Instance-based Counterfactual Explanations for Time Series Classification

Eoin Delaney[1,2,3], Derek Greene[1,2,3], and Mark T. Keane[1,2,3]

[1] School of Computer Science, University College Dublin, Dublin, Ireland
[2] Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland
[3] VistaMilk SFI Research Centre, Ireland
{eoin.delaney,derek.greene,mark.keane}@insight-centre.org

Abstract. In recent years, there has been a rapidly expanding focus on explaining the predictions made by black-box AI systems that handle image and tabular data. However, considerably less attention has been paid to explaining the predictions of opaque AI systems handling *time series* data. In this paper, we advance a novel model-agnostic, case-based technique – *Native Guide* – that generates counterfactual explanations for time series classifiers. Given a query time series, T_q, for which a black-box classification system predicts class, c, a counterfactual time series explanation shows how T_q could change, such that the system predicts an alternative class, c'. The proposed instance-based technique adapts existing counterfactual instances in the case-base by highlighting and modifying discriminative areas of the time series that underlie the classification. Quantitative and qualitative results from two comparative experiments indicate that Native Guide generates plausible, proximal, sparse and diverse explanations that are better than those produced by key benchmark counterfactual methods.

Keywords: Counterfactual Explanation · XCBR · Time Series

1 Introduction

In recent years, the predictive success of machine learning systems has been undermined by their lack of interpretability and beset by growing public disquiet about the fairness, accountability, and transparency of intelligent systems [1,19]. These challenges have led to major efforts in Explainable AI (XAI), where a raft of techniques has been developed to shed light on opaque predictions. Most of this research focuses on image and tabular data, with less attention being given to the explanation of time series data [39]. Explaining time series predictions, arguably, presents a whole new set of issues for XAI, due to the multi-dimensional nature of the data, strong feature dependencies, and the need to define the contexts where explanations could be used. In this paper, we advance an explainable case-based reasoning (XCBR) solution to this XAI problem.

Recently, a variety of CBR methods for XAI has been proposed (see [51] for a review). For image and tabular data, these XCBR techniques provide factual, example-based explanations (e.g., [23,53]), feature-importance explanations

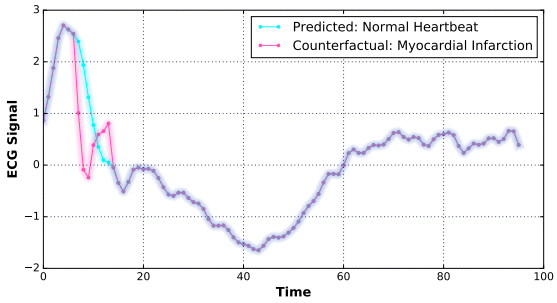
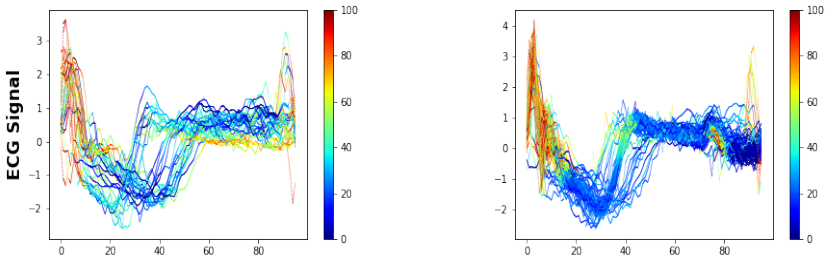


Fig. 1: A counterfactual instance explains the classification of an ECG signal. Here, a black-box's classification of a normal heartbeat is explained with a counterfactual, from *Native Guide*, showing an abnormal, heart-attack signal.

(CBR-LIME; [45]), and counterfactual explanations [25]. In particular, counterfactual explanations have become popular as a post-hoc explanation technique, with over 100 distinct methods being proposed [24]. However, few of these methods consider the explanation of time series [3,8,18,22]. Hence, we advance *Native Guide*, a novel model-agnostic explanation technique for time series classification (TSC) systems that provides counterfactual explanations for their predictions.

XAI's promise for time series classification (TSC). TSC has demonstrated significant promise in a variety of domains, including healthcare and food spectroscopy. However, there is a requirement to explain these decisions to end users. In healthcare, one practical application involves the classification of electrocardiogram signals, where explainable insights can aid medical practitioners in determining what portions of the time series are most informative for detecting abnormalities [42] (e.g. myocardial infarction). Figure 1 shows one such example, where a cardiologist might be shown the normal heartbeat of a patient along with a counterfactual signal as an explanation, basically saying "for this patient, their normal profile looks like this (purple-blue line), but if it changes to this counterfactual profile (purple-pink line), then they are experiencing an infarction" (see also Figure 2). Similar examples can be found in spectroscopy analyses when determining the provenance of different foods. For example, near-infrared spectrographs can distinguish between Arabica and Robusta coffee beans or honey from different regions [5] (see also Figure 7). By identifying portions of the time series that are discriminative for classification, cheaper sensors can be designed that only consider a small portion of the wider spectra. Similarly, in deep learning systems, explainable insights can uncover the portions of a time series that may be most prone to adversarial attacks and show them to model developers [11].

Outline of paper. In the remainder of this paper we first review the related work on time series XAI (Section 2). We then discuss the untapped promise of counterfactual explanations in TSC and the potential properties of good counterfactual explanations in this context (see Section 3). Next we describe the



(a) Normal Heartbeat (b) Myocardial Infarction

Fig. 2: Class Activation Maps (CAM) generated for the ECG200 dataset highlighting (in red) those areas of the time series which are most discriminative for a CNN Classifier. Here, the initial portion of the time series is most discriminative for both classes.

proposed technique (Section 4) and conduct comparative experiments to evaluate the quality of the explanations produced before discussing our results and suggesting promising avenues for future work (see Sections 5 & 6).

2 Related Work

The XAI literature on explaining time series classification has progressed along similar lines to XAI, in general; initial techniques focused on explanation through visualization and feature-importance, rather than on instance-based methods, such as factual or counterfactual explanations [26, 33].

Saliency methods typically visualize an extracted explanation weight vector $\boldsymbol{\omega}$ that captures discriminative areas of a time series for classification [39]. For example, Class Activation Maps (CAMs) [60] utilize these weight vectors to highlight areas of a time series that are most informative for classification decisions of deep neural networks (DNNs) [12, 57] (see Figure 2). Similarly, *shapelets* can also find discriminative subsequences of a time series that can either be directly extracted from a set of time series [58] or learned by minimizing an objective function [16]. Shapelets can capture relationships between features and are closely related to saliency maps as both techniques offer visual explanations for classification tasks. Some have considered using shapelets for contrastive explanation [18]. However, concerns have been raised about the interpretability of shapelets produced by the deployed *learning-shapelets* algorithm [56].

Feature-importance analyses are another method used to find relevant portions of a time series for use in explanation. Many state-of-the-art time series classifiers (e.g. Mr-SEQL [30]) transform the input data and deploy a linear model for classification, where $\boldsymbol{\omega}$ can be directly extracted from the regression coefficients of the classifier. Indeed, model-agnostic techniques such as LIME [46] and SHAP [35], can be used to compute $\boldsymbol{\omega}$ if it is not readily provided by the

base classifier [39]. However, concerns about the stability of these methods have been raised through examining how small perturbations can change the explanation [2, 39]. Schlegel et al. [50] tested the informativeness and robustness of different feature-importance techniques in time series classification. LIME was found to produce poor results across all evaluated datasets (a problem attributed to the high dimensionality of the data); in contrast, saliency-based approaches and SHAP were found to be more robust across different architectures.

More recently, a handful of *instance-based techniques* have been proposed to explain time series classification. *Prototypes* are instances that are maximally representative of a class and have demonstrated promise in producing global insights for time series classification in the healthcare domain [14] but they do not provide insights into the most discriminative areas of the time series. Case-based approaches using twin systems [23, 27, 49] have also been extended to time series data; Leonardi et al. [32] suggested mapping features from a DNN to a CBR system for interpretable haemodialysis classification. However, these techniques do not consider very popular counterfactual explanations. In an earlier unpublished version of the present paper [8], we considered how instances from the case-base could be retrieved for counterfactual explanation. However, we did not retrieve and integrate discriminative feature information in counterfactual generation, a significant novelty in the current method. Here, we advance a new XCBR method for generating good explanatory counterfactuals for any black-box time series classifier.

3 Good Counterfactuals for Time Series: Key Properties

There is a growing consensus that counterfactual explanations are causally informative [33, 43], psychologically effective [6, 9, 25, 36, 37], and legally compliant with respect to GDPR [55]. Arguably, counterfactuals provide more robust and informative explanations than feature-importance methods, such as LIME or SHAP [17]. Although it can be difficult to visualize counterfactual explanations for tabular data [38], in the time series domain their visualization is more straight-forward (see Figure 1). However, counterfactual XAI solutions for time series classification are rare (see e.g [3, 18, 22] for closest works) and we know of no existing XCBR solutions. Indeed, it is unclear if (i) existing counterfactual techniques for tabular/image data can be applied to time series data, (ii) the properties of good counterfactuals from tabular and image data transfer to the time series domain. In Section 4, we present the details of our novel XCBR method, but before that we first consider four potential properties of good counterfactual explanations for time series: namely, proximity, sparsity, plausibility, and diversity.

Proximity. Proximity refers to how close the to-be-explained query is to the generated counterfactual instance. Typically, closeness is measured using predefined distance metrics; close counterfactuals measured using Manhattan distance have been found to be informative [25, 38]. Following recent recommendations on evaluation [10, 21, 24], we use several different distance metrics and a rela-

tive counterfactual distance measure, to monitor the proximity of the generated counterfactual with respect to existing in-sample counterfactual solutions [25].

Sparsity. As noted by [38], counterfactual instances that change fewer features are preferred for informative explanations. Keane and Smyth [25] suggested that a sparsity of $\leq$2 feature differences was preferable for tabular data, on psychological grounds (that have been confirmed in recent user studies). However, the multi-dimensional nature of time series data means that a simple application of this idea is untenable. For image data, it has been argued that counterfactuals need to modify "semantically-meaningful" features instead of small, pixel-level features that may not be humanly-perceptible [28,48]. For time series data, it has been proposed that semantically-meaningful/discriminative information is contained in contiguous subsequences of the series [30,58]. So, by analogy, we argue that "good", sparse counterfactuals need to modify a single discriminative portion of the time series (i.e., a contiguous subsequence), rather than distributed, discrete time-points in series (e.g., see Figures 1 & 7).

Plausibility. Informative counterfactual explanations also need to be plausible [36]. Many suggest that proximity is a good proxy for plausibility [37], though others argue that falling within the data distribution is a better proxy [28,29,44,54]. Poyiadzi et al. [44] argue that plausible counterfactuals are representative of the underlying data distribution. Figure 6 shows some examples of implausible counterfactuals that are out-of-distribution, even though they have high proximity. Hence, in our evaluations we explore several novelty-detection algorithms to find better measures of plausibility/implausibility (see e.g. [20]).

Diversity. Mothal et al. [38] advanced the idea that a system should be able to produce multiple *diverse* explanations for a single query case. One advantage of this is that different users may find different explanations helpful [51]. So, our proposed method generates multiple explanations for a single test instance. However, we explicitly ensure that diversity should not come at the cost of either (i) plausibility or (ii) the loss of semantically meaningful information.

In the next section, we describe the proposed Native Guide method, before we consider a series of tests of it on several different datasets.

4 Native Guide: Counterfactual XAI for Time Series

Like other case-based XAI methods [25,27,31,41], at its core Native Guide relies upon existing instances in the training data, so-called *native guides* or nearest unlike neighbors (NUNs), that it retrieves and adapts to generate counterfactual explanations (see Figure 3). In this section, we outline the two main steps in the algorithm, after first describing the notation adopted.

Notation. Staying consistent with the notation of [15, 18], a time series $T = \{< t_1, t_2, ..., t_m >\}$ is an ordered set of real values, where m is the length. A time series data set $\mathbf{T} = \{T_1, T_2..., T_n\} \in \mathbb{R}^{n \times m}$ is a collection of such time series where each time series has a class label c forming a vector of class labels $\mathbf{Y} \in \mathbb{Z}$. Consider a black-box classifier $b(T)$ that takes a time series T as an input and predicts a probability output $P(\mathbf{Y}|T)$ over the label output space. Given a

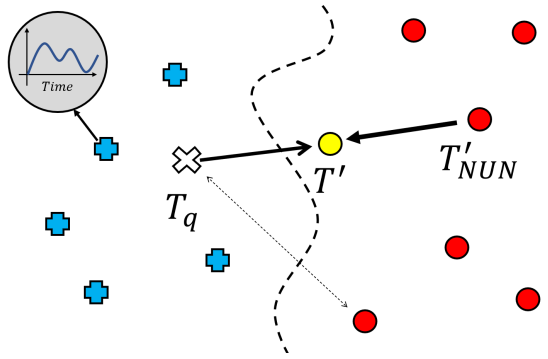


Fig. 3: A query time series T_q (X with solid arrow) and a nearest-unlike neighbor, T'_{NUN} (red circle with solid arrow) are used to guide the generation of counterfactual T' (see yellow circle) in a binary classification task. Another in-sample counterfactual (i.e., the *next* NUN; other red circle with dashed arrow) could also be used to generate another counterfactual for diverse explanations.

to-be-explained query time series T_q, with predicted label c from the black-box classifier (formally $b(T_q) = c$), a counterfactual explanation aims to find how T_q needs to change for the system to classify it alternatively, as c'. We refer to T' as a counterfactual explanation for T_q such that $b(T') = c'$. Although there are many candidate solutions for T', the method prioritizes those that meet the four key properties of proximity, sparsity, plausibility and diversity.

Step 1: Retrieve native guide. Given a query time series, T_q, find a counterfactual instance, T'_{Native}, that exists in the case-base. An example of one such instance is the query's nearest unlike neighbor (T'_{NUN}). In using these "native counterfactual" cases the method guarantees the explanation's *plausibility* as it is, by definition, within the distribution. However, such instances are not guaranteed to be sufficiently proximate to the query or, indeed, sparse, so an adaption step is necessary to generate the "explanatory counterfactual", T' (see Figure 3).

Step 2: Adapt native guide to generate counterfactual. To produce a more proximate explanatory counterfactual, T', the native guide, T'_{Native} is perturbed towards the to-be-explained query-case, T_q (see Figure 3). Typically, counterfactual methods use some L_p distance metric to guide this perturbation (such as Manhattan distance, [55]) and in time series where dynamic time warping (DTW) distance is often more appropriate an analogous averaging technique known as weighted dynamic barycentre averaging can be used [13]. In cases where we are explaining a deep-learner's predictions, the feature-weight vectors of the classifier, $\boldsymbol{\omega}$, can be used to perturb "semantically-meaningful" features of the time series, rather than the "raw" time series data, to guarantee sparsity[4].

[4] Note, SHAP can also be used to generate such vectors, if we are directly explaining any given model, rather than twinning.

Accordingly, using the feature-weights, the method seeks to modify contiguous, subsequences, rather than the whole time series, as follows:

$$T_q = \{< t_1, t_2, t_3, t_4, t_5..., tn >\} \text{ s.t. } b(T_q) = c$$
$$T' = \{< t_1, t'_2, t'_3, t'_4, t_5..., tn >\} \text{ s.t. } b(T') = c'$$

Specifically, the feature-weight vector, $\boldsymbol{\omega}$, can be extracted using techniques such as Class Activation Mapping in the case of DNNs (see e.g. Figure 2). Given T'_{Native} and $\boldsymbol{\omega}$, the most influential contiguous subsequence (measured by the magnitude of weights in $\boldsymbol{\omega}$) is identified and the corresponding region in T_q is replaced with these values. This process can be initialized using a small subsequence and the length of this subsequence can be iteratively incremented until $b(T') = c'$. In the very worst case scenario, the size of the subsequence will be equal to the length of T_q and the native counterfactual T'_{Native} is returned. This adaptation step improves the *proximity* and *plausibility* of the generated counterfactuals. Finally, *diversity* can also be met, as other in-sample instances can be used as guides (e.g., the *next* nearest unlike neighbor), to produce alternative counterfactual explanations for the original query (see Figure 3).

Table 1: Summary of TSC datasets used to evaluate counterfactual explanations

Dataset	**Train Size**	**Test Size**	**Length**	**Type**	**No.Classes**
CBF	30	900	128	Simulated	3
Chinatown	20	343	24	Traffic	2
Coffee	28	28	286	Spectro	2
ECG200	100	100	96	ECG	2
GunPoint	50	150	150	Motion	2

5 Testing Native Guide: Two Comparative Experiments

We test the Native Guide counterfactual method in two experiments evaluating how it meets the properties of good explanatory counterfactuals relative to two benchmark methods on 5 representative datasets. Our focus is on explaining a black-box fully convolutional neural network classifier (FCN). Experiment 1 assesses the proximity and sparseness of the counterfactuals generated. Experiment 2 examines the plausibility and diversity of the counterfactuals generated. Here, we describe the setup for these experiments in terms of the datasets, comparative benchmark methods and black-box classification system.

(I) Datasets. Five diverse datasets (binary and multiclass) from the UCR archive [7] (see Table 1) were used for the classification task. To encourage reproducibility we use the default train-test splits provided by the archive and provide all experimental code, fully detailing hyper-parameters[5].

[5] `https://github.com/e-delaney/Instance-Based_CFE_TSC`

(II) Baseline models. The performance of Native Guide was compared to two baseline models: the w-counterfactual and NUN-CF methods. The *w-counterfactual method* (w-CF) proposed by Wachter et al., [55] is a key benchmark method; it is the most cited counterfactual XAI method in the literature and many other methods are variants of it[6] [24]. It proposes that that counterfactuals can be generated by minimizing a loss function;

$$L(x, x', y', \lambda) = \lambda(b(x') - c')^2 + d(x, x') \tag{1}$$

$$\underset{x'}{argmin}\underset{\lambda}{max}L(x, x', c', \lambda) \tag{2}$$

The first collection of terms in this loss function encourage the output of the classifier b, to be close to the desired class c'. The λ parameter acts as a balancing term. The distance metric $d(x, x')$ measures the amount of change between the to-be explained instance x and the counterfactual candidate x'. A Manhattan distance weighted feature-wise with the inverse median absolute deviation (MAD) is typically used here in order to ensure the generation of sparse solutions that are robust to outliers [55]. One noted weakness of the λ parameter is that it tends to infinity raising stability issues in counterfactual generation [47]. The second method used, the *NUN-CF method*, can be viewed as a simplified variant of Native Guide; it simply uses the NUN for the query case directly, without any adaptation of deep or discriminative features (e.g., see [41]). This model represents a good comparison point as it allows us to see the contributions of the adaptation steps in Native Guide.

(III) Time series classifier. The black-box classifier used was a fully convolutional neural network (FCN)[7], by [57], a state of the art DNN architecture for time series classification (Figure 4). Notably, the Global Average Pooling (GAP) layer reduces the number of parameters in a neural network while enabling the use of the Class Activation Map (CAM) [60]; the latter highlights parts of the input time series that contribute the most to a given classification, enabling the extraction of $\boldsymbol{\omega}$. For each test/query instance, counterfactuals were generated using (i) *NUN-CF*, where the NUN, using a Euclidean distance measure, was selected [40], (ii) *w-Counterfactual* method (w-CF), as proposed by Wachter et al. [55], initializing it with $\lambda = 0.1$ and termination condition $P(c'|T) \geq 0.5$, minimizing the loss function in (see EQ1) with adaptive Nelder-Mead optimization (iii) *Native Guide*, using the closest in-sample counterfactual and the feature-importance vector, $\boldsymbol{\omega}$, given by the Class Activation Map (CAM) [39].

[6] We tried and failed in these tests, to use DiCE [38], a variant of w-CF with added constraints for diversity. We found that DiCE did not generate diverse counterfactuals within reasonable time-limits, suggesting that it is not well suited to high-dimensional time series data (even for shallower ANNs).

[7] Counterfactuals for other classifiers, such as MR-SEQL, were found but not reported.

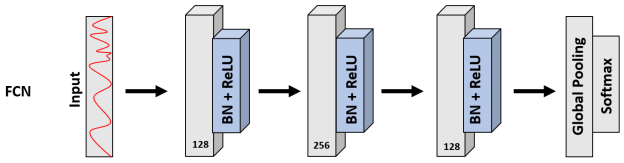
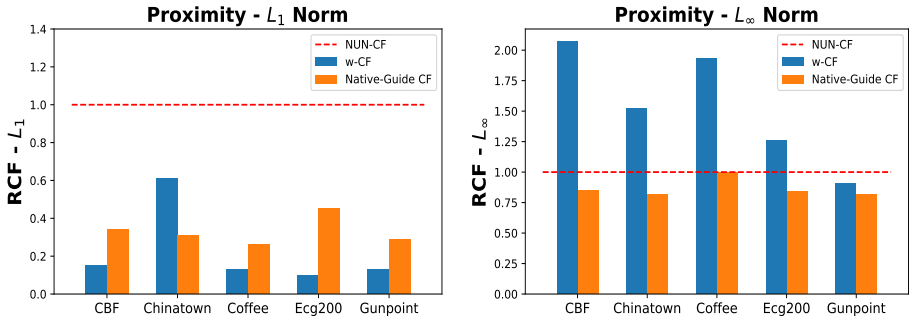


Fig. 4: A fully convolutional neural network (FCN) with three convolutional layers, batch normalization, ReLu activations and global average pooling preceding the final softmax layer enabling the use of a Class Activation Map (CAM) [57].

5.1 Experiment 1: Probing Proximity and Sparsity

This experiment compares the three counterfactual techniques on the five datasets, evaluating the counterfactuals produced in terms of proximity and sparsity. Proximity was evaluated using the relative counterfactual distance ($RCF = \frac{d(T_q,T')}{d(T_q,T'_{NUN})}$) enabling explicit comparisons to in-sample counterfactual instances (as suggested by [24, 25]). Basically, this measure determines whether the distance between the query and the generated counterfactual is closer than that between the query and its "naturally-occurring" NUN. As in some other studies [21], three distance metrics were used; (i) Manhattan Distance (ℓ_1 norm), (ii) Euclidean Distance (ℓ_2 norm), and (iii) Chebyshev Distance (ℓ_∞ norm).



(a) L_1 (b) L_∞

Fig. 5: A comparison of the proximity of query-counterfactual pairs relative to query-NUN pairs for five datasets. In (a) the generated counterfactual explanations are closer to the the query compared to the in-sample NUNs, in terms of ℓ_1 distance. Perhaps more interesting is the fact that the w-counterfactuals are consistently less close than the NUNs, in terms of ℓ_∞ norm. This effect may be due to erroneous spikes in the counterfactual explanations generated by this method.

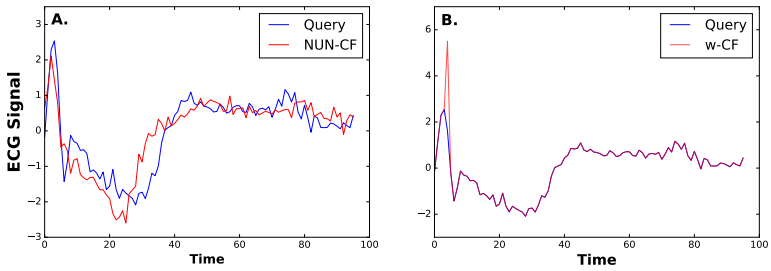


Fig. 6: Comparing counterfactuals (red line) for an ECG200 classification (blue line) generated by (a) NUN-CF and (b) w-CF. Here, NUN-CF fails to generate a proximate/sparse solution and w-CF's erratic spikes raise concerns about whether the counterfactual is out-of-distribution (see Figure 1 for comparison).

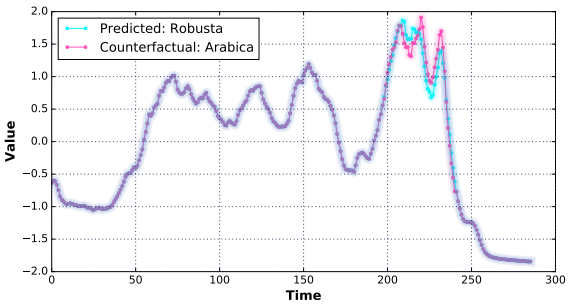


Fig. 7: A Native Guide counterfactual explanation for the coffee dataset. The method perturbs a contiguous subsequence corresponding to a semantically-meaningful and discriminative area of the spectrograph; this area provides information about the caffeine content of the coffee beans. Arabica coffee beans have a lower caffeine and chlorogenic acid content contributing to their finer taste and higher market value [5].

Results and discussion. Both Native Guide and the w-CF counterfactual explanations produce proximate explanations that are significantly closer to the query instance compared to the existing NUNS, on both ℓ_1 and ℓ_2 norms (Wilcoxon, $p<0.01$). In the case of w-CF this is somewhat unsurprising as it minimizes an ℓ_1-based distance-metric optimizing to generate close, sparse counterfactuals. w-CF is known to sometimes produce implausible counterfactual explanations [25, 28]. It is interesting to find that many of the perturbed features in its counterfactuals can be erratic spikes in the time series, reflecting out-of-distribution occurrences (see Figure 6b). Moreover, the explanations produced by w-CF often perturb several different features in non-contiguous locations of the time series, considering these values to be independent. Conversely, Native Guide constrains its perturbations to selected, contiguous subsequences produc-

ing counterfactual explanations that are more plausible and more meaningful (see e.g. Figures 1 and 7). These results indicate that good counterfactual explanations in the time series domain are not necessarily instances that are closest to the query, reflecting previous findings by Downs *et al* for tabular data [10]. Notably, the ℓ_∞ norm seems to be able to diagnose counterfactual instances with erratic feature values. Counterfactual explanations produced by Native Guide are more proximate in terms of ℓ_∞ norm, further suggesting that the w-counterfactuals may not be realistic. Admittedly, a more robust evaluation of plausibility should be considered when evaluating these methods. We turn to this issue in Expt. 2.

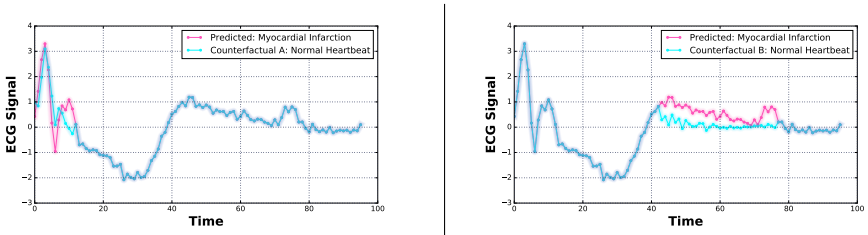
5.2 Experiment 2: Exploring Plausibility and Diversity

In this experiment, we aim to evaluate the *plausibility* of generated counterfactual explanations in time series using novelty detection algorithms to detect out-of-distribution (OOD) explanations. We implement the Local Outlier Factor Method [4, 20], Isolation Forest (IF) [34] and OC-SVM [52] (on both raw time and matrix profile [59] representations of the time series). We test if Native Guide can generate *diverse* explanations, when it uses alternative counterfactual instances as guides (see the other red instances shown in Figure 3). The datasets, classifier, and methods tested are identical to those in Experiment 1.

Results and discussion. The counterfactuals produced by the Native Guide are consistently more plausible than those generated by the benchmark, w-counterfactual method (w-CF; see Table 2). Results also confirm the hypothesis that proximity to the query is a poor heuristic for plausibility. One possible reason why case-based solutions produce more plausible counterfactual explanations is that they are grounded in the training data echoing previous findings by Laugel *et al* [29]. Unlike Native Guide, w-CF fails to perturb discriminative, meaningful subsequences [30, 58]. It is also interesting to note that different novelty detection algorithms produce very different results. For example the local outlier factor method was considerably less sensitive than the kernel-based

Table 2: Comparing the Native Guide (NG-CF) and w-Counterfactual (w-CF) models on plausibility using four OOD metrics (IF, LOF, OC-SVM, OC-SVM MP). Results indicate the percentage of generated counterfactuals that are out-of-distribution (n.b., lower scores are better and the best are highlighted in bold).

	Fully Convolutional Neural Network							
	IF		**LOF**		**OC-SVM**		**OC-SVM MP**	
Dataset	w-CF	NG-CF	w-CF	NG-CF	w-CF	NG-CF	w-CF	NG-CF
CBF	0.15	**0.09**	0.09	**0.00**	0.69	**0.50**	0.61	**0.34**
Chinatown	0.48	**0.37**	0.11	**0.00**	0.44	**0.07**	0.87	**0.22**
Coffee	0.41	**0.37**	0.04	0.04	0.25	**0.14**	0.43	**0.21**
ECG200	0.28	**0.26**	0.22	**0.02**	0.50	**0.16**	0.44	**0.13**
Gunpoint	0.23	**0.20**	**0.19**	0.23	0.18	**0.11**	0.57	**0.3**



(a) Counterfactual A (b) Counterfactual B

Fig. 8: Two diverse counterfactual explanations, generated by Native Guide, for the same query case based on perturbing different in-sample counterfactual cases.

techniques in detecting OOD explanations (see Table 2). Unlike many blind perturbation techniques, Native Guide has the ability to generate diverse counterfactual explanations (see Figure 8). This is particularly useful because (i) different users may prefer different explanations [51] and (ii) counterfactual explanations can also help humans to identify meaningful regions for classification (of which there may be many) [15]. For example, in the electrocardiogram domain we hypothesize that retaining counterfactual cases could help cardiologists to identify abnormalities that are useful for future problem scenarios. While one can evaluate diversity by monitoring feature wise distances between counterfactuals [38], the generated explanations may fail to satisfy domain constraints. Indeed, extensive user testing with experts will be an important avenue for future evaluation as novelty detection can be an imperfect proxy for plausibility.

6 Conclusion and Future Directions

In this paper a novel case-based technique, *Native Guide*, was proposed to provide proximate, sparse, plausible, and diverse counterfactual explanations for time series classification tasks. The method uses existing instances in the case-base to generate better counterfactual candidates. The technique is grounded in relevant evidence from the psychological and social sciences [6, 36] and can integrate explanation weight-vectors extracted from techniques such as Class Activation Mapping [60]. Comparative tests on diverse datasets from the UCR archive using a fully convolutional neural network, demonstrate that the explanatory counterfactuals produced by Native Guide are significantly better than (i) explanations that already existed in the case-base (from NUN-CF) and (ii) explanations produced by constraint-based optimisation techniques (from w-CF). The experiments also indicated that techniques designed for tabular data often failed to produce meaningful explanations in the time series domain. Native Guide generates new time series data which holds promise for data augmentation purposes [13]. Given the ubiquitous nature of time series data and the frequent requirement for explanation, it is clear that experiments with human users and CBR solutions have much to offer in future work.

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

1. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: NeurIPS. pp. 9505–9515 (2018)
3. Ates, E., Aksar, B., Leung, V.J., Coskun, A.K.: Counterfactual explanations for machine learning on multivariate time series data. arXiv preprint arXiv:2008.10781 (2020)
4. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. In: ACM SIGMOD. pp. 93–104 (2000)
5. Briandet, R., Kemsley, E.K., Wilson, R.H.: Discrimination of arabica and robusta in instant coffee by fourier transform infrared spectroscopy and chemometrics. Journal of agricultural and food chemistry **44**(1), 170–174 (1996)
6. Byrne, R.M.: Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In: IJCAI-19. pp. 6276–6282 (2019)
7. Dau, H.A., Bagnall, A., Kamgar, K., Yeh, C.C.M., Zhu, Y., Gharghabi, S., Ratanamahatana, C.A., Keogh, E.: The UCR time series archive. IEEE/CAA Journal of Automatica Sinica **6**(6), 1293–1305 (2019)
8. Delaney, E., Greene, D., Keane, M.T.: Instance-based counterfactual explanations for time series classification. arXiv preprint arXiv:2009.13211 (2020)
9. Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K., Dugan, C.: Explaining models: an empirical study of how explanations impact fairness judgment. In: International Conference on Intelligent User Interfaces. pp. 275–285 (2019)
10. Downs, M., Chu, J.L., Yacoby, Y., Doshi-Velez, F., Pan, W.: Cruds: Counterfactual recourse using disentangled subspaces. In: ICML workshop proceedings (2020)
11. Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Adversarial attacks on deep neural networks for time series classification. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2019)
12. Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Deep learning for time series classification: a review. Data Mining and Knowledge Discovery **33**(4), 917–963 (2019)
13. Forestier, G., Petitjean, F., Dau, H.A., Webb, G.I., Keogh, E.: Generating synthetic time series to augment sparse datasets. In: ICDM. pp. 865–870. IEEE (2017)
14. Gee, A.H., Garcia-Olano, D., Ghosh, J., Paydarfar, D.: Explaining deep classification of time-series data with learned prototypes. CEUR Workshop Proceedings **2429**, 15–22 (2019)
15. Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., Lee, S.: Counterfactual visual explanations. In: ICML. pp. 2376–2384. PMLR (2019)
16. Grabocka, J., Schilling, N., Wistuba, M., Schmidt-Thieme, L.: Learning time-series shapelets. In: ACM SIGKDD. pp. 392–401 (2014)
17. Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., Turini, F.: Factual and counterfactual explanations for black box decision making. IEEE Intelligent Systems **34**(6), 14–23 (2019)

18. Guidotti, R., Monreale, A., Spinnato, F., Pedreschi, D., Giannotti, F.: Explaining any time series classifier. In: CogMI 2020. pp. 167–176. IEEE (2020)
19. Gunning, D., Aha, D.: Darpa's explainable artificial intelligence (xai) program. AI Magazine **40**(2), 44–58 (2019)
20. Kanamori, K., Takagi, T., Kobayashi, K., Arimura, H.: Dace: Distribution-aware counterfactual explanation by mixed-integer linear optimization. In: IJCAI-20. pp. 2855–2862 (2020)
21. Karimi, A.H., Barthe, G., Balle, B., Valera, I.: Model-agnostic counterfactual explanations for consequential decisions. In: AISTATS. pp. 895–905 (2020)
22. Karlsson, I., Rebane, J., Papapetrou, P., Gionis, A.: Explainable time series tweaking via irreversible and reversible temporal transformations. In: ICDM (2018)
23. Keane, M.T., Kenny, E.M.: 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: Proc. ICCBR'19. pp. 155–171. Springer (2019)
24. Keane, M.T., Kenny, E.M., Delaney, E., Smyth, B.: If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. In: IJCAI-21 (2021)
25. Keane, M.T., Smyth, B.: Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai). In: Proc. ICCBR'20. pp. 163–178. Springer (2020)
26. Kenny, E.M., Delaney, E.D., Greene, D., Keane, M.T.: Post-hoc explanation options for xai in deep learning: The insight centre for data analytics perspective. In: International Conference on Pattern Recognition. Springer (2020)
27. Kenny, E.M., Keane, M.T.: Twin-systems to explain artificial neural networks using case-based reasoning: comparative tests of feature-weighting methods in ann-cbr twins for xai. In: IJCAI-19. pp. 2708–2715 (2019)
28. Kenny, E.M., Keane, M.T.: On generating plausible counterfactual and semi-factual explanations for deep learning. In: AAAI-21. pp. 11575–11585 (2021)
29. Laugel, T., Lesot, M.J., Marsala, C., Renard, X., Detyniecki, M.: The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In: Proc. IJCAI-19. pp. 2801–2807 (2019)
30. Le Nguyen, T., Gsponer, S., Ilie, I., O'Reilly, M., Ifrim, G.: Interpretable time series classification using linear models and multi-resolution multi-domain symbolic representations. Data mining and knowledge discovery **33**(4), 1183–1222 (2019)
31. Leake, D., Mcsherry, D.: Introduction to the special issue on explanation in case-based reasoning. The Artificial Intelligence Review **24**(2), 103 (2005)
32. Leonardi, G., Montani, S., Striani, M.: Deep feature extraction for representing and classifying time series cases: towards an interpretable approach in haemodialysis. In: Flairs-2020. AAAI Press (2020)
33. Lipton, Z.C.: The mythos of model interpretability. Queue **16(3)**, 30 (2018)
34. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: ICDM. pp. 413–422 (2008)
35. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. pp. 4765–4774 (2017)
36. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
37. Molnar, C.: Interpretable machine learning. Lulu.com (2020)
38. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: ACM FAccT. pp. 607–617 (2020)
39. Nguyen, T.T., Le Nguyen, T., Ifrim, G.: A model-agnostic approach to quantifying the informativeness of explanation methods for time series classification. In: AALTD Workshop 2020. Springer (2020)

40. Nugent, C., Cunningham, P.: A case-based explanation system for black-box systems. Artificial Intelligence Review **24**(2), 163–178 (2005)
41. Nugent, C., Doyle, D., Cunningham, P.: Gaining insight through case-based explanation. Journal of Intelligent Information Systems **32**(3), 267–295 (2009)
42. Olszewski, R.T.: Generalized feature extraction for structural pattern recognition in time-series data. Tech. rep., Carnegie-Mellon Univ, Pittsburgh (2001)
43. Pearl, J., Mackenzie, D.: The Book of Why. Basic Books, New York (2018)
44. Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., Flach, P.: FACE: feasible and actionable counterfactual explanations. In: AIES. pp. 344–350 (2020)
45. Recio-García, J.A., Díaz-Agudo, B., Pino-Castilla, V.: CBR-LIME: A Case-Based Reasoning Approach to Provide Specific Local Interpretable Model-Agnostic Explanations. In: ICCBR. pp. 179–194. Springer (2020)
46. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: Proc. SIGKDD'16. pp. 1135–1144. ACM (2016)
47. Russell, C.: Efficient search for diverse coherent explanations. In: Conference on Fairness, Accountability, and Transparency. pp. 20–28 (2019)
48. Samangouei, P., Saeedi, A., Nakagawa, L., Silberman, N.: Explaingan: Model explanation via decision boundary crossing transformations. In: European Conference on Computer Vision (ECCV). pp. 666–681 (2018)
49. Sani, S., Wiratunga, N., Massie, S.: Learning deep features for knn-based human activity recognition. In: ICCBR-17 Workshop Proceedings. Springer (2017)
50. Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., Keim, D.A.: Towards a rigorous evaluation of xai methods on time series. arXiv preprint arXiv:1909.07082 (2019)
51. Schoenborn, J.M., Weber, R.O., Aha, D.W., Cassens, J., Althoff, K.D.: Explainable case-based reasoning: A survey. In: AAAI-21 Workshop Proceedings (2021)
52. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural computation **13**(7), 1443–1471 (2001)
53. Sørmo, F., Cassens, J., Aamodt, A.: Explanation in case-based reasoning–perspectives and goals. Artificial Intelligence Review **24**(2), 109–143 (2005)
54. Van Looveren, A., Klaise, J.: Interpretable counterfactual explanations guided by prototypes. arXiv preprint arXiv:1907.02584 (2019)
55. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: automated decisions and the gdpr. Harv.J.Law Tech. **31**, 841 (2017)
56. Wang, Y., Emonet, R., Fromont, E., Malinowski, S., Menager, E., Mosser, L., Tavenard, R.: Learning interpretable shapelets for time series classification through adversarial regularization. arXiv preprint arXiv:1906.00917 (2019)
57. Wang, Z., Yan, W., Oates, T.: Time series classification from scratch with deep neural networks: A strong baseline. In: IJCNN. pp. 1578–1585. IEEE (2017)
58. Ye, L., Keogh, E.: Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. Data mining and knowledge discovery **22**(1-2), 149–182 (2011)
59. Yeh, C.C.M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H.A., Silva, D.F., Mueen, A., Keogh, E.: Matrix profile i: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In: ICDM (2016)
60. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: IEEE CVPR. pp. 2921–2929 (2016)