Post-Hoc Explanation Options for XAI in Deep Learning: The *Insight Centre for Data Analytics* Perspective

Eoin M. Kenny[1,2,3(✉)], Eoin D. 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, Ireland

eoin.kenny@insight-centre.org
eoin.delaney4@ucdconnect.ie
derek.greene@ucd.ie
mark.keane@ucd.ie

Abstract. This paper profiles the recent research work on eXplainable AI (XAI), at the *Insight Centre for Data Analytics.* This work concentrates on *post-hoc* explanation-by-example solutions to XAI as one approach to explaining black box deep-learning systems. Three different methods of *post-hoc* explanation are outlined for image and time-series datasets: that is, factual, counterfactual, and semi-factual methods). The future landscape for XAI solutions is discussed.

Keywords: Explainable AI, Interpretable AI, Trust, Artificial Neural Networks, Convolutional Neural Networks, Case-Based Reasoning, *k*-Nearest Neighbors

1 Introduction

In the last five years, the problem of eXplainable AI (XAI) has been highlighted as the public, business and government face AI-based decision-making in people's everyday lives, jobs, and leisure time [11]. In the European Union, the urgency behind this research area has, in part, being driven by GDPR proposals on explaining automated decisions [40]. However, more broadly, it also arises from a deep concern in the academic community that some AI technologies rely on dubious ethical standards and/or unethical design decisions, decisions that may result AI systems coming to be perceived as unfair, unaccountable, and untrustworthy. For instance, consider the issues around bias and consent in prominent datasets [24, 25, 35]; MIT recently apologized for the Tiny Images dataset, when it was revealed to contain verifiably pornographic images shot in non-consensual settings [35].

Facing these challenges, Ireland's national Artificial Intelligence and Data Analytics centre – the *Insight Centre for Data Analytics* (www.insight-centre.org) – has developed an extensive program of engagement with government and business in the field of XAI, as well as advancing research in the area. On the regulatory side, *Insight* has engaged with initiatives at national and international levels in championing a Magna Carta for Data [32], contributing to the European's Commission's High-Level Expert

Group in AI [1], and the Joint Strategic Research Innovation and Deployment Agenda for the Artificial Intelligence, Data and Robotics Partnership [12]. On the research side, the *Insight* centre has an extensive program that aims to formulate a coherent solution to the XAI problem (e.g., [7, 20, 26, 30, 41]). In this paper, we present a slice of this work directed at image and time-series data, focussing on how opaque, black-box AI systems can be explained with reference to more interpretable, white-box AI systems; what has been termed the *Twin-Systems* approach to XAI [8, 14, 15, 20] (see Fig. 1).

In the remainder of this introduction we make some of key taxonomic distinctions in XAI for different algorithmic approaches to the problem, before showing a taxonomic matrix for the research area. Then, in later sections, we sketch the algorithmic techniques advanced by *Insight* and initial user tests of these techniques.

1.1 A Fundamental Distinction: *Pre-hoc* Versus *Post-Hoc*

Many definitional and taxonomic issues arise in XAI, not least because "explanation" has for decades proven to be very hard concept to define across many disciplines, from Philosophy, to the Philosophy of Science, and Psychology [39]. It is, therefore, not surprising that Computer Science and Artificial Intelligence has struggled too [39]. Arguably, we still really do not have precise definitions for the terms "explanation"", "interpretable" and transparent"; though this does not stop us using them on a regular basis. However, notwithstanding these definitional issues, there has been some agreement on a fundamental distinction between "explanation proper" and "explanation as justification". For example, Sørmo *et al*. [39] point out the philosophical distinction between explaining how the system reached some answer (what they call *transparency*) and explaining why the system produced a given answer (*post-hoc justification*). Lipton [27] makes a similar distinction between *transparency* (i.e., "*How does the model work"*) and *post hoc explanation* (i.e., *"What else can the model tell me?"*). The key idea here is that one can causally explain a model directly, in some sense, (e.g., "it optimizes this function using such-and-such a detailed method") or one can explain/justify how it reached some decision with reference to other information (e.g., "the model did this because it used such-and-such data"). The problem XAI faces is that the former may be accurate but can only be comprehended by a handful of people (i.e., how can the general public understand a deep learning algorithm) and the latter may be comprehensible but is too approximate to really explain what happens (e.g., saying certain data was used may also be uninformative or unclear). At present, these two options for explanation have been set somewhat in opposition to one another [37], even though on occasion they shade into one another [11]. Next, we consider these two opposing positions on "model transparency" and "*post-hoc* explanation" in more detail.

Model Transparency. This explanation position has been terminologically cast as "transparency", "simulatability" or "interpretable machine learning". The key idea here is that one can causally explain a model directly, in some sense, (e.g., "it optimizes this function using such-and-such a detailed method"). In this approach, one understands how the whole model works given some representation of it [27] or via some simplified proxy model that "behaves similarly to the original model, but in a way that is easier to explain" [11] (e.g., [10]). Rudin [37] argues that the use of inherently transparent models is the only appropriate solution to XAI in sensitive domains; pointing to her own use of prototypes [5]. There are two major problems with this approach to XAI. First, to date, few pure instances of good proxy models have been proposed to characterize

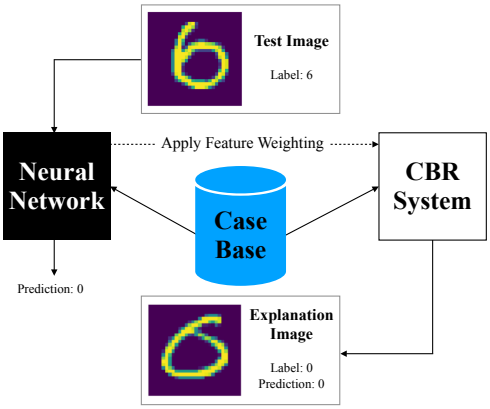

Fig. 1. *The Twin-Systems Explanation Framework*: A deep learning model (Neural Network) produces a miss-classification for an MNIST test image, wrongly labelling a "6" as a "0". This prediction is explained by analysing the feature-weights of the network for that prediction and applying these to a twinned *k*-NN (Case Based Reasoner/CBR System) to retrieve a nearest neighbor to the test-image in the training set. This explanatory image shows that the model used an image of a "0" that looks very like a "6" to make its prediction of a "0". So, though it miss-classifies the item, it is quite faithful to the data it was given.

black box systems. Second, when proxy models have been proposed (e.g., decision trees) very little evidence is provided for why they are more interpretable than the original black box; that is, the researchers typically just assert they are more interpretable without supporting user tests. As Lipton [27] points out "neither linear models, rule-based systems, nor decision trees are intrinsically interpretable…Sufficiently high-dimensional models, unwieldy rule lists, and deep decision trees could all be considered less transparent than comparatively compact neural networks". Finally, it should be said, that it is not wholly clear when a proxy model actually becomes an identifiably separate model; for instance, Frosst & Hinton [10] argue that their model is a stand-alone one, not an interpretable proxy to work "alongside" a deep learner. Presumably, at some (as yet undefined) point a proxy model is no longer a facsimile of the original.

***Post-Hoc* Explanation.** The other explanation position has been terminologically cast as *justification* or *explainable machine learning*. The key idea here is that one can explain/justify how a model reached some decision with reference to other information (e.g., "the model did this because it used such-and-such data"). Lipton [27] has further divided *post-hoc* explanations into (i) textual explanations of system outputs, (ii) visualizations of learned representations or models (e.g., heat/saliency maps; [42]), and (iii) explanations by example (i.e., the classic case-based reasoning approach). This type of "explanation by justification", is an after-the-prediction explanation step where some evidence is given to elucidate the predictions made by the AI system; though, some techniques, such as visual analytics may operate right across the deep learning pipeline [42]. As we shall see, it has recently become clear that explanation-by-example can be

Table 1: The *Insight* Taxonomic Matrix for Explanation-Types X Datasets

	Explanation Type		
Data Sets	***Factual***	***Counterfactual***	***Semi-factual***
Tabular	Kenny & Keane [20] Keane & Kenny [14] Keane & Kenny [15] Kenny *et al.* [17] Kenny *et al.* [18]	Keane & Smyth [16]	----
Image	Keane & Kenny [20] Ford *et al.* [8]	Kenny & Keane [19]	Kenny & Keane [19]
Time-Series	Nguyen *et al.* [30] Delaney *et al*. [7]	Delaney *et al.* [7]	----

divided into three distinct flavors (i.e., factual, counterfactual, and semi-factual). This paper recounts the systematic work that has been done by researchers at the *Insight Centre for Data Analytics* to explore these *post-hoc* solutions to XAI in image and time-series datasets.

1.2 *Post-Hoc* Explanation: Factual, Counterfactual, and Semi-Factuals

Traditionally, *post-hoc* explanations were viewed as explanations-by-example where some factual (i.e., nearest neighboring) case was produced to explain some target query [18]. However, there are other explanatory options based on the type of example used in the explanation. Consider a typical scenario where we are trying to explain a black-box classifier giving loan decisions, operating off a traditional tabular dataset with defined features (e.g., gender). Assume you are refused your loan application and, under your GDPR rights, ask for an explanation. The system could give you a *factual example-based explanation* saying "you were refused the loan because your profile is similar to person-x who was also refused the loan". Alternatively, the system could give you a *counterfactual explanation* saying "if you had a higher salary, you would have the profile of person-x who got the loan". Finally, one could also be given a *semi-factual explanation* saying "even if you had a higher salary, you would still *not* have the profile of person-x who got the loan". Of course, computing these alternative explanations is non-trivial and, as such, the lion's share of research has been on factual *post-hoc* explanations (as in CBR), but there is a growing interest in counterfactual [23, 40], and semi-factual explanations [19, 31]. Finally, almost all of this research has focused on tabular datasets, so here we consider, arguably harder, image and time-series datasets.

1.3 A Taxonomic Matrix for *Post-Hoc* Explanations

In the previous subsection, we saw how *post-hoc* explanations can be divided into three distinct types – factual, counterfactual, and semi-factual – and noted that while some progress has been made in implementing these strategies for tabular datasets, it is only

very recently that researchers have started to consider them for non-tabular datasets. In *Insight*, we have a research program designed to flesh out these *post-hoc* alternatives for explaining different datasets. Accordingly, our research program aims to fill the cells of a matrix created by crossing explanation-types by datasets (see Table 1).

In the remainder of this paper we will sketch some of the solutions we have found when exploring these different explanation-types for image and time-series datasets. So, for image datasets we first consider methods for factual (Section 2), counterfactual (Section 3), and semi-factual explanations (Section 4). Then in the remainder of the paper, we consider counterfactuals explanations for time-series (Section 5) before looking to future directions for this work (see Section 6).

2. *Post-Hoc* Factual Explanations: Images

A "factual *post-hoc* explanation-by-example" is a long name for the case-based explanations used in CBR[1]. Traditionally, these models deploy a *k*-NN to solve some classification or regression problem and then use the nearest-neighbors in *k* to explain the prediction made; where, typically, the prediction is made from some averaging or aggregation of the instances in *k* [17, 18]. The current use of factual explanations extends this approach to explain black-box deep learning models (Artificial Neural Networks or ANNs) where the nearest neighboring cases from a *k*-NN twinned with the ANN are selected based on analyzing the feature-weights of the ANN. Recently, Kenny & Keane [20] generalized this explanation option in the *Twin Systems* approach, where the feature-weights for a test-instance in a deep learning model are applied to a *k*-NN, operating over the same dataset, to find factual explanations (see Fig. 1).

Kenny & Keane [20] also competitively tested several feature-weighting, methods from a literature going back to the 1990s, to determine the most accurate method for capturing ANNs (including, multi-layered perceptrons and convolutional neural networks); these experiments found that a contributions-based method performed best. Recently, Papernot and MacDaniel [33] proposed D*k*NN as a method for finding factual explanation cases, although they did not consider weighting the *k*-NN abstraction, which has been found to be crucial [20]. Also, Chen *et al.* [5] replaced the last layer of a CNN with a CBR system to force the black-box to be more transparent. In the present section, we sketch our contributions-based method (see section 2.1) and show how it can be applied to a CNN dealing with the MNIST and CIFAR datasets before considering some of its explanatory results and user-tests.

2.1 The Method: COLE

A contributions-based feature-weighting method has been found to offer the most accurate analysis of black-box ANNs, with a view to finding factual example-based explanations [20]. This feature-weighting method -- *Contributions Oriented Local Explanations* (COLE) – can be applied to both multi-layered perceptrons (MLPs) and convolutional neural networks (CNNs) to find explanatory cases from the twinned *k*-NN/CBR model (i.e., a CNN-CBR twin) applied to the same dataset. COLE fits a *k*-

[1] Here, we consider factual *examples* as explanations; but LIME [36] gives factual information about the current test instance via feature importance scores also.

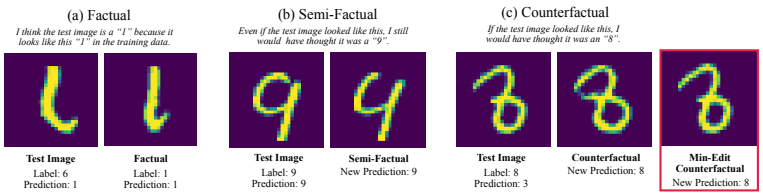

Fig. 2. *Post-hoc* factual, semi-factual, and counterfactual explanations on MNIST showing: (a) a *factual explanation* for a miss-classification of "6" as "1", that uses a nearest-neighbor in latent-space classed as "1", (b) a *semi-factual explanation* for the correct classification of a "9", that shows a synthetic instance with meaningful feature changes that would *not* alter its classification, and (c) a *counterfactual explanation* for the miss-classification of an "8" as a "3", that shows a synthetic test-instance with meaningful feature changes that *would have been* classified as an "8" (n.b., for comparison a counterfactual using a *Min-Edit* method is shown with its human-undetectable feature-changes; from [19]).

NN model with feature contributions to abstract the ANN function, that are calculated by multiplying a data-instance by weights it used in the final prediction.

To implement this in a CNN there are two possible options. Firstly, the CNN may have several fully connected layers post feature-extraction, in which case we have shown how saliency map techniques can be used to implement COLE [20]. Secondly, there may be a linear classifier post feature-extraction (e.g., the ResNet architectures), in which case contributions can be calculated by taking the Hadamard product of an instance's penultimate activations with the weight vector connected to its final classification (henceforth called C-HP). In both approaches it is possible to highlight the most positively contributing features via a feature activation map (FAM) [20].

2.2 Results: Factual Image-based Explanations

Fig. 2a and 3 shows two examples of factual explanations found using a CNN-CBR twin system approach on the MNIST and CIFAR-10 datasets, for correct and incorrect classifications. In Fig. 2a an incorrect classification is made by the system, where a "6" is miss-classified as a "1" and the explanatory nearest-neighbors tell the user that this occurs because the dataset contains data which looks like the test image and was labelled as "1". Fig. 3 shows an example using the CIFAR-10 dataset involving C-HP. It shows the miss-classification of an automobile as a truck. This incorrect prediction is justified by essentially saying to the user *"I think this is a truck because it looks like these trucks I saw before".* In addition, the FAMs highlight the most important (i.e., the most positively contributing) feature in the classification, which clearly focusses on the vehicle wheels in all images. Since these are a central aspect of both automobiles and trucks, it makes the miss-classification more reasonable.

C-HP has been extensively tested on seventeen classification/regression datasets, which consistently showed C-HP to be the best for both MLPs and CNNs. Furthermore, Ford *et al*. [8] have performed a series of user studies using its explanations for MNIST; they asked people to judge the correctness/reasonableness of the predictions made by the CNN in the presence/absence of explanations. These studies showed that explanations impacted people's perceptions of the correctness of the CNN's predictions. However, these studies also showed that the explanations did *not* improve people's overall trust/satisfaction in the system when it produced miss-classifications (i.e., it did not "explain away" error behaviour). This work also found that people have a low-tolerance for error in such automated systems (i.e., algorithmic aversion).

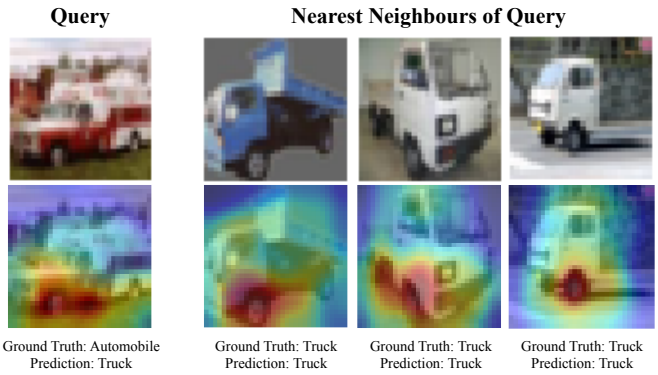

Fig. 3. A CNN-CBR twin miss-classifies an image of an automobile as a truck. The nearest-neighbours are all trucks, justifying the prediction. Also, a FAM shows the CNN is focused on the wheels in the prediction, features indicative of both automobiles and trucks.

3. *Post-Hoc* Counterfactual Explanations: Images

Although factual explanations have traditionally been the focus for example-based explanations, recently there has been an expanding interest in contrastive-example explanations [19, 23, 40]. Indeed, some have argued that contrastive explanations are much more causally-informative than factual ones, as well as being GDPR-compliant [40]. Most current counterfactual methods only apply to tabular data [16, 28], but some recent work has begun to consider images. To deal with images, generative models have been used to produce counterfactual images with large featural-changes for XAI [38]. Recently however, *Insight* researchers have developed a different approach to generating counterfactuals for image datasets, called PlausIble Exceptionality-Based Contrastive Explanations (PIECE) [19]; it generates counterfactual images by focusing on *exceptional* features (an approach inspired by strategies humans use when generating

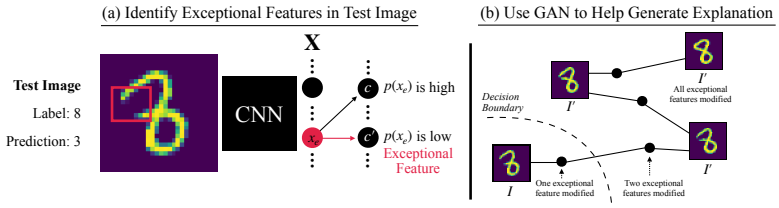

Fig. 4. PIECE Explains an Incorrect Prediction Using a Counterfactual: The test image labelled as "8" is miss-classified as a "3" by the CNN. To show how the image would have to change for the CNN to classify it as an "8", PIECE generates a counterfactual by (a) identifying the features that have a low probability of occurrence in the counterfactual class c' (i.e., "8" class) before modifying them to be the expected feature values for c', and (b) using a GAN to visualize the image I' (here we show progressive exceptional-feature changes that gradually produce a plausible counterfactual image of an "8").

counterfactuals [4]). The algorithm generates counterfactuals by identifying "exceptional" features in the test image, and then modifying these to be "normal".

3.1 The Method: PIECE

PIECE involves two distinct systems, a CNN that is generating predictions to be explained, and a GAN that helps generate explanatory images. This algorithm will work with any trained CNN, provided there is a GAN trained on the same dataset as the CNN. PIECE has three main steps: (i) "exceptional" features are identified in the CNN for a test image from the perspective of the counterfactual class, (ii) these are then modified to be their expected values, and (iii) the resulting latent-feature representation of the explanatory counterfactual is visualized in the pixel-space with help from the GAN.

Fig. 4 illustrates how PIECE works in practice to generate a counterfactual image-explanation. Here, the counterfactuals to a test image I, in class c, with latent features x, are denoted as I', c' and x', respectively. Fig. 4 shows a test image labelled as class "8" (i.e., c) is miss-classified as class "3" (i.e., c'). Exceptional features are identified using mathematical probability in the extracted feature layer $\boldsymbol{X}$ which have a low chance of occurrence in c'; these are then modified to be their expected feature values for class c' which modify the latent representation x to be x'. This new latent counterfactual representation x' is then visualized in the pixel space as the explanation I' using a GAN.

3.2 Results: Counterfactual Image-based Explanations

Kenny & Keane [19] have compared PIECE to a simple *Min-Edit* method in a series of experiments (along with several other methods in the literature) to highlight the difference it finds. Fig. 2c shows the counterfactual explanations for the miss-classification of an "8" as a "3" for PIECE and *Min-Edit*. PIECE shows a plausible counterfactual which fully removes all irregularities from the perspective of the counterfactual class "8", whilst the Min-Edit counterfactual does not convey meaningful information to help

a user understand the difference between the two classes. Furthermore, [19] compared PIECE to a *Min-Edit* approach, generating 193 counterfactual explanations for right and wrong classifications on the MNIST and CIFAR-10 datasets. The evaluation measures assessed the plausibility of the generated instances by virtue of their proximity to the underlying data distribution. On most measures, PIECE was significantly better than the *Min-Edit* approach and other popular methods [19].

4. *Post-Hoc* Semi-Factual Explanations: Images

The last missing piece of the puzzle for post-hoc explanations is the largely under researched semi-factual explanations. To understand semi-factuals computationally, it is interesting to contrast them with counterfactuals; whilst counterfactuals are typically described as the minimum distance an instance must travel to cross a decision boundary, a semi-factual can be seen as the *maximal distance* an instance can travel *without* changing its classification (n.b., while still being a plausible instance). An AI loan application system might explain its decision semi-factually by saying "Even if you had asked for a slightly lower amount, you still would have been refused the loan". We have found only one decade-old paper related to semi-factual explanations (see [31] on *a-fortiori* reasoning that was only on tabular data).

4.1 Method & Results: PIECE for Semi-Factuals

To implement semi-factual explanations for images, we used the PIECE algorithm, but stop the modification of exceptional features before the decision boundary is crossed. As we shall see, this results in a large, plausible change to the image that does *not* change the classification. For comparison, we compared it again against the *Min-Edit* method; although this time, the method is stopped not after crossing the decision boundary, but one optimization step before, so the classification remains.

We measure "good semi-factuals" for images with the L_1 distance between the test image and synthetic explanatory semi-factual in the pixel-space (n.b., the greater the distance the better). Kenny & Keane [19] compared PIECE against the *Min-Edit* method, finding significant differences between the two in terms for how much of the image is modified before reaching the decision boundary. This result shows that that the "blind perturbation" *Min-edit* method is suboptimal for generating semi-factuals close to the decision boundary. Figs. 2 and 5 show some examples of semi-factual explanation for the MNIST and CelebA datasets, respectively. Fig 2b shows a semi-factual explanation for the correct classification of a "9" on MNIST. Glossed, the explanation is saying *"Even if the test image looked like this (i.e., closer to a 4), the model*

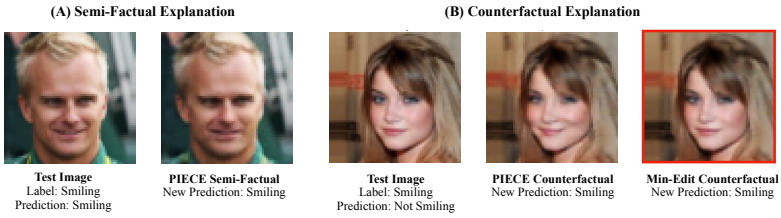

Fig. 5 (A) A semi-factual explanation justifying why the initial classification was definitely correct, in that, even if the image was smiling much less, it still would have classified it as "smiling. (B) A counterfactual explanation conveying to a user why the CNN made a mistake, and how the image would need to look for it to have classified it correctly (as computed by PIECE and Min-Edit)

would still have thought it was a 9, ergo, the initial classification is definitely correct". A similar explanation is conveyed in Fig. 5a for the CelebA dataset.

5 *Post-Hoc*, Counterfactual Explanations: Time-Series

We have now demonstrated how nearest neighbor techniques in twin systems can explain the predictions of black-box deep learners (such as CNN's). Next, we focus on explanations in the time-series domain. The current focus in XAI for time series mainly focusses on saliency-based approaches where important sub-sequences or features are highlighted [30]. However, given the immense success of nearest-neighbor classifiers using a variety of distance measures [3, 29], instance-based counterfactual explanations also seem like an exciting avenue to pursue for XAI in time series.

Earlier we highlighted that counterfactual explanations with image-data is a recent development in XAI; however, until very recently, counterfactuals explanations of time-series models have been largely ignored. The best way to understand how counterfactual explanations might be used for explaining time-series data is to explore how they differ from factual explanations. Consider a binary classification system which decides whether a city has Oceanic Climate or a Mediterranean Climate based on weekly temperatures over some historical period. The system explain a prediction factually saying "Amsterdam has an Oceanic Climate because it is most similar to London (a city in the training data) which also has an Oceanic Climate". In contrast, the system might explain its decision counterfactually by saying "If Amsterdam had slightly hotter summers the system would predict the city to have a Mediterranean climate". Tabular methods for counterfactuals [40], quickly become intractable for time-series data because of the number of possible feature dimensions and the domain-specific distance measures (such as DTW). In response to this gap in the literature, some recent proposals have been made to use contrastive methods to explain time-series predictions. Karlsson *et al.* [13] implement explainable time-series tweaking, using an opaque shapelet-based classifier, where they find the minimum number of changes to be performed to the given time series that changes the classification decision. Also, by modifying the original loss function [59] to generate counterfactuals, Ates *et al*. [2]

have explored generating counterfactual explanations for multivariate time series classification problems. Also, Labaien *et al*. [21] have progressed contrastive explanations for the predictions of recurrent neural networks in time-series prediction. Recently researchers at *Insight* proposed an instance-based approach, called Native-Guide, for counterfactual generation in time series [7]. This approach has been shown to work with any classifier, using both DTW and Minkowski distance measures.

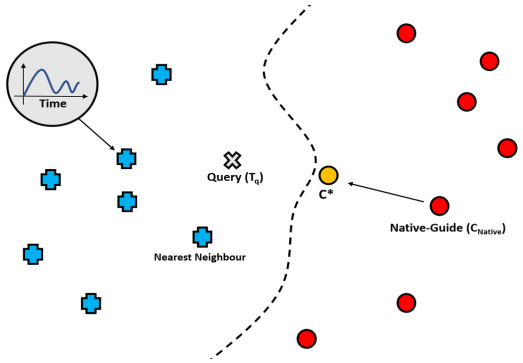

Fig. 6: A time series data set for a binary classification task with two class labels. A query time series T_q (represented as X) and it's Native-Guide C_{Native}. The generated counterfactual C^* is represented in yellow.

5.1 The Method: Native-Guide for Time-series Counterfactuals

The current method – Native Guide – incorporates a strategy where the closest in-sample counterfactual instance to the test-instance is adapted to form a new counterfactual explanation [16, 22, 31]. Here the "Native-Guide" is a counterfactual instance that already exists in the dataset, it is the nearest-neighbor time-series to the query that involves a class change (see Fig. 6). We can retrieve this in-sample counterfactual instance using a simple 1-NN search. Once this instance is found it is perturbed towards the query until just before the decision boundary. The generated counterfactual instance C* (the yellow point in Fig.6), should offer better explanations than the original in-sample counterfactual as it is in closer to the query whilst still staying within the distribution of the data. Fig. 7 shows a specific example in the climate domain for a counterfactual explanation of a time-series.

When using Euclidean distance perturbation a simple weighted perturbation strategy works well. But, when working with DTW a technique known as weighted Dynamic Barycenter Averaging (DBA) is required to implement the perturbation [9, 34]:

Definition: Weighted average of time series under DTW. *Given a weighted set of time series* $\mathrm{D} = (\mathrm{T}_1,\beta_1)$, ..., $(\mathrm{T}_\mathrm{N},\beta_\mathrm{N})$*, the average time series under DTW,* $\overline{\mathrm{T}}$*, is the time series that minimizes:*

$$argmin\ \bar{T} = \sum_{i=1}^{N} \beta_i \cdot DTW^2(\bar{T}, T_i)$$

5.2 Results: Native-Guide for Counterfactuals

Native Guide was tested on a climate case-study and over 35 diverse datasets from the UCR archive [6]. In these experiments, a specialized distance-measure (called RCF, see [16]) was used to assess if the generated counterfactuals were in close proximity to the query, along with novelty detection algorithms, to assess if the generated counterfactuals were within the distribution of the data. The generated counterfactual instances that are not within the distribution of the data are referred to as being Out-of-Distribution (OOD). A subset of our results are shown in Table 2 with a full analysis and discussion of results in the original paper [7]. The results highlight that Native-Guide gen-

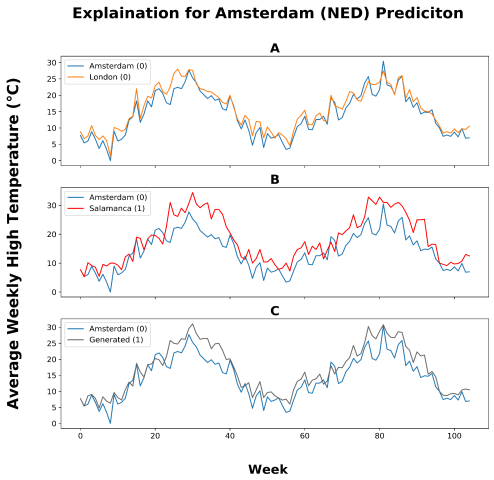

Fig. 7: Different explanations for a queried city (Amsterdam) in the climate prediction task **(A)** Factual Explanation: "Amsterdam has an Oceanic climate because it is most similar to London, which has an Oceanic climate too". **(B)** In-sample Counterfactual Explanation: "If Amsterdam had the same weather profile as Salamanca the system would classify it as having a Mediterranean climate". Salamanca's weather profile is quite different to Amsterdam's (noticeably hotter summers and warmer winters). An explanation that is more similar to the query might be more informative and this motivates the generation of a new counterfactual using Salamanca as a "Native-Guide" **(C)** If Amsterdam had a weather profile like the Generated-Instance then system would classify it as having a Mediterranean climate. This is a better explanation than **B** because the generated time series is much closer to the original query by comparison to Salamanca and is also within the data distribution.

erates proximal and plausible counterfactual explanations for a diverse range of datasets. The generated counterfactual instances are significantly closer to the query when compared to the existing in-sample counterfactual instances.

Table 2: Subset of results for UCR Data Experiment (Dynamic Time Warping Implementation). #CF indicates the number of counterfactual instances generated.

Data Set	Train Size	Test Size	#CF	OOD	RCF
ECG200	100	100	23	1	0.304
GunPoint	50	150	14	0	0.151
ItalyPowerDemand	67	1029	51	1	0.323
PhalangesOutlinesCorrect	1800	858	233	6	0.408

6 Future Directions

This paper has briefly summarized the *Insight Centre for Data Analytics'* engagement and contributions to the rapidly evolving and increasingly important field of XAI. The immediate avenue for future work is to fill out the matrix detailed above (in Table 1). This plan underscores the need to explore semi-factual explanations in tabular and time-series datasets. Additionally, we have already considered and published work on natural language counterfactual explanations which focused on the issue of grammatical plausibility [26]; future work in this area will also extend this to factual and semi-factual explanations. Other interesting avenues exist in applying the PIECE algorithm to tabular, text, and time-series datasets, to see if the modification of *exceptional* features to generate contrastive explanations carries over into these other domains. Finally, it is important to reiterate that whatever explanation strategy bears the most fruit computationally will need to be psychologically verified in user tests, such as those in [8].

Acknowledgements. This paper emanated from research funded by (i) Science Foundation Ireland (SFI) to the Insight Centre for Data Analytics (12/RC/2289-P2), (ii) SFI and DAFM on behalf of the Government of Ireland to the VistaMilk SFI Research Centre (16/RC/3835).

References

1. Ala-Pietilä, P.: Landline - 10/10/20: High-Level Expert Group on Artificial Intelligence, https://ec.europa.eu/digital-single-market/en/high-level-expert-group -artificial-intelligence
2. Ates, E. et al.: Counterfactual Explanations for Machine Learning on Multivariate Time Series Data. arXiv:2008.10781. (2020)
3. Bagnall, A. et al.: The Great Time Series Classification Bake Off: An Experimental Evaluation of Recently Proposed Algorithms. Extended Version. arXiv:1602.01711. (2016)

4. Byrne, R.M.J.: Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19). (2019)
5. Chen, C. et al.: This Looks Like That. In: NeurIPS. (2020)
6. Dau, H.A. et al.: The UCR Time Series Archive. arXiv:1810.07758. (2019)
7. Delaney, E. et al.: Instance-Based Counterfactual Explanations for Time Series Classification. arXiv:2009.13211. (2020)
8. Ford, C. et al.: Play MNIST For Me! User Studies on the Effects of Post-Hoc, Example-Based Explanations & Error Rates on Debugging a Deep Learning, Black-Box Classifier. In: IJCAI-20 XAI workshop. (2020)
9. Forestier, G. et al.: Generating Synthetic Time Series to Augment Sparse Datasets. In: 2017 IEEE International Conference on Data Mining. (2017)
10. Frosst, N., Hinton, G.: Distilling a Neural Network Into a Soft Decision Tree. arXiv:1711.09784. (2017)
11. Gilpin, L.H. et al.: Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. arXiv:1806.00069. (2018)
12. Hahn, T.: Landline - 10/10/20: Strategic Research, Innovation and Deployment Agenda, https://ai-data-robotics-partnership.eu/wp-content/uploads/2020/09/AI-Data-Robotics-Partnership-SRIDA-V3.0.pdf.
13. Karlsson, I. et al.: Explainable time series tweaking via irreversible and reversible temporal transformations. arXiv:1809.05183. (2018)
14. Keane, M.T., Kenny, E.M.: How Case-Based Reasoning Explains Neural Networks. In: Case-Based Reasoning Research and Development. (2019)
15. Keane, M.T., Kenny, E.M.: The Twin-System Approach as One Generic Solution for XAI. In IJCAI-19 XAI workshop. (2019)
16. Keane, M.T., Smyth, B.: Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI). In: Case-Based Reasoning Research and Development. (2020)
17. Kenny, E.M. et al.: Bayesian Case-Exclusion and Personalized Explanations for Sustainable Dairy Farming. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20). (2020)
18. Kenny, E.M. et al.: Predicting Grass Growth for Sustainable Dairy Farming. In: Case-Based Reasoning Research and Development. (2019)
19. Kenny, E.M., Keane, M.T.: On Generating Plausible Counterfactual and Semi-Factual Explanations for Deep Learning. arXiv:2009.06399. (2020)
20. Kenny, E.M., Keane, M.T.: Twin-Systems to Explain Artificial Neural Networks using Case-Based Reasoning. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19) (2019)
21. Labaien, J. et al.: Contrastive Explanations for a Deep Learning Model on Time-Series Data. In: Big Data Analytics and Knowledge Discovery. (2020)
22. Laugel, T. et al.: Defining Locality for Surrogates in Post-hoc Interpretablity. arXiv:1806.07498. (2018)
23. Laugel, T. et al.: The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19) (2019)
24. Leavy, S. et al.: Data, Power and Bias in Artificial Intelligence. arXiv:2008.0734. (2020)

25. Leavy, S. et al.: Mitigating Gender Bias in Machine Learning Data Sets. arXiv:2005.06898. (2020)
26. Linyi Y. et al.: Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification. In: Proceedings of the 28th International Conference on Computational Linguistics. (2020)
27. Lipton, Z.C.: The Mythos of Model Interpretability. arXiv:1606.03490. (2017)
28. Mittelstadt, B. et al.: Explaining Explanations in AI. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. (2019)
29. Mueen, A., Keogh, E.: Extracting Optimal Performance from Dynamic Time Warping. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2016)
30. Nguyen, T.T. et al.: A Model-Agnostic Approach to Quantifying the Informativeness of Explanation Methods for Time Series Classification. In Proceedings of the 5th Workshop on Advanced Analytics and Learning on Temporal Data at ECML. (2020)
31. Nugent, C. et al.: Gaining insight through case-based explanation. J Intell Inf Syst. 32, 3, 267–295. (2009).
32. O'Sullivan, B.: Landline - 10/10/20: Towards a Magna Carta for Data: Expert Opinion Piece: Engineering and Computer Science Committee, https://www.ria.ie/sites/default/files/ria_magna_carta_data.pdf.
33. Papernot, N., McDaniel, P.: Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning. arXiv:1803.04765. (2018)
34. Petitjean, F. et al.: A global averaging method for dynamic time warping, with applications to clustering. Pattern Recognition. (2011)
35. Prabhu, V.U., Birhane, A.: Large image datasets: A pyrrhic win for computer vision? arXiv:2006.16923. (2020)
36. Ribeiro, M.T. et al.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. (2016)
37. Rudin, C.: Please Stop Explaining Black Box Models for High Stakes Decisions. arXiv:1811.10154. (2018)
38. Seah, J.C.Y. et al.: Chest Radiographs in Congestive Heart Failure: Visualizing Neural Network Learning. Radiology. 290, 2, 514–522 (2019)
39. Sørmo, F. et al.: Explanation in Case-Based Reasoning–Perspectives and Goals. Artificial Intelligence Review. (2005)
40. Wachter, S. et al.: Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. SSRN Journal. (2017)
41. Horta, V.A. and Mileo, A.: Towards Explaining Deep Neural Networks Through Graph Analysis. In International Conference on Database and Expert Systems Applications. (2019)
42. Hohman, F., Kahng, M., Pienta, R. and Chau, D.H.: Visual analytics in deep learning. IEEE transactions on visualization and computer graphics, 25, 2674-2693 (2018)