Counterfactual Explanations for Time Series Should be Human-Centered and Temporally Coherent in Interventions

Emmanuel C. Chukwu[1], Rianne M. Schouten[1], Monique Tabak[2], and Mykola Pechenizkiy[1]

[1] Data Mining Group, Eindhoven University of Technology, the Netherlands
[2] Biomedical Signals and Systems Group, University of Twente, the Netherlands

Abstract. Counterfactual explanations are increasingly proposed as interpretable mechanisms to achieve algorithmic recourse. However, current counterfactual techniques for time series classification are predominantly designed with static data assumptions and focus on generating minimal input perturbations to flip model predictions. This position paper argues that such approaches are fundamentally insufficient in clinical recommendation settings, where interventions unfold over time and must be causally plausible and temporally coherent. We advocate for a shift towards counterfactuals that reflect sustained, goal-directed interventions aligned with clinical reasoning and patient-specific dynamics. We identify critical gaps in existing methods that limit their practical applicability, specifically, temporal blind spots and the lack of user-centered considerations in both method design and evaluation metrics. To support our position, we conduct a robustness analysis of several state-of-the-art methods for time series and show that the generated counterfactuals are highly sensitive to stochastic noise. This finding highlights their limited reliability in real-world clinical settings, where minor measurement variations are inevitable. We conclude by calling for methods and evaluation frameworks that go beyond prediction flips. We emphasize the need for actionable, purpose-driven interventions that are feasible in real-world contexts for the users of such applications.

1 Introduction

In an ideal world, AI systems would be useful to all users in varying contexts and conditions. Yet, in many real-world applications, particularly in healthcare, users may seek not only to understand an unfavorable outcome but also to explore how they might achieve a more desirable one. For example, AI-powered digital health interventions have been shown to effectively support lifestyle changes in hypertension management [32], offering personalized guidance on activity, sleep, stress, and diet. Such systems illustrate the shift from static decision support to adaptive, goal-directed recommendations. This shift is central to the field of *algorithmic recourse*, which focuses on producing explanations and actionable suggestions for individuals impacted by automated decisions [27]. A distinction

arises between *contrastive explanations* (CEs), which describe why an outcome occurred instead of an alternative, and *consequential recommendations* (CRs), which suggest specific interventions that would alter the outcome. While both fall under the umbrella of counterfactual reasoning, CRs rely on stronger assumptions, such as causal invariance and stationarity: the idea that past actions would yield consistent results if repeated now [27, footnote p.95:4]. These assumptions raise practical concerns, especially in dynamic domains like healthcare.

A Counterfactual Explanation (CFE) is a post hoc interpretability method that identifies the minimal changes to input features required to produce the desired prediction [54]. CFEs hold the promise of explaining model behavior and supporting decision-making by answering contrastive (*Why A, not B?*) and counterfactual (*What if it had been B instead of A?*) queries [47,29,52]. For instance, a CFE might show a physician what ECG signal pattern would result in a different diagnostic outcome or demonstrate to a developer how specific input modifications influence model confidence. However, CFEs are increasingly framed not just as explanations, but as recommendations. In e-health applications such as AI coaching systems [32], suggestions of increasing physical activity or adjusting the diet can be interpreted as counterfactuals to achieve better results. In this case, CFEs are expected to serve as both a model interpretability tool and a behavior-guiding recommendation mechanism. This extension demands additional constraints such as *actionability*, *plausibility*, *sparsity*, and *diversity*, which have led to the development of various evaluation metrics [52,17,12].

We argue that the growing multiplicity of purposes, users, and evaluation metrics in CFE research creates ambiguity: it becomes unclear which methods are appropriate for which use cases. An effective CFE technique for system debugging may be misleading in a user-facing recommendation system. This ambiguity is particularly problematic in time series classification (TSC), where temporal dependencies make the generation of valid counterfactuals inherently more complex. Changing a past value can alter the plausibility of future values, and many counterfactual sequences, such as unrealistic changes to an ECG, may be infeasible for the user to act upon. These challenges require a careful re-evaluation of both the assumptions and evaluation criteria underpinning CFEs. This position paper argues that CFEs for time series must reflect temporally coherent, causally plausible, and user-centered interventions rather than just input perturbations to change predictions. At a minimum, CFEs intended for recommendation should be robust to realistic variability, such as noise from human behavior or imperfect execution of suggested actions. We critically assess the current state of CFE evaluation, particularly in time series settings, and describe how new metrics and frameworks should better capture the demands of user-centered, real-world recommendations.

2 Background

Consider a fixed black-box classifier $f : \mathcal{X} \to \mathcal{Y}$ that takes an observation $\mathbf{x} \in \mathcal{X}$ as input and returns a probability vector $[0, 1]^k$ for discrete output labels k. We

write $f_c(\mathbf{x})$ to denote the probability for class $c \in [1, k]$. The class label of the highest score is simply written as $\hat{y}$; the true label is y. In the tabular data case, $\mathbf{x} = (x_1, x_2, ..., x_d)$ is a d-dimensional vector taking values in $\mathcal{X}_1 \times ... \times \mathcal{X}_d$. Alternatively, we may consider $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_d) \in \mathbb{R}^{m \times d}$ to be a univariate ($d = 1$) or multivariate ($d > 1$) time series, where $\mathbf{x}_j = (x_j^1, x_j^2, ..., x_j^m)$ is a sequence of values of length m taken from the same feature space $\mathcal{X}_j$. The time in between successive measurements may be regular or irregular, may differ between features (channels), and could range from nanoseconds to several minutes, depending on the origin of the data. Obviously, incorporating temporal structure is of utmost importance in TSC tasks.

The field of algorithmic recourse is concerned with providing explanations and recommendations to individuals who are unfavorably impacted by the outcomes of black-box classifiers. Given an original input $\mathbf{x} \in \mathcal{X}$, a CFE is a modified input $\mathbf{x}'$ that changes the prediction from the current class y to a different but desired class y', with $y \neq y'$. Naturally, in generating CFEs, the model prediction $f(\mathbf{x}')$ should be valid, the modified input $\mathbf{x}'$ should be close in distance to the original instance $\mathbf{x}$. Following [24], a CFE is then computed as follows:

$$\underset{\mathbf{x}' \in \mathcal{X}}{\arg\min}\ cost(\mathbf{x}, \mathbf{x}')\ s.t. f(\mathbf{x}) \neq f(\mathbf{x}'), \tag{1}$$

where $cost(\mathbf{x}, \mathbf{x}') : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ is a distance metric in the input space. For differentiable classifiers, a constrained optimization approach cf. [54] can be used:

$$\underset{\mathbf{x}' \in \mathcal{X}}{\arg\min}\ loss(f(\mathbf{x}), f(\mathbf{x}')) + \lambda \cdot cost(\mathbf{x}, \mathbf{x}'). \tag{2}$$

Equation (2) expresses CFE generation as a differentiable loss, balancing two key properties: validity (left term) and proximity to the original input (right term). Section 3 expands this by discussing additional properties relevant for the use of recommendations, such as actionability, feasibility, sparsity, diversity, and robustness, each with associated evaluation metrics.

Equations (1) and (2) originate from the tabular setting, where instances $\mathbf{x} \in \mathbb{R}^d$. In TSC, Equation (1) is often retained to define CFEs as modified inputs that change the predicted class. However, instead of using constrained optimization as in Equation (2), most TSC methods rely on pattern mining approaches, such as nearest unlike neighbors [13,3], subsequence mining [6,4], or genetic algorithms [21,43]. However, cost metrics like $cost(\mathbf{x}, \mathbf{x}')$ are inherited from the tabular setting.

2.1 Experimental setup

We support our position with a small-scale evaluation of four representative CFE methods for time series classification (TSC): NG-CF [13], CoMTE [3], AB-CF [36], and TSEvo [21] (summarized in Section 3.1). These methods were selected for their methodological diversity, citation impact, and prior rigorous evaluations. CFEs for NG-CF, CoMTE, and TSEvo were generated using the TSInterpret

Table 1: Dataset characteristics.

Dataset	**Training**	**Test**	**Length** m	**#Classes** k	**#Features** d
ECG200	100	100	96	2	1
ECG5000	500	4500	140	5	1
TLECG	23	1139	82	2	1
Epilepsy	137	138	207	4	3

library [22], with outputs verified against their original repositories. AB-CF was executed directly using its official implementation. All methods were run with default parameters, except for NG-CF (using the Nearest Unlike Neighbor (NUN) option) and TSEvo (limited to 200 epochs).

We evaluate each method on four datasets from the UEA-UCR archive[3], varying in dimensionality (univariate vs. multivariate), number of classes (2–5), and class distribution. Two datasets represent clinical signals: ECG and neurological activity. Table 1 provides an overview of dataset characteristics. For each dataset-method pair, we randomly sampled $n = 100$ test instances, generating one CFE per instance targeting a class different from the predicted label (the second-highest probability in multiclass cases). CFEs were considered invalid if they predicted the original class, and not all methods discovered a CFE for every instance. Table 2 reports the number of discovered, invalid, and valid CFEs. All subsequent analyses use the valid subset, with $n_{\text{valid}} \leq 100$.

For each test instance $\mathbf{x}_i$ and its counterfactual $\mathbf{x}'_i$, we evaluate the confidence of the predicted class and its validity, defined as 1 if the predicted class matches the true (for $\mathbf{x}_i$) or target label (for $\mathbf{x}'_i$), and 0 otherwise. To assess robustness, we introduce Gaussian noise to both $\mathbf{x}_i$ and $\mathbf{x}'_i$, and re-evaluate confidence and validity scores. This scenario, known as *robustness against noisy execution (NE)* [24], simulates deviations in input due to imprecise user action or noisy data acquisition. Specifically, we apply the perturbation function $\mathbf{x}_{\text{gauss}} = \mathbf{x} + \epsilon \cdot \mathcal{N}(0, \sigma_x)$, where σ_x is the mean of the per-feature standard deviation of $\mathbf{x}$, and ϵ controls noise intensity. Noise is sampled to match the shape of $\mathbf{x}$ and added element-wise, approximating stochastic real-world variability. This aligns with recent formulations of robustness that consider both stochastic noise [42] and adversarial perturbations [14] as part of a broader robustness spectrum.

In our experiments, ϵ is varied from 0.0 to 1.2 to simulate increasing noise levels. We report the average confidence ($avgConf \in [0, 1]$) and validity ($avgVal \in [0, 1]$) across n_{valid} test instances. This evaluation captures the model's resilience and the reliability of its counterfactual explanations under perturbations. The change in validity under noise reflects the *invalidation rate* (IR) [42], quantifying robustness as the average divergence between clean and noisy predictions. Metric definitions are provided in Section A. Code and results are available at `https://github.com/Healthpy/cfe_tsc_pos`.

[3] `https://www.timeseriesclassification.com/`

Table 2: An overview of the number of invalid CFEs. Starting number is $n = 100$ instances.

Method	ECG200	ECG5000	Epilepsy	TLECG
AB-CF	40	37	9	0
COMTE	10	5	3	0
NG-CF	28	10	–	0
TSEVO	10	5	3	0

3 Related work

Interpretability and explainability are critical to building trust in machine learning, particularly in high-stakes domains like eHealth[9,1]. CFEs provide local interpretability by identifying how an input must change to alter a model's prediction [29]. However, in this paper, we emphasize that explaining model behavior is not equivalent to providing actionable recommendations for recourse. Depending on the system and the user context, CFEs must satisfy more than just validity (producing the desired prediction) and proximity (similarity to the original input) [52]. Additional properties become essential, including sparsity (minimal changes in characteristics), diversity (generating multiple and varied CFEs), and plausibility (changes that remain realistic or adhere to the datamanifold) [52]. Critically, *actionability*: whether users can interpret and implement the recommended changes must also be considered [17,50].

Furthermore, when using CFEs in real-world applications, it is crucial that they retain their validity under minor variations, such as small changes in the original input or slight deviations in the proposed change vector. Data are rarely pristine, and users may not be able to precisely execute the recommended changes. In this context, [24] evaluate the CFE methods in four types of robustness. First, robustness to *model changes* (MC) ensures that CFEs remain valid even when the model f is retrained or slightly altered. Second, robustness to *model multiplicity* (MM) addresses the variability between models that perform similarly but produce different predictions for the same input CFEs that should remain stable across this diversity. Third, robustness to *noisy execution* (NE) reflects the imperfect implementation of the recommendations; small deviations should still produce the desired result. Fourth, robustness to *input changes* (IC) requires similar inputs to produce similar CFEs, supporting fairness and interpretability when users have nearly identical profiles. An alternative categorization of [18] distinguishes the robustness to modifications in MC, IC, and CFE (related to NE in [24]). To assess these properties, [53] and [24] list several evaluation metrics, primarily developed for tabular data. Proximity is commonly measured using ℓ_1 and ℓ_2 norms, while sparsity is quantified using the ℓ_0 norm [25]. Validity and robustness are evaluated through changes in predicted probabilities $f_c(\mathbf{x})$, such as *validity after retraining* (VaR) or *validity after perturbation* (VaP) [24]. Additional techniques include adversarial perturbation analysis, noise-based sampling, and

diversity evaluations [33]. However, these metrics, rooted in static data, may not fully address real-world temporal and practical challenges.

In the context of time series, evaluation strategies are generally similar to those for tabular data. Instead, in Section 4, we argue that these metrics may not be sufficient for evaluating CFE methods for TSC, since time series are inherently sequential and an input change in one value may not be achieved without changing the previous or the next value. To the best of our knowledge, no survey on CFE methods for TSC has addressed these implications, despite the growing recognition that contributions in this area must extend beyond algorithmic performance to real-world applicability. In this paper, we provide an overview of the existing TSC CFE methods in Table 3. The methods vary greatly in whether they modify single time points [21], segments [13,36], or entire channels [3]. Many methods build on pattern mining methodology such as motif discovery [38,6], shapelets [35,4,20,39], nearest unlike neighbors [3,13] and saliency maps [37,46]. Some methods rely on genetic algorithms [21,43]; others use gradient-based modifications of the input space [55,56].

Table 3: Overview of Counterfactual Explanation Methods for Time Series Classification. We distinguish methods based on Distance metrics (DiOp), Shapelets (ShOp), Gradients (GrOp), Adversarial learning (AlOp), Evolutionary algorithms (EvOp), Reinforcement learning (RLOp), and Causality (Causal). U/M refers to applicability in univariate or multivariate time series.

Paper	**Method**	**Mechanism**	**Category**	**U/M**
[3]	CoMTE	Channel substitution via greedy search	DiOp	M
[13]	NG-CF	Segment substitution based on NUNs	DiOp	U
[35]	SG-CF	Shapelet-guided transformation	ShOp	U, M
[4]	SETS	Shapelet coefficient modification	ShOp	M
[16]	TimeX	Barycenter averaging and saliency maps	DiOp	U, M
[37]	CELS	Saliency map-guided perturbations	GrOp	U
[34]	M-CELS	Saliency map-guided perturbations	GrOp	M
[55]	Glacier	Gradient search in original or latent space	GrOp	U
[20]	Time-CF	Shapelet extraction and TimeGAN generation	ShOp, AlOp	U, M
[56]	LatentCF++	Latent space perturbation with autoencoders	GrOp	U, M
[6]	DiscoX	Matrix Profile discord replacement	DiOp	U, M
[36]	AB-CF	Attention-based segment modification	DiOp	U, M
[5]	TeRCE	Temporal rule mining with shapelets	ShOp	M
[51]	G_{CF}	Conditional generative model	AlOp	U, M
[38]	MG-CF	Motif-based subsequence replacement	ShOp	U, M
[21]	TSEvo	Multi-objective evolutionary search	EvOp	U, M
[48]	CFWoT	RL-based sequential decision-making	RLOp	M
[31]	SPARCE	GAN-based sparse counterfactuals	AlOp	M
[46]	LASTS	Saliency maps, instance-based	Hybrid	U, M
[57]	CounTS	Variational Bayesian causal modeling	Causal	M
[43]	Sub-SpaCE	Genetic algorithms	EvOp	U

3.1 Detailed summary of 4 distinguished CFE methods for TSC

In this paper, we experimentally evaluate the performance of 4 CFE generation methods for TSC. First, NG-CF[13] is among one of the early strategies, including those based on K-Nearest Neighbors (KNN) or NUN, have inspired local perturbation methods such as Native Guides [13]. These methods are confined largely to univariate settings and are heavily reliant on simplistic heuristics like proximity and sparsity. It follows two steps: (i) retrieving the Native Guide, where the closest instance from a different class is selected, and (ii) adapting the Native Guide, where the instance is iteratively perturbed toward the query using distance metrics like dynamic time warping (DTW) or feature weight vectors (e.g., Class Activation Mapping) to modify key subsequences.

Second, CoMTE [3] is a multivariate extension. Although efficient, such methods ignore broader temporal and semantic structures, producing counterfactuals that may appear valid locally but lack global interpretability. CoMTE produces counterfactuals in a multivariate context by perturbing the channels of a time series using a heuristic method. They modified the initial approach of Wachter et al[54]. by substituting the point-wise distance function d with one that operates on a channel-by-channel basis.

Third, AB-CF [36] focuses on identifying and perturbing the most important segments of a time series using an attention mechanism. By narrowing attention to a small set of influential segments, AB-CF ensures that the generated counterfactuals are valid, sparse, and interpretable. The method first generates a pool of candidate subsequences and then selects one or more per time series to replace, based on their contribution to the model prediction, quantified using the Shannon entropy algorithm. This targeted modification strategy enables AB-CF to produce efficiently computed counterfactuals. Focusing on locality, the method struggles to represent overall temporal relationships, leading to counterfactuals valid locally but implausible globally.

Fourth, TSEvo[21] extends the concept of CFEs to both univariate and multivariate time series classification. It formulates an optimization problem that balances three key objectives: proximity, sparsity, and output distance. This multiobjective optimization problem is solved using a genetic algorithm. The algorithm employs crossover and mutation operations, applied both at the level of individual time series values and larger segments, allowing it to explore counterfactuals tailored for time series data. Unfortunately, the method is computationally expensive.

The effectiveness of CFE methods ultimately depend on the utility of their CFEs from a human-centered perspective. We argue that beyond technical correctness, counterfactuals are valuable only to the extent that they are interpretable, actionable, and robust in practical decision-making contexts.

4 CFEs for TSC have a temporal blind spot

It is obvious that in TSC tasks, the incorporation of temporal structure is of utmost importance, and valuable methods have been developed specifically for

this task [40]. However, we observe that such time awareness is lacking in methods generating CFEs for those same classification tasks. In addition, we observe that many existing evaluation metrics are borrowed from the tabular data case: These metrics suffer from a temporal blindspot which limits their real-world utility, particularly in longitudinal decision-making scenarios.

Consider Diameter App [19], an e-health application that aims to support patients with Type 2 diabetes to understand blood glucose fluctuations and the relationship with certain lifestyle behavior choices [45,8,10]. Here, it would be valuable to have CFEs suggesting a reduced carbohydrate intake at time t to prevent hyperglycemia. Although such a CFE may be locally valid, the intervention can inadvertently affect the glucose level at $t + 1$. A CFE should be temporally aware and consider downstream effects rather than instantaneous classification shifts. It thus seems more appropriate to first change the input sequence at one time point, then model its effect on the rest of the input sequence, and then evaluate the potential change in prediction.

A step in the right direction would be to consider modifying subsequences rather than changing the entire time series. Three of the four CFE methods that we consider in this paper take this approach (NG-CF [13], AB-CF [36] and TSEvo [21]). The fourth method, CoMTE [3], modifies entire channels (although it aims to reduce the number of channels). Nevertheless, it is questionable whether these kind of input changes could be considered feasible for end users like patients and individual persons: one could probably directly influence the blood glucose value at time t, but not again at time $t + 1$, $t + 2$, and so on, without taking into account the effect of earlier changes. In this regard, robustness becomes a critical property in time series CFEs, since minor perturbations at any point could affect consequent events and disturb the semantic integrity of the sequence and its classification accuracy. Sections 4.1 and 5.1 provide demonstrations.

In addition, we argue that CFEs for TSC should not be evaluated solely with evaluation metrics developed for the tabular data case. Indeed, in addition to evaluating the validity and confidence of CFE predictions, the formal definition of a CFE as given in Equation (1) persuades us to evaluate their $cost(\mathbf{x}, \mathbf{x}') \in \mathcal{X} \times \mathcal{X}$. Many existing CFE methods for TSC use common distance metrics such as the ℓ_1 and ℓ_2 norm to evaluate $cost(\mathbf{x}, \mathbf{x}')$ as proximity. For multivariate time series data, these metrics have high dimensionality. In addition, they do not reflect other aspects that strongly relate to the concept of proximity, such as whether two sequences have a similar periodicity or amplitude. We contend that other evaluation metrics are needed, not just for evaluating proximity but also for evaluating other important CFE properties such as sparsity. In Section 4.1, we demonstrate that small adjustments to existing metrics are insufficient.

4.1 Demonstration

Figure 1 compares the confidence of the predicted classes between CFEs and their original instances for the ECG200 dataset. Recall from Table 1 that ECG200 is a small dataset with $N = 100$ training examples, $k = 2$ classes, $d = 1$ feature and a sequence length of $m = 96$. Table 2 displays that many CFEs were not

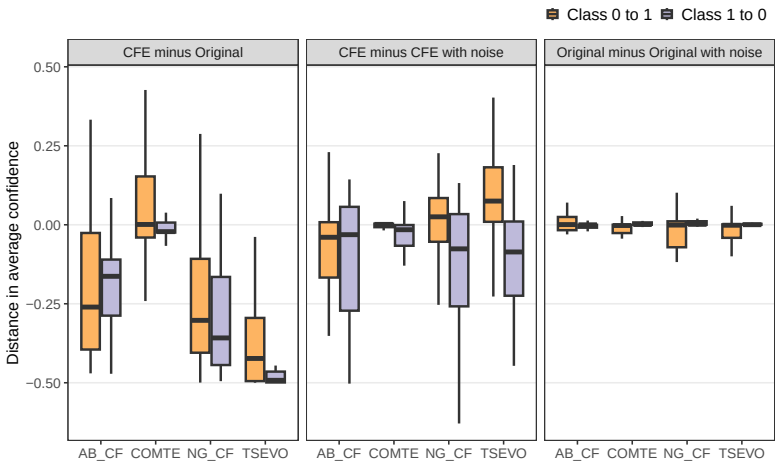
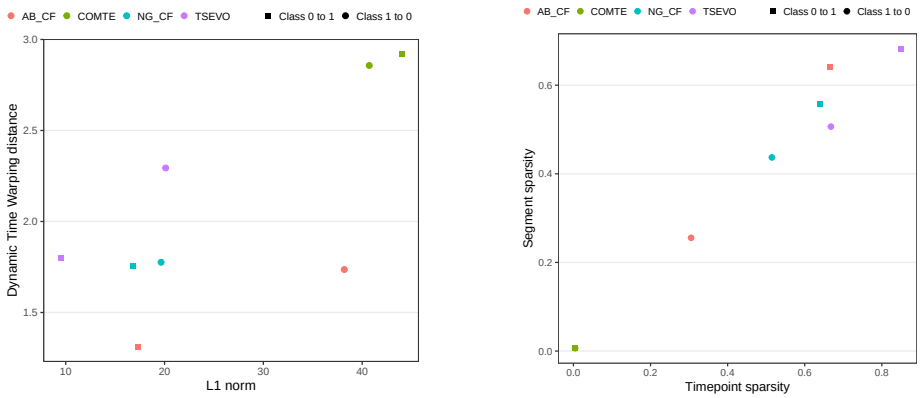


Fig. 1: Confidence distance between original and CFE (left), CFE with/without $\epsilon = 0.2$ Gaussian noise (center), and original input with/without noise (right), for n_{valid} CFEs from 4 methods on ECG200. Results are shown separately for CFEs targeting class 1 (orange) and class 0 (purple).

generated at all (for the AB-CF and NG-CF methods), or considered invalid (for all 4 CFE methods). We now discuss results in Figure 1 (see Table 4 in Section B.1 for detailed information) for the n_{valid} instances only.

First, we find that the classification model is robust to Gaussian noise with $\epsilon = 0.2$. This can be seen in the third, most right panel, where the difference in prediction probabilities between instances with and without noise are close to zero. Differences between CFE methods may occur due to randomness in the generated noise. In contrast to original input, adding Gaussian noise to CFEs notably affects the predicted class probability (center panel of Figure 1). For all four methods, prediction confidence generally decreases under noise, except in two cases. First, for NG-CF and TSEvo, confidence in predicting class 1 *increases* after noise is added, as seen in the orange boxplots on the right of the center panel (values greater than 0). This indicates that CFE robustness may vary between groups of instances, possibly due to training-related issues such as class imbalance. Such disparities are problematic in practice, as they suggest that some users may receive more (or less) stable counterfactuals. Evaluation protocols should therefore go beyond global metrics and assess fairness between subgroups, as discussed in further detail in Section 5.

Second, CoMTE exhibits minimal change in prediction confidence when Gaussian noise ($\epsilon = 0.2$) is added, clearly shown in both the center and left panels of Figure 1. Unlike other methods, CoMTE's CFEs retain similar confidence to the original inputs. This is likely because CoMTE relies on NUN substitution [3],



(a) Proximity (b) Sparsity

Fig. 2: Comparison of the average proximity and average sparsity for 4 CFE methods on the ECG200 dataset. (a) Proximity is measured with ℓ_1 norm (x-axis) and the DTW distance (y-axis). (b) Sparsity is measured by evaluating changes in time points (x-axis) and changes in entire segments (y-axis). Results are displayed separately for CFEs generated for class 1 (rather than 0, with square) and class 0 (rather than 1, with circle).

directly replacing the input with a near-instance from another class. Although this produces high-confidence counterfactuals by design, such CFEs may lack meaningful interpretability or personalization. Despite its robustness performance, the reliance of CoMTE on NUNs raises concerns. As shown by the green dots in Figure 2, these CFEs exhibit low sparsity and high proximity. We measure proximity (Figure 2a) using both ℓ_1 norm and Dynamic Time Warping (DTW), and sparsity (Figure 2b) using two metrics: the traditional count of changed input values and an adapted metric counting modified segments. The results are consistent across both formulations and reveal a near-linear trade-off between proximity and sparsity. This shows that high scores on traditional robustness or proximity metrics do not necessarily imply useful or interpretable counterfactuals.

4.2 Alternative Views

Unlike tabular data, time series are inherently sequential; modifying one time step may disrupt temporal dependencies or produce unrealistic patterns, even if the perturbation is sparse. Consequently, standard pointwise proximity metrics are insufficient: evaluation must consider global shape, alignment, and temporal dynamics. While several works adapt existing metrics for time series [13,3,5,43], we argue that current CFE evaluation strategies for TSC are fragmented: too many metrics exist, and too few account for temporal structure meaningfully.

A growing body of work calls for a more standardized and holistic evaluation of explainability methods. Bhattacharya et al. [7] advocate a framework that spans model, data, prediction, and user dimensions, stressing that evaluation must move beyond isolated metrics. Artelt et al. [2] recommend prioritizing plausibility over proximity to improve fairness and robustness, while Nguyen et al. [41] propose the AMEE framework, assessing the robustness of the explanation via perturbations between classifiers. Without such standardization, evaluation inconsistencies can lead to misleading conclusions and unreliable deployment [52,11,44]. Tools like TSInterpret [22] support multiple CFE methods (e.g., TSEvo [21], NG-CF [13], CoMTE [3]) and offer visualizations, while XTSC-Bench [23] provides benchmarking across models and datasets. Kan et al. [25] further benchmark five TSC CFE methods, proposing new metrics tailored for time series. Yet, even with such tools, fixed evaluation sets, regardless of their temporal design, are insufficient on their own.

We argue that CFEs for TSCs must be developed and evaluated within the context of their intended application. For example, CFEs supporting ML debugging differ significantly from those offering behavioral recommendations to patients or treatment insights to clinicians. Time series inputs from wearables (e.g., ECG, HR, step counts) vary in modifiability: while physical activity can be altered, users cannot change their ECG. Thus, clinical feasibility cannot be assumed from algorithmic success. It is essential to ask: who is the user, what can they act upon, and in what context will CFEs be used? Without these considerations, even CFEs that score well on existing metrics may fail in practice.

5 CFEs Should be User-Centered and Recourse-Aware

Classic CFE algorithms are typically task-centered, focused on flipping model predictions with minimal input changes. However, CFEs are increasingly used not only for interpretation, but also as a foundation for recommendations, guiding users on how to achieve a desired outcome [52,17,28]. This requires that CFEs align with user goals, prior knowledge, and domain-specific constraints. Effective recommendations must account for the user's ability to interpret and implement changes. Without this user-centered perspective, CFEs may remain technically valid, yet practically irrelevant or misleading.

5.1 Demonstration

We illustrate this issue using results from the ECG200 dataset. Additional results for ECG5000, Epilepsy and TwoLeadECG are provided in the Appendix B, with highlights in Section B.2 only when deviations occur. Figure 3 presents a robustness analysis under increasing Gaussian noise. Across all four methods, the original examples consistently maintain a higher validity under noise than their counterfactuals. TSEVO and NG-CF show sharp drops in counterfactual validity (below 0.4 at $\epsilon = 0.4$), while their originals remain above 0.65 even at $\epsilon = 1.2$. COMTE shows a gradual decline, though originals still outperform

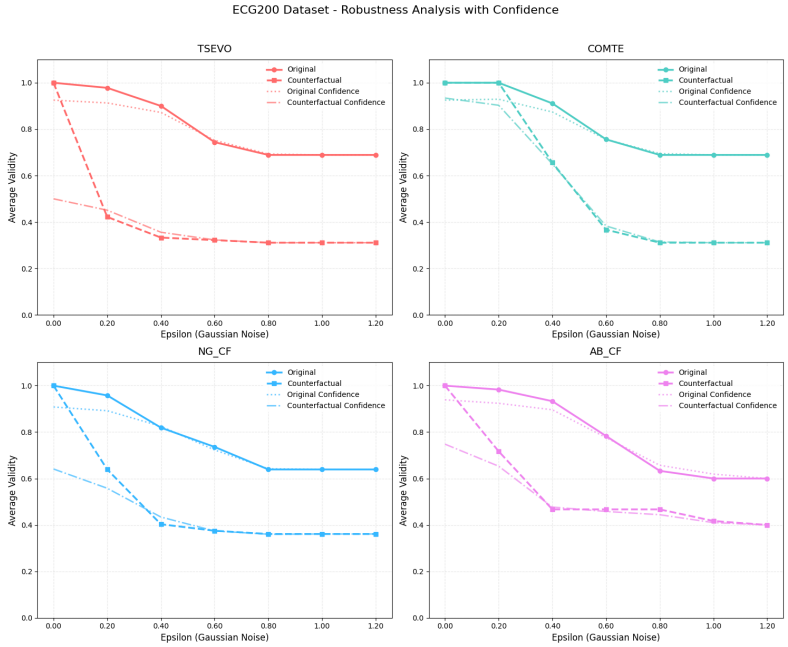


Fig. 3: Robustness analysis on the ECG200 dataset showing average validity under increasing Gaussian noise for four counterfactual methods. Solid lines represent original instances; dashed lines represent counterfactuals. Confidence levels are shown with dotted and dash-dotted lines. Original examples consistently show greater robustness to noise than their counterfactual counterparts.

counterfactuals. AB-CF CFEs degrade rapidly and the originals again show greater resilience. The confidence curves further confirm this robustness gap between the originals and CFEs.

Figure 1 reveals another concern: The counterfactual robustness varies depending on the starting class of the user. For all methods, confidence differences between original and counterfactual instances show consistent asymmetries when split by recourse direction (class 0 to 1 vs. 1 to 0). This implies that the user experience with CFEs is not uniform; some users receive more robust and actionable recourse than others. This asymmetry raises concerns about the consistency of the model and CFEs and highlights the need for evaluation protocols that account for such directional biases.

5.2 Alternative views

This raises a critical question: are CFEs even suitable for achieving algorithmic recourse in the first place? It may be more appropriate to reconsider the goal of recourse from the perspective of practical utility, rather than theoretical optimality or elegance. While many CFE methods have been proposed, their

real-world applicability remains underexplored. CFEs may be useful in certain scenarios but should not be treated as a universal solution for recourse. Recent work has shown that when unknown attributes of the user are introduced, CFEs may offer limited benefit compared to simpler explanations like reason codes [49]. This highlights the importance of evaluating interpretability methods from a user-centered perspective, rather than relying solely on model-centric metrics.

To better formalize the distinction between explanations and actionable recommendations, Karimi et al. [27] differentiate between input-based CEs and recourse through feasible actions (CR). While CEs are changes in the input space defined as $\mathbf{x}^{CE} = \mathbf{x} + \boldsymbol{\delta}$ using a distance function like $dist(\mathbf{x}, \mathbf{x}^{CE})$, CRs are based on interventions in a set of feasible actions $\mathcal{A}(\mathbf{x})$:

$$\underset{\mathbf{a} \in \mathcal{A}(\mathbf{x})}{\arg\min}\ cost(\mathbf{x}, \mathbf{a}) \quad \text{s.t.} \quad f(\mathbf{x}) \neq f(\mathbf{x}^{CE}). \tag{3}$$

Here, actions $\mathbf{a}$ are modeled as interventions in a structural causal model, where modifying one variable may affect others. Thus, the cost of changing the input vector ($cost(\mathbf{x}, \mathbf{x}')$) can differ significantly from the cost of taking actions in the real world ($cost(\mathbf{x}, \mathbf{a})$). In causal settings, the counterfactual instance becomes *structural counterfactual* $\mathbf{x}^{SCF}(\mathbf{a}, \mathbf{x})$, where actions $\mathbf{a}$ are grounded in the causal graph.

However, defining actions through structural models requires access to causal knowledge, which may not always be available. This motivates alternative formulations that encode feasibility directly into the optimization objective. For example, Ates et al. [3] introduce a binary diagonal matrix A in the cost function, where $A_{jj} = 1$ if feature j is allowed to change. SG-CF [35], another optimization-based method for TSC, incorporates a shapelet-guided loss that balances validity, proximity, sparsity, and contiguity. These approaches embed constraints into $cost(\mathbf{x}, \mathbf{x}')$, enabling flexible but structured recourse generation. However, for multivariate time series, the change vector grows from $\boldsymbol{\delta} \in \mathbb{R}^d$ to $\mathbb{R}^{d \times m}$, posing the question whether traditional formulations of $cost(\cdot, \cdot)$ remain meaningful.

Ultimately, whether by redefining the input space or integrating richer feasibility constraints, we argue that algorithmic recourse must be designed in collaboration with the end user. In this direction, Knijnenburg et al. [30] proposed the User-Centric Evaluation Framework for recommender systems, highlighting six key dimensions: objective and subjective system aspects, user experience, interaction, personal characteristics, and situational context. Recently, Donoso et al. [15] extended this framework to the evaluation of explainable AI (XAI) systems. Following this line of reasoning, we contend that the development and evaluation of CFEs, especially for time series, should be tightly coupled with empirical studies of user experience to ensure that recourse methods are not only technically sound but also practically usable.

6 Conclusion

This position paper identifies two key limitations in current CFE methods for time series: the lack of user-centered design and a temporal blind spot in the

evaluation. We show that strong performance on standard metrics does not imply practical utility. A method may be robust and valid, but it produces CFEs with high proximity or implausible suggestions that are unsuitable for end-users, such as patients. In addition, class-specific disparities in CFE performance reveal issues of fairness that could be overlooked. Current evaluations also ignore the sequential nature of time series in a real-world context. Local input changes may disrupt temporal coherence, leading to unrealistic or unfeasible recommendations. To be actionable, CFEs must reflect user context and respect temporal structure. We call for methods and evaluation frameworks that integrate both: moving beyond prediction flips to feasible, goal-directed interventions.

References

1. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F.: Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion **99**, 101805 (Nov 2023). https://doi.org/10.1016/j.inffus.2023.101805, https://www.sciencedirect.com/science/article/pii/S1566253523001148
2. Artelt, A., Vaquet, V., Velioglu, R., Hinder, F., Brinkrolf, J., Schilling, M., Hammer, B.: Evaluating Robustness of Counterfactual Explanations (Jul 2021). https://doi.org/10.48550/arXiv.2103.02354, http://arxiv.org/abs/2103.02354, arXiv:2103.02354 [cs]
3. Ates, E., Aksar, B., Leung, V.J., Coskun, A.K.: Counterfactual Explanations for Multivariate Time Series. In: 2021 International Conference on Applied Artificial Intelligence (ICAPAI). pp. 1–8 (May 2021). https://doi.org/10.1109/ICAPAI49758.2021.9462056, https://ieeexplore.ieee.org/document/9462056
4. Bahri, O., Boubrahimi, S.F., Hamdi, S.M.: Shapelet-Based Counterfactual Explanations for Multivariate Time Serie s. arXiv.org (2022). https://doi.org/10.48550/ARXIV.2208.10462, https://arxiv.org/abs/2208.10462
5. Bahri, O., Li, P., Boubrahimi, S.F., Hamdi, S.M.: Temporal Rule-Based Counterfactual Explanations for Multivariate Time Series. In: 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 1244–1249 (Dec 2022). https://doi.org/10.1109/ICMLA55696.2022.00200, https://ieeexplore.ieee.org/document/10069254
6. Bahri, O., Li, P., Filali Boubrahimi, S., Hamdi, S.M.: Discord-based counterfactual explanations for time series classificati on. Data Mining and Knowledge Discovery **38**(6), 3347–3371 (Aug 2024). https://doi.org/10.1007/s10618-024-01028-9, http://dx.doi.org/10.1007/s10618-024-01028-9
7. Bhattacharya, A., Verbert, K.: "How Good Is Your Explanation?": Towards a Standardised Evaluation Approach for Diverse XAI Methods on Multiple Dimensions of Explainability. In: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization. pp. 513–515 (2024)
8. den Braber, N., Vollenbroek-Hutten, M.M.R., Westerik, K.M., Bakker, S.J.L., Navis, G., van Beijnum, B.J.F., Laverman, G.D.: Glucose Regulation Beyond HbA1c in Type 2 Diabetes Treated With Insulin: Real-World Evidence From the DIALECT-2 Cohort. Diabetes Care **44**(10), 2238–2244 (Jul 2021). https://doi.org/10.2337/dc20-2241

9. Chou, Y.L., Moreira, C., Bruza, P., Ouyang, C., Jorge, J.: Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. Information Fusion **81**, 59–83 (May 2022). https://doi.org/10.1016/j.inffus.2021.11.003, https://www.sciencedirect.com/science/article/pii/S1566253521002281
10. Danne, T., Nimri, R., Battelino, T., Bergenstal, R.M., Close, K.L., DeVries, J.H., Garg, S., Heinemann, L., Hirsch, I., Amiel, S.A., Beck, R., Bosi, E., Buckingham, B., Cobelli, C., Dassau, E., Doyle, F.J., Heller, S., Hovorka, R., Jia, W., Jones, T., Kordonouri, O., Kovatchev, B., Kowalski, A., Laffel, L., Maahs, D., Murphy, H.R., Nørgaard, K., Parkin, C.G., Renard, E., Saboo, B., Scharf, M., Tamborlane, W.V., Weinzimer, S.A., Phillip, M.: International Consensus on Use of Continuous Glucose Monitoring. Diabetes Care **40**(12), 1631–1640 (Dec 2017). https://doi.org/10.2337/dc17-1600
11. Daza, E.J.: Causal Analysis of Self-tracked Time Series Data Using a Counterfactual Framework for N-of-1 Trials*. Methods of Information in Medicine **57**(S 1), e10–e21 (May 2018). https://doi.org/10.3414/ME16-02-0044, http://www.thieme-connect.de/DOI/DOI?10.3414/ME16-02-0044, publisher: Schattauer GmbH
12. Del Ser, J., Barredo-Arrieta, A., Díaz-Rodríguez, N., Herrera, F., Saranti, A., Holzinger, A.: On generating trustworthy counterfactual explanations. Information Sciences **655**, 119898 (Jan 2024). https://doi.org/10.1016/j.ins.2023.119898, https://www.sciencedirect.com/science/article/pii/S0020025523014834
13. Delaney, E., Greene, D., Keane, M.T.: Instance-Based Counterfactual Explanations for Time Series Classification. In: Sánchez-Ruiz, A.A., Floyd, M.W. (eds.) Case-Based Reasoning Research and Development. pp. 32–47. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-86957-1_3
14. Dominguez-Olmedo, R., Karimi, A.H., Schölkopf, B.: On the Adversarial Robustness of Causal Algorithmic Recourse. In: Proceedings of the 39th International Conference on Machine Learning. pp. 5324–5342. PMLR (Jun 2022), https://proceedings.mlr.press/v162/dominguez-olmedo22a.html, iSSN: 2640-3498
15. Donoso-Guzmán, I., Ooge, J., Parra, D., Verbert, K.: Towards a comprehensive human-centred evaluation framework for explainable AI. In: World Conference on Explainable Artificial Intelligence. pp. 183–204. Springer (2023)
16. Filali Boubrahimi, S., Hamdi, S.M.: On the Mining of Time Series Data Counterfactual Explanations using Barycenters. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 3943–3947. CIKM '22, Association for Computing Machinery, New York, NY, USA (Oct 2022). https://doi.org/10.1145/3511808.3557663, https://dl.acm.org/doi/10.1145/3511808.3557663
17. Guidotti, R.: Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery (Apr 2022). https://doi.org/10.1007/s10618-022-00831-6, https://doi.org/10.1007/s10618-022-00831-6
18. Guyomard, V., Fessant, F., Guyet, T., Bouadi, T., Termier, A.: Generating robust counterfactual explanations (Apr 2023). https://doi.org/10.48550/arXiv.2304.12943, http://arxiv.org/abs/2304.12943, arXiv:2304.12943 [cs] version: 1
19. Hietbrink, E.A.G., Nijeweme-d'Hollosy, W.O., Middelweerd, A., Konijnendijk, A.A.J., Schrijver, L.K., Voorde, A.S.t., Fokkema, E.M.S., Laverman, G.D., Vollenbroek-Hutten, M.M.R.: A Digital Coach (E-Supporter 1.0) to Support Physical Activity and a Healthy Diet in People With Type 2 Diabetes: Acceptability and

Limited Efficacy Testing. JMIR Formative Research **7**(1), e45294 (Jul 2023). https://doi.org/10.2196/45294, https://formative.jmir.org/2023/1/e45294, company: JMIR Formative Research Distributor: JMIR Formative Research Institution: JMIR Formative Research Label: JMIR Formative Research Publisher: JMIR Publications Inc., Toronto, Canada

20. Huang, Q., Chen, W., Bäck, T., van Stein, N.: Shapelet-based Model-agnostic Counterfactual Local Explanations for Time Series Classification (Feb 2024). https://doi.org/10.48550/arXiv.2402.01343, http://arxiv.org/abs/2402.01343, arXiv:2402.01343 [cs]
21. Höllig, J., Kulbach, C., Thoma, S.: TSEvo: Evolutionary Counterfactual Explanations for Time Series Classification. In: 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 29–36 (Dec 2022). https://doi.org/10.1109/ICMLA55696.2022.00013, https://ieeexplore.ieee.org/document/10069160
22. Höllig, J., Kulbach, C., Thoma, S.: TSInterpret: A unified framework for time series interpretability (Aug 2022), http://arxiv.org/abs/2208.05280, arXiv:2208.05280 [cs]
23. Höllig, J., Thoma, S., Grimm, F.: XTSC-Bench: Quantitative Benchmarking for Explainers on Time Series Classification (2023). https://doi.org/10.48550/ARXIV.2310.14957, https://arxiv.org/abs/2310.14957, publisher: [object Object] Version Number: 1
24. Jiang, J., Leofante, F., Rago, A., Toni, F.: Robust Counterfactual Explanations in Machine Learning: A Survey. vol. 9, pp. 8086–8094 (Aug 2024). https://doi.org/10.24963/ijcai.2024/894, https://www.ijcai.org/proceedings/2024/894, iSSN: 1045-0823
25. Kan, Z., Rezaei, S., Liu, X.: Benchmarking Counterfactual Interpretability in Deep Learning Models for Time Series Classification (Oct 2024). https://doi.org/10.48550/arXiv.2408.12666, http://arxiv.org/abs/2408.12666, arXiv:2408.12666 [cs]
26. Karim, F., Majumdar, S., Darabi, H., Chen, S.: LSTM Fully Convolutional Networks for Time Series Classification. IEEE Access **6**, 1662–1669 (2018). https://doi.org/10.1109/ACCESS.2017.2779939, http://arxiv.org/abs/1709.05206, arXiv:1709.05206 [cs]
27. Karimi, A.H., Barthe, G., Schölkopf, B., Valera, I.: A survey of algorithmic recourse: contrastive explanations and consequential recommendations. ACM Computing Surveys **55**(5), 1–29 (2022)
28. Karimi, A.H., Schölkopf, B., Valera, I.: Algorithmic recourse: from counterfactual explanations to interventions. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. p. 353–362. FAccT '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445899, https://doi.org/10.1145/3442188.3445899
29. Keane, M.T., Smyth, B.: Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI). In: Watson, I., Weber, R. (eds.) Case-Based Reasoning Research and Development. pp. 163–178. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58342-2_11
30. Knijnenburg, B.P., Willemsen, M.C.: Evaluating recommender systems with user experiments. In: Recommender systems handbook, pp. 309–352. Springer (2015)
31. Lang, J., Giese, M.A., Ilg, W., Otte, S.: Generating Sparse Counterfactual Explanations for Multivariate Time Series. In: Iliadis, L., Papaleonidas, A., Angelov, P., Jayne, C. (eds.) Artificial Neural Networks and Machine Learning –

ICANN 2023. pp. 180–193. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-44223-0_15
32. Leitner, J., Chiang, P.H., Agnihotri, P., Dey, S.: The Effect of an AI-Based, Autonomous, Digital Health Intervention Using Precise Lifestyle Guidance on Blood Pressure in Adults With Hypertension: Single-Arm Nonrandomized Trial. JMIR Cardio **8**, e51916 (May 2024). https://doi.org/10.2196/51916, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11167324/
33. Leofante, F., Potyka, N.: Promoting Counterfactual Robustness through Diversity (Dec 2023). https://doi.org/10.48550/arXiv.2312.06564, http://arxiv.org/abs/2312.06564, arXiv:2312.06564 [cs]
34. Li, P., Bahri, O., Boubrahimi, S.F., Hamdi, S.M.: M-CELS: Counterfactual Explanation for Multivariate Time Series Data Guided by Learned Saliency Maps (Nov 2024). https://doi.org/10.48550/arXiv.2411.02649, http://arxiv.org/abs/2411.02649, arXiv:2411.02649 [cs]
35. Li, P., Bahri, O., Boubrahimi, S.F., Hamdi, S.M.: SG-CF: Shapelet-Guided Counterfactual Explanation for Time Series Classification. In: 2022 IEEE International Conference on Big Data (Big Data). pp. 1564–1569 (Dec 2022). https://doi.org/10.1109/BigData55660.2022.10020866, https://ieeexplore.ieee.org/document/10020866
36. Li, P., Bahri, O., Boubrahimi, S.F., Hamdi, S.M.: Attention-Based Counterfactual Explanation for Multivariate Time Series. In: Wrembel, R., Gamper, J., Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) Big Data Analytics and Knowledge Discovery. pp. 287–293. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-39831-5_26
37. Li, P., Bahri, O., Boubrahimi, S.F., Hamdi, S.M.: CELS: Counterfactual Explanations for Time Series Data via Learned Saliency Maps. In: 2023 IEEE International Conference on Big Data (BigData). pp. 718–727 (Dec 2023). https://doi.org/10.1109/BigData59044.2023.10386229, https://ieeexplore.ieee.org/document/10386229
38. Li, P., Boubrahimi, S.F., Hamdi, S.M.: Motif-guided Time Series Counterfactual Explanations (Feb 2024). https://doi.org/10.48550/arXiv.2211.04411, http://arxiv.org/abs/2211.04411, arXiv:2211.04411 [cs]
39. Li, P., Hosseinzadeh, P., Bahri, O., Boubrahimi, S.F., Hamdi, S.M.: Reliable Time Series Counterfactual Explanations Guided by ShapeDBA. In: 2024 IEEE International Conference on Big Data (BigData). pp. 1574–1579 (Dec 2024). https://doi.org/10.1109/BigData62323.2024.10825447, https://ieeexplore.ieee.org/document/10825447, iSSN: 2573-2978
40. Middlehurst, M., Schäfer, P., Bagnall, A.: Bake off redux: a review and experimental evaluation of recent time series classification algorithms (Apr 2023), http://arxiv.org/abs/2304.13029, arXiv:2304.13029 [cs]
41. Nguyen, T.T., Nguyen, T.L., Ifrim, G.: Robust Framework for Explanation Evaluation in Time Series Classification (Oct 2023). https://doi.org/10.48550/arXiv.2306.05501, http://arxiv.org/abs/2306.05501, arXiv:2306.05501 [cs]
42. Pawelczyk, M., Datta, T., van-den Heuvel, J., Kasneci, G., Lakkaraju, H.: Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse (Oct 2023). https://doi.org/10.48550/arXiv.2203.06768, http://arxiv.org/abs/2203.06768, arXiv:2203.06768 [cs]
43. Refoyo, M., Luengo, D.: Sub-SpaCE: Subsequence-Based Sparse Counterfactual Explanations for Time Series Classification Problems. In: Longo, L., Lapuschkin, S., Seifert, C. (eds.) Explainable Artificial Intelligence. pp. 3–17. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-63800-8_1

44. Runge, J., Gerhardus, A., Varando, G., Eyring, V., Camps-Valls, G.: Causal inference for time series. Nature Reviews Earth & Environment **4**(7), 487–505 (Jul 2023). https://doi.org/10.1038/s43017-023-00431-y, https://www.nature.com/articles/s43017-023-00431-y, publisher: Nature Publishing Group
45. Schouten, R.M., Bueno, M.L.P., Duivesteijn, W., Pechenizkiy, M.: Mining sequences with exceptional transition behaviour of varying order using quality measures based on information-theoretic scoring functions. Data Mining and Knowledge Discovery **36**(1), 379–413 (Jan 2022). https://doi.org/10.1007/s10618-021-00808-x, https://doi.org/10.1007/s10618-021-00808-x
46. Spinnato, F., Guidotti, R., Monreale, A., Nanni, M., Pedreschi, D., Giannotti, F.: Understanding Any Time Series Classifier with a Subsequence-based Explainer. ACM Transactions on Knowledge Discovery from Data **18**(2), 36:1–36:34 (Nov 2023). https://doi.org/10.1145/3624480, https://dl.acm.org/doi/10.1145/3624480
47. Stepin, I., Alonso, J.M., Catala, A., Pereira-Fariña, M.: A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. Ieee Access **9**, 11974–12001 (2021)
48. Sun, X., Aoki, R., Wilson, K.H.: Counterfactual Explanations for Multivariate Time-Series without Training Datasets (May 2024). https://doi.org/10.48550/arXiv.2405.18563, http://arxiv.org/abs/2405.18563, arXiv:2405.18563 [cs]
49. Upadhyay, S., Lakkaraju, H., Gajos, K.Z.: Counterfactual Explanations May Not Be the Best Algorithmic Recourse Approach. In: Proceedings of the 30th International Conference on Intelligent User Interfaces. pp. 446–462. IUI '25, Association for Computing Machinery, New York, NY, USA (Mar 2025). https://doi.org/10.1145/3708359.3712095, https://dl.acm.org/doi/10.1145/3708359.3712095
50. Ustun, B., Spangher, A., Liu, Y.: Actionable Recourse in Linear Classification. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 10–19. FAT* '19, Association for Computing Machinery, New York, NY, USA (Jan 2019). https://doi.org/10.1145/3287560.3287566, https://dl.acm.org/doi/10.1145/3287560.3287566
51. Van Looveren, A., Klaise, J., Vacanti, G., Cobb, O.: Conditional Generative Models for Counterfactual Explanations (Jan 2021). https://doi.org/10.48550/arXiv.2101.10123, http://arxiv.org/abs/2101.10123, arXiv:2101.10123 [cs, stat]
52. Verma, S., Boonsanong, V., Hoang, M., Hines, K., Dickerson, J., Shah, C.: Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. ACM Computing Surveys p. 3677119 (Jul 2024). https://doi.org/10.1145/3677119, https://dl.acm.org/doi/10.1145/3677119
53. Verma, S., Boonsanong, V., Hoang, M., Hines, K.E., Dickerson, J.P., Shah, C.: Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review (Nov 2022). https://doi.org/10.48550/arXiv.2010.10596, http://arxiv.org/abs/2010.10596, arXiv:2010.10596 [cs, stat]
54. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR (Mar 2018). https://doi.org/10.48550/arXiv.1711.00399, http://arxiv.org/abs/1711.00399, arXiv:1711.00399 [cs]
55. Wang, Z., Samsten, I., Miliou, I., Mochaourab, R., Papapetrou, P.: Glacier: guided locally constrained counterfactual explanations for time series classification. Machine Learning (Mar 2024). https://doi.org/10.1007/s10994-023-06502-x, https://doi.org/10.1007/s10994-023-06502-x

56. Wang, Z., Samsten, I., Papapetrou, P.: Counterfactual Explanations for Survival Prediction of Cardiovascular ICU Patients. pp. 338–348 (Jun 2021). https://doi.org/10.1007/978-3-030-77211-6_38
57. Yan, J., Wang, H.: Self-Interpretable Time Series Prediction with Counterfactual Explanations. arXiv (2023). https://doi.org/10.48550/ARXIV.2306.06024, https://arxiv.org/abs/2306.06024, version Number: 3

A More on experimental setup

A.1 Classification model

Given a training dataset D of N input-label pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ with $\mathbf{x}_i$ a d-dimensional time series and y_i a true class label, we train an LSTM-FCN classifier f following [26]. LSTM-FCN is a hybrid model that combines Long Short-Term Memory (LSTM) networks, which are effective at modelling temporal dependencies in sequential data, with Fully Convolutional Networks (FCNs), which capture local and hierarchical patterns through convolutional layers. This architecture has shown competitive results in TSC tasks due to its ability to capture both sequential and spatial features [26]. The LSTM-FCN implemented in this study is shown in Figure 4, with parameter settings.

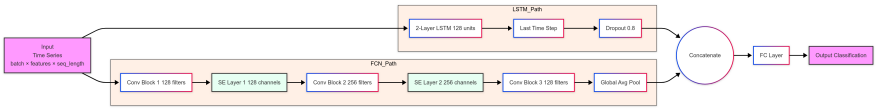


Fig. 4: LSTM-FCN following [26].

A.2 Evaluation metrics

Given a dataset of N input-label pairs $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$, where $y^{(i)}$ is the true class label, we define the following metrics:

Average Confidence: The confidence of prediction for sample i is defined as the probability assigned to the predicted class:

$$Conf^{(i)} = f_{y^{(i)}}(\mathbf{x}^{(i)}), \quad \text{where } y^{(i)} = \arg\max_{c \in [1,k]} f_c(\mathbf{x}^{(i)}).$$

Then, the average confidence across all samples is:

$$\text{Average Confidence}(avgConf) = \frac{1}{N} \sum_{i=1}^{N} f_{y^{(i)}}(\mathbf{x}^{(i)}).$$

Average Validity: The validity of a prediction is defined as 1 if the predicted class matches the true label, and 0 otherwise. When $\mathbf{x}$ is altered by noise, the validity is defined as 1 if the new predicted class matches the previous prediction.

$$\text{Val}^{(i)} = \begin{cases} 1 & \text{if } y^{(i)} = \tilde{y}^{(i)} \\ 0 & \text{otherwise} \end{cases} = \mathbb{1}\left(y^{(i)} = \tilde{y}^{(i)}\right), \text{where } \tilde{y} \text{ is the predicted class}$$

Then, the average validity is:

$$\text{Average Validity}(avgVal) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left(y^{(i)} = \tilde{y}^{(i)}\right).$$

A.3 Other evaluation metrics

Sparsity (ℓ_0): This implementation measures sparsity between a time series x and its counterfactual $\mathbf{x}$ at two levels: point-wise, using `np.isclose()` to compare individual values within a small tolerance, and segment-wise, by dividing the series into 10% segments, averaging each, and comparing those means. Both return a score between 0 and 1 via `np.mean()`, with higher values indicating greater similarity.

$$\ell_0 = \frac{1}{m} \sum_{h=1}^{m} \mathbb{1} \left(|\mathbf{x}_i - \mathbf{x}'_i| \leq e \right) \tag{4}$$

where $e =$ `1e-3`

ℓ_1**-norm**: (Manhattan) distance: Calculates the sum of absolute differences between corresponding points in the original time series $\mathbf{x}$ and counterfactual $\mathbf{x}'$.

$$\ell_1(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{m} |\mathbf{x}_i - \mathbf{x}'_i| \tag{5}$$

ℓ_2**-norm**: (Euclidean) distance: Computes the square root of the sum of squared differences between $\mathbf{x}$ and $\mathbf{x}'$, representing the straight-line distance. It emphasizes larger deviations.

$$\ell_2(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{i=1}^{m} (\mathbf{x}_i - \mathbf{x}'_i)^2} \tag{6}$$

DTW: Measures similarity between sequences that may vary in speed/time by finding the optimal alignment between the points in $\mathbf{x}$ and $\mathbf{x}'$.

$$d_{\mathrm{DTW}}(\mathbf{x}, \mathbf{x}') = \min_{\pi} \sqrt{\sum_{(i,j) \in \pi} (\mathbf{x}_i - \mathbf{x}'_j)^2} \tag{7}$$

where π is the optimal alignment path minimizing the total cost.

B More on experimental results

B.1 Detailed results

This section presents the results underlying Figures 1 and 2 in Tables 4 and 5.

B.2 Additional results

Table 4: Table underlying Figure 1. The table contains the mean and standard deviation of the distance in confidence between CFE and original instance (left), CFE with and without $\epsilon = 0.2$ Gaussian noise (center), and original instance with and without $\epsilon = 0.2$ Gaussian noise (right), for n_{valid} CFEs generated with 4 CFE methods (x-axis) on the ECG200 dataset. Results are displayed separately for CFEs generated for class 1 (rather than 0, in orange) and class 0 (rather than 1, in purple).

			CFE minus Orig.		CFE w/wo noise		Orig. w/wo noise	
Method	to-class	n_{valid}	mean	sd	mean	sd	mean	sd
AB-CF	1	24	-0.19	0.24	-0.00	0.08	-0.09	0.18
	0	36	-0.19	0.13	-0.02	0.06	-0.10	0.18
CoMTE	1	28	0.05	0.18	-0.01	0.05	-0.00	0.06
	0	62	-0.01	0.13	0.01	0.04	-0.04	0.08
NG-CF	1	26	-0.22	0.24	-0.05	0.13	-0.00	0.14
	0	46	-0.29	0.17	0.01	0.05	-0.13	0.19
TSEVO	1	28	-0.37	0.14	-0.04	0.10	0.08	0.19
	0	62	-0.45	0.11	-0.00	0.03	-0.11	0.16

Table 5: Table underlying Figure 2. The table contains the mean and standard deviation for 3 proximity metrics (ℓ_1, ℓ_2 and DTW) and 2 sparsity metrics (ℓ_0 and segment-based sparsity), over n_{valid} CFEs (see Table 2). Results are displayed for CFEs generated for class 1 (rather than 0) and class 0 (rather than 1) separately.

		ℓ_1		ℓ_2		DTW		Tp sparsity ℓ_0		Segment sparsity	
Method	to-class	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
AB_CF	1	17.31	14.00	3.37	1.84	1.31	0.74	0.67	0.21	0.64	0.23
	0	38.19	20.29	5.86	2.75	1.74	0.69	0.31	0.21	0.26	0.23
CoMTE	1	43.98	20.80	6.13	2.74	2.92	1.11	0.00	0.01	0.01	0.04
	0	40.72	6.79	5.45	0.87	2.86	0.90	0.00	0.01	0.01	0.02
NG_CF	1	16.86	17.04	3.50	2.41	1.75	1.39	0.64	0.26	0.56	0.30
	0	19.65	12.98	3.43	1.74	1.78	0.98	0.51	0.25	0.44	0.25
TSEVO	1	9.50	8.57	2.91	1.92	1.80	1.03	0.85	0.08	0.68	0.14
	0	20.12	14.17	4.52	2.14	2.29	1.24	0.67	0.19	0.51	0.20

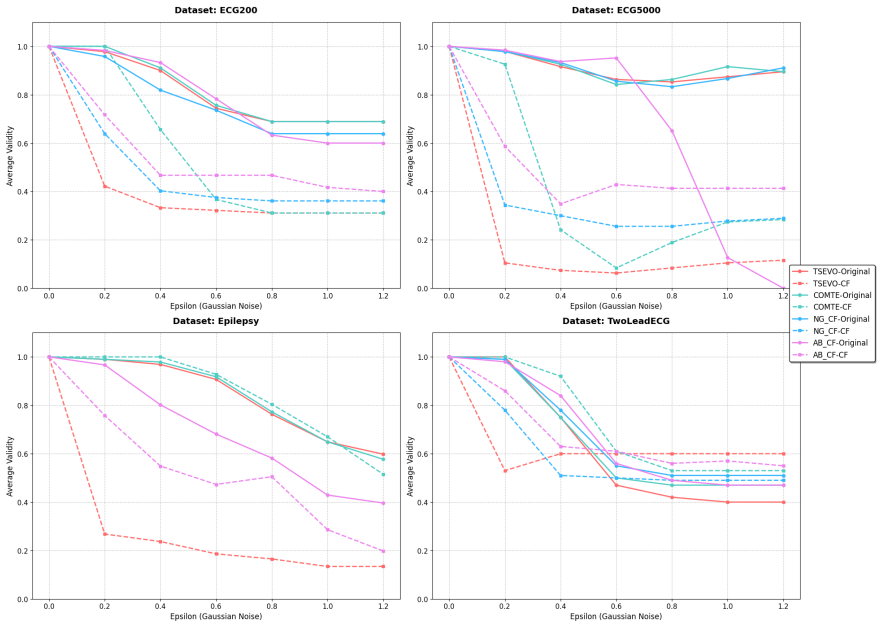


Fig. 5: Robustness analysis of the CFE methods for all the datasets used