Shapelet-based Model-agnostic Counterfactual Local Explanations for Time Series Classification

Qi Huang[*][†]**, Wei Chen**[*]**, Thomas Bäck, Niki van Stein**

LIACS, Leiden University
Niels Bohrweg 1, 2333 CA, Leiden
The Netherlands
q.huang@liacs.leidenuniv.nl

Abstract

In this work, we propose a model-agnostic instance-based post-hoc explainability method for time series classification. The proposed algorithm, namely **Time-CF**, leverages *shapelets* and *TimeGAN* to provide counterfactual explanations for arbitrary time series classifiers. We validate the proposed method on several real-world univariate time series classification tasks from the *UCR Time Series Archive*. The results indicate that the counterfactual instances generated by **Time-CF** when compared to state-of-the-art methods, demonstrate better performance in terms of four explainability metrics: closeness, sensibility, plausibility, and sparsity.

Introduction

In recent years, the performance of machine learning models has become increasingly impressive. Advances in the capability of data computation, the development of sophisticated algorithms, and the proliferation of high-quality datasets have all contributed to this upward trend (Sarker 2021a,b). Machine learning models and methods are being leveraged in various aspects of daily life and science, such as automated driving, banking, healthcare, physics, and geoscience, where intelligible explanations for their decision-making processes are essential for establishing public trust over utilized opaque models (Rudin 2019). In addition, these methods can also generate new insights for researchers in various domains. Especially in safety-critical applications such as medical diagnosis and self-driving cars, mistakes in predictions could be fatal; thus, there is an urgent need either to provide **post-hoc** explanations based on the decisions made by the models or to build inherently explainable models before predictions **(ante-hoc)**.

In response to the opacity property of machine learning models, the field of eXplainable AI (XAI) has emerged to address this issue. Although XAI methods have been proposed and effectively applied to *static* data such as images, text, and tabular data, there has been relatively less work on time series data (Schlegel et al. 2019). The existing XAI approaches for time series analysis can be categorized into three types regarding the level of explanation: point-level,

---

[*]These authors contribute equally to the paper

[†]Corresponding author

subsequence-level, and instance-level. Among the existing methods, counterfactual explanations for instance-level have recently gained wider popularity (Theissler et al. 2022).

As its name suggests, an instance-level counterfactual explanation for a to-be-explained (target) time series is an artificially generated instance, derived from the target, that does not exist in original observations but contradicts the original facts e.g., predicted to be a different class in classification. Perturbing the target time series, where minimal changes are preferred, is one of the most common techniques for creating counterfactual explanations for time series instances (Li et al. 2022; Li, Boubrahimi, and Hamd 2022). Explainability is hence achieved by highlighting the difference between the original instance and its generated counterpart. The focus of this research is to use a Generative Adversarial Network (GAN) to fabricate counterfactual explanations that can explain the predictive behavior of time series classifiers on an instance level.

Related Works

In this work, we focus on counterfactual-based XAI methods for Time Series Classification (TSC). TSC stands for a type of machine learning task where a machine learning model (classifier) shall be learned to maximally correctly assign labels to a set of time series instances. We refer the readers to the survey paper regarding XAI for time series classification (Theissler et al. 2022) for additional information.

Wachter *et al.* are pioneers in generating counterfactual instances to elucidate anomalies (Wachter, Mittelstadt, and Russell 2017). This involves simultaneously minimizing a prediction loss function that drives the prediction on counterfactuals towards a different class and a distance loss function that ensures the similarity between counterfactuals and to-be-explained instances. Recent efforts to use counterfactual instances as explanations are mostly derived from this work. **Native-Guide** (Delaney, Greene, and Keane 2021a) is another counterfactual-based method, introduced by *Delaney et al*, following four desirable properties: Sparsity, Proximity, Plausibility, and Diversity. In essence, they find the nearest-unlike neighbor of the to-be-explained instance and generate a counterfactual instance for explanation. **MG-CF** (Li, Boubrahimi, and Hamd 2022), **SG-CF** (Li et al. 2022) are also extensions of (Wachter, Mittelstadt, and Russell 2017), but incorporate shapelets into consideration.
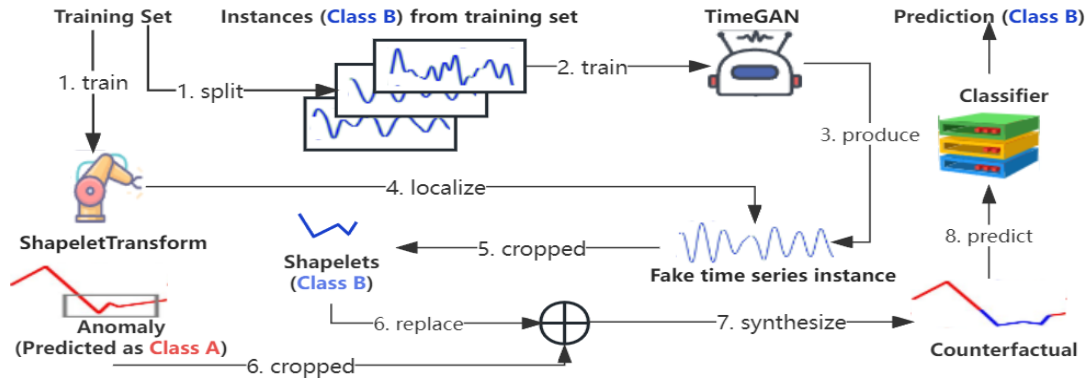
Figure 1: The flowchart of Time-CF in binary classification. The process describes how the counterfactual explanation of an instance labeled as Class A, is synthesized by TimeGAN and shapelet transform, based on instances from Class B.

Specifically, **MG-CF** generates counterfactual instances by using **shapelet Transform** (Hills et al. 2014) to extract all potential shapelet candidates and then replace the same part of the to-be-explained instance. Moreover, **SG-CF** adds an additional term to the loss function used in (Wachter, Mittelstadt, and Russell 2017), which ensures high Euclidean-based similarity between the counterfactual instance and the shapelet, thus providing more robust explanations. Besides directly perturbing the original sequence which is essentially injecting random noise to the region-of-interest, (Lang et al. 2022) and (Dantas, Marcos, and Ienco 2023) propose to employ the generative adversarial networks (Goodfellow et al. 2020) to *fabricate* entire counterfactual instances, which also enlightens us in developing our method.

Methodology

Having discussed related work on counterfactual-based instance-level XAI for time series classification, we now introduce our proposed algorithm, **Time-CF**. A flowchart of the algorithm under binary classification is shown in Figure 1. An example outcome of Time-CF is displayed in Figure 2, where only a subsequence (the blue curve) of the to-be-explained instance is altered (to the orange curve).

The process starts with extracting a designated number of shapelet candidates of various lengths from the time series instances. Shapelets are subsequences of a time series instance that can be extracted by shapelet transform (Hills et al. 2014; Bostrom and Bagnall 2017). When it comes to instance-level explanations, shapelets can be directly projected onto the original time series, thus providing more informative and intuitive explanations that require minimal domain knowledge. A typical Shapelet Transform (ST) takes as input the training set and outputs extracted shapelets. The transform is composed of three steps: candidate extraction, computation of information gain per candidate, and shapelet selection. As a consequence of the selection, shapelets with less discriminative information (w.r.t. class labels) are filtered out, and those with relatively high discriminative power are retained for the perturbation steps. In this study, we utilize the Random Shapelet Transform (RST), which accommodates time limitations and random
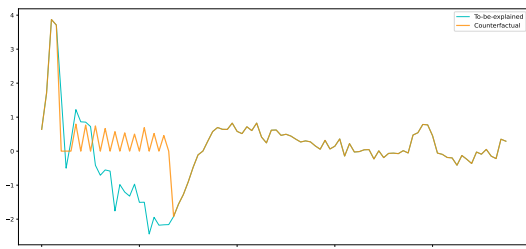

Figure 2: An example of a time series counterfactual instance, where the blue curve represents the to-be-explained shapelet of the electrocardiogram instance from the UCR ECG200 (Chen et al. 2015). The orange curve stands for its generated counterfactual subsequence, which drives the classifier to alter its prediction.

candidate extraction, thus providing time-saving and flexibility advantages in comparison to the complete shapelet transform. After shapelet extraction, the collected candidates are sorted by their respective information gain in descending order and then filtered to the top N (the number of retained shapelets). Inspired by the uprising interest in utilizing generative models for XAI in time series analysis (Van Looveren et al. 2021; Lang et al. 2022; Dantas, Marcos, and Ienco 2023), we likewise alternatively apply generative adversarial networks dedicated for time series data namely, *TimeGAN* (Yoon, Jarrett, and Van der Schaar 2019), as our counterfactual generation model.

Algorithm 1 depicts how the counterfactual instances are generated. With a given to-be-explained time series instance T with class label l, we first train a TimeGAN using all instances from the training data except for those with label l. The trained TimeGAN is then used to generate artificial instances. The final output sequences of TimeGAN can closely resemble the sequence-level distribution of the training data while preserving the within-sequence stepwise transition (Yoon, Jarrett, and Van der Schaar 2019). As depicted in lines 5 to 7 of the pseudo-code, a total of M fake instances, each of length N, are obtained based on M initial

Algorithm 1: Generating counterfactual explanations of a classifier f for a time series instance

> **Data:** shapelet candidates: S, a to-be-explained instance with class label l: T, the length of T: N, a to-be-explained classifier f, the maximum number of generated sequences: M
>
> **Result:** a collection of counterfactual instances: C

```
1  C ← ∅ ;                                      ▷ A set of counterfactual instances for T
2  F ← ∅ ;                                             ▷ A set of generated fake instances
3  Train a TimeGAN with instances of other labels;
4  for i ← 1 to M do
5  │    x ← [x₁, x₂, ..., x_N]ᵀ ∼ Uniform[0, 1] ;              ▷ Sample the initial vector
6  │    T_fake ← TimeGAN(x) ;                            ▷ Transform x to a fake sequence
7  │    F ← F ∪ {T_fake}
8  for s in S do
9  │    start_pos, s_len ← s.info ;               ▷ Get the starting index and length of s
10 │    if s.class == l then
   │           ▷ Retrieve shapelets that belong to the same interval as s from each T_fake in F
11 │    │    S_fake ← crop(start_pos, s_len, F);
12 │    │    for s_j in S_fake do
13 │    │    │    T_fake ← replace(T, s_j, start_pos, s_len) ;  ▷ Replace the target subsequence in T with s_j
14 │    │    │    if f(T_fake) ≠ s.class then
15 │    │    │    │    C ← C ∪ {T_fake}

16 return C
```

vectors. These initial vectors are uniformly sampled from $[0, 1]^N$. Next, the algorithm iterates over the shapelet candidate set S. For each candidate s, it is essential to identify its start and end positions within the corresponding instance. Further, the $crop()$ function is executed to extract the artificial shapelets S_{fake} by cropping each interval with the same time range (start and end positions) in each generated fake instance T_{fake} from F, that is, an instance from the same class. The final process includes iterating through the set of fake shapelets and using the $replace()$ method to replace the contiguous time points (ranging from the start position and end position stored previously) of the to-be-explained time series instance T, with one of the (fake) shapelets generated by TimeGAN. A synthetic time series instance T_{fake} is considered as a counterfactual only if it flips the prediction of the to-be-explained classifier, says, the predictive label of T_{fake} is no longer l. Moreover, this step reserves all the fake counterfactual time series instances for diverse explanations. In case of finding multiple counterfactual instances, the final recommended one is determined by:

$$T_c \leftarrow \underset{T_f \in C}{\operatorname{argmin}} \, d(T_f, T) \qquad (1)$$

where d is a (distance) metric defined over the given time series. Specifically, we minimize the Hamming distance in this study to ensure minimum perturbation actions.

Experimental Setup

Time-Cf is evaluated on four types of classifiers (Convolutional Neural Networks (CNNs), Diverse Representation Canonical Interval Forest (DrCIF), Catch22 with Random Forest (RF), and K Nearest Neighbor (KNN) Classifier. Each classifier is trained with four binary classifica-

tion time series data from the UCR time series classification archive (Chen et al. 2015). They are ECG200, an electrocardiogram; Wafer, sensor signals during silicon wafer processing for semiconductor fabrication; FordA, readings of engine noise for outlier detection; MoteStrain, sensor measurements of humidity or temperature. The results obtained by Time-CF on these four datasets are compared with those of Native Guide (Delaney, Greene, and Keane 2021a) and Mlxtend (Raschka 2018), where the former is a well-established counterfactual-based method (see related works) and the latter is an open-source library tailored for interpreting machine learning classifiers.

Analysis of Results

Although there is no consensus on the evaluation of counterfactual-based methods, closeness, sparsity, plausibility, and diversity have been widely used in recent studies (Delaney, Greene, and Keane 2021a; Mothilal, Sharma, and Tan 2020; Li, Boubrahimi, and Hamd 2022; Li et al. 2022). In a similar vein, we utilize the first three measures and replace diversity with sensibility in our evaluation since the existence of instance-wise explanations is more crucial and questionable, as observed in our study (see Fig. 4).

Closeness It measures the *shape* similarity between the counterfactual instance and the original instance in Manhattan distances (Delaney, Greene, and Keane 2021b). Figure 3 shows that the counterfactual instances made by our method are the closest to the instances being explained in all datasets. This reveals Time-CF's proficiency in perturbing the minimal, contiguous segment of the original instances.

Sensibility Sensibility serves as a metric to determine the extent to which counterfactual-based methodologies are
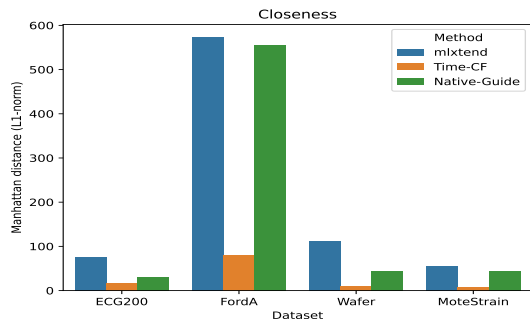
Figure 3: The results of L1-norm closeness. A low value is preferred, as it indicates less scale change.
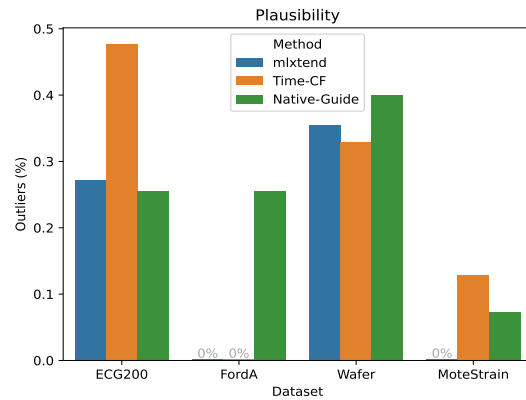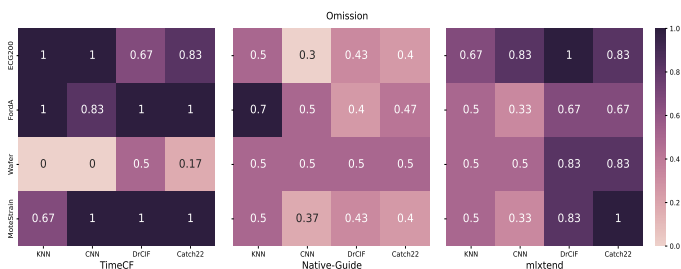

Figure 4: Assessment of sensibility, where **higher** values (darker in plot) are preferable. The figure displays the ratios of each post-hoc explainer in **successfully** finding counterfactual instances across four datasets.

Figure 5: Assessment of plausibility, where a lower value is preferable. The y-axis quantifies the outlier ratios of the generated counterfactual instances.
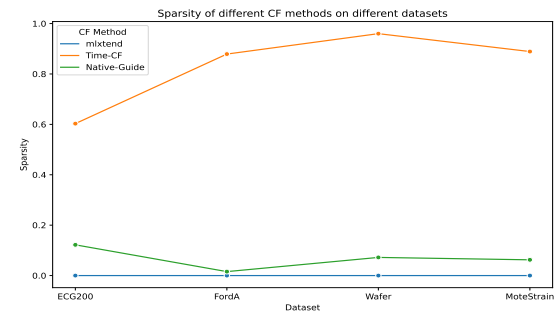

Figure 6: Measurement for sparsity (the higher, the better). This metric indicates the extent to which the number of time steps is altered. A value of 0 means that the counterfactual instances are derived from the original instances that have alterations at every time point.

model-agnostic or can explain any classifier. It records the proportion of instances per classifier that are explained by the given XAI approach. The results in Fig. 4 demonstrate that both **mlxtend** and **Native-Guide** fail to effectively handle a variety of classifiers across these datasets. In contrast, with the exception of the extremely imbalanced dataset *Wafer*, our method can aptly discern the informative features in time series data, thereby positioning itself as a model-agnostic tool that can, in theory, clarify any classifier.

Plausibility In this experiment, we delve deep into the percentage of generated instances identified as out of distribution. Given our emphasis on the importance of plausibility when generating counterfactual instances, we exclusively retain those counterfactuals with the lowest outlier rates to ensure heightened plausibility. As depicted in Fig. 5, our method outperforms others in terms of plausibility in dataset Wafer. However, it shows worse performance than the others when using an imbalanced dataset. Note that this metric is dependent on the outlier detection algorithm, which in this study is the Isolation Forest.

Sparsity Lastly, curve graph in Fig. 6 contrasts the sparsity level of each XAI method. When comparing the sparsity levels among **mlxtend** (blue), **Native-Guide** (green) and Time-CF (yellow), our approach consistently showcases superior performance. It is evident that both **mlxtend** and **Native-Guide** require modifications to nearly all the time

steps to alter the class of the to-be-explained instance from one to another, with values approaching a sparsity of 0%. In contrast, our approach only perturbs a minor segment, thereby offering not only clearer visualization for analytical objectives but also a more intuitive comprehension of which portions primarily influence the classifier.

Conclusion

In this work, a model-agnostic, counterfactual-based XAI method, **Time-CF** is introduced to explain any prediction for any classifier concerning time series classification tasks. **Time-CF** chiefly utilizes shapelets and TimeGAN to deliver meaningful counterfactual explanations, exhibiting exceptional performance in terms of closeness, sensibility, plausibility, and sparsity measures. The incorporation of shapelets indicates that a contiguous segment of the time series instance is perturbed, thus offering intuitive insights for the analysis of anomalies for end-users. In addition, TimeGAN ensures that at least one counterfactual explanation is gen-

erated, provided the training set is both ample and not extremely imbalanced. TimeGAN also guarantees that the explanations produced adhere to the data distribution.

In future work, we aim to refine our methodology to cater to multi-class and multivariate time series classification tasks. Moreover, a more profound exploration of the synergies between TimeGAN and Shapelet Transform (ST) will be undertaken to address the suboptimal performance observed with imbalanced datasets, and a wider set of experiments will be conducted to assess the behavior and performance of the proposed method against other purely GAN-based methods (Lang et al. 2023).

Acknowledgement

References

Bostrom, A.; and Bagnall, A. 2017. Binary Shapelet Transform for Multiclass Time Series Classification. In Hameurlain, A.; Küng, J.; Wagner, R.; Madria, S.; and Hara, T., eds., *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXXII: Special Issue on Big Data Analytics and Knowledge Discovery*, 24–46. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-662-55608-5.

Chen, Y.; Keogh, E.; Hu, B.; Begum, N.; Bagnall, A.; Mueen, A.; and Batista, G. 2015. The UCR Time Series Classification Archive. www.cs.ucr.edu/~eamonn/time_series_data/.

Dantas, C. F.; Marcos, D.; and Ienco, D. 2023. Counterfactual Explanations for Land Cover Mapping in a Multi-class Setting. *arXiv preprint arXiv:2301.01520*.

Delaney, E.; Greene, D.; and Keane, M. T. 2021a. Instance-based counterfactual explanations for time series classification. In *Case-Based Reasoning Research and Development: 29th International Conference, ICCBR 2021, Salamanca, Spain, September 13–16, 2021, Proceedings 29*, 32–47. Springer.

Delaney, E.; Greene, D.; and Keane, M. T. 2021b. Uncertainty estimation and out-of-distribution detection for counterfactual explanations: Pitfalls and solutions. *arXiv preprint arXiv:2107.09734*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Hills, J.; Lines, J.; Baranauskas, E.; Mapp, J.; and Bagnall, A. 2014. Classification of time series by shapelet transformation. *Data mining and knowledge discovery*, 28: 851–881.

Lang, J.; Giese, M.; Ilg, W.; and Otte, S. 2022. Generating Sparse Counterfactual Explanations For Multivariate Time Series. *arXiv preprint arXiv:2206.00931*.

Lang, J.; Giese, M. A.; Ilg, W.; and Otte, S. 2023. Generating Sparse Counterfactual Explanations for Multivariate Time Series. In Iliadis, L.; Papaleonidas, A.; Angelov, P.; and Jayne, C., eds., *Artificial Neural Networks and Machine Learning – ICANN 2023*, 180–193. Cham: Springer Nature Switzerland. ISBN 978-3-031-44223-0.

Li, P.; Bahri, O.; Boubrahimi, S. F.; and Hamdi, S. M. 2022. SG-CF: Shapelet-Guided Counterfactual Explanation for Time Series Classification. In *2022 IEEE International Conference on Big Data (Big Data)*, 1564–1569. IEEE.

Li, P.; Boubrahimi, S. F.; and Hamd, S. M. 2022. Motif-guided Time Series Counterfactual Explanations. *arXiv preprint arXiv:2211.04411*.

Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 607–617.

Raschka, S. 2018. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *The Journal of Open Source Software*, 3(24).

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5): 206–215.

Sarker, I. H. 2021a. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, 2(6): 420.

Sarker, I. H. 2021b. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3): 160.

Schlegel, U.; Arnout, H.; El-Assady, M.; Oelke, D.; and Keim, D. A. 2019. Towards a rigorous evaluation of XAI methods on time series. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 4197–4201. IEEE.

Theissler, A.; Spinnato, F.; Schlegel, U.; and Guidotti, R. 2022. Explainable AI for Time Series Classification: A review, taxonomy and research directions. *IEEE Access*.

Van Looveren, A.; Klaise, J.; Vacanti, G.; and Cobb, O. 2021. Conditional generative models for counterfactual explanations. *arXiv preprint arXiv:2101.10123*.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31: 841.

Yoon, J.; Jarrett, D.; and Van der Schaar, M. 2019. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32.