Shapelet-Based Counterfactual Explanations for Multivariate Time Series

Omar Bahri
omar.bahri@usu.edu
Utah State University
Logan, Utah, USA

Soukaina Filali Boubrahimi
soukaina.boubrahimi@usu.edu
Utah State University
Logan, Utah, USA

Shah Muhammad Hamdi
shamdi1@nmsu.edu
New Mexico State University
Las Cruces, New Mexico, USA

ABSTRACT

As machine learning and deep learning models have become highly prevalent in a multitude of domains, the main reservation in their adoption for decision-making processes is their black-box nature. The Explainable Artificial Intelligence (XAI) paradigm has gained a lot of momentum lately due to its ability to reduce models opacity. XAI methods have not only increased stakeholders' trust in the decision process but also helped developers ensure its fairness. Recent efforts have been invested in creating transparent models and post-hoc explanations. However, fewer methods have been developed for time series data, and even less when it comes to multivariate datasets. In this work, we take advantage of the inherent interpretability of shapelets to develop a model agnostic multivariate time series (MTS) counterfactual explanation algorithm. Counterfactuals can have a tremendous impact on making black-box models explainable by indicating what changes have to be performed on the input to change the final decision. We test our approach on a real-life solar flare prediction dataset and prove that our approach produces high-quality counterfactuals. Moreover, a comparison to the only MTS counterfactual generation algorithm shows that, in addition to being visually interpretable, our explanations are superior in terms of proximity, sparsity, and plausibility.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; *Causal reasoning and diagnostics.*

KEYWORDS

counterfactual explanations, multivariate time series, shapelets

ACM Reference Format:
Omar Bahri, Soukaina Filali Boubrahimi, and Shah Muhammad Hamdi. 2018. Shapelet-Based Counterfactual Explanations for Multivariate Time Series. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/1122445.1122456

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

During the last decade, machine learning and deep learning models have established themselves as the state-of-the-art in multiple scientific and data-centric domains. Their high performance in prediction and classification tasks has allowed them to be adopted in a plethora of fields such as banking, insurance, healthcare, and meteorology (Angra and Ahuja, 2017, Sarker, 2021, Shinde and Shah, 2018). Despite this huge success, one of the main issues still hindering their full deployment is the black-box nature of most algorithms, and the limitations it imposes in terms of interpretability and explainability. This in turn brings on questions about the fairness of such algorithms, making them difficult to include in critical decision-making processes. In this context, the EU General Data Protection Regulation (GDP, 2018), a law that insists on the importance of fairness, trustworthiness, and privacy, and that urges companies to provide explanations to users and consumers, has been introduced in 2016. In addition, initiatives such as DARPA's Explainable AI (XAI) (Gunning, 2016) have been started by different agencies to develop interpretable machine learning solutions. As a result, more transparent machine learning algorithms such as decision trees and linear regression have seen increased interest. In parallel, a novel line of research that focuses on providing post hoc interpretable explanations generated by extra modules on top of the complex black-box models was met with great success. While most of these methods focused on image data (Van Looveren and Klaise, 2019), tabular data (Lundberg and Lee, 2017, Schlegel et al., 2019, Van Looveren and Klaise, 2019), and text data (Ribeiro et al., 2016), more recent works started tackling univariate time series (Ates et al., 2021, Delaney et al., 2020, Guidotti et al., 2020, Parvatharaju et al., 2021). However, because of their high-dimensionality, multivariate time series (MTS) data remain under-explored, with CoMTE (Ates et al., 2021) being the only counterfactual explanation method specifically designed for MTS. In this work, we capitalize on the inherent interpretability of shapelet-based algorithms in time series classification (TSC) to develop a shapelet explainer for time series (SETS), a model-agnostic counterfactual generation algorithm for MTS data. We evaluate our results on an open-source solar-flare dataset. The rest of this paper is organized as follows. In Section 2, we present the state-of-the-art machine learning explanation methods, with a particular focus on time series. In Section 3, we introduce SETS and discuss some important counterfactual evaluation measures. In Section 4, we present the experimental setup. In Section 5, we discuss the results. And in Section 6, we conclude with a summary.

2 RELATED WORK

One of the most popular model-agnostic explanation methods is LIME, developed by Ribeiro et al. (2016). To explain a dataset sample, LIME generates random neighboring instances around it by performing small perturbations. Then, a surrogate linear model that mimics the behavior of the original black-box model is trained on the generated instances, and its feature importance values are used to explain the model decision. One of the main drawbacks of LIME is its assumption of linearity, which rarely holds true when it comes to high-dimensional and complex time series datasets (Parvatharaju et al., 2021). Inspired by game theory, SHAP (Lundberg and Lee, 2017) is another feature-based explanation method that overcomes this linearity limitation. It computes the importance of features by deriving their additive Shapley values. Although more robust than LIME, both methods raise concerns about their stability, as slight perturbations can totally change the model decision (Adebayo et al., 2018, Delaney et al., 2020, Nguyen et al., 2020). Instance-based explanation methods were developed as an alternative to feature-based approaches. In particular, counterfactuals, artificial instances generated as close as possible to the original dataset sample in such a way that the model prediction changes have gained increased popularity. In general, counterfactuals are generated by introducing perturbations from a representative sample of the target class or guided by an objective function (or both). Wachter et al. (2018) were among the first to generate counterfactual explanations. Their method consists in minimizing a loss function consisting of a prediction term to reach the target class label, and a distance term to ensure the counterfactual lies close to the original instance. Mothilal et al. (2020) extend this approach by adding a diversity constraint to allow the generation of different counterfactuals. Dhurandhar et al. (2018) introduced an autoencoder-based loss term to enforce interpretability by keeping the counterfactuals within the target class data distribution. In addition, Van Looveren and Klaise (2019) add a prototype loss term to ensure interpretability and speed up the search process. Even though none of the techniques mentioned so far have been proposed for time series data, they might in theory be used in this context. For example, an apriori segmentation of the data –although it results in some information loss– has made it possible for SHAP and LIME to be applied to time series (Schlegel et al., 2019). In the work by Ates et al. (2021), 11 statistical features have been extracted from the data. However, the results have not been satisfactory (Ates et al., 2021). Similarly, instance-based explanation methods do not yield the best results when employed for time series (Delaney et al., 2020). More recently, a few techniques designed for time series datasets have made their way to the literature. Guidotti et al. (2020) build a decision tree using shapelets (Grabocka et al., 2014) extracted from the dataset. Then, they extract shapelet-based rules from the decision tree to explain the black-box model decisions. In native guide (NG) (Delaney et al., 2020), the original instance's nearest-unlike-neighbor (nun) is extracted from the target class and used to perturb the original instance. When possible, the most important contiguous subsequence is found using Class Activation Mapping (Zhou et al., 2015), and the perturbations are introduced at its level. Otherwise, dynamic time warping barycenter averaging (DBA) (Petitjean et al., 2011) is used. Ates et al. (2021) introduced CoMTE, a counterfactual multivariate time series explainability method. CoMTE extracts a nun for each feature variable using KD-trees, and replaces entire dimensions to build a counterfactual. The authors proposed a heuristic search method based on hill climbing to select the feature variables to be modified. In case the heuristic fails to provide a counterfactual, a greedy search is performed to find the optimal feature set. To our knowledge, this is the first counterfactual generation algorithm developed for MTS.

3 METHODOLOGY

3.1 Shapelet Transform

A shapelet is a phase-independent, characteristic subsequence that occurs repeatedly in a time series dataset. Since the first shapelet-based classification algorithm was introduced by Ye and Keogh (2011), multiple works followed suit (Fang et al., 2018, Grabocka et al., 2014, Lines et al., 2012, Rakthanmanon and Keogh, 2013). In a recent time series classification benchmark study by Bagnall et al. (2017), shapelet transform (ST) (Bostrom and Bagnall, 2015, Hills et al., 2014, Lines et al., 2012) proved to be among the best algorithms. Therefore, we use it to extract shapelets for our algorithm. Moreover, shapelet discovery and classification in ST are performed in two separate steps, which makes its choice even more convenient.

The first step involved in mining shapelets using ST is to extract all candidate shapelets S_i from each dataset sample, i.e. all subsequences of the desired predefined lengths. In the next step, each S_i is slided across each dataset sample, and the minimum distance that separates it from all subsequences w of similar length is recorded as the distance to that dataset sample. Equation 1 defines the sliding window function, such as the set of all subsequences of length equal to that of S is represented by W. Then, the final set of shapelets is selected based on their information gain.

$$sDist(S, T) = min_{w \in W}(dist(S, w)) \tag{1}$$

3.2 Shapelet Explainer for Time Series

In this section, we present SETS, an instance-based, model-agnostic counterfactual generation algorithm.

3.2.1 Model Fitting. Given the inherent interpretability of shapelets and the success of shapelet-based algorithms, we consider the shapelets extracted using ST as the building blocks of our counterfactual generation algorithm. During the computation of the distances separating shapelets from dataset instances, we store the distances between every shapelet and its potential occurrences (subsequences of the same length) and retain the closest ones according to a predefined threshold. These retained shapelets occurrences are then used to select class-shapelets: shapelets characteristic of each dataset class, i.e. those that happen under that class only. The remaining shapelets are discarded. Then, for each shapelet, the occurrence distribution is computed as the average of all its occurrences.

3.2.2 Counterfactual Generation. Given a dataset instance x of class A, generating a counterfactual instance x_{cf} of class B starts with the extraction of the nearest-neighbor x_{nn} of x from class B

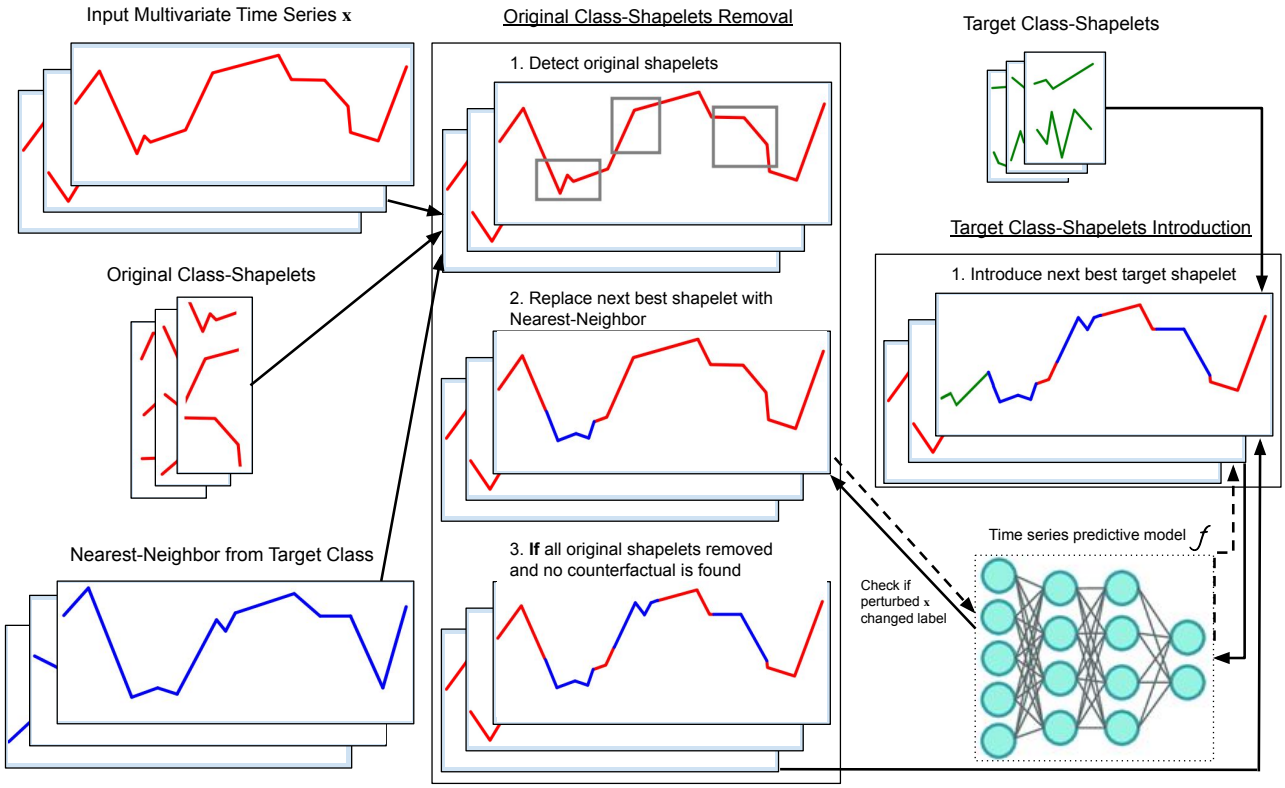


Figure 1: SETS single dimension perturbation process

instances using the k-Nearest-Neighbor (kNN) algorithm. Next, the dataset dimensions are sorted in descending order according to their highest scoring shapelets. Then, steps *a* and *b* below are performed for each dimension, class-shapelet by class-shapelet and until a valid counterfactual x_{cf} is found, i.e. until $argmax(f(x_{cf})) = B$, where f is the prediction function of the black-box model. Figure 1. illustrates the process for a single dimension. If none of the dimensions succeed in creating a valid counterfactual, the perturbations from all possible subsets of dimensions are combined until a valid counterfactual is found.

Original Class-Shapelets Removal. x_{cf} is constructed by replacing each A class-shapelet –by descending order of information gain– contained in x by the values of x_{nn} at the same time steps, scaled to the original feature range using min-max scaling. The motivation behind this step is to remove the shapelets that swayed the model prediction $f(x_{cf})$ toward label A.

Target Class-Shapelets Introduction. x_{cf} is constructed by introducing each B class-shapelet –by descending order of information gain– into the time steps defined by its occurrence distribution, scaled to the original feature range using min-max scaling. The motivation behind this step is to influence the model prediction $f(x_{cf})$ towards label B.

3.3 Evaluation Measures

By relying on shapelets in the counterfactual generation process, we are able to build highly interpretable explanations. On the one hand, performing changes at the level of shapelets only creates meaningful perturbations. This adds value to the counterfactuals by making them easily interpretable by stakeholders. On the other hand, being able to visualize the shapelets and see their impact on tasks such as classification and prediction, independently of the counterfactuals, increases the confidence in the SETS algorithm. However, there is still a need for quantitative evaluation, particularly when it comes to comparing different counterfactual generation approaches. Although there is no standard to evaluate explanation methods (Ates et al., 2021, Lipton, 2016, Schmidt and Biessmann, 2019), proximity, interpretability, and sparsity have been used repeatedly in the literature (Ates et al., 2021, Delaney et al., 2020, Karimi et al., 2019, Mothilal et al., 2020, Van Looveren and Klaise, 2019). In this section, we introduce the three measures and describe how we compute them to evaluate our approach.

3.3.1 Proximity. Also referred to as distance or closeness, the proximity measure ensures that the counterfactual is close to the original sample. Ideally, the perturbations should be as small as possible. However, proximity is not the only criteria sought in a good counterfactual. Therefore, a balance has to be found with the two measures below. Following (Delaney et al., 2020, Downs et al., 2020, Karimi

et al., 2019, Keane et al., 2021), we decided to use three distance metrics to evaluate the proximity of counterfactuals. We use the Manhattan distance (L_1-norm, equation 2) and the Euclidian distance (L_2-norm, equation 3) to measure the distance between the counterfactual x_{cf} and its original instance x, and the L_{inf}-norm (equation 3) to get the magnitude of the highest perturbation at a single time step. D is the number of dimensions in the multivariate dataset, and T is the length of the series.

$$\left\|x - x_{cf}\right\|_{L_1} = \sum_{i}^{D} \sum_{j}^{T} \left|x_{ij} - x_{cf\,ij}\right| \tag{2}$$

$$\left\|x - x_{cf}\right\|_{L_2} = \sqrt{\sum_{i}^{D} \sum_{j}^{T} \left(x_{ij} - x_{cf\,ij}\right)^2} \tag{3}$$

$$\left\|x - x_{cf}\right\|_{L_{inf}} = \sum_{i}^{D} \sum_{j}^{T} max \left|x_{ij} - x_{cf\,ij}\right| \tag{4}$$

3.3.2 Sparsity. Another important quality in a good counterfactual is the sparsity of the perturbed features and in the case of time series, of the perturbed time steps. Introducing changes to several features and at different time steps will not only make the explanation harder to comprehend by stakeholders but might even make it impossible to perform. Therefore, perturbed features should be kept to a minimum. In addition, time series data should be changed at the level of short, contiguous intervals, as perturbations that affect single, dispersed time steps are not meaningful (Delaney et al., 2020, Parvatharaju et al., 2021).

3.3.3 Plausibility. Being close to the original sample and having sparse perturbations might not be enough to warrant the plausibility of a counterfactual. Counterfactual explanations need to be easily interpretable by humans. Thus, they must be realistic. One way to quantitatively study the plausibility of a counterfactual is to check whether it belongs to the same data manifold of the dataset (Delaney et al., 2020, Karimi et al., 2019, Poyiadzi et al., 2019, Van Looveren and Klaise, 2019). This can be achieved by applying novelty detection techniques, which detect out-of-distribution (OOD) instances.

Similarly to the work by Delaney et al. (2020), we adopt three novelty detection approaches to assess the plausibility of counterfactuals. In particular, we perform (1) the local outlier factor (LOF) method (Breuniq et al., 2000, Kanamori et al., 2020) which calculates the local density deviation of each instance with respect to its neighbors and flags the ones with lower densities as outliers, (2) isolation forest (IF) (Liu et al., 2008) which considers how far a dataset instance is to the rest of the dataset, and the one class support vector machine (OC-SVM) (Schölkopf et al., 2001) method (on the raw time and the matrix profile (OC-SVM MP) (Yeh et al., 2017) representations of the time series).

4 EXPERIMENTAL SETUP

4.1 Dataset

A solar flare is an extremely powerful burst of electromagnetic radiation originating from the surface of the sun. Because of their sporadicity, forecasting them remains a big challenge. Furthermore, solar flares are highly dangerous for astronauts, space equipment, and can even damage infrastracture at ground surface such as electric power grids and navigational signals (Boubrahimi et al., 2018, Martens and Angryk, 2017). Therefore, forecasting solar flare events is very important to perform critical preventive measures and save billions of dollars worth of damage (Angryk et al., 2020a). Since solar flares are rare events, the datasets available for training forecasting models are small, which makes their outputs less dependable. Thus, generating post hoc explanations in the form of counterfactuals has the potential of increasing their trustworthiness in the eyes of stakeholders, including physicists and decision makers.

We evaluate SETS on a real-life, MTS solar flare dataset created by a research group from Georgia State University (Angryk et al., 2020a). The dataset contains 1354 samples of 60 time steps, recorded at 12 minutes intervals for a total of 12 hours. The class distribution of the dataset samples is shown in Table 1. The four class labels represent the classification of the most powerful solar flare recorded in the following 12 hours. The data was captured by the Helioseismic Magnetic Imager (HMI) (Angryk et al., 2020b, Bobra et al., 2014, Schou et al., 2012) on the Solar Dynamics Observatory (SDO) (Pesnell et al., 2012) run by NASA.

Table 1: Dataset Class Distribution

Class	X	M	B/C	Q	All
Number of elements	303	350	356	345	1354

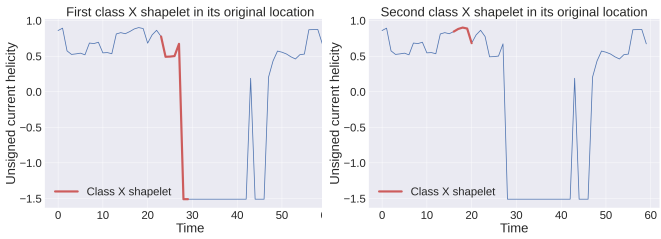
4.2 Implementation Details

We used the sktime (Löning et al., 2019) implementation of ST, and introduced a slight modification to extract the indices of the occurrences of each shapelet along with their distances as described in Section 3.2. Because of the high-dimensionality of the dataset, we ran the contracted shapelet transform implementation, which does not significantly hurt the performance of ST (Bostrom and Bagnall, 2017). This approach consists in randomly selecting shapelets for a user-defined amount of time, instead of trying all possible subsequences. For this experiment, we ran the algorithm for a total of 4 hours, including all dimensions. In order to achieve sparse perturbations, the length of the shapelets was restricted to a maximum of 50% of the length of the time series. We provide access to our code and to the solar-flare dataset on our GitHub repository.

4.3 Compared Methods

4.3.1 NG. It extracts the nearest neighbor to the original instance from the target class and introduces perturbations from its time steps. We employ the model agnostic version of the algorithm, which uses DBA to guide the perturbations. The implementation is provided by the authors in the original publication Delaney et al. (2020). Since NG was developed for univariate time series datasets, we adapt it to the multivariate case by simply reshaping the entire dataset under one dimension.

4.3.2 CoMTE. To our knowledge, CoMTE (Ates et al., 2021) is the only counterfactual explanation method developed specifically for



(a) (b)

Figure 2: Class X Shapelets

(a) (b)

Figure 3: Counterfactual generation using target class-shapelet introduction

MTS. First, CoMTE finds a distractor from the target class by constructing its KD-tree and considering the original instance's nearest neighbor. Then, it picks a small set of feature dimensions from the original sample and replaces it with the distractor's variables, resulting in a counterfactual with the target class label. Ideally, the set of feature dimensions should be optimal, which can be achieved using a greedy search. To speed up the search process, the authors proposed a heuristic method based on hill climbing, followed by a post hoc trimming step. In case the heuristic fails to provide a counterfactual, the greedy search is performed. The implementation we use is provided in the original publication.

5 EXPERIMENTAL RESULTS

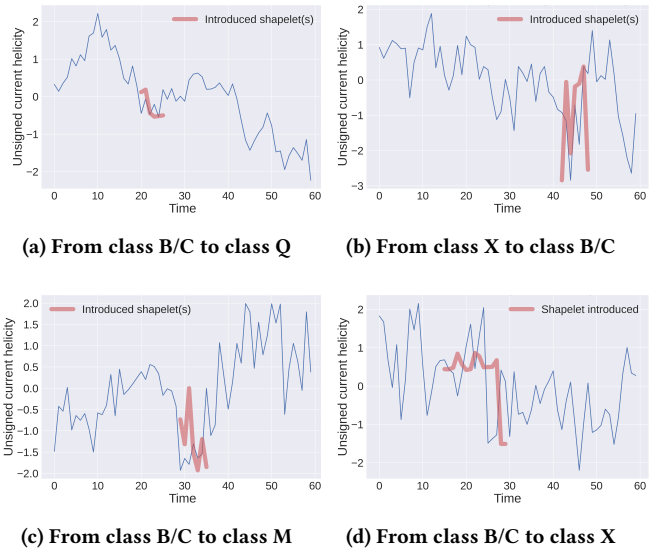
We split the solar flare dataset into a training set with 70% of the instances and a testing set with the remaining 30%. Then, we run ST on the training set as described in section 4.2, and use the extracted shapelets to create our explanations as detailed in section 3.2. Since the dataset contains 4 classes, we generate 3 counterfactuals for each instance –one for each of the possible target classes– for a total of 1221 counterfactuals. While we use a simple neural network as the black-box model, NG, CoMTE, SETS are model agnostic and can be applied to any other machine learning model.

5.1 Qualitative Evaluation

SETS starts by finding class-shapelets, i.e. shapelets that occur under one class only. By first examining these shapelets, the user can have a better understanding of their discriminative power and their role in the dataset. Domain experts can even go a step further and draw important insights into the problem at hand. Figure 2 shows two shapelets that happen uniquely under class X, at the level of the total unsigned current helicity feature. This means that whenever one of these two shapelets is detected, an extremely powerful solar-flare event is going to burst.

Then, SETS exploits the class-shapelets to create meaningful counterfactual explanations. For example, considering the unsigned current helicity dimension of the original time series sample in Figure 3.a, SETS generates a counterfactual of class X (from the original class M) by simply introducing the shapelet from Figure 2.a. The perturbed dimension of the counterfactual is plotted in Figure 3.b.

In Figure 4, we show 4 more counterfactuals generated using SETS. For visualization purposes, we select counterfactuals that required perturbations at the level of one time series dimension only.



(a) From class B/C to class Q (b) From class X to class B/C

(c) From class B/C to class M (d) From class B/C to class X

Figure 4: Counterfactual explanations generated using SETS

However, SETS perturbed an average of 2.62 dimensions through the testing set. On the other hand, CoMTE substitutes entire dimensions to generate counterfactuals, which makes them less qualitatively interpretable. Moreover, in this experiment, significantly more dimensions were perturbed compared to SETS. The original univariate version of NG perturbs all time steps. Therefore, it suffers from the same problems as CoMTE in the multivariate adaptation.

5.2 Quantitative Evaluation

Out of the 1221 counterfactuals generated by NG, 11 are largely out-of-distribution. This is an expected outcome, as NG keeps incrementing the DBA weights until a counterfactual found. Therefore, the average evaluation measure values are highly affected by these 11 samples. In order, to fairly compare the three methods, we present both the means and medians of the the proximity and sparsity measures, and we show the plausibility results with and without these 11 outliers. However, it is important to note that 0.9% of NG counterfactuals are significantly worse than average.

Table 2: Proximity comparison

	L1-norm		L2-norm		Linf-norm	
	Mean	Median	Mean	Median	Mean	Median
NG	1.42 x 10^{12}	836.68	1.38 x 10^{12}	28.50	1.37 x 10^{12}	4.17
CoMTE	2132.17	2148.37	60.81	6.95	6.96	6.96
SETS	**90.64**	**86.41**	**10.11**	**10.95**	**2.64**	**2.47**

Table 3: Sparsity comparison

	Sparsity	
	Mean	Median
NG	135.69	130.01
CoMTE	159.38	159.65
SETS	**9.29**	**8.11**

Table 4: Plausibility comparison

		NG	CoMTE	SETS
IF	w/ NG outliers	0.28	0.004	**0.0**
	w/o NG outliers	0.20	0.005	**0.0**
LOF	w/ NG outliers	0.05	**0.0**	**0.0**
	w/o NG outliers	0.04	**0.0**	**0.0**
OC-SVM	w/ NG outliers	0.08	0.82	**0.06**
	w/o NG outliers	0.06	0.82	**0.06**
OC-SVM MP	w/ NG outliers	0.076	0.816	**0.057**
	w/o NG outliers	0.065	0.816	**0.058**

5.2.1 Proximity. As shown in Table 2, the counterfactuals generated by SETS are significantly closer to the original instances than those created using NG and CoMTE. This can be observed in all three metrics: the Manhattan distance (L_1-norm), the Euclidian distance (L_2-norm), and the L_{inf}-norm. This means that not only are the counterfactuals closer to the original instances but that the perturbations are also much smaller in magnitudes; which suggests that NG and CoMTE counterfactuals contain more pronounced spikes. Table 2 also shows that NG produces closer counterfactuals than CoMTE, except for the 11 outliers discussed above (very high mean and lower median compared to CoMTE's).

5.2.2 Sparsity. We compute sparsity as the total number of perturbed time steps throughout all time series dimensions. Again, Table 3 shows that the counterfactual explanations generated by SETS are superior to NG's and CoMTE's.

5.2.3 Plausibility. Table 4 shows that all plausibility measures consistently point to the fact that SETS generates significantly more plausible counterfactuals. The OC-SVM method on both the raw time space and the matrix profile favorises NG's counterfactuals over those generated by CoMTE. However, the IF and LOF methods have detected fewer outliers among CoMTE's.

6 CONCLUSION

In this work, we proposed SETS, a model agnostic MTS counterfactual explanation algorithm. SETS makes use of shapelets to introduce meaningful perturbations to the original dataset instance, and to create highly interpretable counterfactuals. In particular, the perturbations introduced by SETS are contiguous, which makes the counterfactual explanations more plausible than the sparse ones generated by CoMTE. In addition, the use of shapelets provides SETS with the added quality of visual explainability. Indeed, plotting the counterfactuals and visualizing the perturbations can provide important insight as to which shapelets influenced the black-box model decision. We tested our approach on a real-life solar flare prediction dataset using a neural network as the black-box model and compared it to two state-of-the-art time series counterfactual generation algorithms, including the only one specifically developed for MTS. The results show that SETS' counterfactuals are superior in terms of proximity, sparsity, and plausibility, with the additional visual interpretability edge. In the future, we would like to experiment with an optimization-based approach, guided by a shapelet-based loss function.

REFERENCES

NoEU General Data Protection RegulationTitle. Technical report, 2018. URL https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rulesen.

J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity Checks for Saliency Maps. *Advances in Neural Information Processing Systems*, 2018-Decem: 9505–9515, oct 2018. ISSN 10495258. doi: 10.48550/arxiv.1810.03292. URL https://arxiv.org/abs/1810.03292v3.

S. Angra and S. Ahuja. Machine learning and its applications: A review. *Proceedings of the 2017 International Conference On Big Data Analytics and Computational Intelligence, ICBDACI 2017*, pages 57–60, oct 2017. doi: 10.1109/ICBDACI.2017.8070809.

R. A. Angryk, P. C. Martens, B. Aydin, D. Kempton, S. S. Mahajan, S. Basodi, A. Ahmadzadeh, X. Cai, S. Filali Boubrahimi, S. M. Hamdi, M. A. Schuh, and M. K. Georgoulis. Multivariate time series dataset for space weather data analytics. *Scientific Data*, 7(1):1–13, dec 2020a. ISSN 20524463. doi: 10.1038/s41597-020-0548-x. URL www.nature.com/scientificdata.

R. A. Angryk, P. C. Martens, B. Aydin, D. Kempton, S. S. Mahajan, S. Basodi, A. Ahmadzadeh, X. Cai, S. Filali Boubrahimi, S. M. Hamdi, M. A. Schuh, and M. K. Georgoulis. Multivariate time series dataset for space weather data analytics. *Scientific Data*, 7(1):227, 2020b. ISSN 2052-4463. doi: 10.1038/s41597-020-0548-x. URL https://doi.org/10.1038/s41597-020-0548-x.

E. Ates, B. Aksar, V. J. Leung, and A. K. Coskun. Counterfactual Explanations for Multivariate Time Series. In *International Conference on Applied Artificial Intelligence (ICAPAI)*, pages 1–8. Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICAPAI49758.2021.9462056. URL http://arxiv.org/abs/2008.10781http://dx.doi.org/10.1109/ICAPAI49758.2021.9462056.

A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3):606–660, may 2017. ISSN 1573756X. doi: 10.1007/s10618-016-0483-9.

M. G. Bobra, X. Sun, J. T. Hoeksema, M. Turmon, Y. Liu, K. Hayashi, G. Barnes, and K. D. Leka. The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARPs – Space-Weather HMI Active Region Patches. *Solar Physics*, 289(9):3549–3578, 2014. ISSN 1573-093X. doi: 10.1007/s11207-014-0529-3. URL https://doi.org/10.1007/s11207-014-0529-3.

A. Bostrom and A. Bagnall. Binary shapelet transform for multiclass time series classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9263, pages 257–269. Springer Verlag, 2015. ISBN 9783319227283. doi: 10.1007/978-3-319-22729-0_20.

A. Bostrom and A. Bagnall. Binary shapelet transform for multiclass time series classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10420 LNCS:24–46, dec 2017. ISSN 16113349. doi: 10.1007/978-3-662-55608-5_2. URL http://arxiv.org/abs/1712.06428.

S. F. Boubrahimi, R. Ma, B. Aydin, S. M. Hamdi, and R. Angryk. Scalable kNN Search Approximation for Time Series Data. In *Proceedings - International Conference on Pattern Recognition*, volume 2018-Augus, pages 970–975. Institute of Electrical and Electronics Engineers Inc., nov 2018. ISBN 9781538637883. doi: 10.1109/ICPR.2018.8546103.

M. M. Breuniq, H. P. Kriegel, R. T. Ng, and J. Sander. LOF. *ACM SIGMOD Record*, 29(2):93–104, may 2000. ISSN 01635808. doi: 10.1145/335191.335388. URL https://dl.acm.org/doi/abs/10.1145/335191.335388.

E. Delaney, D. Greene, and M. T. Keane. Instance-based Counterfactual Explanations for Time Series Classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12877 LNAI:32–47, sep 2020. ISSN 16113349. doi: 10.1007/978-3-030-86957-1_3. URL https://arxiv.org/abs/2009.13211v2.

A. Dhurandhar, P. Y. Chen, R. Luss, C. C. Tu, P. Ting, K. Shanmugam, and P. Das. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. *Advances in Neural Information Processing Systems*, 2018-Decem: 592–603, feb 2018. ISSN 10495258. doi: 10.48550/arxiv.1802.07623. URL https://arxiv.org/abs/1802.07623v2.

M. Downs, J. Chu, Y. Yacoby, F. Doshi-Velez, and P. WeiWei. CRUDS: Counterfactual Recourse Using Disentangled Subspaces. *ICML Workshop on Human Interpretability in Machine Learning*, 2020.

Z. Fang, P. Wang, and W. Wang. Efficient learning interpretable shapelets for accurate time series classification. *Proceedings - IEEE 34th International Conference on Data Engineering, ICDE 2018*, pages 497–508, oct 2018. doi: 10.1109/ICDE.2018.00052.

J. Grabocka, N. Schilling, M. Wistuba, and L. Schmidt-Thieme. Learning time-series shapelets. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 392–401, New York, NY, USA, aug 2014. Association for Computing Machinery. ISBN 9781450329569. doi: 10.1145/2623330.2623613. URL https://dl.acm.org/doi/10.1145/2623330.2623613.

R. Guidotti, A. Monreale, F. Spinnato, D. Pedreschi, and F. Giannotti. Explaining any time series classifier. *Proceedings - 2020 IEEE 2nd International Conference on Cognitive Machine Intelligence, CogMI 2020*, pages 167–176, oct 2020. doi: 10.1109/COGMI50398.2020.00029.

D. Gunning. Broad Agency Announcement Explainable Artificial Intelligence (XAI). Technical report, , Defense Advanced Research Projects Agency (DARPA), 2016.

J. Hills, J. Lines, E. Baranauskas, J. Mapp, and A. Bagnall. Classification of time series by shapelet transformation. *Data Mining and Knowledge Discovery*, 28(4): 851–881, may 2014. ISSN 13845810. doi: 10.1007/s10618-013-0322-1. URL https://link.springer.com/article/10.1007/s10618-013-0322-1.

K. Kanamori, T. Takagi, K. Kobayashi, and H. Arimura. DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization. *IJCAI International Joint Conference on Artificial Intelligence*, 3:2855–2862, jul 2020. ISSN 1045-0823. doi: 10.24963/IJCAI.2020/395. URL https://www.ijcai.org/proceedings/2020/395.

A.-H. Karimi, G. Barthe, B. Balle, and I. Valera. Model-Agnostic Counterfactual Explanations for Consequential Decisions. may 2019. doi: 10.48550/arxiv.1905.11190. URL https://arxiv.org/abs/1905.11190v5.

M. T. Keane, E. M. Kenny, E. Delaney, and B. Smyth. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. *IJCAI International Joint Conference on Artificial Intelligence*, pages 4466–4474, feb 2021. ISSN 10450823. doi: 10.48550/arxiv.2103.01035. URL https://arxiv.org/abs/2103.01035v1.

J. Lines, L. M. Davis, J. Hills, and A. Bagnall. A shapelet transform for time series classification. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 289–297, New York, New York, USA, 2012. ACM Press. ISBN 9781450314626. doi: 10.1145/2339530.2339579. URL http://dl.acm.org/citation.cfm?doid=2339530.2339579.

Z. C. Lipton. The Mythos of Model Interpretability. *Communications of the ACM*, 61 (10):35–43, jun 2016. ISSN 15577317. doi: 10.48550/arxiv.1606.03490. URL https://arxiv.org/abs/1606.03490v3.

F. T. Liu, K. M. Ting, and Z. H. Zhou. Isolation forest. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 413–422, 2008. ISSN 15504786. doi: 10.1109/ICDM.2008.17.

M. Löning, A. Bagnall, S. Ganesh, V. Kazakov, J. Lines, and F. J. Király. sktime: A Unified Interface for Machine Learning with Time Series. sep 2019. URL https://arxiv.org/abs/1909.07872v1.

S. M. Lundberg and S. I. Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 2017-Decem:4766–4775, may 2017. ISSN 10495258. doi: 10.48550/arxiv.1705.07874. URL https://arxiv.org/abs/1705.07874v2.

P. C. Martens and R. A. Angryk. Data Handling and Assimilation for Solar Event Prediction. *Proceedings of the International Astronomical Union*, 13(S335):344–347, 2017. ISSN 17439221. doi: 10.1017/S1743921318000510. URL https://doi.org/10.1017/S1743921318000510.

R. K. Mothilal, A. Sharma, and C. Tan. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. 2020. doi: 10.1145/3351095.3372850. URL https://doi.org/10.1145/3351095.3372850.

T. T. Nguyen, T. Le Nguyen, and G. Ifrim. A model-agnostic approach to quantifying the informativeness of explanation methods for time series classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12588 LNAI:77–94, 2020. ISSN 16113349. doi: 10.1007/978-3-030-65742-0_6/FIGURES/9.

P. S. Parvatharaju, R. Doddaiah, T. Hartvigsen, and E. A. Rundensteiner. Learning Saliency Maps to Explain Deep Time Series Classifiers. *International Conference on Information and Knowledge Management, Proceedings*, pages 1406–1415, oct 2021. doi: 10.1145/3459637.3482446.

W. D. Pesnell, B. J. Thompson, and P. C. Chamberlin. The Solar Dynamics Observatory (SDO). *Solar Physics*, 275(1-2):3–15, jan 2012. ISSN 00380938. doi: 10.1007/s11207-011-9841-3. URL https://ui.adsabs.harvard.edu/abs/2012SoPh..275....3P/abstract.

F. Petitjean, A. Ketterlin, and P. Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693, mar 2011. ISSN 0031-3203. doi: 10.1016/J.PATCOG.2010.09.013.

R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach. FACE: Feasible and Actionable Counterfactual Explanations. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, sep 2019. doi: 10.1145/3375627.3375850. URL http://arxiv.org/abs/1909.09369http://dx.doi.org/10.1145/3375627.3375850.

T. Rakthanmanon and E. Keogh. Fast shapelets: A scalable algorithm for discovering time series shapelets. In *Proceedings of the 2013 SIAM International Conference on Data Mining, SDM 2013*, pages 668–676. Siam Society, 2013. ISBN 9781611972627. doi: 10.1137/1.9781611972832.74.

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, pages 97–101, feb 2016. doi: 10.48550/arxiv.1602.04938. URL https://arxiv.org/abs/1602.04938v3.

I. H. Sarker. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science 2021 2:3*, 2(3):1–21, mar 2021. ISSN 2661-8907. doi: 10.1007/S42979-021-00592-X. URL https://link.springer.com/article/10.1007/s42979-021-00592-x.

U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, and D. A. Keim. Towards a Rigorous Evaluation of XAI Methods on Time Series. *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*, pages 4197–4201, sep 2019. doi: 10.48550/arxiv.1909.07082. URL https://arxiv.org/abs/1909.07082v2.

P. Schmidt and F. Biessmann. Quantifying Interpretability and Trust in Machine Learning Systems. jan 2019. doi: 10.48550/arxiv.1901.08558. URL https://arxiv.org/abs/1901.08558v1.

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13(7):1443–1471, jul 2001. ISSN 08997667. doi: 10.1162/089976601750264965. URL https://dl.acm.org/doi/abs/10.1162/089976601750264965.

J. Schou, P. H. Scherrer, R. I. Bush, R. Wachter, S. Couvidat, M. C. Rabello-Soares, R. S. Bogart, J. T. Hoeksema, Y. Liu, T. L. Duvall, D. J. Akin, B. A. Allard, J. W. Miles, R. Rairden, R. A. Shine, T. D. Tarbell, A. M. Title, C. J. Wolfson, D. F. Elmore, A. A. Norton, and S. Tomczyk. Design and Ground Calibration of the Helioseismic and Magnetic Imager (HMI) Instrument on the Solar Dynamics Observatory (SDO). *Solar Physics*, 275(1-2):229–259, jan 2012. ISSN 00380938. doi: 10.1007/s11207-011-9842-2. URL https://ui.adsabs.harvard.edu/abs/2012SoPh..275..229S/abstract.

P. P. Shinde and S. Shah. A Review of Machine Learning and Deep Learning Applications. *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*, jul 2018. doi: 10.1109/ICCUBEA.2018.8697857.

A. Van Looveren and J. Klaise. Interpretable Counterfactual Explanations Guided by Prototypes. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12976 LNAI:650–665, jul 2019. ISSN 16113349. doi: 10.48550/arxiv.1907.02584. URL https://arxiv.org/abs/1907.02584v2.

S. Wachter, B. Mittelstadt, and C. Russell. COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR. *Harvard Journal of Law & Technology*, 31(2), 2018. doi: 10.1177/1461444816676645. URL https://perma.cc/3HF6-G9DS];.

L. Ye and E. Keogh. Time series shapelets: A novel technique that allows accurate, interpretable and fast classification. *Data Mining and Knowledge Discovery*, 22 (1-2):149–182, jan 2011. ISSN 13845810. doi: 10.1007/s10618-010-0179-5. URL https://link.springer.com/article/10.1007/s10618-010-0179-5.

C.-C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh. Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets. pages 1317–1322, feb 2017. doi: 10.1109/ICDM.2016.0179.

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:2921–2929, dec 2015. ISSN 10636919. doi: 10.48550/arxiv.1512.04150. URL https://arxiv.org/abs/1512.04150v1.