Temporal Rule-Based Counterfactual Explanations for Multivariate Time Series

Omar Bahri*, Peiyu Li†, Soukaina Filali Boubrahimi‡ and Shah Muhammad Hamdi§
Department of Computer Science, Utah State University
Logan, UT, USA
Email: *omar.bahri@usu.edu, †peiyu.li@usu.edu, ‡soukaina.boubrahimi@usu.edu, §s.hamdi@usu.edu

2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA) | 978-1-6654-6283-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICMLA55696.2022.00200

***Abstract*—The black-box nature of machine learning models is the main reason impeding their full adoption in decision-making processes. In order to reduce models' opacity and overpass this challenge, major efforts that aim to increase stakeholders' trust and ensure the fairness of decisions are being made by the data mining community under the Explainable Artificial Intelligence (XAI) paradigm. The two main categories of solutions are 1) developing fully transparent algorithms and 2) providing post hoc explanations. However, the literature is rather scarce when it comes to time series data, and even more so in the context of multivariate time series. In this work, we aim to exploit the discriminative power of shapelets and temporal rules in time series mining and capitalize on their inherent interpretability to develop a model-agnostic, temporal rule counterfactual explainer (TeRCE) for multivariate time series datasets. Counterfactual explanations indicate how should the input change such that the decision output changes too. Thus, they can highly increase the interpretability of black-box models. We test TeRCE on five benchmark datasets from the UEA archive and prove that it produces high-quality counterfactuals. Moreover, we show that in addition to being visually and conceptually interpretable, our approach performs better than the state-of-the-art algorithms in terms of proximity, sparsity, and second in terms of plausibility.**

***Index Terms*—counterfactual explanations, multivariate time series, shapelets, temporal association rules**

I. Introduction

Fueled by the data explosion and the tremendous development of storage and processing capabilities, machine learning systems have indisputably become the first choice for predictive tasks. In many cases, they are capable to perform even better than physics-based models, whilst being less expensive and faster to run. However, their full integration into decision-making processes requires improvements at the level of their explainability and interpretability. Indeed, besides a few algorithms such as linear regression and decision trees, machine learning algorithms are of a black-box nature: their inner workings are hard to comprehend by humans. In particular, the high complexity of deep neural network-based methods —which have achieved breakthroughs in domains such as computer vision and natural language processing— raises concerns regarding their fairness and trustworthiness. In this context, in response to several initiatives that aim to reduce the opacity of machine learning models and to provide fair and trustworthy explanations to end users [1], [2], EXplainable Artificial Intelligence (XAI) received a lot of attention from the research and industry communities alike. XAI is concerned with providing intrinsic explanations by developing interpretable models with simple, human-understandable logic and transparent inference processes. Such models include decision trees, naive bayes, and linear regression. However, because of the imposed simplicity and interpretability constraints, these models are not able to achieve high performance in more complex problems. Therefore, another major focus of XAI is to develop post hoc explanation algorithms. Such algorithms can be used on top of the highly complex black-box models to generate human understandable explanations. For example, feature attribution methods assess the contribution of each input feature to the model decision, while counterfactual explanation methods indicate the change to the input needed in order to result in a different model decision output. In recent years, post hoc explanation methods have achieved important success on tabular, image, and text datasets [3]–[6]. In addition, a few methods have also been developed for univariate time series data [7]–[9]. However, multivariate time series methods have been very sparse because of their challenging high-dimensional nature. In this work, we exploit the discriminative power of shapelets and temporal rules in time series mining and capitalize on their inherent interpretability to develop a model-agnostic, temporal rule counterfactual explainer (TeRCE) for multivariate time series data. We evaluate the performance of our approach on five benchmark datasets from the University of East Anglia (UEA) archive [10], and compare it to state-of-the-art counterfactual generation methods. The rest of this paper is organized as follows. Section II contains a brief description of state-of-the-art post hoc explanation methods. Section III introduces the counterfactual generation problem, describes the proposed algorithm, and discusses evaluation measures. Section IV describes the experimental setup. Section V presents the results. And finally, Section V concludes with a summary.

II. Related Works

Feature attribution methods were the first to generate post hoc explanations. For example, LIME [6] works by slightly perturbing the original instance and examining the effect of the perturbations on a surrogate linear model, while SHAP [4] derives the additive Shapley values of all features in order to compute their importance. In 2018, instance-based counterfactual explanation methods have seen the light with the work of Wachter et al. [11]. This algorithm aims to

978-1-6654-6283-9/22/$31.00 ©2022 IEEE
DOI 10.1109/ICMLA55696.2022.00200

generate explanations by optimizing a loss function containing a prediction term and a distance term. Then, several optimization-based algorithms that add new terms to the loss function to improve the quality of the counterfactuals and to speed up the search were proposed as an extension [3], [12], [13]. Since none of the methods mentioned so far has been designed for time series data, efforts to adapt them to the time series context through an apriori data segmentation or some other techniques have yielded poor results [5], [9], [14]. Therefore, new algorithms specifically developed for time series have been recently proposed. [7] extract shapelets from the dataset and use them to build a decision tree. The tree is then used to generate explanation rules. Delaney et al. [9] developed native guide (NG), an algorithm that extracts the nearest-unlike-neighbor (nun) of the original instance and uses it to introduce perturbations at the level of the most important contiguous time interval, selected based on the class activation mapping (CAM) of the black-box model. In case CAM cannot be applied to the machine learning model at hand, NG defaults to perturbations using dynamic time warping barycenter averaging (DBA). In 2021, CoMTE, the first counterfactual explanation algorithm for multivariate time series was developed by [14]. CoMTE perturbs the original instance by replacing entire feature dimensions extracted from its nun.

III. Methodology

A. Problem Definition

We describe the multivariate time series counterfactual explanation problem as follows. Consider a multivariate time series dataset $\mathcal{D} = \{MTS_1, MTS_2, ..., MTS_n\}$, where each instance MTS_i is assigned to a class $C_l \in \{C_1, C_2, ..., C_L\}$ and represented by a d dimensional list of time series vectors TS_j of T time steps: $MTS_i = (TS_1, TS_2, ..., TS_d)$ and $TS_j = (TS_{j1}, TS_{j2}, ...TS_{jT})$. f is the prediction function of a black-box classification model trained on $\mathcal{D}$. Given the class prediction $f(MTS_i) = C_l$, the goal is to generate a counterfactual instance MTS_{cf} by introducing a perturbation to MTS_i such that $f(MTS_{cf}) \neq C_l$.

B. Shapelet Transform

Several shapelet-based classification algorithms have been recently introduced in the literature [15]–[18]. However, shapelet transform (ST) [19]–[21] has been pointed out among the most powerful time series classification algorithms [22], in addition to being fully transparent and interpretable. Thus, we adopt it in this work to extract shapelets as the first step in our algorithm.

A shapelet is defined as a phase-independent, discriminative time series interval that occurs frequently in a dataset. In ST, shapelets are mined by first iterating through all possible candidate shapelets S_i of predefined lengths in a time series dataset. Then, the distance separating each S_i from each dataset sample TS_i is recorded as the minimum distance between S_i and each subsequence of TS_i of the same length. Based on these distances, the information gains of shapelets are computed and the ones with the highest values are retained. Since ST was initially developed for univariate time series, it can be adapted to multivariate datasets by extracting shapelets from each dimension separately.

TABLE I: Allen's Interval relationships

precedes	meets	overlaps	finished by	contains	starts	equals
p	m	o	F	D	s	
preceded by	met by	overlapped by	finishes	during	started by	
P	M	O	f	d	S	e

C. TeRCE

In this section, we present TeRCE, a model-agnostic, temporal rule counterfactual explainer for multivariate time series data.

1) Model Fitting:

a) Temporal Rule Mining: A temporal rule $A \rightarrow B$ defines a relationship between an antecedent time interval A and a subsequent time interval B. In its most basic form, the relationship is a $precedes$ operator, indicating that A occurs before B. We mine temporal rules from a multivariate time series dataset $\mathcal{D}$ by considering all possible temporal relationships between the shapelets extracted using ST. The set of possible relationships defined in Allen's Interval Algebra [23] is shown in Table I. Since the 13 relationships in Table I are symmetric, we only consider 7 of them, namely: $precedes$, $meets$, $overlaps$, $finishes$, $contains$, $starts$, and $equals$. Then, we record the occurrence of each rule in a binary vector $R = (r_1, r_2, ..., r_n)$ where $r_i = 0$ if the rule happens in MTS_i and $r_i = 1$ if it does not happen in MTS_i. Figure 2. shows two example rules extracted from the Libras dataset.

b) Class-Rules Extraction: We compute the fisher scores [24], [25] of the occurrence vectors of the rules extracted in the previous step and retain the most discriminative rules, i.e. those with the highest scoring occurrence vectors. Then, we select class-rules —rules that occur under one class label only— and discard the others. Next, we compute the occurrence distribution of each class-rule as the average of all the occurrence indices of its antecedent shapelet and the average of all the occurrence indices of its subsequent shapelet. The motivation behind using these temporal rules is that, in addition to their discriminative power, they are highly interpretable. Moreover, they add a notion of phase dependence to the phase-independent shapelets by relating them across different dimensions.

2) Counterfactual Generation: For a dataset sample MTS_{orig} of class $f(MTS_{orig}) = C_{orig}$, a counterfactual explanation MTS_{cf} such that $f(MTS_{cf}) = C_{target}$ where $C_{target} \neq C_{orig}$ is generated as follows. First, a k-Nearest-Neighbor model with $k = 1$ is trained on the dataset intances of class C_{target} and is used to find MTS_{nun}, the nearest-unlike-neighbor of MTS_{orig}. Then, step a and step b are performed until a valid counterfactual MTS_{cf} is found as shown in Figure 1.

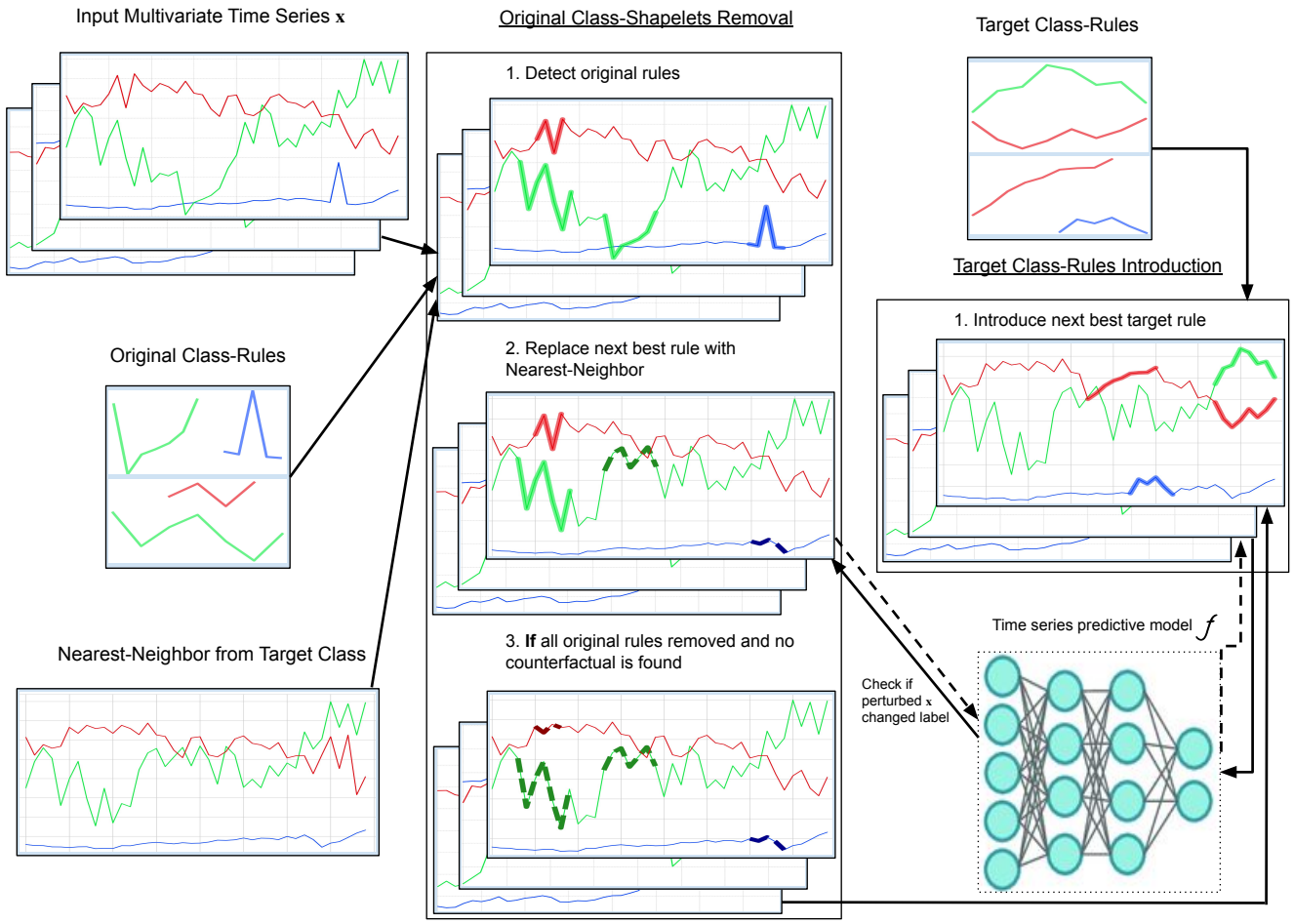


Fig. 1: TeRCE counterfactual generation. First, the two original class-rules (ApB and AdB) are replaced from the nun. Then, the two target class-rules (AeB and AoB) are introduced.

a) Original Class-Rule Removal: First, the original instance MTS_{orig} is searched for the presence of original class-rules. Then, each detected original class-rule is replaced —in descending order of Fisher score— by the values of MTS_{nun} at the same time steps, after scaling to the original range of magnitudes using min-max scaling. This step aims to discard the rules that characterize class C_{orig} and that play a role in the original model prediction.

b) Target Class-Rule Introduction: The C_{target} class-rules are sorted in descending order of Fisher score and introduced to MTS_{orig} according to their occurrence distribution intervals, after scaling to MTS_{orig}'s original range of magnitudes using min-max scaling. If none of them leads to a valid counterfactual, perturbations made of two or more class-rules are introduced until $f(MTS_{cf}) = C_{target}$.

D. Evaluation Measures

Introducing perturbations at the level of the rules allows TeRCE to generate highly meaningful counterfactual explanations. Visualizing the generated counterfactuals and highlighting the perturbations such as in Figure 4 increases the interpretability of the explanations and helps understand the process. In addition, considering the discriminative power of the temporal rules in a classification problem context will potentially increase the stakeholders' trust in TeRCE. However, comparing our algorithm to other counterfactual generation approaches requires performing a quantitative evaluation. In this section, we present three counterfactual evaluation measures that have been repeatedly used in the literature and describe their use in this paper.

1) Proximity: (or closeness, distance). Ideally, the generated counterfactual explanation should be as close as possible to the original instance. However, since other measures have to be considered, the proximity measure ensures that the perturbation remains of small magnitude. Similar to [9], we use the $L_1 - norm$ (or Manhattan distance), the $L_2 - norm$ (or Euclidian distance), and the $L_{inf} - norm$ as measures of proximity, as described in equations (1), (2), and (3) respectively. The first two measures compute the distance between MTS_{orig} and MTS_{cf} and the third measure gets the magnitude of the highest perturbation time step.

$$||MTS_{orig} - MTS_{cf}||_{L_1} = \sum_{j}^{d}\sum_{i}^{T}\left|MTS_{ij} - MTS_{cf\,ij}\right| \quad (1)$$

$$||MTS_{orig} - MTS_{cf}||_{L_2} = \sqrt{\sum_{j}^{d}\sum_{i}^{T}\left(MTS_{ji} - MTS_{cf\,ji}\right)^2} \quad (2)$$

$$||MTS_{orig} - MTS_{cf}||_{L_{inf}} = \sum_{j}^{d}\sum_{i}^{T} max\left|MTS_{ji} - MTS_{cf\,ji}\right| \quad (3)$$

2) Sparsity: The perturbation introduced to generate a counterfactual explanation should affect the fewest number of features possible, and time steps in the case of time series data, to generate more informative explanations [13]. The more dimensions and time steps are changed, the harder it is to understand by stakeholders and the less likely it is to be feasible. Moreover, time series perturbations should be constrained to short, contiguous intervals for the counterfactuals to be meaningful [8], [9].

3) Plausibility: (or interpretability). Counterfactual explanations must be realistic and easily interpretable by humans, which is not guaranteed by the proximity and sparsity measures. Therefore, plausibility is considered as the third criterion. It can be measured by considering whether the counterfactual belongs to the training data manifold by using novelty detection techniques or other approaches [9], [26], [27].

Following the work of Van Looveren and Klaise [3], we use the IM measure to compute the plausibility of counterfactual MTS_{cf}. First, we train autoencoder AE on the entire dataset and autoencoder AE_{target} on instances of class C_{target}. Then, we compute IM2 as the distance between the reconstructions of MTS_{cf} using AE and AE_{target}, scaled by the $L_1 - norm$ of MTS_{cf} to allow comparisons across all classes as shown in equation (5). A lower value of IM is desirable, as it indicates that the distribution of C_{target} instances describes MTS_{cf} as well as the distribution of the entire dataset, i.e. MTS_{cf} is as close to the data manifold of C_{target} instances as it is close to the date manifold of all dataset classes.

$$\text{IM}(MTS_{cf}) = \frac{\|AE_{target}(MTS_{cf}) - AE(MTS_{cf})\|\,|_2^2}{\|MTS_{cf}\|\,|_1 + \epsilon} \quad (4)$$

IV. Experimental Setup

A. Datasets

We evaluate our approach on the five classification datasets from the University of East Anglia (UEA) MTS archive [10]. BasicMotions and Epilepsy contain accelerometer and gyroscope data describing four activities, collected by a smartwatch. FingerMovements contains Electroencephalogram (EEG) data recorded while pressing a keyboard using the index and pinky fingers only and labeled as left or right hand. Libras contains the trajectories of hand movements from the Brazilian sign language projected on a 2D coordinate space. And RacketSports contains smartwatch data recorded while playing one of the two strokes of badminton and squash. The characteristics of each dataset are described in Table II.

TABLE II: Datasets Characteristics

Dataset	Train size	Test size	Dimensions	TS length	Classes
BasicMotions	40	40	6	100	4
Epilepsy	137	138	3	206	4
FingerMovements	316	100	28	50	2
Libras	180	180	2	45	15
RacketSports	151	152	6	30	4

B. Implementation Details

We used the sktime [28] implementation of ST, and modified it to extract the indices of the occurrences of each shapelet along with their distances to be able to mine rules. Since the multivariate time series datasets are high-dimensional, we ran the contracted shapelet transform implementation. Instead of going through all possible subsequences, this approach randomly selects shapelets for a user-defined amount of time. It has been shown that it does not significantly affect the performance of ST [29]. We limited the time contract to 30 minutes per dataset. To the perturbation sparse, we restricted the length of the shapelets to a maximum of 25% of the length of the time series. In this work, we adopt a one-vs-all approach for generating counterfactual explanations for multi-class datasets. However, TeRCE is also able to generate a counterfactual for each target class separately. We provide free access to our code and the solar-flare dataset in the GitHub repository .

C. Compared Methods

1) Alibi: [3] This algorithm was originally developed for image and tabular data. It is a model agnostic, optimization-based counterfactual generation approach. The main novelty of Alibi is the introduction of a prototype term to the loss function, in addition to the prediction, proximity, and autoencoder (plausibility) term. This allows for speeding up the optimization process and increases the interpretability of the generated counterfactual explanations.

2) NG: [9] The model agnostic version we use in this paper extracts the nearest-unlike-neighbor of the instance to be perturbed from the target class and used in conjunction with dynamic time warping barycenter averaging (DBA) to guide the perturbations. NG was initially proposed for univariate time series datasets. In order to adapt it to multivariate data, we reshape the datasets so that all dimensions form a single 1-D feature vector.

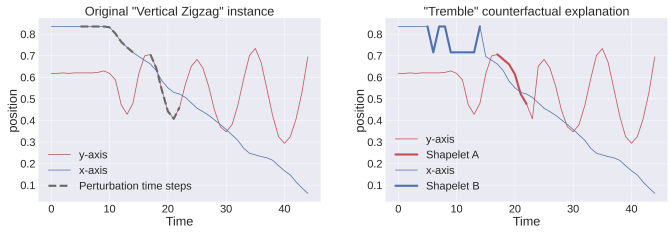
3) CoMTE: [14] CoMTE was originally designed for multivariate time series. It extracts the nun of the original instance using the KD-tree of the target class and uses it to replace a set of entire time series dimensions. A heuristic search based on hill climbing and a post hoc trimming step are employed to find a minimal set of dimensions to be perturbed.

TABLE III: Proximity Comparison: L_1-norm, L_2-norm and L_{inf}-norm

	BasicMotions			Epilepsy			FingerMovements			Libras			RacketSports			Average Rank		
	L1	L2	Linf	L1	L2	Linf	L1	L2	Linf	L1	L2	Linf	L1	L2	Linf	L1	L2	Linf
Alibi	846.62	61.18	15.21	191.68	12.26	2.00	7523	284	28.62	2.17	0.49	0.19	351.95	51.79	19.76	2.2	2.4	2.0
NG	947.18	59.25	**13.59**	175.64	**9.77**	**1.49**	7774	305	34.06	3.87	0.56	**0.15**	374.72	**49.30**	**18.78**	2.8	2.0	1.4
CoMTE	1917.83	116.28	21.45	489.18	24.66	3.06	55735	1900	**15.81**	15.97	2.05	0.48	869.33	100.74	31.18	4.0	4.0	3.4
TeRCE	**345.91**	**53.76**	16.26	**84.92**	11.34	2.95	**318**	**76**	34.22	**1.36**	**0.42**	0.20	**306.82**	66.41	25.58	**1.0**	**1.6**	3.2

(a) (b)

Fig. 2: Class-rules in their original locations (Libras dataset)



(a) (b)

Fig. 3: Counterfactual generation using target class-rule introduction

If this approach fails to generate a valid counterfactual, a greedy search is performed.

V. Results

We evaluate TeRCE on the five datasets described in Section IV.A and discuss the results in this section. TeRCE is model agnostic, meaning that it can generate explanations for any machine learning model, regardless of its architecture or transparency. Since ROCKET [30] was the highest ranked and fastest classifier in [31], we used is as the black-box model.

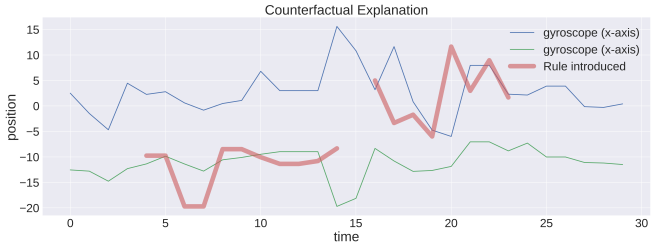


Fig. 4: Counterfactual explanation generated by introducing an ApB Rule (RacketSports Dataset)

TABLE IV: Sparsity Comparison

	BasicMotions	Epilepsy	FingerMovements	Libras	RacketSports	Average Rank
Alibi	597.75	607.86	1300	88.83	179.89	2.2
NG	580.11	563.37	1145	82.64	165.88	2.0
CoMTE	598.88	615.14	1395	89.73	179.00	3.8
TeRCE	**60.62**	**91.84**	**26**	**17.10**	**31.32**	**1.0**

TABLE V: Plausibility Comparison: IM

	BasicMotions	Epilepsy	FingerMovements	Libras	RacketSports	Average Rank
Alibi	**0.027**	0.017	0.00207	0.0107	**0.0405**	**1.6**
NG	0.035	0.022	**0.00201**	**0.0100**	0.0462	2.6
CoMTE	0.032	0.019	0.00523	0.0129	0.0506	3.6
TeRCE	0.031	**0.009**	0.00209	0.0115	0.0409	2.2

A. Qualitative Evaluation

Not only do the class-rules mined by TeRCE represent the building blocks of the counterfactual generation process, they also contain important information when considered separately. Indeed, they can be used for classification purposes where each rule occurrence vector represents a dataset feature. In addition, visualizing the temporal rules can help stakeholders achieve a better understanding of the explanation process and even assist domain experts in acquiring significant insights from the dataset. In Figure 2, we show two class-rules extracted from the Libras dataset. The first rule characterizes ”tremble” hand movements and the second rule is characteristic of ”face-up curve” hand movements. In Figure 3, we illustrate how the first rule from Figure 2.a was used by TeRCE to generate a counterfactual explanation of class ”tremble”, by simply introducing it to a dataset instance from the ”vertical zigzag” class. In addition, a counterfactual explanation created from the RacketSports dataset using TeRCE is shown in Figure 4. For visualization purposes, the explanations we chose require the removal or introduction of one rule only.

B. Quantitative Evaluation

In this section, we evaluate the proximity, sparsity, and plausibility of the TeRCE counterfactual explanations generated from the five benchmark datasets, and compare the results to baseline algorithms discussed in section IV.C.

1) Proximity: Table III shows that the counterfactuals generated by TeRCE are significantly closer to the original instances. However, the perturbations might include some high-magnitude spikes, as suggested by the 3.2 L_{inf} average rank. Given that Alibi and NG are based on loss optimization and DBA respectively, the perturbations they introduced are of consistently lower magnitudes; however, this indicates that their sparsity is high.

2) Sparsity: We computed sparsity as the total number of perturbed time steps throughout all time series dimensions. Again, Table IV proves that the perturbations introduced by TeRCE are significantly more contiguous than those of Alibi,

NG, and CoMTE. The latter produced the least desirable counterfactuals in terms of sparsity since it replaces entire feature dimensions.

3) Plausibility: When it comes to plausibility, Table V shows that Alibi performed the best compared to the other algorithms, while TeRCE ranked second. However, as discussed in the previous sections, TeRCE has the advantage of being visually and conceptually interpretable.

VI. Conclusion

In this work, we proposed TeRCE, a model-agnostic MTS counterfactual explanation algorithm. By capitalizing on the inherent interpretability of shapelets and temporal rules, TeRCE ensures the meaningfulness of the introduced perturbations and the high interpretability of the resulting counterfactual explanations. Moreover, using temporal rules makes the algorithm visually interpretable and easily understandable by end-users. By visualizing the counterfactuals and considering the perturbations, the rules with the most influence on the classification process can be determined. In addition, plotting the temporal rules can help domain experts learn new insights from the dataset. We evaluated TeRCE on five datasets from the UEA archive and compared the generated explanations to three state-of-the-art counterfactual generation algorithms.

References

[1] D. Gunning, "Broad Agency Announcement Explainable Artificial Intelligence (XAI)," , Defense Advanced Research Projects Agency (DARPA), Tech. Rep., 2016.

[2] "NoEU General Data Protection RegulationTitle," Tech. Rep., 2018.

[3] 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)*, vol. 12976 LNAI, pp. 650–665, jul 2019.

[4] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, pp. 4766–4775, may 2017.

[5] 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*, pp. 4197–4201, sep 2019.

[6] 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*, pp. 97–101, feb 2016.

[7] 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*, pp. 167–176, oct 2020.

[8] 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*, pp. 1406–1415, oct 2021.

[9] 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)*, vol. 12877 LNAI, pp. 32–47, sep 2020.

[10] A. G. Bostrom, "Shapelet Transforms for Univariate and Multivariate Time Series Classification," 2018, arXiv:1409.0876.

[11] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, no. 2, 2018.

[12] 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*, vol. 2018-Decem, pp. 592–603, feb 2018.

[13] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona, Spain: Association for Computing Machinery, 2020.

[14] 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)*. Institute of Electrical and Electronics Engineers Inc., 2021, pp. 1–8.

[15] L. Ye and E. Keogh, "Time series shapelets: A novel technique that allows accurate, interpretable and fast classification," *Data Mining and Knowledge Discovery*, vol. 22, no. 1-2, pp. 149–182, jan 2011.

[16] 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*. Siam Society, 2013, pp. 668–676.

[17] 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*. New York, NY, USA: Association for Computing Machinery, aug 2014, pp. 392–401.

[18] 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*, pp. 497–508, oct 2018.

[19] 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*. New York, New York, USA: ACM Press, 2012, pp. 289–297.

[20] J. Hills, J. Lines, E. Baranauskas, J. Mapp, and A. Bagnall, "Classification of time series by shapelet transformation," *Data Mining and Knowledge Discovery*, vol. 28, no. 4, pp. 851–881, may 2014.

[21] 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)*, vol. 9263. Springer Verlag, 2015, pp. 257–269.

[22] 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*, vol. 31, no. 3, pp. 606–660, may 2017.

[23] J. F. Allen, "Maintaining Knowledge about Temporal Intervals," *Communications of the ACM*, vol. 26, no. 11, pp. 832–843, nov 1983.

[24] R. O. Duda, P. E. P. E. Hart, and D. G. Stork, *Pattern classification*, 2nd ed. Wiley, 1973.

[25] X. He, D. Cai, and P. Niyogi, "Laplacian Score for Feature Selection," Tech. Rep., 2005.

[26] A.-H. Karimi, G. Barthe, B. Balle, and I. Valera, "Model-Agnostic Counterfactual Explanations for Consequential Decisions," in *International Conference on Artificial Intelligence and Statistics*, PMLR, Ed., may 2020.

[27] 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*, pp. 344–350, sep 2019.

[28] 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.

[29] 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)*, vol. 10420 LNCS, pp. 24–46, dec 2017.

[30] A. Dempster, F. Petitjean, and G. I. Webb, "ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels," *Data Mining and Knowledge Discovery*, vol. 34, no. 5, pp. 1454–1495, sep 2020.

[31] A. Pasos Ruiz, M. Flynn, J. Large, ·. M. Middlehurst, and ·. A. Bagnall, "The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances," *Data Mining and Knowledge Discovery*, vol. 35, pp. 401–449, 2021.