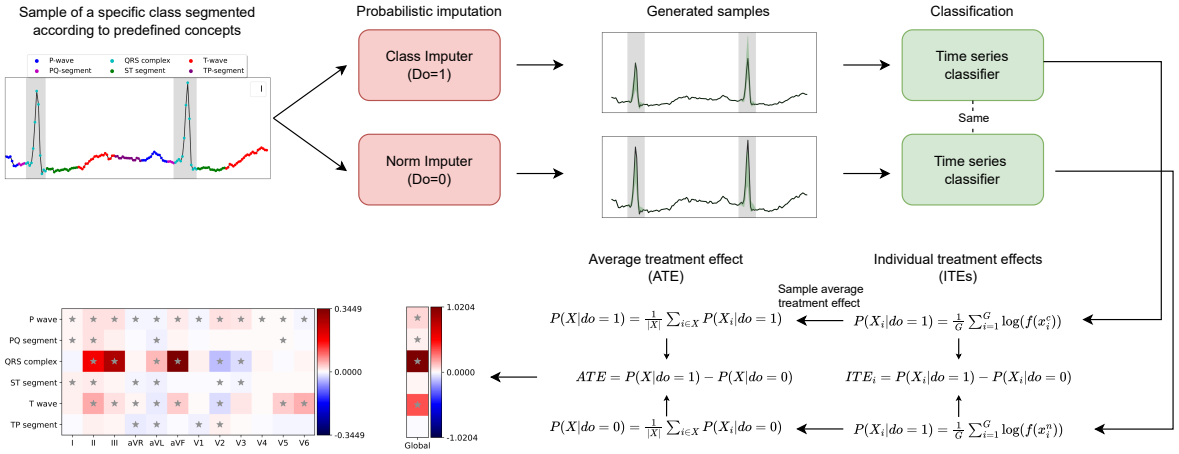


Figure 1: Schematic representation of the proposed *CausalConceptTS* approach: We start from a sample corresponding to a specific class, segmented according to predefined concept, which can either be expert-defined (such as ECG segments) or simply inferred by clustering. For a chosen concept, we impute corresponding segments using two different imputation models, one trained on samples corresponding to the original class and one corresponding to a baseline class of choice typically associated with healthy controls, yielding two sets of imputed samples. These two sets are passed through a predefined classifier of our choice that we aim to investigate. The log difference of the corresponding mean output probabilities yields an individual treatment effect or causal attribution quantifying the causal effect of the concept in question on a specific classifier output. Sample-averaged ITEs yield corresponding average treatment effects (ATEs), which we visualize in terms of channel-agnostic as well as channel-specific causal attribution maps.

Causal data generating process Building on work on causal attributions in the context of image data (Goyal et al., 2019), we adopt the causal data-generating process proposed in (Schölkopf et al., 2012). We phrase the following discussion in a medical language but stress that the framework applies to time series in general and even other data domains with concepts defined based on segmentation masks of the original input.

We assume that a patient's disease state, in our case parametrized through several binary indicator variables D is generated through some noise variable ϵ_D, together with static patient data such as demographic data, which we do not model explicitly but only through (unobserved) noise variables ϵ_S. The processed proceeds involving further (unobserved) noise variables ϵ_M, ϵ_X^c $(c = 1, \ldots, C)$. More specifically, we assume that the data-generating process proceeds in several stages, which we formulate in the language of structural causal models (SCMs) (Pearl, 2009):

1. We assume that the disease state is generated from two noise variables ϵ_D and ϵ_S through a SCM g, i.e. , $D = g(\epsilon_D, \epsilon_S)$.

2. Rather than assuming that the signal is generated directly, we assume that the generation process proceeds via a semantic segmentation mask M (with entries in $[1, \ldots, C]$) of the same shape as the eventual signal. Again, we assume that M is generated through an SCM h_M, i.e. $M = h_M(D, \epsilon_S, \epsilon_M)$ from the disease state D and two noise variables ϵ_S and ϵ_M. As definite examples, M could represent ECG-segments in the case of ECG data or microstates in the case of EEG data.

3. The signal X is subsequently generated from the mask M and the disease state D, i.e., $X = h_X(M, D, \epsilon_X^1, \ldots, \epsilon_X^C$. More explicitly, $X \equiv X(X^1, \ldots, X^C, M)$, where X^c denotes the subset of X where the mask M takes the value c, i.e, the actual numerical values X takes in segment c. As before,

we assume that $X^c = h_X^c(D, \epsilon_S, \epsilon_X^c)$ for a SCM h_X^c, i.e., the actual signal corresponding to segment c is generated based on the disease state D and additional noise variables.

4. Eventually the signal X is passed through the fixed classifier f (output probability of specific class) under consideration to estimate counterfactual outcomes.

Individual and average treatment effects We now aim to investigate the causal effect of the disease state D on the classifier f by intervening on D. As a simplifying assumption, we assume that the underlying segmentation map M remains unchanged under this intervention, i.e., we only intervene on the level of h_X. We intervene by setting the disease state to a specific value D^* (which in our case coincides with the label of the sample X). As reference value we consider a baseline state D^0 (typically associated with healthy control samples). Then the *individual treatment effect (ITE)* for sample X of segment $c \in [1, \ldots, C]$ on the classifier f is defined as (Shalit et al., 2017)

$$\begin{aligned} \mathrm{ITE}(X, f, c, D^*, D^0) = &\log_2 E_{h_X^c} f(X(X_c^{\complement}, (X_c|\mathrm{do}(D = D^*)), M)) \\ &- \log_2 E_{h_X^c} f(X(X_c^{\complement}, (X_c|\mathrm{do}(D = D^0)), M)) \,, \end{aligned} \tag{1}$$

where we use, in contradistinction to the conventional definition of the ITE, logarithmic differences instead of ordinary differences since we aim to compare output probabilities, see (Blücher et al., 2022) for a discussion in the context of associational attributions. The expectation value in Eq. (1) refers to the data-generating process h_X^c. Here and in the following, we use a shorthand notation where $X_c^{\complement}$ refers to the complement of X_c in the set of all features, i.e., $X_c^{\complement} \equiv \{X_1, \ldots, X_{c-1}, X_{c+1}, \ldots, X_C\}$. Below, we will use a high-fidelity generative model to sample from h_X^c. By averaging over samples, we obtain the *average treatment effect* (over the set of samples with disease state D^*), i.e.,

$$\mathrm{ATE}(f, c, D^*, D^0) = E_{X \sim \mathcal{D}(D^*)} \mathrm{ITE}(X, f, c, D^*, D^0) \,, \tag{2}$$

where $\mathcal{D}(\mathcal{D}^*)$ refers to the data distribution of samples with label D^*.

Individual associational effect Note that the individual treatment effect shows a strong structural resemblance to the previously proposed PredDiff attribution method (Blücher et al., 2022), which can be considered as a special case of the Shapley value formalism where only a single coalition (the complement of the feature set X_c under consideration) contributes. In analogy to Eq. 1, we define an *individual associational attribution (IAA)*

$$\mathrm{IAA}(X, f, c, D^*, D^0) = \log_2 f(X) - \log_2 E_{X_c \sim k_X^c} f(X(X_c^{\complement}, X_c, , M)) \,, \tag{3}$$

where the expectation value refers to the conditional distribution $k_X^c \equiv p(X_c|X_c^{\complement})$. The IAA coincides with the PredDiff attribution for X_c.

Relation between causal and associational attributions We can now compare Eq. 1 and Eq. 3 to identify differences and similarities between causal and associational attributions. The first term in Eq. 1 refers to the observed outcome. We therefore expect that $E_{h_X^c} f(X(X_c^{\complement}, (X_c|\mathrm{do}(D = D^*), M)) \approx f(X)$ if D^* coincides with the true label of the sample X. The second term in Eq. 1 refers to the counterfactual outcome. The main difference between the *causal* ITE from Eq.1 and the *associational* attribution from Eq.3 boils down to the use of a class-conditional imputer (conditioned on the background state D^0) in the case of the causal ITE,

$$E_{h_X^c} f(X(X_c^{\complement}, (X_c|\mathrm{do}(D = D^0), M)) \approx \int \mathrm{d}X_c f(X(X_c^{\complement}, X_c, M)) p(X_c|D^0, X_c^{\complement}) \,, \tag{4}$$

compared to using a (class-)unconditional imputer in the case of the associational IAA,

$$E_{X_c \sim k_X^c} f(X(X_c^{\complement}, X_c, M)) := \int \mathrm{d}X_c f(X(X_c^{\complement}, X_c, M)) p(X_c|X_c^{\complement}) \,, \tag{5}$$

where we omitted the dependence of the probability weight on the segmentation mask M to simplify the notation. The insights from this paragraph allow us to empirically compare causal and associational attributions on the level of individual samples. The relation in Eq. 4 is only approximate as it only captures the dependence of the generative distribution on D^0 but neglects a dependence on other possibly confounding variables such as static patient metadata. For a visual overview of our proposed pipeline workflow, see Figure 1. The causal graph underlying our approach is shown in Figure 2.

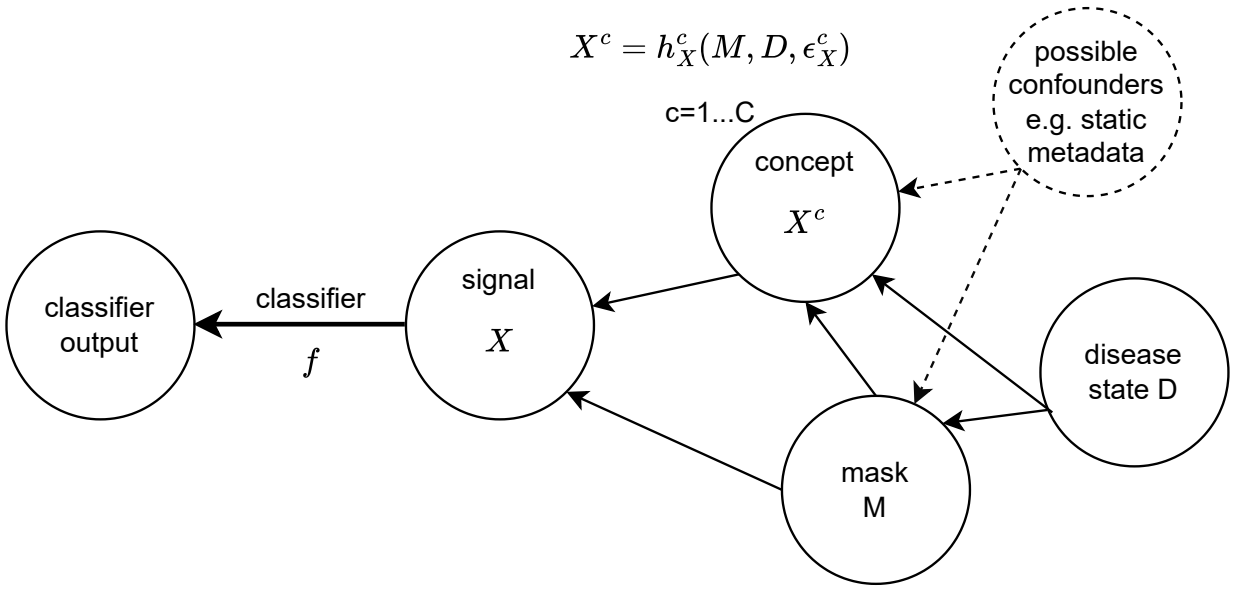


Figure 2: Causal graph underlying our approach. The data generating process is rooted in a disease state D, which causes a (segmentation) mask M. The disease state D in combination with the mask M define the specific numerical values X^c (for concept c), which in combination leads to the input signal X. X is passed through a predefined classifier f. We investigate the causal effect of X^c on the classifier output by intervening on the disease state D. In our experiments, we neglect the causal effect of the disease state M, i.e., keep the segmentation mask M unchanged. Similarly, we do not take into account possible confounders such as static metadata that could influence X^c or M which are expressed as dashed lines.

Generative model architecture Here we elaborate on the specification of the generative model utilized for sampling from either the interventional distribution h_X^c or the conditional distribution k_X^c. This can be read off most explicitly from Eq. 4 and Eq. 5, where we approximate the respective right-hand side by sampling from an imputation model. For our specific implementation, we leverage the recently proposed structured state-space diffusion (SSSD) model for time series imputation (Alcaraz & Strodthoff, 2023a). This model, a diffusion model, extends the popular DiffWave architecture (Kong et al., 2021) by employing two S4 layers instead of bidirectional dilated convolutions, thereby enhancing its capability to capture long-term dependencies. Alongside a modified diffusion procedure wherein noise is applied solely to the input segments to be imputed, this approach yielded state-of-the-art results for time series imputation across various domains. To train a class-conditional diffusion model for a specific class, we simply subsample the training set to include only samples of the desired label, proceeding as in the class-unconditional case.

Generative model details The imputation model employed within *CausalConceptTS* incorporates 36 residual layers and 256 residual and skip channels, while keeping further hyperparameters unchanged compared to (Alcaraz & Strodthoff, 2023a). We optimize the mean squared error (MSE) using the Adam optimizer, with the model undergoing 200 diffusion steps via a linear schedule. We approximate the expectation values in Eq. 4 and Eq. 5 through sampling from an appropriate generative model. The number of considered samples is an important hyperparameter. Our experiments showed convergence after around 15 samples on average due to the generative model's probabilistic nature. Consequently, we maintain generating 40 samples

per real sample to ensure robustness. Training details and additional details on the computational complexity can be found in the supplementary material.

Channel-specific attributions When assessing channel-specific attributions, we do not condition on inputs from other channels captured at the same time as the channel to be imputed, to avoid issues with correlated channels at identical time steps, see also the discussion of interaction effects for associational attributions in (Blücher et al., 2022). Consequently, we consistently utilize an imputer trained in a blackout-missing manner. Subsequently, we substitute channels not intended for imputation with their respective values from the original dataset.

Classifier model architecture Building on recently successful applications in the context of physiological time series (Strodthoff et al., 2024; Wang & Strodthoff, 2023; Saab et al., 2024; Alcaraz & Strodthoff, 2024), we also leverage structured state space models (with four layers) as classifier models (Gu et al., 2022). For optimization, the Adam optimizer is utilized with a learning rate and weight decay both set to 0.001. The learning rate schedule is maintained constant throughout training. A batch size of 64 samples is used for each training iteration, spanning a total of 20 epochs. The training objective is to minimize the binary cross-entropy loss. During training, we apply a model selection strategy on the best performance (AUROC) on the validation set which usually converges before the total epochs. For the test set, we report the 95% confidence intervals obtained through bootstrapping over 1000 iterations. For additional details on the classifier model, we refer to the supplementary material.

Concept discovery and concept validation At first sight, the proposed approach may seem to rely crucially on predefined concepts. However, many time series lack such predefined concepts. While the discovery of concepts and their evaluation lies beyond the scope of this work, it should not be seen as a constraint for this work. Therefore, in the absence of expert-annotated concepts, we identify concepts by k-mean clustering using the raw time series as input and the squared Euclidean distance as distance measure. We determine the number of clusters using the elbow method. To assess if the identified clusters are class-discriminate, we use a simple concept validation step. To this end, we conduct classification using gradient-boosted decision trees (XGBoost), using simple statistical features extracted from the cluster segmentations as input. These employing six sample-wise and channel-wise concept statistics namely, minimum, maximum, mean, standard deviation, median, and number of time steps. Ideally, higher model performance indicates that these concepts effectively distinguish between classes.

Uncertainty quantification in ATEs The fact that we approximate the expectation values for causal/associational effects in Eq. 4 and Eq. 5 through finite samples from a corresponding imputation model allows us to infer not only point estimates of the corresponding effects from the corresponding sample means but also gives us access to the uncertainty estimate at the level of ITEs or IAAs and then correspondingly also at the level of average causal effects. Specifically, we conduct 1,000 bootstrap iterations by sampling with replacement from the test set to compute 95% ATEs prediction intervals. We claim a statistically significant causal effect if the prediction interval inferred in this way does not include the value 0.

4 Experiments

We conduct our experiments using a diverse range of time series classification tasks. Specifically, we present results for three tasks derived from various qualitative time series data sourced from the meteorological and the physiological domain. We present our primary experimental findings through figures, each illustrating either the associational or causal attributions. In these visualizations, we provide two attribution: on the right, we present the 'global' causal effect, encompassing the impact across all channels collectively; on the left, we delineate the channel-specific computation of the treatment effect for each concept. When considering uncertainty quantification, a star symbol indicates statistically significant causal effect in the sense of a 95% prediction interval that does not the value 0. We focus the comparison of associational against causal effects mainly on such significant effects. To visualize the considered concepts, we present an exemplary plot of a time series from the dataset under consideration superimposed with corresponding segmentation maps/concept assignments. To foster more research in this field and enhance usability for applications, we are making the source code used in our investigations available in a suitable repository (Anonymous, 2024).

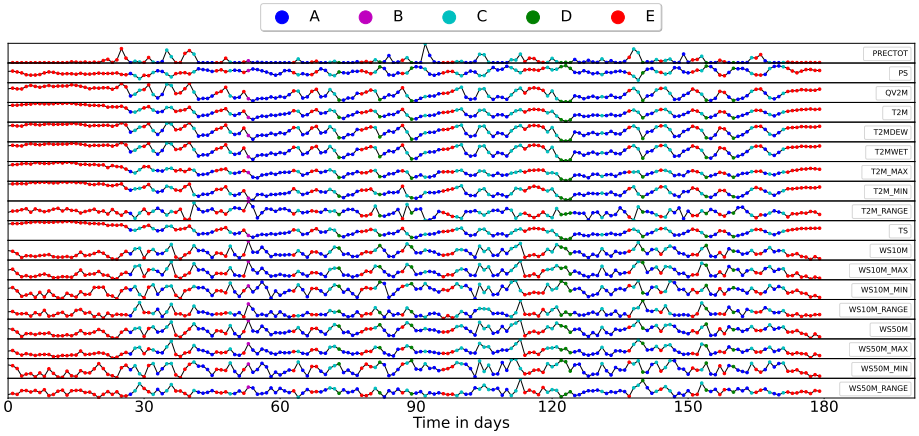


Figure 3: Schematic representation of the concepts for the drought dataset

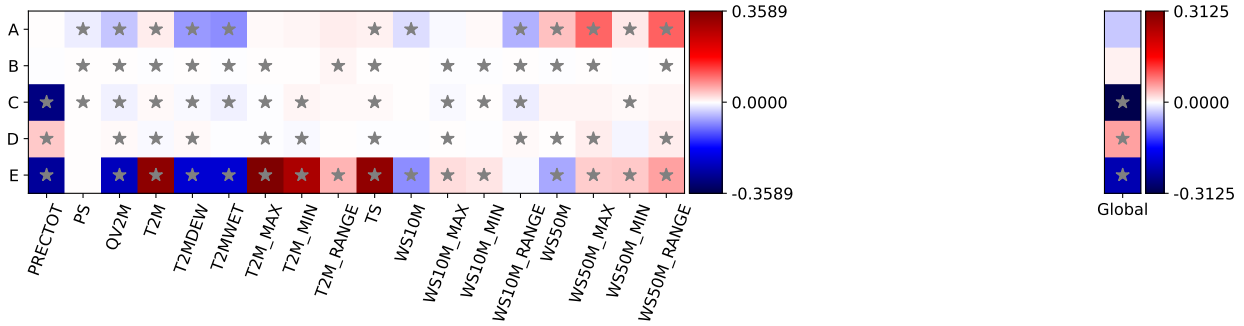
Drought prediction

As first task, we explore the drought dataset (Minixhofer, 2021), sourced from the U.S. Drought Monitor. This publicly available dataset involves classifying, in a binary manner, whether the upcoming week will experience drought conditions based on six months of daily sampled meteorological data. The dataset contains 18 features (Precipitation PRECOT, surface pressure PS, humidity, temperature, Dew/Frost point, wet bulb, as well as minimum and maximum temperature all at 2 meters QV2M, T2M, T2MDEW, T2MWET, T2M_MAX, T2M_MIN, T2M_RANGE. Earth skin temperature TS. Wind speed at 10 and 50 meters with their corresponding maximums, minimums, and ranges respectively WS10M, WS10M_MAX, WS10M_MIN, WS10M_RANGE, WS50M, WS50M_MAX, WS50M_MIN, and WS50M_RANGE). In the absence of expert concepts, we identify five concepts (A-E) through k-means clustering leading to an AUROC 0.7447 (95% PI 0.7406-0.7483) during concept validation. We report a classification performance for the S4 model of 0.8941 (95% PI 0.8919- 0.8962). Figure 3 visualizes concept assignments for a drought sample.

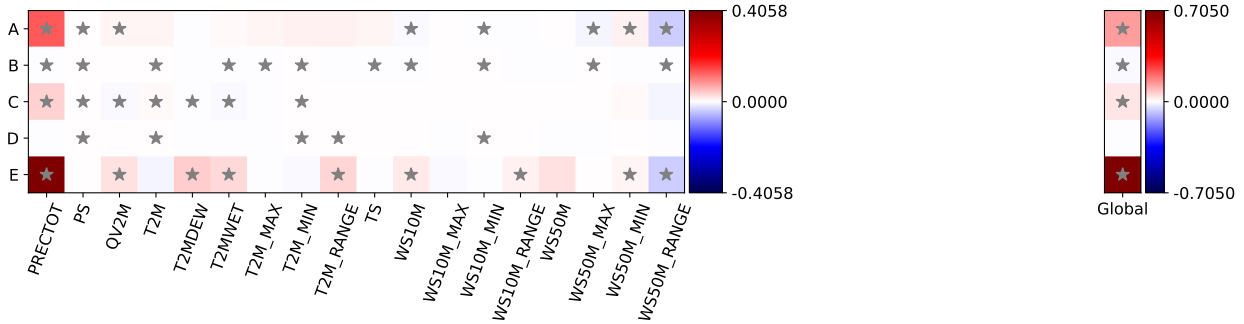
Figure 4 shows (A) associational and (B) causal attributions for the drought prediction task. Interestingly, both channel-wise attribution maps reveal a diverse range of variables with significant effects, yet they sometimes disagree on whether the effects are positive or negative. One notable observation is precipitation, which shows the highest positive effect in the causal setting but appears negative in the associational setting. Extensive research has validated the positive significant impact of precipitation on drought prediction (Cancelliere et al., 2007; Anshuka et al., 2019) which is the largest positive attribute for causal, whereas associational effect is negative across several concepts. Similarly, in concept E, a group of variables at 2 meters have been shown to have positive effects, including humidity and dew/frost point temperatures (Behrangi et al., 2015), as well as wet bulb readings, which causal attributions properly account for them while associational do not. Additionally, for concept A, factors such as the minimum, maximum, and range of wind speed at 50 meters have been shown to have a positive influence (Štěpánek et al., 2018), which again causal unlike associational attributions properly attribute to.

ECG classification

As the second dataset, we leverage the PTB-XL dataset (Wagner et al., 2020; Goldberger et al., 2000), which is a publicly available dataset of clinical 12-lead ECG data (I, II, III, aVR, aVL, aVF, V1-V6). Although PTB-XL provides annotations in terms of diverse hierarchical levels of ECG statements in a multi-label setting, we keep the setup simply by restricting ourselves to investigation of the causal concept effects of inferior myocardial infarction (IMI) in a binary classification setting against healthy controls (NORM+SR). We utilize a sample length of 248 time steps and for the predefined segmentation of the signal into channel-specific ECG segments, we leverage segmentation maps provided by (Wagner et al., 2024). Here, we consider six concepts: P-wave, PQ-segment, QRS complex, ST-segment, T-wave, and TP-segment, which reach an AUROC score of 0.9287 (95% PI 0.913-0.9435) during concept validation. The classifier reaches an AUROC classification



(A) Associational attributions



(B) Causal attributions

Figure 4: Illustration of the (A) associational and (B) causal attributions on the drought dataset

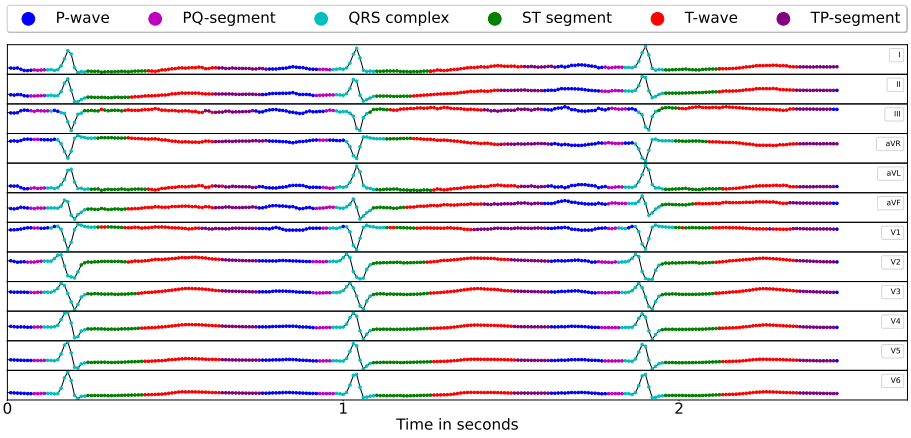
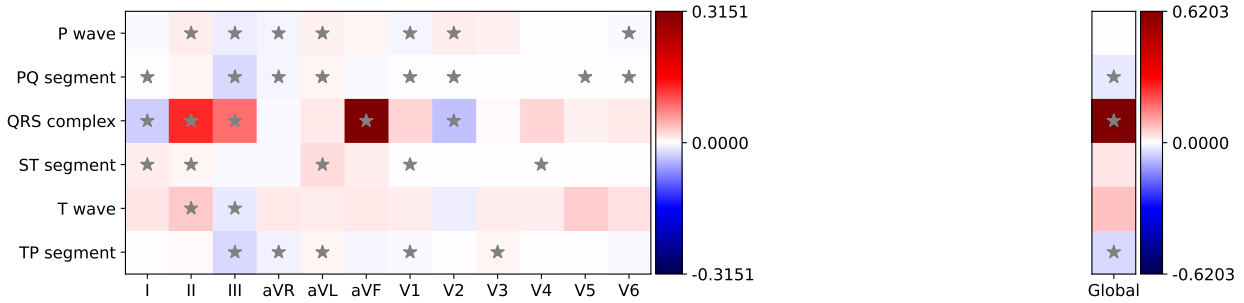


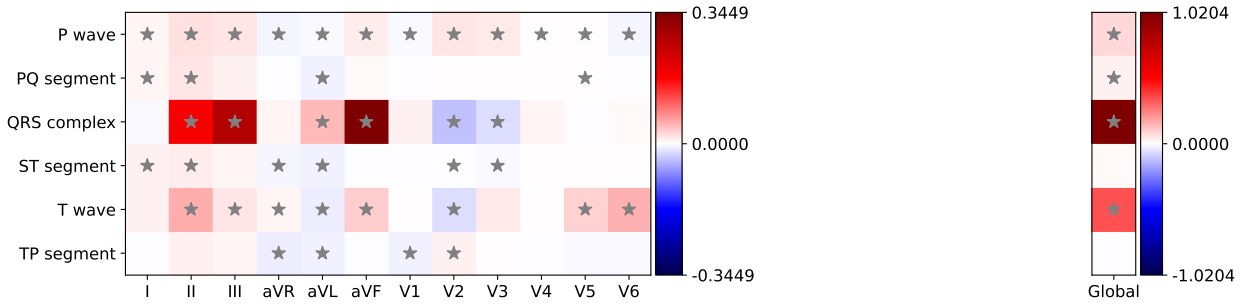
Figure 5: Schematic representation of the concepts for the PTB-XL dataset

performance of 0.9722 (95% PI 0.9621-0.9797). Figure 5 shows a visual representation of these concepts for a myocardial infarction sample.

Figure 6 presents both associational and causal attributions for the ECG classification task. The literature extensively covers this task, allowing us to draw conclusions on the channel level. Both attribution maps appropriately highlight positive effects for the QRS complex in leads II, III, and aVF, which have been linked to pathological longer and deeper Q-waves (Thygesen et al., 2018). In the associational attribution map, a negative significant effect is observed in the T-wave for lead III, while the causal attribution indicates a positive significant effect. Literature works align in this case rather with the causal attribution in the sense



(A) Associational attributions



(B) Causal attributions

Figure 6: Illustration of the (A) associational and (B) causal attributions on the PTB-XL dataset

that high T-waves exhibit a positive pattern (Dressler & Hugo, 1947). Similarly, literature results suggest a positive effect for the P-wave in leads I, II, and III (Grossman & Delman, 1969), which are recognized as significant and positive effects from causal attributions, while associational attributions only show significant positive effects in II and a negative effect in III.

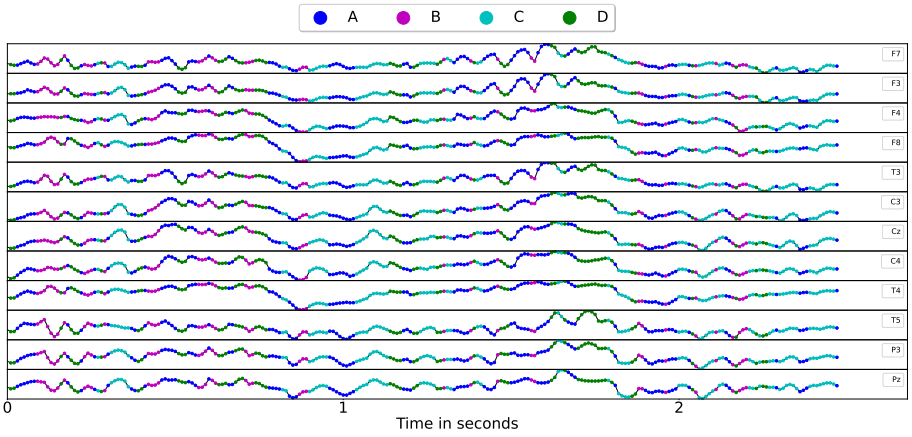


Figure 7: Schematic representation of the concepts for the schizophrenia dataset

EEG classification

As the third dataset, we analyze the schizophrenia dataset (Borisov et al., 2005), which includes EEG signals from a study involving paranoid schizophrenia patients and healthy controls. This dataset comprises 16 EEG channels (F7, F3, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, O2), with each channel spanning 248 time steps. Further details on the dataset and preprocessing are available in the supplementary material.

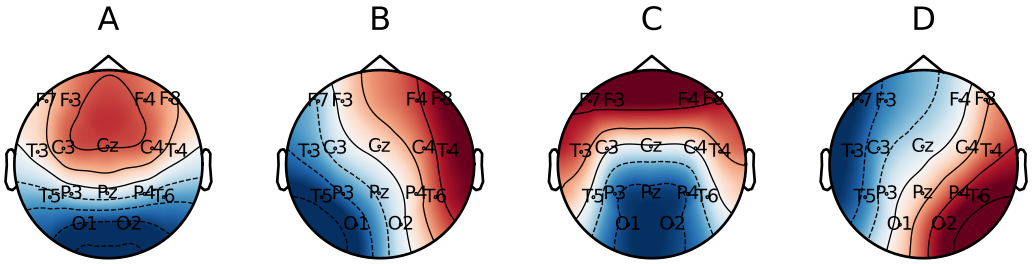
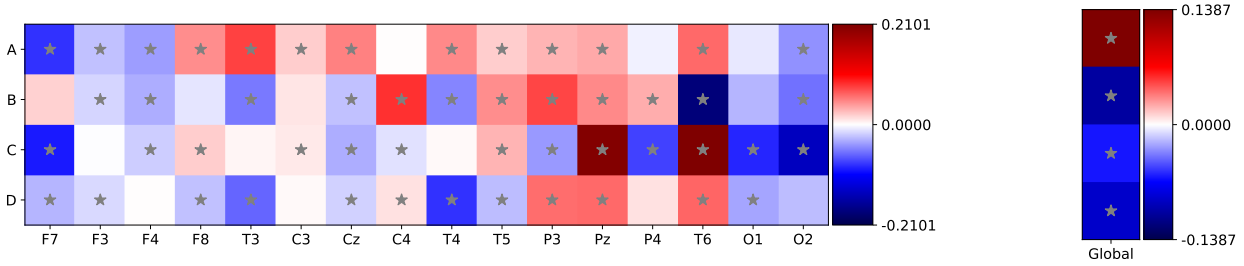
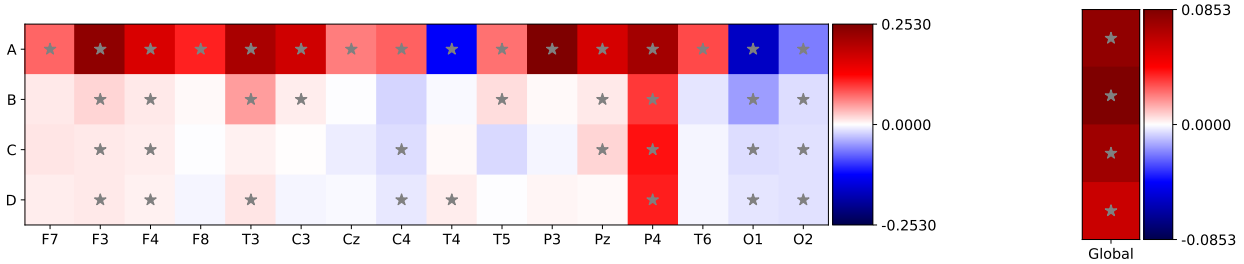


Figure 8: Spatial distribution of brain activity patterns during different states of brain processing. Dark red indicates increased activity, while dark blue signifies decreased activity.

To extract meaningful concepts, we employ an EEG microstates segmentation (Pascual-Marqui et al., 1995) through open-source software (von Wegner, 2017; Gramfort et al., 2013). These microstates capture transient brain states reflecting underlying neural dynamics, often linked to specific cognitive processes. Our analysis identifies four distinct concepts (A-D) leading to a concept validation score (AUROC) of 0.8249 (95% PI 0.7682-0.8793). As a supporting illustration to compare our findings with the literature, we present in Figure 8 a topographic map illustrating the overall brain activity during each investigated EEG microstate. We report an AUROC classification performance for the S4 model of 0.9671 (95% PI 0.9432-0.9849). Figure 7 shows an exemplary visualization of the concepts for a schizophrenia sample.



(A) Associational attributions



(B) Causal attributions

Figure 9: Illustration of the (A) associational and (B) causal attributions on the schizophrenia dataset

Figure 9 presents the associational and causal attributions for the EEG classification task. Several studies in the literature have identified specific patterns associated with schizophrenia. From a global perspective, B exhibits statistically significant differences between patients and controls in numerous studies, considering both duration (Kikuchi et al., 2007; Koenig et al., 1999; Nishida et al., 2013) and occurrence (Koenig et al., 1999; Nishida et al., 2013). Moreover, other studies have highlighted the importance of A and C based on features such as occurrence, coverage, and duration (Keihani et al., 2022), as well as D due to increased mean duration (Sun et al., 2021). Thus, while associational attributions do not adequately cover all expert knowledge attributions globally, causal attributions do. From a channel-wise perspective to the best of our

knowledge, we are the first work to investigate any effect of single leads microstates for schizophrenia detection using EEG. In the two previous datasets, the concepts typically exhibit a consistent pattern across channels, however, here the associational plot appears to show random behavior.

5 Discussion and conclusion

Limitations At this stage, *CausalConceptTS* faces several limitations, which we briefly discuss in the following. First, our method does not account for intervening on the segmentation mask M but relies on a predefined mask from the original sample. This could pose issues, especially for pathologies, like the left bundle branch block in the ECG case, which is characterized by a wide QRS complex, i.e., altering the segmentation mask significantly. To mitigate this, one could consider combinations of adjacent segments instead of individual segments. Second, the generative model for imputation is trained solely on real samples, assuming it generalizes well to unseen classes when conditioned on segments from other classes. Third, intervening on specific segments with a different disease inevitably requires evaluating the model slightly outside its model scope, blending characteristics of the original disease and the intervened state. Fourth, the proposed approach only focuses on the causal effect of the considered concepts but does not incorporate other possible confounding factors, such as static patient metadata like demographic data, see the discussion below Eq.5. Their impact could be investigated by explicitly conditioning the generative model on these variables.Finally, an extensive analysis of channel correlations, which is closely related to the question of interaction effects (Blücher et al., 2022), both from an associational as well as from a causal point of view, is beyond the scope of this work but represents a pressing direction for future research.

Use-cases for XAI and broader impact statement As described in (Wagner et al., 2024), one has to distinguish different use-cases for XAI such as providing side-information for end-users, model auditing, and knowledge discovery. While the first use-case can only be assessed with extensive user studies, the two latter use-cases rely on the fact that the used attribution method faithfully captures the model behavior. In this case basing auditing decisions or claiming discoveries on potentially misleading associtational attributions represents a danger. In any case, it is worth acknowledging the difference between causal and associational attributions and taking this aspect into account when assessing the suitability of particular attribution methods in particular in safty-critical application domains.

Conclusion The paper proposes a framework to assess the causal effect of label/disease-specific manifestation of predefined segments of a time series on a given fixed time series classifier. Its key component is a high-fidelity diffusion model, which is used to infer counterfactual manifestations of segments under consideration. This allows us to compute individual and average treatment effects. Furthermore, we demonstrate that the main difference between such causal attributions and purely associational, perturbation-based attributions lies in the use of a class-conditional as opposed to an unconditional imputation model. These insights allow for a direct comparison of causal and associational attributions. The differences between causal and associational attributions hint at the danger of drawing misleading conclusions from associational attributions. We showcase our approach for a diverse set of three time series classification tasks and find a good alignment of the identified causal effects with expert knowledge.

References

Juan Lopez Alcaraz and Nils Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models. *Transactions on Machine Learning Research*, 2023a. ISSN 2835-8856.

Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based conditional ECG generation with structured state space models. *Computers in Biology and Medicine*, pp. 107115, June 2023b. doi: 10.1016/j.compbiomed.2023.107115.

Juan Miguel Lopez Alcaraz and Nils Strodthoff. Mds-ed: Multimodal decision support in the emergency department–a benchmark dataset for diagnoses and deterioration prediction in emergency medicine. *arXiv preprint arXiv:2407.17856*, 2024.

Anonymous. Code repository CausalConceptTS. `https://anonymous.4open.science/r/CausalConceptTS-8055/README.md`, 2024. [Accessed 22-05-2024].

Anshuka Anshuka, Floris F van Ogtrop, and R Willem Vervoort. Drought forecasting through statistical models using standardised precipitation index: a systematic review and meta-regression analysis. *Natural Hazards*, 97:955–977, 2019.

Emre Ates, Burak Aksar, Vitus J Leung, and Ayse K Coskun. Counterfactual explanations for multivariate time series. In *2021 international conference on applied artificial intelligence (ICAPAI)*, pp. 1–8. IEEE, 2021.

Md Abul Bashar and Richi Nayak. Tanogan: Time series anomaly detection with generative adversarial networks. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1778–1785. IEEE, 2020.

Ali Behrangi, Paul C Loikith, Eric J Fetzer, Hai M Nguyen, and Stephanie L Granger. Utilizing humidity and temperature data to advance monitoring and prediction of meteorological drought. *Climate*, 3(4):999–1017, 2015.

Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell systems*, 12(6):654–669, 2021.

Stefan Blücher, Johanna Vielhaben, and Nils Strodthoff. PredDiff: Explanations and interactions from conditional expectations. *Artificial Intelligence*, 312:103774, 2022. doi: 10.1016/j.artint.2022.103774.

SV Borisov, A Ya Kaplan, NL Gorbachevskaya, and IA Kozlova. Analysis of eeg structural synchrony in adolescents with schizophrenic disorders. *Human Physiology*, 31:255–261, 2005.

Antonino Cancelliere, G Di Mauro, Brunella Bonaccorso, and G Rossi. Drought forecasting using the standardized precipitation index. *Water resources management*, 21:801–819, 2007.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

Jonathan Crabbé and Mihaela Van Der Schaar. Explaining time series predictions with dynamic masks. In *International Conference on Machine Learning*, pp. 2166–2177. PMLR, 2021.

Eoin Delaney, Derek Greene, and Mark T Keane. Instance-based counterfactual explanations for time series classification. In *International conference on case-based reasoning*, pp. 32–47. Springer, 2021.

Houtao Deng, George Runger, Eugene Tuv, and Martyanov Vladimir. A time series forest for classification and feature extraction. *Information Sciences*, 239:142–153, 2013.

William Dressler and Roesler Hugo. High t waves in the earliest stage of myocardial infarction. *American heart journal*, 34(5):627–645, 1947.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

Ben D Fulcher and Nick S Jones. hctsa: A computational framework for automated time-series phenotyping using massive feature extraction. *Cell systems*, 5(5):527–531, 2017.

D. Gillies. *Causality, Probability, and Medicine*. Taylor & Francis, 2018. ISBN 9781317564287.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215.

Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint 1907.07165*, 2019.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13, 2013. doi: 10.3389/fnins.2013.00267.

James I Grossman and Abner J Delman. Serial p wave changes in acute myocardial infarction. *American Heart Journal*, 77(3):336–341, 1969.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Jon Hills, Jason Lines, Edgaras Baranauskas, James Mapp, and Anthony Bagnall. Classification of time series by shapelet transformation. *Data mining and knowledge discovery*, 28:851–881, 2014.

Murtadha D Hssayeni. Imbalanced time-series data regression using conditional generative adversarial networks. In *International Conference on Machine Learning and Applications*, 2022.

Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. *Advances in neural information processing systems*, 33: 6441–6452, 2020.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.

Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Shun Chen. Lstm fully convolutional networks for time series classification. *IEEE access*, 6:1662–1669, 2017.

Ahmadreza Keihani, Seyed Saman Sajadi, Mahsa Hasani, and Fabio Ferrarelli. Bayesian optimization of machine learning classification of resting-state eeg microstates in schizophrenia: A proof-of-concept preliminary study based on secondary analysis. *Brain Sciences*, 12(11):1497, Nov 2022. ISSN 2076-3425. doi: 10.3390/brainsci12111497.

Mitsuru Kikuchi, Thomas Koenig, Yuji Wada, Masato Higashima, Yoshifumi Koshino, Werner Strik, and Thomas Dierks. Native eeg and treatment effects in neuroleptic-naïve schizophrenic patients: Time and frequency domain approaches. *Schizophrenia Research*, 97(1):163–172, 2007. ISSN 0920-9964. doi: https://doi.org/10.1016/j.schres.2007.07.012.

Thomas Koenig, Dietrich Lehmann, Marco CG Merlo, Kieko Kochi, Daniel Hell, and Martha Koukkou. A deviant eeg brain microstate in acute, neuroleptic-naive schizophrenics at rest. *European archives of psychiatry and clinical neuroscience*, 249:205–211, 1999.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *9th International Conference on Learning Representations, ICLR 2021*, 2021.

Peiyu Li, Soukaïna Filali Boubrahimi, and Shah Muhammad Hamdi. Motif-guided time series counterfactual explanations. In *International Conference on Pattern Recognition*, pp. 203–215. Springer, 2022.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

Christoph Minixhofer. Predict droughts using weather & soil data, Mar 2021. URL `https://www.kaggle.com/datasets/cdminix/us-drought-meteorological-data`.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

Keiichiro Nishida, Yosuke Morishima, Masafumi Yoshimura, Toshiaki Isotani, Satoshi Irisawa, Kay Jann, Thomas Dierks, Werner Strik, Toshihiko Kinoshita, and Thomas Koenig. Eeg microstates associated with salience and frontoparietal networks in frontotemporal dementia, schizophrenia and alzheimer's disease. *Clinical Neurophysiology*, 124(6):1106–1114, 2013. ISSN 1388-2457. doi: https://doi.org/10.1016/j.clinph.2013.01.005.

Roberto D Pascual-Marqui, Christoph M Michel, and Dietrich Lehmann. Segmentation of brain electrical activity into microstates: model estimation and validation. *IEEE Transactions on Biomedical Engineering*, 42(7):658–665, 1995.

J. Pearl. *Causality*. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2009. ISBN 9780521895606.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

Thanawin Rakthanmanon and Eamonn Keogh. Fast shapelets: A scalable algorithm for discovering time series shapelets. In *proceedings of the 2013 SIAM International Conference on Data Mining*, pp. 668–676. SIAM, 2013.

Vikas C. Raykar, Arindam Jati, Sumanta Mukherjee, Nupur Aggarwal, Kanthi Sarpatwar, Giridhar Ganapavarapu, and Roman Vaculin. Tsshap: Robust model agnostic feature-based explainability for time series forecasting, 2023.

Thomas Rojat, Raphaël Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Díaz-Rodríguez. Explainable artificial intelligence (xai) on timeseries data: A survey. *arXiv preprint arXiv:2104.00950*, 2021.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019.

Marc Rußwurm and Marco Körner. Self-attention for raw optical satellite time series classification. *ISPRS journal of photogrammetry and remote sensing*, 169:421–435, 2020.

Khaled Saab, Siyi Tang, Mohamed Taha, Christopher Lee-Messer, Christopher Ré, and Daniel L Rubin. Towards trustworthy seizure onset detection using workflow notes. *npj Digital Medicine*, 7(1):42, 2024.

Patrick Schäfer. The boss is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery*, 29:1505–1530, 2015.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 459–466, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085. PMLR, 06–11 Aug 2017.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

Sulaiman Somani, Adam J Russak, Felix Richter, Shan Zhao, Akhil Vaid, Fayzan Chaudhry, Jessica K De Freitas, Nidhi Naik, Riccardo Miotto, Girish N Nadkarni, Jagat Narula, Edgar Argulian, and Benjamin S Glicksberg. Deep learning and the electrocardiogram: review of the current state-of-the-art. *EP Europace*, 23(8):1179–1191, February 2021. doi: 10.1093/europace/euaa377.

Petr Štěpánek, Miroslav Trnka, Filip Chuchma, Pavel Zahradníček, Petr Skalák, Aleš Farda, Rostislav Fiala, Petr Hlavinka, Jan Balek, Daniela Semerádová, et al. Drought prediction system for central europe and its validation. *Geosciences*, 8(4):104, 2018.

Nils Strodthoff, Juan Miguel Lopez Alcaraz, and Wilhelm Haverkamp. Prospects for AI-Enhanced ECG as a Unified Screening Tool for Cardiac and Non-Cardiac Conditions – An Explorative Study in Emergency Care. *European Heart Journal - Digital Health*, pp. ztae039, 05 2024. ISSN 2634-3916. doi: 10.1093/ehjdh/ztae039.

Qiaoling Sun, Jiansong Zhou, Huijuan Guo, Ningzhi Gou, Ruoheng Lin, Ying Huang, Weilong Guo, and Xiaoping Wang. Eeg microstates and its relationship with clinical symptoms in patients with schizophrenia. *Frontiers in Psychiatry*, 12, 2021. ISSN 1664-0640. doi: 10.3389/fpsyt.2021.761203.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems*, 34: 24804–24816, 2021.

Kristian Thygesen, Joseph S Alpert, Allan S Jaffe, Bernard R Chaitman, Jeroen J Bax, David A Morrow, Harvey D White, and Executive Group on behalf of the Joint European Society of Cardiology (ESC)/American College of Cardiology (ACC)/American Heart Association (AHA)/World Heart Federation (WHF) Task Force for the Universal Definition of Myocardial Infarction. Fourth universal definition of myocardial infarction (2018). *Circulation*, 138(20):e618–e651, 2018.

Frederic von Wegner. GitHub - Frederic-vW/eeg_microstates: EEG microstate analysis — github.com, 2017. URL `https://github.com/Frederic-vW/eeg_microstates/tree/master`. [Accessed 28-04-2024].

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7 (1):154, 2020. doi: 10.1038/s41597-020-0495-6.

Patrick Wagner, Temesgen Mehari, Wilhelm Haverkamp, and Nils Strodthoff. Explaining deep learning for ecg analysis: Building blocks for auditing and knowledge discovery. *Computers in Biology and Medicine*, 176:108525, June 2024. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2024.108525.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, August 2023. doi: 10.1038/s41586-023-06221-2.

Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3–11, mar 2019. doi: 10.1016/j.patrec.2018.02.010.

Tiezhi Wang and Nils Strodthoff. S4sleep: Elucidating the design space of deep-learning-based sleep stage classification models, 2023.

Zhendong Wang, Isak Samsten, Rami Mochaourab, and Panagiotis Papapetrou. Learning time series counterfactuals via latent space representations. In *Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings 24*, pp. 369–384. Springer, 2021.

Ziqi Zhao, Yucheng Shi, Shushan Wu, Fan Yang, Wenzhan Song, and Ninghao Liu. Interpretation of time-series deep models: A survey. *arXiv preprint arXiv:2305.14582*, 2023.