

Explaining Time Series Classifications Counterfactually: A Human-Centered Evaluation

Eoin Delaney¹, Greta Warren², Derek Greene³, Mark T. Keane³

¹Trinity College Dublin

²University of Copenhagen

³University College Dublin
eoin.delaney@tcd.ie

Abstract

Counterfactual explanations can promote accountability, fairness, and transparency in AI-supported decision-making by demonstrating how changes in input data can alter outcomes, offering actionable feedback for users to revise unfavorable decisions. However, the majority of counterfactual algorithms focus on tabular and image data, ignoring many high-stakes scenarios that rely on time series data. Furthermore, user studies of counterfactuals for time series are exceedingly rare. We address this research gap by proposing a novel methodology that advances carefully-controlled tests of counterfactual explanations for time series classification. In two experiments, participants were asked to classify time series plots and provide confidence ratings, before receiving feedback on the correct classification along with an explanation. The studies manipulated the explanation strategy used and the visualization used. Results indicate that counterfactual explanations elicit improved domain understanding and higher confidence in people in comparison to factual explanations. Wider implications for explanation in time series domains are discussed.

Data — <https://github.com/e-delaney/ts-user-study>

1 Introduction

In recent years, there has been an explosion of research in eXplainable AI (XAI) designed to address the opacity of current AI systems and to establish appropriate reliance in predictions. Consequently, there is now a panoply of XAI strategies from ante-hoc transparent systems (Rudin 2019; Chen et al. 2019) to post-hoc explanation methods such as saliency maps (Selvaraju et al. 2017; Zhou et al. 2016), feature-importance analyses (e.g., SHAP (Lundberg and Lee 2017) and LIME (Ribeiro, Singh, and Guestrin 2016)), and instance-based methods deploying examples, prototypes and counterfactuals as explanations (Wachter, Mittelstadt, and Russell 2017; Kenny and Keane 2019; Kim et al. 2016). Although these efforts often aim to enhance users’ understanding of AI systems, most of this research concentrates on algorithmic development rather than assessing what users actually understand (Adadi and Berrada 2018; Lipton 2017; Anjomshoae et al. 2019; Pičulin et al. 2025). This has been countered by *human-centric approaches* to XAI, which

place a greater emphasis on designing explanations for human explanation goals, stakeholder perspectives and the broader contexts in which these algorithmic explanations are delivered (Wang et al. 2019; Liao, Gruen, and Miller 2020; Langer et al. 2021; Tahaei et al. 2023; Lai et al. 2023; Ehsan et al. 2024; Schuff et al. 2022). In XAI, post-hoc counterfactual explanations can offer actionable “algorithmic recourse” (Verma, Dickerson, and Hines 2020). For instance, a rejected loan applicant might be told: “If you had requested \$2k less, you would have been approved” (McGrath et al. 2018). Beyond providing a clear path to a different outcome, these explanations have been proposed as psychologically intuitive (Miller 2019; Byrne 2019) and in alignment with legal transparency requirements (Wachter, Mittelstadt, and Russell 2017). Although upwards of 150+ distinct counterfactual algorithms exist in the literature, few studies evaluate the extent to which people actually understand them as claimed (one review found only < 7% of papers performed any sort of user testing (Keane et al. 2021)). In this paper, we report user studies of counterfactual explanations for time series data, an area with even poorer user-study coverage.

The Promise of Time Series XAI for Transparency

A majority of XAI research, including counterfactual XAI, has primarily focused on tabular data (Wachter, Mittelstadt, and Russell 2017; Gomez et al. 2020; Mothilal, Sharma, and Tan 2020), with growing work on computer vision and image data (Goyal et al. 2019; Akula, Wang, and Zhu 2020; Kenny and Keane 2020) and text data (Mayne et al. 2025). Counterfactual explanations of time series data have only been addressed more recently (Delaney, Greene, and Keane 2021a; Ates et al. 2021; Guidotti 2024) (for surveys see (Rojat et al. 2021; Šimić, Sabol, and Veas 2021; Theissler et al. 2022; DeMartino 2020)). This underdevelopment is somewhat surprising considering the ubiquity of time series in many rich use-cases across diverse domains where transparency is critical, including medicine, environmental sensing, finance, and manufacturing (Dau et al. 2019; Fawaz et al. 2019a; Gee et al. 2019; Schulam and Saria 2017; Ghassemi et al. 2020; Lucic, Haned, and de Rijke 2020b). For instance, XAI for time series classification can reveal complex morphological and physical features that clinicians may not readily spot, aiding in the detection of apnea (cessation of breathing) and bradycardia (slowing of heart rate) in pre-

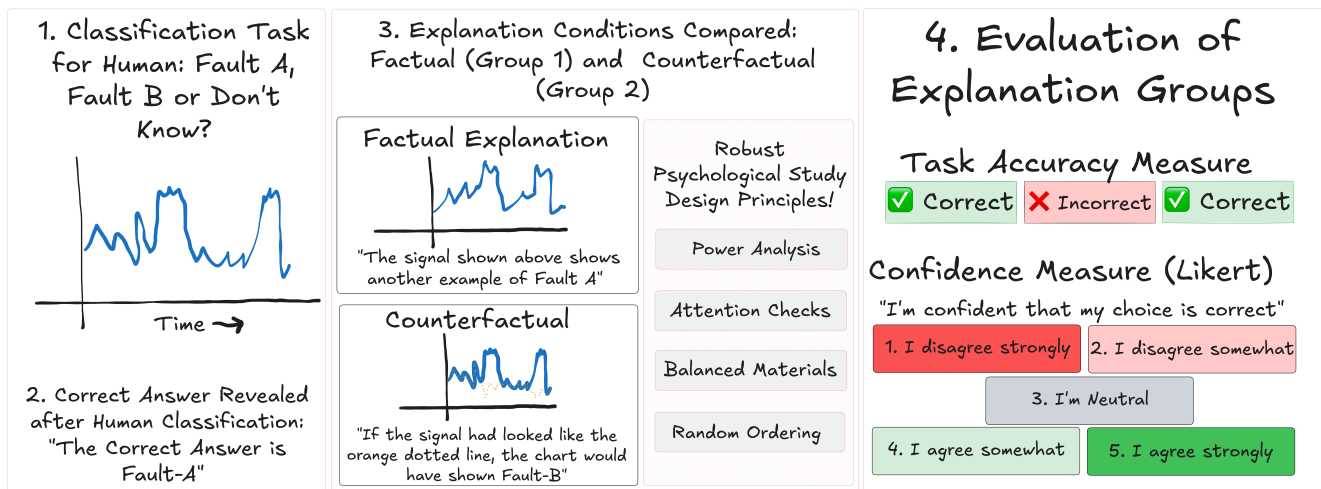


Figure 1: Sketch outline of the study design. Participants are randomly assigned to a group and given a time series classification task. Upon submitting an answer, they are provided with the correct answer and feedback through either (i) a factual explanation or (ii) a counterfactual explanation. The process is repeated for 40 classifications. Task performance and human confidence in their own predictions are evaluated. Robust psychological study design principles are followed including attention checks, randomization of materials and power analysis to determine the minimum number of participants required to find a meaningful statistical effect.

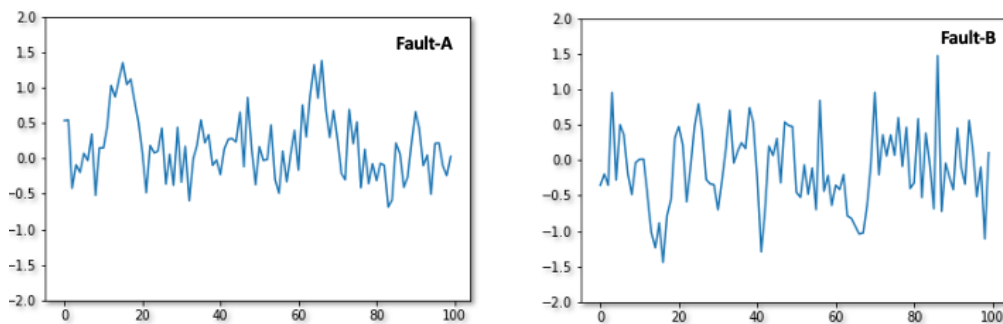


Figure 2: Time Series Examples From Current Human Studies with Non-Experts: Charts of sensor readings plotting two engine faults (i.e., Fault-A and Fault-B). The positive sine-like behavior between time points 10–20 and 60–70 form a noisy $\cap$ shape indicating Fault-A and the negative sine-like behavior between the same time points form a noisy $\cup$ shape indicating Fault-B.

term infants (Gee et al. 2019). In manufacturing, time series of sensor readings can detect critical machine failures (Cook, Mısırlı, and Fan 2020) (e.g., akin to those shown in Figure 2). As AI decision-support systems are increasingly deployed in industry, there is an increasing need to explain decisions and to provide actionable feedback to both expert and non-expert users. Ideally, a good explanation method should help users learn more about the domain, the decisions that are being made, and the AI system that makes those decisions (Doshi-Velez and Kim 2017). This paper specifically focuses on explanation methods for the task of *time series classification* (TSC) which involves mapping an unlabelled time series – a test instance – to a predefined class label (e.g., Fault-A or Fault-B).

Challenges: Developing Controlled & Human-Centered Evaluation Methodologies

Time series domains pose distinct challenges for both designing and evaluating human-centered explanation methods due to their high dimensionality and the ambiguity of what constitutes a meaningful feature. For example, non-expert users may struggle to interpret time series data, despite the fact that they encounter it regularly (e.g., polling data or energy consumption) (Zhang and Lim 2022). Furthermore, while models may extract latent features, these are often uninterpretable to humans, and the notion of a “feature” is less well-defined than in tabular or image domains (Delaney, Greene, and Keane 2021a). We therefore argue for controlled evaluation strategies that begin with simplified datasets with well-defined properties, enabling transparent, human-centered benchmarks to guide future work. With regard to evaluation, a frequently overlooked step in XAI user

studies is the fair selection of test materials. This requires balanced, representative, and unbiased instance sets, yet many studies rely on hand-picked or unbalanced examples that can favor particular explanation methods. Such practices risk introducing spurious correlations, unequal noise, or uneven condition sizes, undermining the validity of results. Additional methodological issues persist. Many XAI user studies are statistically underpowered due to insufficient participant numbers, highlighting the need for a-priori power analyses (Erdfelder, Faul, and Buchner 1996). Attention checks, essential for data quality control (Oppenheimer, Meyvis, and Davidenko 2009), are often omitted, as is randomization of material order to mitigate order effects.

Next, we describe the time series test items used in this work. Figure 2 shows two instances from a binary fault-classification dataset (Fault-A and Fault-B), each defined by objectively specified properties: positive and negative sine-wave motifs. These patterns are representative of real-world time series, such as electrocardiograms, where elevated sine-wave activity can indicate conditions like severe hyperkalemia (Dau et al. 2019; Pluijmen and Hersbach 2007; Loubser et al. 2023). Each instance consists of a 100-point time series with Gaussian noise, into which discriminative features were embedded at time points 10–20 and 60–70. Fault-A contains a positive sine-wave motif, while Fault-B contains a negative one. This controlled design allows us to assess whether non-expert human users can learn to discriminate time series signals, and whether explanations support or hinder this learning.

Evaluating Explanation Strategies & Visualisation Methods for Time Series Data

Using our proposed methodology for tackling time series data, we examined two key variables, (i) explanation strategies (factual versus counterfactual explanations), and (ii) visualisation methods (overlaid versus non-overlaid explanations) in two separate experiments. Our goal was to establish human-centered benchmarks for explaining time series, to determine if reproducible effects could be found for classes of instances with objectively-defined properties. If such benchmarks can be found, then we have a framework for progressively exploring more complex scenarios (e.g., adding more features or more diverse features).

Explanation Strategy. The main variable compared two different post-hoc, instance-based explanation strategies: a factual example-based approach versus a counterfactual example-based one. Factual-example explanations for time series classifications inform the recipient that “this instance was classified as X because a very similar instance was previously classified as X”; more technically, the current test instance is explained by this nearest neighbour (e.g., (Cunningham and Delany 2021; Caruana et al. 1999)). Counterfactual explanations for time series classifications express that “this instance was classified as X because it would have to look like this very different instance, to be classified as Y”; more technically, the current test instance is explained by this *nearest unlike neighbour* or some generated contrasting instance (e.g., (Nugent, Doyle, and Cunningham 2009; Ates et al. 2021; Guidotti 2024)). In the current studies, we

determine which explanation strategy – factual or counterfactual – most impacts people’s understanding of time series classification.

Visualisation Strategy. The second variable of interest compared different visualisations of the presented explanations. Time series explanations have been displayed as separate plots to be compared (Ghods and Cook 2022; Hsu, Liu, and Tseng 2019), represented in tabular form (Fahse, Blohm, and van Giffen 2022; Saluja et al. 2021), described in text (Fahse, Blohm, and van Giffen 2022; Ming et al. 2019; Munir et al. 2019), visualised as heat-mapped saliency plots (Serramazza et al. 2023; Turbé et al. 2023; Wang, Yan, and Oates 2017) or highlighted regions (Schlegel et al. 2019; Saluja et al. 2021; Guillemé et al. 2019; Strodthoff and Strodthoff 2019; Guidotti et al. 2020), and as overlaid plots (Ming et al. 2019; Karlsson et al. 2018; Maweu et al. 2021; Cho et al. 2020; Guidotti et al. 2020; Han et al. 2022; Van Looveren et al. 2021; Filali Boubrahimi and Hamdi 2022). Many XAI time series papers adopt the latter visualisation (though few have tested its efficacy). In Experiment 1 we compare both explanation strategies (factual and counterfactual) using overlaid visualisations, whereas in Experiment 2, the factual explanation is presented separately (not overlaid) to test whether presentational choices influence how people learn from explanations in time series classification.

Outline of Paper & Contributions

In Section 2, we review the related computational and psychological work on explaining time series classifications. Then, in Sections 3 and 4, we implement our evaluation methodology in two user studies, before concluding with a reflection on the results and future implications (Section 5). In summary, this paper makes three key contributions:

- **Evaluation method:** A novel evaluation methodology for testing XAI in time series classification using objectively-defined features in a controlled setting, adequate statistical power, and attention checks to enable the identification of controlled and psychologically robust effects.
- **Explanation tests:** The first user studies comparing factual and counterfactual explanations for time series classification with results indicating that counterfactual explanations elicit faster and more accurate learning.
- **Visualisation tests:** The first user tests exploring different explanation presentation-modes for time series classification.

2 Related Work: Methods & User Tests of Time Series XAI

Many computational methods have been proposed to explain time series classification. Much of this work has centered on visual explanations using saliency mapping (Wang, Yan, and Oates 2017; Zhou et al. 2016; Selvaraju et al. 2017) and feature importance vectors (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017; Villani, Lockhart, and Magazzini 2022). However, more recent work has focused on factual example-based methods (Fahse, Blohm, and van Giffen

2022) and counterfactual methods (Ming et al. 2019; Karlsson et al. 2018; Maweu et al. 2021; Cho et al. 2020; Guidotti et al. 2020; Han et al. 2022; Van Looveren et al. 2021; Filali Boubrahimi and Hamdi 2022)¹. As is the case in XAI more broadly, there has been a significant lag in human-centered evaluations of these methods (Adadi and Berrada 2018). For instance, a review by Theissler et al. (2022) reports only a handful of papers with user evaluations, of the 70+ algorithmic time series XAI papers that they survey. In the following subsections, we review the relevant computational and psychological literature covering: (i) methods using saliency maps, feature importance & shapelets, (ii) methods using factual and counterfactual explanations, (iii) user evaluations of factual explanations, counterfactual explanations and user studies comparing factual to counterfactual explanations.

Feature Importance in Time Series

Saliency maps and Class Activation Maps (CAMs) highlight areas of a time series that are most informative for making classification decisions (Adebayo et al. 2018; Zhou et al. 2016; Fawaz et al. 2019b; Zhou et al. 2016). Schlegel et al. (2019) conducted a computational comparison experiment for XAI methods in time series. Interestingly, LIME yielded poor results across all evaluated datasets, saliency maps were found to be most informative when paired with RNNs, and SHAP’s results were the most robust for all models, though its computational complexity is a drawback (see also (Turbé et al. 2023)). A closely-related technique to saliency maps involves highlighting discriminative sub-series known as shapelets (Ye and Keogh 2011) as visual explanations for classifications. One of the advantages of shapelets is that they can capture relationships between features, unlike a simple feature relevance vector (Grabocka et al. 2014; Schäfer 2016; Karlsson et al. 2018).

Factual & Counterfactual Explanations in Time Series Classification

Time series can also be explained using both factual example-based explanations and counterfactual explanations. *Factual explanations* draw on long-standing work in case-based reasoning (CBR) using k -NNs (Duda, Hart et al. 1973; Cover and Hart 1967; Cunningham and Delany 2021), where a factual example or case is used to explain a prediction (Caruana et al. 1999; Sørmo, Cassens, and Aamodt 2005; Leake 1996; Keane and Kenny 2019).

Counterfactual explanations for time series have emerged, following algorithmic advances in other data modalities. Wachter et al.’s (Wachter, Mittelstadt, and Russell 2017) seminal *Min-Edit* method introduced the major innovation of generating *synthetic* counterfactual-instances to provide much better explanatory coverage. However, such perturbation based techniques can risk generating counterfactuals that are out-of-distribution and not perceptibly different from the query instance (Delaney,

Greene, and Keane 2021b; Russell 2019; Laugel et al. 2019; Guidotti 2024). As such, many counterfactual techniques have supplemented proximity constraints with plausibility and sparsity constraints. Delaney, Greene, and Keane (2021a) introduced the first model-agnostic method for generating counterfactual explanations for time series classification. Their instance-based approach identifies discriminative subsequences (e.g., via Class Activation Mapping or SHAP) and replaces them with subsequences from counterfactual-class instances. The method outperforms benchmark approaches such as Min-Edit (Wachter, Mittelstadt, and Russell 2017) and largely avoids out-of-distribution issues. Related work in multivariate time series classification similarly constructs counterfactuals from training instances (Ates et al. 2021). Collectively, this literature shows that meaningful counterfactuals modify contiguous subsequences rather than entire time series (Delaney, Greene, and Keane 2021a; Labaien, Zugasti, and De Carlos 2020; Karlsson et al. 2018). Accordingly, we define features as motifs in contiguous regions in the present work.

Human Studies on Factual & Counterfactual Explanations

Factual Explanations have been studied in case-based reasoning for over two decades (Leake 1996; Sørmo, Cassens, and Aamodt 2005), and have been extended to explain the predictions of deep learning models (Kenny and Keane 2019). However, one survey found that only 1% of papers performed any user evaluations. Existing results show that factual examples impact correctness judgments of misclassifications and that trust in an AI system decreases rapidly when a model makes multiple errors (Kenny et al. 2021; Kenny, Delaney, and Keane 2023). Buçinca et al. (2020) examined factual explanations in a calorie-assessment task and found that they aided people’s task performance relative to no-explanation controls. van der Waa et al. (2021) tested factual examples explaining the decisions of a simulated diabetes app and found that they improved understanding over contrastive examples, but played more of a role in convincing people (even of incorrect decisions!). However, none of these tests explored time series data.²

Counterfactual Explanations have also encountered similar problems with the majority of user testing focused on tabular data, with mixed results regarding efficacy of this explanation strategy (Lim, Dey, and Avrahami 2009; Binns et al. 2018; Cai, Jongejan, and Holbrook 2019; van der Waa et al. 2021; Warren, Byrne, and Keane 2023; Warren, Keane, and Byrne 2022; Dai et al. 2022; Celar and Byrne 2023;

¹We note that some of this prior work is related to time series forecasting (Fahse, Blohm, and van Giffen 2022; Schulam and Saria 2017), distinct from time series classification.

²Several XAI studies on time series have explored feature-importance analyses (such as LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017)), as well as techniques that emphasize prediction-relevant regions (Mokhtari, Higdon, and Başar 2019; Schlegel et al. 2019; García and Aznarte 2020). Some of these papers report user studies showing benefits for explanations using LIME and SHAP, along with improvements in users’ satisfaction and trust (Saluja et al. 2021), though many suffer from low sample-sizes (see also (Munir et al. 2019) tests using textual explanations).

Task 1: Please click on the **Don't Know** button:

Fault-A **Don't Know** **Fault-B**

Task 2: You **do not** have to make a confidence choice below:

I disagree strongly **I disagree somewhat** **I'm neutral about it** **I agree somewhat** **I agree strongly**

Figure 3: *Attention Check Example*. An example of an attention check item in the user study. Note how users are explicitly asked to select the *Don't Know* option here. There are four attention checks throughout the study. If a user fails more than half of the attention checks they are excluded from the analysis. Attention check items are considerably less noisy than other items and they contain no class distinguishing motifs. Such attention checks aid design robustness in evaluation.

You were instructed to select Don't Know

The correct answer for the above chart is **Don't Know**

Next

16%

Figure 4: *Attention Check Answer*. The screen presented to users after they complete an attention check item in the user study. Note that users are reminded that they were instructed to select the *Don't Know* option.

Lage et al. 2019; Lucic, Haned, and de Rijke 2020a; Förster et al. 2020, 2021; Kirfel and Liefgreen 2021; Kuhl, Artelt, and Hammer 2022). Lucic, Haned, and de Rijke (2020b) found that contrastive explanations increased self-reported understanding of an AI system’s decisions in a forecasting domain. Conversely, studies by van der Waa et al. (2021) demonstrated that factual-example explanations elicited better understanding than contrastive explanations. Similar patterns of mixed findings are also found in the image domain (Singla et al. 2019; Goyal et al. 2019; Larasati, De Liddo, and Motta 2020; Cai, Jongejan, and Holbrook 2019; Zhao,

Oyama, and Kurihara 2021; Delaney et al. 2023). Previous studies in a variety of domains found that counterfactuals elicit higher task accuracy than a no-explanation control (Warren, Byrne, and Keane 2023; Goyal et al. 2019).

Unfortunately, there are almost no user studies using counterfactual explanations with time series (Theissler et al. 2022). The few existing user studies that exist in the time series domain tend to focus on presentational issues, comparing line graphs to novel alternatives (e.g., radial charts, circular heatmaps) (Adnan, Just, and Baillie 2016; Heer, Kong, and Agrawala 2009; Heer and Bostock 2010) and feature attribution (Baer et al. 2026). We have found only one user study on time series, albeit on forecasting rather than classification, comparing factual and counterfactual explanations. Fahse, Blohm, and van Giffen (2022) compared factual and counterfactual explanations in a time series forecasting task where participants predicted bakery sales (N=115). While all explanation types improved accuracy (RMSE) over the control, factual explanations significantly outperformed both counterfactual and hybrid approaches. However, these findings are tempered by several methodological limitations. The study appears under-powered, likely requiring a significantly larger sample size to validate its claims. Furthermore, the statistical reporting is opaque; the authors employed a 2x2 ANOVA to analyze five conditions without clarifying the design mapping, attention checking or reporting essential F-values, degrees of freedom, and post-hoc procedures. Additionally, because the study did not identify which time series features influenced participant predictions, it remains unclear why factual examples provided superior support in this specific forecasting context. Next, we revisit the comparison of factual and counterfactual explanations from the literature. However, unlike the majority of user evaluation, we consider this comparison in the context

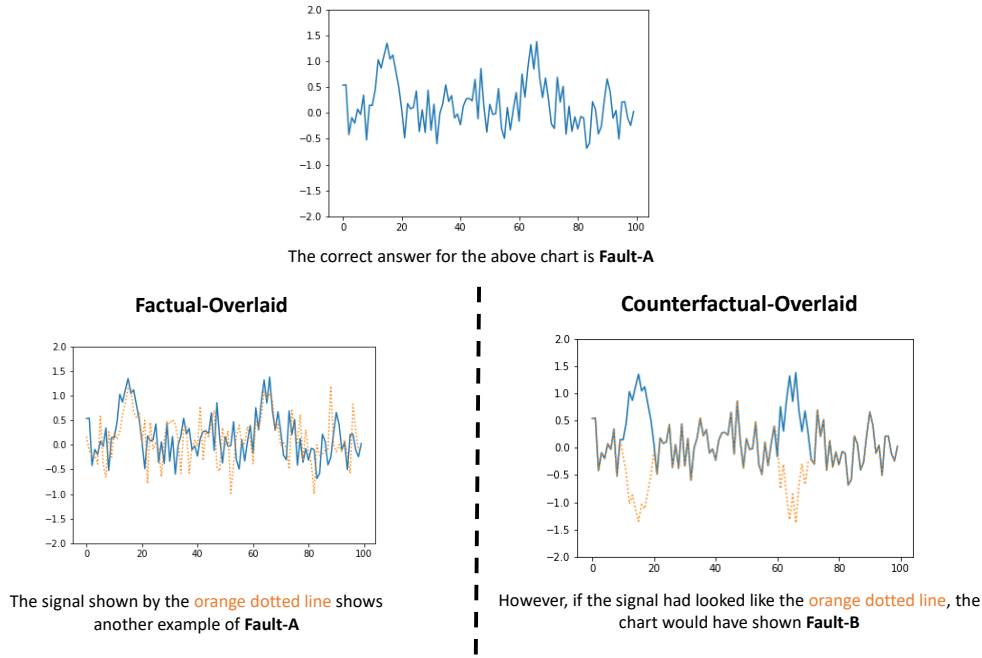


Figure 5: Experiment 1 Explanations: After they classified a test instance (see top chart), people were told the correct answer (e.g., its Fault-A) along with either (i) a factual explanation using a nearest neighbour (Factual-Overlaid group), or (ii) a counterfactual explanation (Counterfactual-Overlaid group), as color-coded overlays on the test instance.

of time series data.

3 Experiment 1: Overlaid Factual & Counterfactual Explanations

Our review of related work on the explanation of time series reveals a thin literature with inconclusive user findings on the efficacy of factual and counterfactual XAI strategies. Importantly, this literature does not tell us much about how these explanations should be presented to users (e.g., whether they should be textual, highlightings or overlaid plots). This state of affairs, perhaps reflects the very real difficulties in systematically tackling time series domains, in the selection of suitable test materials and the determination of features. Unless we start from a scenario, where we define features precisely, it is hard to see how we can accurately assess how people comprehend these explanations due to spurious effects.

Hence, in the current experiments, we aimed to control the complexity of materials by generating a representative and balanced dataset using synthetic (but plausible) time series data. The test instances showed one of two engine faults – Fault A or Fault B (see Figure 2 for examples) – based on defined contiguous regions (at time-points 10–20 and 60–70) where the key features that distinguished the classes occurred (i.e., using positive/negative sine waves) with the remaining regions of the time series being non-feature-regions showing Gaussian noise. The classification task was verified as tractable using a state-of-the-art sequence learning classifier (see Mr-SEQ (Le Nguyen et al. 2019)) that could ac-

curately classify the 40 test instances presented to people in the studies. For a discussion about the limitations of using synthetic data for generalizability see Section 5.

Both studies compared factual and counterfactual explanations using matched material sets. For the Factual-Overlaid group, the explanatory example was the nearest-like-neighbor to the test instance in the training data using Euclidean distance (n.b., temporal axes were already aligned; see Figure 5 bottom-left plot). For the Counterfactual-Overlaid group, a contrasting explanation-instance was generated by flipping the values in the feature regions of the time series (by reflecting the motif over the x-axis such that a $\cap$ motif resembles a $\cup$ and vice versa; see Figure 5 bottom-right plot). This method is inspired by, and operates in a similar fashion to, Native-Guide (Delaney, Greene, and Keane 2021a). It modifies contiguous regions of the time series with the additional constraint that the number of features modified is fixed to maintain control for psychological effects of sparsity in user testing. Note, all the encoded feature-regions also had added noise to make similarities in the shape of the features less obvious. In each study, people received 40 time series instances to classify, their task being to predict the class of each test instance (i.e., as Fault-A, Don't Know, Fault-B) and indicate their confidence in their choice (on a 5-point Likert-type scale). They were then given feedback on the correct classification along with an explanation (i.e., a factual or counterfactual explanation). Previous user studies show that participants are more engaged with a task, if they make their own predictions before seeing those of the AI model. The key performance measure

was *accuracy*, the proportion of instances people classified correctly over the course of the study. Figure 5 shows one of the materials used; the top-chart shows the presented instance with the feedback given after the participant made their classification; the two bottom panels show the alternative explanations given to the two separate groups (i.e., the overlaid factual or counterfactual explanation on the original instance). In Experiment 1, both explanations were visualised as overlays on the original instance, shown as an orange dotted line (see Figure 5). In Experiment 2, we consider a different mode of presentation for the factual explanations.

Method

Participants, Design & Ethical Review Sixty-two participants were randomly assigned to one of two independent groups, the Factual-Overlaid group (N=30, after excluding 2 participants who failed attention checks) and Counterfactual-Overlaid group (N=32). The sample size required was determined by an a priori power analysis, designed to balance the probability of Type I and Type II errors, using G*Power (Erdfeuler, Faul, and Buchner 1996); it indicated that 56 participants (28 in each group) were required for a one-way between-participant design with two groups, assuming a large effect size ($d = 0.8$), for an alpha of .05 and power of .90. Participants were recruited using the Prolific Academic platform (<https://www.prolific.co>). Respondents were pre-screened to select native English speakers from the United Kingdom, United States, Australia, Canada, Ireland and New Zealand. Ethics approval for both studies was granted in advance, with the reference code LS-E-21-215-Delaney-Keane. The study was explained to participants at the outset and their consent was requested; they were informed that they could withdraw from the study at any time and, after the study, could withdraw their data too, if they had any concerns. After completing all items, the rationale of the study was explained to them and they were given the contact details of the PI (should a follow-up be needed).

Materials A base time series of length 100 was generated by adding Gaussian noise to the entire time range, based on a normal distribution with a mean of zero and a standard deviation of 0.5. Distinguishing motifs were introduced into specific, selected time-intervals (namely, the 10–20 and 60–70 regions). In these regions, sine-like motifs were created that exhibit oscillatory behavior. Each motif had a duration of 10 time points and was generated using a sine wave equation. The frequency of the sine wave is inversely proportional to the motif duration. Fault-A was characterized by positive sine waves in these regions whilst Fault-B was characterized by negative sine waves (see Figure 2 for samples). To incorporate noise into the motifs, an additional noise scale parameter was used. Independent Gaussian noise was added to each motif region, following a normal distribution with a mean of zero and a scale determined by the noise scale parameter (set to 0.25, slightly less noisy than the rest of the time series). Using these formulae 200 time series instances were generated, 50/50 divided between the Fault-A and Fault-B classes. Forty instances were randomly se-

lected as a test set and the remainder were used to train the Mr-SEQL classifier (Le Nguyen et al. 2019). These 40 test instances were presented to participants in the experiment (randomly re-ordered for each person).

Measures: Accuracy and Confidence A combination of objective and subjective measures was used (see (Hoffman et al. 2018; Warren, Byrne, and Keane 2023; Kim, Maathuis, and Sent 2024) for a discussion). The objective measure was *accuracy*, which measured the proportion of test instances that users correctly classified. The subjective measure was people's self-assessment of their *confidence* in their own classifications. Participants rated their confidence in their classification using a 5-point Likert scale, ranging from 1 ("Not at all confident") to 5 ("Extremely confident"). Participants gave this judgment for each test instance before they were shown the correct model-classification and its associated overlaid explanation.

Procedure Both groups were given the same general instructions for the tasks to be carried out. They were told that: *In this study, you are asked to examine charts of data about an engine's performance. Each chart measures the vibrations made by the engine. Some patterns of vibration show common faults in the engine. Only one of the two faults will occur in any chart. You will be shown roughly 40 charts and your task is to decide whether Fault A or Fault B has occurred in each one. You will get feedback on whether you have chosen the correct fault. Your aim is to choose the correct fault in each chart. Then they were told: In Task 1, you have to indicate which fault is being shown in the chart; indicate Fault-A or Fault-B by clicking on the button. If you really Don't Know then click that button. In Task 2, you should indicate your confidence in the choice you have made. When you click the next button you will be told the fault that is present in the chart.* After participants made their decision about which fault was present in the chart, they were presented with the correct answer along with an explanation. Each group was shown one of two sorts of explanation. In the Factual-Overlaid group they were shown another example of the fault overlaid on the presented test instance, (i.e., the factual explanation) using the nearest neighbor to the test instance in the dataset, along with an explanatory statement (see Figure 5). In the Counterfactual-Overlaid group they were given a counterfactual plot overlaying the original test instance, along with an explanatory statement (see Figure 5). Participants were given a total of 40 items to classify as either Fault-A or Fault-B (20 of each). Items were randomly re-ordered for each participant. Four attention checks were provided using an out-of-distribution signal, for which participants were explicitly told to label the fault as *Don't Know* (See Appendix for an example). Participants who failed more than half of the attention checks were removed from the analysis.

Results & Discussion

Participants receiving the counterfactual explanations were **more accurate** than those receiving factual explanations and expressed greater confidence in their classifications. The Counterfactual-Overlaid group was more accurate

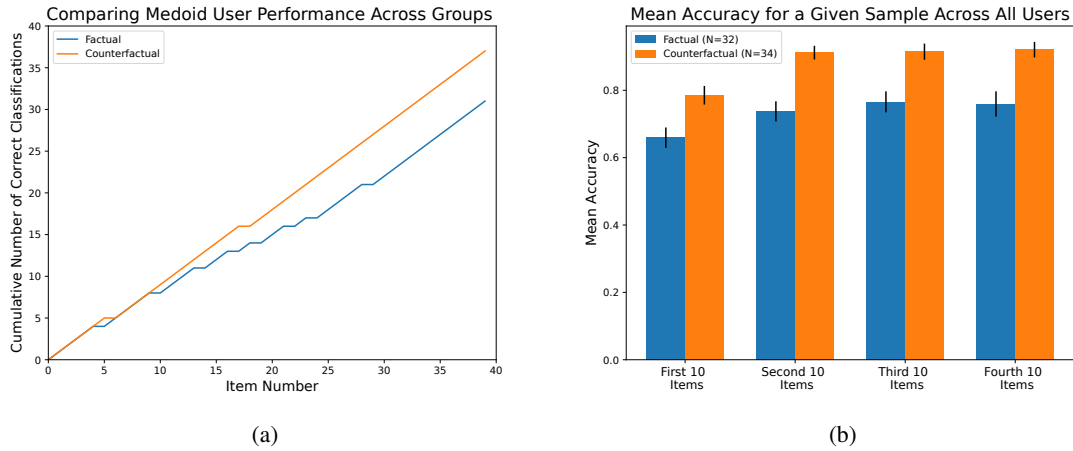


Figure 6: *Comparing Factual and Counterfactual Groups*. Performance across time is demonstrated by considering the medoid performance of users in (a) and the mean performance for the whole sample across different time intervals for (b). **Counterfactual explanations elicit faster and more accurate learning** in comparison to factual explanations. Error bars represent standard error of the mean. Note for (a) there is a different line plot for each user so we display the average (computed with k -medoids) performance in each group.

($M = 0.88$, $SD = 0.10$) in the classification task than the Factual-Overlaid group, ($M = 0.71$, $SD = 0.18$), $t(62) = 4.47$, $p < 0.001$, $d = 1.13$. So, counterfactual explanations provide better feedback to users than factual explanations, helping them to discriminate the two classes. Those individuals who received the counterfactual explanations were also significantly more confident in their classification choices ($M = 4.18$, $SD = 0.46$), relative to those who received factual explanations ($M = 3.60$, $SD = 0.55$), $t(62) = 4.46$, $p < 0.001$, $d = 1.13$. Contrary to the inconsistent results prevalent in the broader XAI literature, this study demonstrates that when presented appropriately, counterfactual explanations in a time series classification task significantly enhance people’s domain understanding compared to factual explanations for the same task. As such, it shows that discriminating features, defined in two contiguous regions of a time series, can be learned by people when they are presented as overlaid counterfactual instances, by comparison to overlaid factual instances. In short, we have a cognitive benchmark for the conditions under which this counterfactual strategy works for complex time series data. However, there remains one issue with respect to the presentation format for the factual explanation to be explored.

4 Experiment 2: Testing non-Overlaid Factual Explanations

In Experiment 1, we saw that when factual and counterfactual explanations were presented as overlays on the original test instances, people were more accurate in making correct classifications for counterfactual over factual explanations. Intuitively, the overlaid counterfactuals seem to “naturally” visualize the contrasting features of the two classes; they present people with the differences between the two instances, analogous to the instructions given for counterfactuals using tabular data (e.g., “if you had asked for a smaller

loan, you would have been successful”). However, for the factual explanations overlays do not seem as “natural”. Traditionally, factual case-based explanations have often been presented *next to* the to-be-explained instance, rather than being overlaid (see e.g., (Valdez-Ávila et al. 2022; Jeyakumar et al. 2020)). Indeed, Jeyakumar et al. (2020) report that users preferred this type of explanation over ones that showed highlighted regions that reflect attributions. Hence, in this experiment, the Factual-Separate group received the explanatory nearest neighbour below the test instance (see Figure 7), whereas the Counterfactual-Overlaid group received explanatory counterfactual overlaid (as in Expt. 1; see Figure 7). So, this experiment was designed to test the optimal “natural” presentation mode for each explanation strategy³.

Method

Participants, Design, Procedure & Ethical Review

Sixty-six participants were randomly assigned to the two groups, the Factual-Separate ($n=32$) to the Counterfactual-Overlaid ($n=34$) groups (after excluding 2 participants that failed attention checks). The design was identical to that of Experiment 1, so the same power analysis apply. As before, the ethical review covered this experiment too and the procedure was identical, as were the measures used. The only difference was in the mode of presentation used for the factual explanations, as a separated plot below the test instance, rather than using the overlay of Experiment 1.

³One could have used a full 2×2 design for this experiment crossing Explanation-type and Visualization-type to examine interactions between these two variables. However, our aim was to find the optimal conditions for presenting these strategies, not the conditions under which they are less optimal.

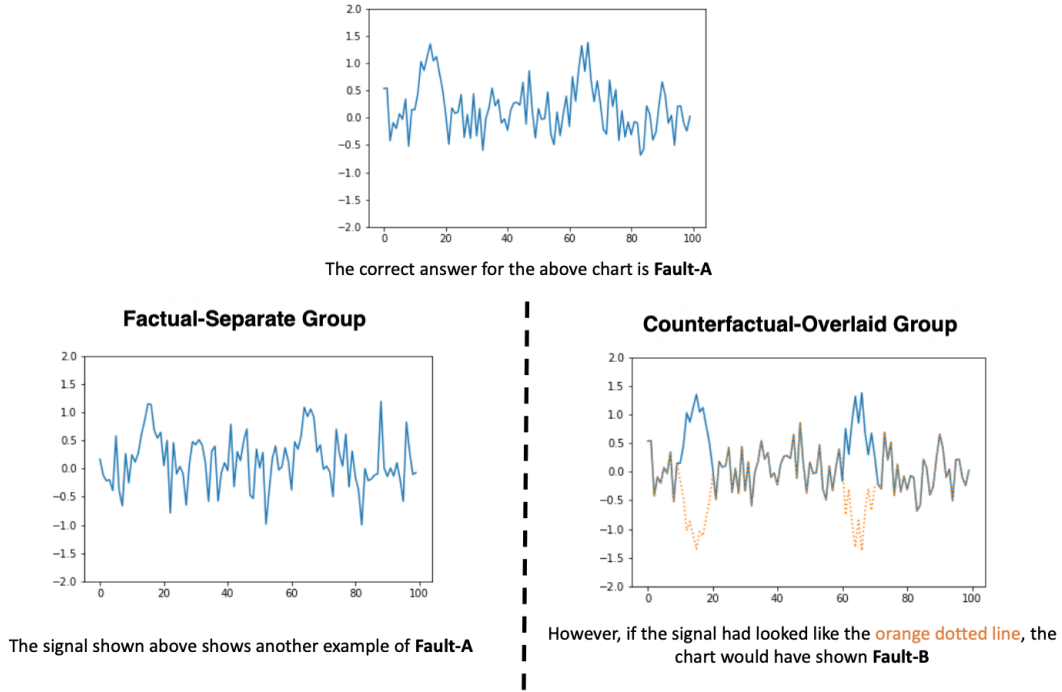


Figure 7: Experiment 2 Explanations: After participants classified a test instance (see top chart), they got the correct answer (e.g., this is Fault-A) and either (i) a factual explanation as a nearest neighbour shown in a separate chart below the test instance (Factual-Separate group), or (ii) a counterfactual explanation as a color-coded overlay on the test instance (Counterfactual-Overlaid group).

Results & Discussion

Again, participants receiving counterfactuals displayed greater accuracy than those receiving factual explanations and expressed greater confidence in their classifications. The Counterfactual-Overlaid group was more accurate ($M = 0.883$, $SD = 0.13$) in the classification task than the Factual-Separate group, ($M = 0.730$, $SD = 0.12$), $t(66) = 4.88$, $p < 0.001$, $d = 1.19$. Indeed, an analysis of the medoid proportion of correct classifications in each group, at each time-step of the presentation, shows that the benefits of the counterfactual occurs at an early stage (by the 10th item of 40). When considering only the first ten tasks that were presented, users in the counterfactual group were also significantly more accurate ($M = 0.79$, $SD = 0.16$) than those who received factual explanations ($M = 0.66$, $SD = 0.17$), $t(66) = 3.03$, $p < 0.05$. Surprisingly, they were even more accurate than those who received 40 factual explanations. These results indicate that counterfactual explanations also facilitate **faster learning** of class distinguishing motifs (see figure 6). People receiving the counterfactual explanations ($M = 4.23$, $SD = 0.44$) were significantly more confident about their classifications, relative to those who received factual explanations with ($M = 3.78$, $SD = 0.62$), $t(66) = 3.41$, $p < 0.001$, $d = 0.84$. These results replicate the benefits seen before for counterfactual explanations in this task over factual explanations for time series classification. They also show that the lower accuracy seen in the factual explanation groups is not a result of the mode of presentation used

for the factual ones. Interestingly, though the counterfactual explanations have equivalent accuracy levels in both experiments (88% correct), the factual explanations do marginally better in the current (73% here versus 71% before), though this improvement is not significantly different. Importantly, these findings confirm the cognitive benchmark for the defined manipulation, as we have replicated the effect seen in Experiment 1, under slightly different conditions.

5 Conclusions, Limitations & Future Directions

Although many important application domains (e.g., in personalised medicine and climate modelling) often depend on time series, the unique challenges that arise in dealing with this type of data, such as high-dimensionality and feature correlation, have hampered progress in XAI. Indeed, the slow progress in algorithmic development has been matched by even less progress in user testing, raising concerns about the reliability and transparency of existing approaches. Accordingly, we have tried to remedy these challenges in the present work. In our methodology, we precisely control feature definitions, enabling the creation of balanced and representative materials that can be used to test different explanation strategies. We have applied this methodology in two studies (i) comparing two popular example-based explanation methods (factual versus counterfactual), and (ii) varying the mode of presentation used (overlaid versus sep-

arate). The results of these studies show that overlaid counterfactual explanations improve people's understanding of time series classifications in comparison to factual explanations. Results indicate that counterfactual explanations facilitate faster and more accurate learning in comparison to factual explanations in a time series classification task.

How generalizable is the proposed approach and task to other time series domains? We accept that one limitation of these tests is ecological validity. The effects found are shown for motifs with specific properties and it remains to be seen how the effects found will generalize to real-world datasets and complex motifs. The main reason for this design choice was to mitigate the effects of confounding variables such as noise and one class being easier to distinguish. Despite our best efforts, we found it difficult to find a real-world time series classification dataset that non-expert humans could meaningfully engage with. Importantly, we are not blindly creating materials as the sort of discriminative and periodic motifs that we tested do occur in many real world datasets (e.g., electrocardiograms in healthcare (Dau et al. 2019), where elevated sine wave patterns in an ECG indicate severe hyperkalemia, or elevated potassium levels in the blood (Pluijmen and Hersbach 2007; Loubser et al. 2023)). In this sense, we are trying to follow the methodology used in cognitive psychology where researchers first find demonstrable effects under controlled “laboratory” conditions before trying to apply those findings more widely. While we rigorously control for confounding variables, we acknowledge that some may still persist. Future directions for this work lie in exploring other systematic properties and motifs in real-world time series data. Other properties such as variance could be extracted from toolkits such as *TSfresh* (Christ et al. 2018) and integrated within our design methodology for real-world datasets. More generally, our methodology could be deployed in other high-stakes areas of time series XAI such as anomaly detection where similar transparency challenges arise (Der et al. 2024).

Ethical Reflections & Concerns

Ethical Considerations & Mitigation of Concerns

The research plans and designs were assessed by the relevant Ethics Committee at our institution (see Method sections). We were concerned that the user studies should be clear and transparent to participants and only conducted with their full consent. As such, instructions ensure informed consent and made it clear that participants could withdraw from the study at any time without prejudice. At the end of the testing period, they were debriefed on the background to the study and its goals. They were given the option at this point to withdraw their data from the study (again without prejudice) and in interests of transparency the name, affiliation and contact details of the PI for the project were provided. There was a concern that the dataset domain chosen for the study should not be a contentious one, politically sensitive or challenging in its content. This concern was mitigated by choosing a neutral topic (e.g., vibration measurements indicating an engine's fault). Finally, we reflected at length on the best methodology for testing the challenges address, opting for

a controlled, hypothesis-testing framework because this approach has received the least attention in the field (i.e., the norm is unsupported anecdotal claims).

Reflection on Adverse & Unintended Impacts

Though it is not commonly acknowledged in the XAI literature, we believe that any explanation strategy that effects cognitive change in users is potentially open to abuse (Sullivan and Kasirzadeh 2024), though the scenarios in which this might occur may be, as yet, unspecified. Much XAI research emphasises the goal of eliciting trust (in the AI system) in end-users. However, trust without understanding is problematic and invites adverse impacts; for example, some recent work on factual and counterfactual XAI reviewed here, has shown that users may learn little about the AI-system/domain and yet will have a greater trust in the system when explanations are provided (even for incorrect predictions). Accordingly, we believe explanation strategies need to show knowledge change (rather than trust *per se*), to improve users understanding of the AI-system/domain (hence, our choice of accuracy as a measure in the studies). If users are given a better understanding of the AI-system/domain and the context for the system's decision making, then adverse and unintended consequences are, we believe, more likely to be avoided (though they may never be completely removed).

Acknowledgements

This publication has emanated from research conducted with the financial support of (i) Science Foundation Ireland (SFI) to the Insight Centre for Data Analytics under Grant Number 12/RC/2289 P2 and (ii) SFI and the Department of Agriculture, Food and Marine on behalf of the Government of Ireland under Grant Number 16/RC/3835 (VistaMilk).

References

- Adadi, A.; and Berrada, M. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6: 52138–52160.
- Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 9525–9536.
- Adnan, M.; Just, M.; and Baillie, L. 2016. Investigating time series visualisations to improve the user experience. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5444–5455.
- Akula, A.; Wang, S.; and Zhu, S.-C. 2020. Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2594–2601.
- Anjomshoe, S.; Najjar, A.; Calvaresi, D.; and Främling, K. 2019. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*, Montreal, Canada, May 13–17, 2019, 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems.

- Ates, E.; Aksar, B.; Leung, V. J.; and Coskun, A. K. 2021. Counterfactual explanations for multivariate time series. In *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*, 1–8. IEEE.
- Baer, G.; Zhang, C.; Grau, I.; and Gorp, P. V. 2026. Does Explanation Correctness Matter? Linking Computational XAI Evaluation to Human Understanding. *arXiv:2603.25251*.
- Binns, R.; Van Kleek, M.; Veale, M.; Lyngs, U.; Zhao, J.; and Shadbolt, N. 2018. 'It's Reducing a Human Being to a Percentage'. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. ISBN 9781450356206.
- Buçinca, Z.; Lin, P.; Gajos, K. Z.; and Glassman, E. L. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 454–464.
- Byrne, R. M. 2019. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2019-Augus, 6276–6282. ISBN 9780999241141.
- Cai, C. J.; Jongejan, J.; and Holbrook, J. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*, 258–262.
- Caruana, R.; Kangaroo, H.; Dionisio, J.; Sinha, U.; and Johnson, D. 1999. Case-based explanation of non-case-based learning methods. In *Proceedings of the AMIA Symposium*, 212. American Medical Informatics Association.
- Celar, L.; and Byrne, R. M. 2023. How people reason with counterfactual and causal explanations for Artificial Intelligence decisions in familiar and unfamiliar domains. *Memory & Cognition*, 1–16.
- Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; and Su, J. K. 2019. This Looks Like That: Deep Learning for Interpretable Image Recognition. *Advances in Neural Information Processing Systems*, 32: 8930–8941.
- Cho, S.; Lee, G.; Chang, W.; and Choi, J. 2020. Interpretation of deep temporal representations by selective visualization of internally activated nodes. *arXiv preprint arXiv:2004.12538*.
- Christ, M.; Braun, N.; Neuffer, J.; and Kempa-Liehr, A. W. 2018. Time series feature extraction on basis of scalable hypothesis tests (tsfresh—a python package). *Neurocomputing*, 307: 72–77.
- Cook, A. A.; Mısırlı, G.; and Fan, Z. 2020. Anomaly Detection for IoT Time-Series Data: A Survey. *IEEE Internet of Things Journal*, 7(7): 6481–6494.
- Cover, T.; and Hart, P. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1): 21–27.
- Cunningham, P.; and Delany, S. J. 2021. k-Nearest neighbour classifiers-A Tutorial. *ACM computing surveys (CSUR)*, 54(6): 1–25.
- Dai, X.; Keane, M. T.; Shalloo, L.; Ruelle, E.; and Byrne, R. M. 2022. Counterfactual explanations for prediction and diagnosis in xai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 215–226.
- Dau, H. A.; Bagnall, A.; Kamgar, K.; Yeh, C.-C. M.; Zhu, Y.; Gharghabi, S.; Ratanamahatana, C. A.; and Keogh, E. 2019. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6): 1293–1305.
- Delaney, E.; Greene, D.; and Keane, M. T. 2021a. Instance-based counterfactual explanations for time series classification. In *International Conference on Case-Based Reasoning*, 32–47. Springer.
- Delaney, E.; Greene, D.; and Keane, M. T. 2021b. Uncertainty Estimation and Out-of-Distribution Detection for Counterfactual Explanations: Pitfalls and Solutions. *arXiv preprint arXiv:2107.09734*.
- Delaney, E.; Pakrashi, A.; Greene, D.; and Keane, M. T. 2023. Counterfactual Explanations for Misclassified Images: How Human and Machine Explanations Differ. *Artificial Intelligence*.
- DeMartino, G. F. 2020. The confounding problem of the counterfactual in economic explanation*. *Review of Social Economy*, 0(0): 1–11.
- Der, A.; Yeh, C.-C. M.; Zheng, Y.; Wang, J.; Zhuang, Z.; Wang, L.; Zhang, W.; and Keogh, E. 2024. Pupae: Intuitive and actionable explanations for time series anomalies. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, 37–45. SIAM.
- Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Duda, R. O.; Hart, P. E.; et al. 1973. *Pattern classification and scene analysis*, volume 3. Wiley New York.
- Ehsan, U.; Passi, S.; Liao, Q. V.; Chan, L.; Lee, I.-H.; Muller, M.; and Riedl, M. O. 2024. The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–32.
- Erdfelder, E.; Faul, F.; and Buchner, A. 1996. GPOWER: A general power analysis program. *Behavior research methods, instruments, & computers*, 28(1): 1–11.
- Fahse, T. B.; Blohm, I.; and van Giffen, B. 2022. Effectiveness of Example-Based Explanations to Improve Human Decision Quality in Machine Learning Forecasting Systems.
- Fawaz, H. I.; Forestier, G.; Weber, J.; Idoumghar, L.; and Muller, P.-A. 2019a. Adversarial attacks on deep neural networks for time series classification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.
- Fawaz, H. I.; Forestier, G.; Weber, J.; Idoumghar, L.; and Muller, P.-A. 2019b. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4): 917–963.
- Filali Boubrahimi, S.; and Hamdi, S. M. 2022. On the Mining of Time Series Data Counterfactual Explanations using Barycenters. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 3943–3947.

- Förster, M.; Hühn, P.; Klier, M.; and Kluge, K. 2021. Capturing Users' Reality: A Novel Approach to Generate Coherent Counterfactual Explanations. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, 1274.
- Förster, M.; Klier, M.; Kluge, K.; and Sigler, I. 2020. Evaluating explainable Artificial intelligence—What users really appreciate. In *Proceedings of the 28th European Conference on Information Systems (ECIS)*.
- García, M. V.; and Aznarte, J. L. 2020. Shapley additive explanations for NO2 forecasting. *Ecological Informatics*, 56: 101039.
- Gee, A. H.; Garcia-Olano, D.; Ghosh, J.; and Paydarfar, D. 2019. Explaining deep classification of time-series data with learned prototypes. *CEUR Workshop Proceedings*, 2429: 15–22.
- Ghassemi, M.; Naumann, T.; Schulam, P.; Beam, A. L.; Chen, I. Y.; and Ranganath, R. 2020. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020: 191.
- Ghods, A.; and Cook, D. J. 2022. PIP: Pictorial interpretable prototype learning for time series classification. *IEEE computational intelligence magazine*, 17(1): 34–45.
- Gomez, O.; Holter, S.; Yuan, J.; and Bertini, E. 2020. Vice: Visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 531–535.
- Goyal, Y.; Wu, Z.; Ernst, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Counterfactual visual explanations. In *International Conference on Machine Learning*, 2376–2384. PMLR.
- Grabocka, J.; Schilling, N.; Wistuba, M.; and Schmidt-Thieme, L. 2014. Learning time-series shapelets. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 392–401.
- Guidotti, R. 2024. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5): 2770–2824.
- Guidotti, R.; Monreale, A.; Spinnato, F.; Pedreschi, D.; and Giannotti, F. 2020. Explaining Any Time Series Classifier. In *IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*, 167–176. IEEE.
- Guillemé, M.; Masson, V.; Rozé, L.; and Termier, A. 2019. Agnostic local explanation for time series classification. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 432–439. IEEE.
- Han, T.; Henderson, J.; Akbarian, P.; and Ghosh, J. 2022. Improving Counterfactual Explanations for Time Series Classification Models in Healthcare Settings. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*.
- Heer, J.; and Bostock, M. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 203–212.
- Heer, J.; Kong, N.; and Agrawala, M. 2009. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 1303–1312.
- Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.
- Hsu, E.-Y.; Liu, C.-L.; and Tseng, V. S. 2019. Multivariate time series early classification with interpretability using deep learning and attention mechanism. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 541–553. Springer.
- Jeyakumar, J. V.; Noor, J.; Cheng, Y.-H.; Garcia, L.; and Srivastava, M. 2020. How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods. *Advances in Neural Information Processing Systems*, 33.
- Karlsson, I.; Rebane, J.; Papapetrou, P.; and Gionis, A. 2018. Explainable time series tweaking via irreversible and reversible temporal transformations. In *2018 IEEE International Conference on Data Mining (ICDM)*, 207–216. IEEE.
- Keane, M. T.; and Kenny, E. M. 2019. How case-based reasoning explains neural networks: A theoretical analysis of XAI using post-hoc explanation-by-example from a survey of ANN-CBR twin-systems. In *International Conference on Case-Based Reasoning*, 155–171. Springer.
- Keane, M. T.; Kenny, E. M.; Delaney, E.; and Smyth, B. 2021. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. In *30th International Joint Conference on Artificial Intelligence (IJCAI)*.
- Kenny, E. M.; Delaney, E.; and Keane, M. T. 2023. Advancing Post-Hoc Case-Based Explanation with Feature Highlighting. In Elkind, E., ed., *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, 427–435. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Kenny, E. M.; Ford, C.; Quinn, M.; and Keane, M. T. 2021. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, 103459.
- Kenny, E. M.; and Keane, M. T. 2019. Twin-systems to explain artificial neural networks using case-based reasoning: Comparative tests of feature-weighting methods in ANN-CBR twins for XAI. In *Twenty-Eighth International Joint Conferences on Artificial Intelligence (IJCAI), Macao, 10-16 August 2019*, 2708–2715.
- Kenny, E. M.; and Keane, M. T. 2020. On generating plausible counterfactual and semi-factual explanations for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Kim, B.; Koyejo, O.; Khanna, R.; et al. 2016. Examples are not enough, learn to criticize! Criticism for Interpretability. In *NIPS*, 2280–2288.
- Kim, J.; Maathuis, H.; and Sent, D. 2024. Human-centered evaluation of explainable AI applications: a systematic review. *Frontiers in Artificial Intelligence*, 7: 1456486.

- Kirfel, L.; and Liefgreen, A. 2021. What If (and How...)? - Actionability Shapes People's Perceptions of Counterfactual Explanations in Automated Decision-Making. In *ICML Workshop on Algorithmic Recourse*.
- Kuhl, U.; Artelt, A.; and Hammer, B. 2022. Keep Your Friends Close and Your Counterfactuals Closer: Improved Learning From Closest Rather Than Plausible Counterfactual Explanations in an Abstract Setting. *arXiv preprint arXiv:2205.05515*.
- Labaien, J.; Zugasti, E.; and De Carlos, X. 2020. Contrastive Explanations for a Deep Learning Model on Time-Series Data. In *International Conference on Big Data Analytics and Knowledge Discovery*, 235–244. Springer.
- Lage, I.; Chen, E.; He, J.; Narayanan, M.; Kim, B.; Gershman, S.; and Doshi-Velez, F. 2019. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*.
- Lai, V.; Zhang, Y.; Chen, C.; Liao, Q. V.; and Tan, C. 2023. Selective explanations: Leveraging human input to align explainable ai. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 1–35.
- Langer, M.; Oster, D.; Speith, T.; Hermanns, H.; Kästner, L.; Schmidt, E.; Sesing, A.; and Baum, K. 2021. What do we want from Explainable Artificial Intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296: 103473.
- Larasati, R.; De Liddo, A.; and Motta, E. 2020. The effect of explanation styles on user's trust. In *Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies*. Cagliari, Italy.
- Laugel, T.; Lesot, M.-J.; Marsala, C.; Renard, X.; and Detryniecki, M. 2019. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In *IJCAI-19*, 2801–2807.
- Le Nguyen, T.; Gsponer, S.; Ilie, I.; O'Reilly, M.; and Ifrim, G. 2019. Interpretable time series classification using linear models and multi-resolution multi-domain symbolic representations. *Data Mining and Knowledge Discovery*, 33(4): 1183–1222.
- Leake, D. B. 1996. *Case-Based Reasoning: Experiences, lessons and future directions*. MIT press.
- Liao, Q. V.; Gruen, D.; and Miller, S. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Conference on Human Factors in Computing Systems - Proceedings*.
- Lim, B. Y.; Dey, A. K.; and Avrahami, D. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Conference on Human Factors in Computing Systems - Proceedings*, 2119–2128. ISBN 9781605582474.
- Lipton, Z. C. 2017. The Doctor Just Won't Accept That! In *31st Conference on Neural Information Processing Systems (NIPS 2017)*.
- Loubser, J.; Bronislowski, L. P.; Fonarov, I.; and Casadesus, D. 2023. Sine-wave electrocardiogram rhythm in a patient on haemodialysis presenting with severe weakness and hyperkalaemia. *BMJ Case Reports*, 16(3): e255007.
- Lucic, A.; Haned, H.; and de Rijke, M. 2020a. Contrastive Explanations for Large Errors in Retail Forecasting Predictions through Monte Carlo Simulations.
- Lucic, A.; Haned, H.; and de Rijke, M. 2020b. Why does my model fail? contrastive local explanations for retail forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 90–98.
- Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 4765–4774.
- Maweu, B. M.; Dakshit, S.; Shamsuddin, R.; and Prabhakaran, B. 2021. CEFES: a CNN explainable framework for ECG signals. *Artificial Intelligence in Medicine*, 115: 102059.
- Mayne, H.; Kearns, R. O.; Yang, Y.; Bean, A. M.; Delaney, E. D.; Russell, C.; and Mahdi, A. 2025. LLMs Don't Know Their Own Decision Boundaries: The Unreliability of Self-Generated Counterfactual Explanations. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 24161–24186. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.
- McGrath, R.; Costabello, L.; Van, C. L.; Sweeney, P.; Kamiab, F.; Shen, Z.; and Lecue, F. 2018. Interpretable credit application predictions with counterfactual explanations. *arXiv preprint arXiv:1811.05245*.
- Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267: 1–38.
- Ming, Y.; Xu, P.; Qu, H.; and Ren, L. 2019. Interpretable and steerable sequence learning via prototypes. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 903–913.
- Mokhtari, K. E.; Higdon, B. P.; and Başar, A. 2019. Interpreting financial time series with SHAP values. In *Proceedings of the 29th annual international conference on computer science and software engineering*, 166–172.
- Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–617.
- Munir, M.; Siddiqui, S. A.; Küsters, F.; Mercier, D.; Dengel, A.; and Ahmed, S. 2019. Tsxplain: Demystification of dnn decisions for time-series using natural language and statistical features. In *Artificial Neural Networks and Machine Learning—ICANN 2019: Workshop and Special Sessions: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings 28*, 426–439. Springer.
- Nugent, C.; Doyle, D.; and Cunningham, P. 2009. Gaining insight through case-based explanation. *Journal of Intelligent Information Systems*, 32(3): 267–295.

- Oppenheimer, D. M.; Meyvis, T.; and Davidenko, N. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, 45(4): 867–872.
- Pičulin, M.; Petek, B.; Ograjenšek, I.; and Strumbelj, E. 2025. Position: Explainable AI Cannot Advance Without Better User Studies. In *Forty-second International Conference on Machine Learning Position Paper Track*.
- Pluijmen, M. J.; and Hersbach, F. M. 2007. Sine-wave pattern arrhythmia and sudden paralysis that result from severe hyperkalemia. *Circulation*, 116(1): e2–e4.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. 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*, 1135–1144. ACM.
- Rojat, T.; Puget, R.; Filliat, D.; Del Ser, J.; Gelin, R.; and Díaz-Rodríguez, N. 2021. Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey.
- Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215.
- Russell, C. 2019. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 20–28.
- Saluja, R.; Malhi, A.; Knapič, S.; Främling, K.; and Cavdar, C. 2021. Towards a rigorous evaluation of explainability for multivariate time series. *arXiv preprint arXiv:2104.04075*.
- Schäfer, P. 2016. Scalable time series classification. *Data Mining and Knowledge Discovery*, 30(5): 1273–1298.
- Schlegel, U.; Arnout, H.; El-Assady, M.; Oelke, D.; and Keim, D. A. 2019. Towards a rigorous evaluation of XAI Methods on Time Series. *arXiv preprint arXiv:1909.07082*.
- Schuff, H.; Jacovi, A.; Adel, H.; Goldberg, Y.; and Vu, N. T. 2022. Human interpretation of saliency-based explanation over text. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 611–636.
- Schulam, P.; and Saria, S. 2017. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems*, 1697–1708.
- Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.
- Serramazza, D. I.; Nguyen, T. T.; Nguyen, T. L.; and Ifrim, G. 2023. Evaluating Explanation Methods for Multivariate Time Series Classification. *arXiv preprint arXiv:2308.15223*.
- Šimić, I.; Sabol, V.; and Veas, E. 2021. XAI Methods for Neural Time Series Classification: A Brief Review. *arXiv preprint arXiv:2108.08009*.
- Singla, S.; Pollack, B.; Chen, J.; and Batmanghelich, K. 2019. Explanation by Progressive Exaggeration. In *International Conference on Learning Representations*.
- Sørmo, F.; Cassens, J.; and Aamodt, A. 2005. Explanation in Case-Based Reasoning: Perspectives and goals. *Artificial Intelligence Review*, 24(2): 109–143.
- Strodthoff, N.; and Strodthoff, C. 2019. Detecting and interpreting myocardial infarction using fully convolutional neural networks. *Physiological measurement*, 40(1): 015001.
- Sullivan, E.; and Kasirzadeh, A. 2024. Explanation Hacking: The perils of algorithmic recourse. *arXiv preprint arXiv:2406.11843*.
- Tahaei, M.; Constantinides, M.; Quercia, D.; Kennedy, S.; Muller, M.; Stumpf, S.; Liao, Q. V.; Baeza-Yates, R.; Aroyo, L.; Holbrook, J.; et al. 2023. Human-Centered Responsible Artificial Intelligence: Current & Future Trends. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–4.
- Theissler, A.; Spinnato, F.; Schlegel, U.; and Guidotti, R. 2022. Explainable AI for time series classification: a review, taxonomy and research directions. *IEEE Access*.
- Turbé, H.; Bjelogrić, M.; Lovis, C.; and Mengaldo, G. 2023. Evaluation of post-hoc interpretability methods in time-series classification. *Nature Machine Intelligence*, 5(3): 250–260.
- Valdez-Ávila, M. F.; Pérez-Pérez, G. A.; Sarabia-Osorio, H.; Bermejo-Sabbagh, C.; and Orozco-del Castillo, M. G. 2022. A Case-based Explanation Method for Weather Forecasting.
- van der Waa, J.; Nieuwburg, E.; Cremers, A.; and Neerincx, M. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291: 103404.
- Van Looveren, A.; Klaise, J.; Vacanti, G.; and Cobb, O. 2021. Conditional generative models for counterfactual explanations. *arXiv preprint arXiv:2101.10123*.
- Verma, S.; Dickerson, J.; and Hines, K. 2020. Counterfactual Explanations for Machine Learning: A Review.
- Villani, M.; Lockhart, J.; and Magazzeni, D. 2022. Feature importance for time series data: Improving kernelshap. *arXiv preprint arXiv:2210.02176*.
- Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31: 841.
- Wang, D.; Yang, Q.; Abdul, A.; and Lim, B. Y. 2019. Designing theory-driven user-centric explainable AI. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. ISBN 9781450359702.
- Wang, Z.; Yan, W.; and Oates, T. 2017. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, 1578–1585. IEEE.
- Warren, G.; Byrne, R. M. J.; and Keane, M. T. 2023. Categorical and Continuous Features in Counterfactual Explanations of AI Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI '23*, 171–187. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701061.

Warren, G.; Keane, M. T.; and Byrne, R. M. 2022. Features of Explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI. In *IJCAI-ECAI'22 Workshop: Cognitive Aspects of Knowledge Representation*.

Ye, L.; and Keogh, E. 2011. Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. *Data mining and knowledge discovery*, 22(1-2): 149–182.

Zhang, W.; and Lim, B. Y. 2022. *Towards Relatable Explainable AI with the Perceptual Process*, volume 1. Association for Computing Machinery. ISBN 9781450391573.

Zhao, W.; Oyama, S.; and Kurihara, M. 2021. Generating natural counterfactual visual explanations. In *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*, 5204–5205.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.