

Fast Counterfactual Explanation for Solar Flare Prediction

Peiyu Li, Omar Bahri, Soukaïna Filali Boubrahimi, Shah Muhammad Hamdi

Department of Computer Science, Utah State University, Logan, UT 84322, USA

Emails: {peiyu.li, omar.bahri, soukaina.boubrahimi, s.hamdi}@usu.edu

Abstract—Solar flare prediction has become essential in space weather research due to its potential adverse space-weather ramifications. Over recent years, a set of machine learning models on solar flare prediction have been proposed and significant improvement has been made over the previous state of the art. However, most existing research work focuses on the prediction task and ignores the interpretability behind the prediction task. In this paper, we provide a post-hoc explanation method based on solar flare prediction, FAST-CF. In particular, we incorporate the nearest unlike neighbor for guiding the counterfactual search, which is fast to search for the optimal result. In addition, FAST-CF encapsulates the desirable properties of a counterfactual explanation for solar flare prediction. We use different evaluation metrics to compare the performance of FAST-CF with the other two baselines and verify the superiority of our method to existing state-of-the-art.

Index Terms—Solar flare prediction, Multivariate Time Series, EXplainable Artificial Intelligence (XAI), Counterfactual Explanation

I. INTRODUCTION

A solar flare is an intense burst of radiation coming from the release of magnetic energy associated with sunspots [1]. X-rays and UV radiation emitted by solar flares can cause electromagnetic disturbances on the earth, as with radio frequency communications and power line transmissions [2].

In recent years, the success of supervised machine learning (ML) methods especially deep neural networks on solar flare prediction have been verified by experts in the space weather domain [3]–[7]. However, the interpretability of the decision-making process behind solar flare prediction can not be guaranteed. Some ML models exhibit high performance but they are opaque in terms of explainability. Some AI researchers argue that the explanation is not essential for all AI applications, since it is too difficult to achieve, and unnecessary in certain applications [8]. However, for critical applications in the space weather domain such as solar flare prediction, it is vital for human beings to understand, trust and apply these AI systems to deal with corresponding problems. Therefore, the insight of interpretability is of crucial importance in predicting a solar flare event involving the potentially hazardous impacts of the solar flares [1].

X- and M-classes of solar flares are most often targeted in intense classes in solar flare prediction. As most flares occur in the Active Regions of the Sun, flare prediction can be modeled as a supervised learning problem of machine learning, specifically the binary classification between flaring and non-flaring Active Regions (AR), where flaring Active Regions are

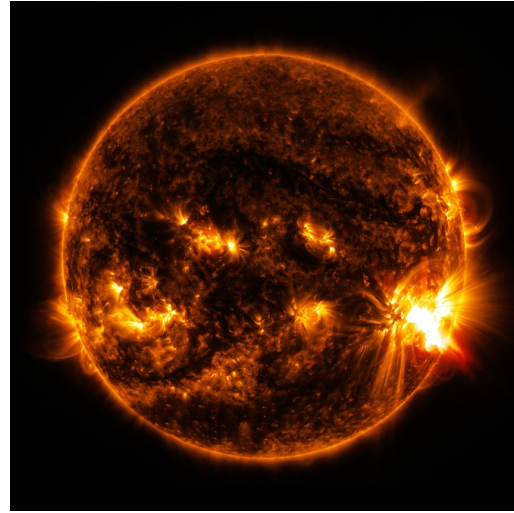


Fig. 1. NASA's Solar Dynamics Observatory captured this image of an X2.0-class solar flare bursting off the lower right side of the sun on Oct. 27, 2014. The image shows a blend of extreme ultraviolet light with wavelengths of 131 and 171 Angstroms. Image Credit: NASA/SDO

considered to be in a positive class and non-flaring Active Regions are considered to be in the negative class [2]. In this work, as positive class examples, we consider the Active Regions that have one or more M-class or X-class flares during their crossing of the observable solar disk. The Active Regions that have never flared during the disk crossing (not even C-class flares) are considered negative class examples. In terms of binary classification between flaring and non-flaring Active Regions, we provide the post-hoc counterfactual explanations for the solar flare prediction. In particular, a counterfactual instance is defined as a synthetic instance for which a trained machine learning model predicts the desired output which is different from the prediction made on the query instance [9]. In the case of solar flare prediction, if a given solar flare is predicted by a classifier as a flaring Active Region, what changes can be made from that solar flare to obtain a different prediction, non-flaring Active Region? In addition, we define the given solar flare as the *query instance*, the instance that has been changed to obtain a different prediction as the *counterfactual instance*, the label of the query instance as the *query label* and the label of the counterfactual instance as the *desired label*. To generate the counterfactual instance

for each query instance, we find the nearest unlike neighbor (nearest neighbor of the desired label from the training dataset) firstly. Then we try to find the most important top k dimensions by comparing the distance between the query instance and its nearest unlike neighbor. Finally, we substitute the top k dimensions from the original query instance such that the classification label changes to the class of desired.

Our paper contributions are summarized below:

- 1) We propose a method that encapsulates the desirable properties of a counterfactual explanation for solar flare prediction.
- 2) Our new method does not require the use of class activation maps to search for the counterfactual explanation, which makes it model-agnostic.
- 3) We incorporate the nearest unlike neighbor for guiding the counterfactual search, which speeds up the search for a counterfactual explanation.
- 4) We conduct experiments on the publicly available solar flare dataset and show the superiority of our methods compared with other baselines.

To the best of our knowledge, this is the first effort to focus on a small set of dimension substitution while generating a counterfactual explanation for solar flare prediction. The rest of this paper is organized as follows: in section II, we lay the ground for our research by introducing the related works. Section III introduces the preliminary concepts. Section IV describes our proposed method in detail. We present the experimental results and evaluations in comparison to other baselines in section V. Finally, we conclude our work in section VI.

II. RELATED WORK

In the post-hoc interpretability paradigm, various approaches have been proposed in the literature for text, image, and tabular data, such as LIME [10], LORE [11], SHAP [12], GeCo [13] and wCF [14]. LIME is a feature-based approach that shows that explanations are useful for a variety of models in trust-related tasks for text data and image data. LORE is an extension work based on LIME, which is a local black box model-agnostic explanation approach based on logic rules, but LORE explanation works mainly on tabular data. SHAP is a unified framework that operates by calculating feature importance values using model parameters. While wCF aims at minimizing a loss function and using adaptive Nelder-Mead optimization to encourage the counterfactual to change the decision outcome and keep the minimum Manhattan distance from the query input instance. Similarly, GeCo has been proposed as another method that is used to deal with the plausibility and feasibility issues of the generated counterfactual explanation has been proposed. The model achieves the desirable counterfactual properties by introducing a new plausibility-feasibility language (PLAF) [13]. Both GeCo and wCF focus on structured tabular datasets. However, experiments have indicated that techniques designed for tabular data often failed to produce meaningful explanations in the time series domain [15].

In the computer vision domain, visualization techniques have been widely applied to provide interpretability for different applications successfully, such as highlighting the most important parts of images to class activation maps (CAM) in convolutional neural networks [16]. Huang et al [3] applied the Convolution Neural Network to the flare forecasting using patches of ARs of solar line-of-sight magnetograms. The authors extract CNN feature maps from the interior layers in the model that show their models pay attention to the area of the PIL. However, the feature map, which is just a result of the calculation between the input image and CNN kernels, does not indicate important areas of the input image for prediction results. Later on, [17] presents a visual explanation of a deep-learning solar flare forecast model. In particular, the authors interpret the model using two CNN attribution methods (guided backpropagation and Gradient-weighted Class Activation Mapping [Grad-CAM]) that provide quantitative information on explaining the model. They show that the polarity inversion line is an important feature of the deep learning flare forecasting model.

Recently, an instance-based counterfactual explanation for time series classification has been proposed [15]. The instance-based counterfactual explanation uses the explanation weight vector (from the Class Activation Mapping) and the in-sample counterfactual (NUN) to generate counterfactual explanations for time series classifiers. The instance-based technique adapts existing counterfactual instances in the case base by highlighting and modifying discriminative areas of the time series that underlie the classification. The success of this method has been verified by comparative tests on diverse datasets from the UCR archive.

Finally, a counterfactual solution for multivariate time series data called CoMTE has been proposed by Etes et al. [18], which focuses on selecting time series from the training set and substituting them in the sample under investigation to obtain different classification results. Since the method is observing the effect of turning off one variable at a time, it takes a long time to generate counterfactual instances with the high-dimension nature of the multivariate time series.

III. PRELIMINARY

A. Notation

We assume an univariate time series $\mathbf{x} = \{x_1, x_2, \dots, x_m\}$ is an ordered set of real values, where m is the length. In the case of multivariate time series, the time series is a list of vectors over d dimension and m observations, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_d]$.

Then we can define a multivariate time series dataset $\mathbf{D} = \{\mathbf{X}_0, \mathbf{X}_1, \dots, \mathbf{X}_n\}$ as a collection of n multivariate time series where each multivariate time series has mapped to a mutually exclusive set of classes $\mathbf{C} = \{c_1, c_2, \dots, c_l\}$. We split the dataset $\mathbf{D}$ into a training set and a test set. The training set aims to train a time series classifier f . For each query instance in the test set which is associated with a class $f(\mathbf{X}_j) = c_j$, a counterfactual explanation model $\mathcal{M}$ generates a perturbed sample with the minimal perturbation that lead to $f(\mathbf{X}'_j) = c'_j$ such that $c_j \neq c'_j$.

B. Desirable properties of a counterfactual instance

According to [19], to provide a useful, plausible alternative for the query instance, a counterfactual instance should obey the following initial desirable properties:

- 1) **Validity:** The prediction of the to-be-explained model f on the counterfactual instance $\mathbf{X}'$ needs to be different from the prediction of the to-be-explained model f on the query instance $\mathbf{X}$ (i.e., if $f(\mathbf{X}) = c_i$ and $f(\mathbf{X}') = c_j$, then $c_i \neq c_j$).
- 2) **Proximity:** The to-be-explained query needs to be close to the generated counterfactual instance, which means the distance between $\mathbf{X}'$ and $\mathbf{X}$ should be minimal.
- 3) **Sparsity:** The perturbation δ changing the query instance $\mathbf{X}$ into $\mathbf{X}' = \mathbf{X} + \delta$ should be sparse, which means fewer number of data points that needs to be changed to get the counterfactual explanation is preferred.
- 4) **Contiguity:** The counterfactual instance $\mathbf{X}' = \mathbf{X} + \delta$ needs to be perturbed in a single contiguous segment which makes the solution semantically meaningful.
- 5) **Interpretability:** The counterfactual $\mathbf{X}'$ needs to be in-distribution. We consider an instance $\mathbf{X}'$ interpretable if it lies close to the model's training data distribution. The $\mathbf{X}'$ should be an inlier with respect to the training dataset and an inlier to the counterfactual class.
- 6) **Model-agnosticism:** The counterfactual explanation model should produce a solution independent of the classification model f , high-quality counterfactuals without prior knowledge of the gradient values derived from optimization-based classification models should be generated.

In our experimental evaluation part, we will take these properties into consideration to verify the superiority of our method to existing state-of-the-art explainability methods.

IV. FAST COUNTERFACTUAL EXPLANATION (FAST-CF) FOR SOLAR FLARE PREDICTION

In this section, we describe our proposed fast counterfactual explanation method for solar flare prediction in detail. In particular, the method includes two main steps: 1. Retrieve the nearest unlike neighbor 2. Adapt the nearest unlike neighbor to generate counterfactual instances. The process of FAST-CF generation is shown in Figure 2. The algorithm is shown in Algorithm 1.

A. Retrieve the nearest unlike neighbor

Given a query instance $\mathbf{X}$, find a counterfactual instance candidate $\mathbf{X}_c$ that exists in the training dataset. An example of one such instance is the query's nearest unlike neighbor. This nearest unlike neighbor is from the training dataset, the label of it is our desired label, which guarantees the explanation's interpretability property as it is, by definition, within the distribution. However, such instances are not guaranteed to satisfy the proximity, sparsity, and contiguity properties. Therefore, an adaptation step is necessary to satisfy the remaining properties.

B. Adapt the nearest unlike neighbor to generate counterfactual instance

To generate the counterfactual instance that satisfies validity, proximity, sparsity, and contiguity properties, we try to find the most important top k dimensions by comparing the distance per dimension between the query instance and its nearest unlike neighbor. If the distance between the query instance and its nearest unlike neighbor is relatively large for specific dimension data, we will consider this dimension as the important dimension that we want to focus on. Then we substitute the top k dimensions from the nearest unlike neighbor such that the classification label changes to the class of desired. To guarantee the proximity and the validity property at the same time, we set the k as a parameter and k will be determined as the minimum value that can make sure the counterfactual instance is classified to the class of desired.

In addition, substituting the top k dimensions from the nearest unlike neighbor guarantees the counterfactual instance we generate is perturbed in several contiguous segments and sparse, instead of changing the whole time series, as follows

$$\mathbf{X} = \langle \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \dots, \mathbf{x}_d \rangle \quad s.t. \quad f(\mathbf{X}) = c \quad (1)$$

$$\mathbf{X}' = \langle \mathbf{x}'_1, \mathbf{x}_2, \mathbf{x}'_3, \mathbf{x}_4, \mathbf{x}'_5, \dots, \mathbf{x}_d \rangle \quad s.t. \quad f(\mathbf{X}') = c' \quad (2)$$

Algorithm 1 Fast Counterfactual Explanation for Solar Flare Prediction

Input: Training set $\mathbf{T}$, query set *samples*, prediction model for solar flare time series data f , the number of dimension of query instance d

Output: CF, counterfactual instances for query instances

```

1: CF =  $\emptyset$ 
2: for  $X \leftarrow \text{samples}$  do
3:    $X_c$  = nearest unlike neighbor from  $\mathbf{T}$ 
4:   Dists = []
5:   for dimension  $i$  from 1 to  $d$  do
6:     Dist = np.sum( $X[i, :] - X_c[i, :]$ )
7:     Dists.append(Dist)
8:    $k = 0$ ,  $X' = X$   $\triangleright k$  is initialized to 0 and the counterfactual instance is initialized as  $X$ 
9:   while  $f(X') \neq c'$  do
10:     $k = k + 1$ 
11:     $\text{idx} = \text{np.argmax}(\text{Dists}, -k)[-k:]$ 
12:     $\text{indices} = \text{idx}[\text{np.argsort}((- \text{np.array}(\text{Dists}))[\text{idx}])]$   $\triangleright$  Find the top  $k$  dimensions where there are top  $k$  maximum distances
13:    for  $\text{index} \leftarrow \text{indices}$  do
14:       $X'[\text{index}, :] = X_c[\text{index}, :]$   $\triangleright$  replace the top  $k$  dimensions
15:    CF.add( $X'$ )
16: return CF

```

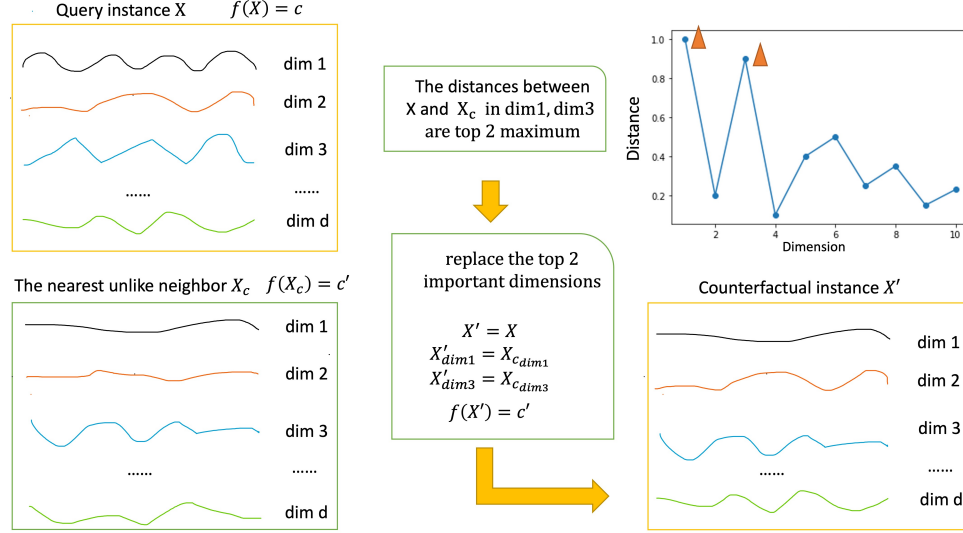


Fig. 2. The process of FAST-CF generation

V. EXPERIMENTAL SETTING

A. Data sets description

In our experiments, we use a solar flare prediction dataset that was published by the Data Mining Lab of Georgia State University [1]. Each sample in this dataset is a multivariate time series, each of them has 60 values representing an observation taken at 12-minute intervals. The first value in the array is the sample observation taken at the furthest point in time from the prediction period and the last value in the array is the sample observation taken at the closest point in time to the prediction period. The period of observations represented by each labeled time series is a 12-hour window of observations sliced from a longer time series. In particular, each multivariate time series includes 33 solar magnetic field parameters. The dataset consists of 5 classes, namely X, M, C, B, and Q, where Q represents the flare-quiet regions where no flare has been detected within the observation period. In particular, to conduct a binary case of counterfactual explanation, the X- and M-classes of solar flares are considered to be positive class to represent the flaring active regions, the C-, B-, and Q-classes of solar flares are considered to be negative class to represent the non-flaring regions. The class distribution of the original dataset is imbalanced. To address the imbalance issue of the dataset, we use the undersampling technique to generate a balanced dataset.

B. Baseline methods

We evaluated our proposed method with the other two baselines, Alibi [20] and Native guide counterfactual [15].

- **Alibi Counterfactual (Alibi):** The Alibi generates counterfactual explanations by optimizing an objective function,

$$L = L_{pred} + \lambda L_{dist}, \quad (3)$$

where the first loss term L_{pred} guides the search towards points X' which would change the model prediction and the second term L_{dist} ensures that X' is close to X . This form of loss has a single hyperparameter λ weighing the contributions of the two competing terms.

- **Native guide counterfactual (NG-CF):** NG-CF is another baseline that we used to compare our proposed FAST-CF methods. NG-CF uses Dynamic Barycenter (DBA) averaging of the query time series x and the nearest unlike neighbor from another class to generate the counterfactual example [15].

C. Prediction Model Details

For fairness purposes, we evaluated all the aforementioned counterfactual baselines on the same predictive model f . In particular, we used a convolutional neural network model that consists of two convolution layers with respectively 128 and 64 one-dimensional filters and ReLU activations. Each convolutional layer is followed by a max-pooling layer. Dropout with a fraction of 30% is applied during training. The output of the second pooling layer is flattened and fed into a fully connected layer of size 256 with ReLU activation and 50% dropout. This dense layer is followed by a softmax output layer over the number of classes. The model is trained using an Adam optimizer with batch size 32.

D. Experimental result

In this section, we utilize different evaluation metrics to compare FAST-CF with the other two baselines with respect to the desirable properties of counterfactual instances discussed in Section III-B. Since the data we use in our experiments is multivariate time series, for better understanding, we flatten the multivariate time series data into one dimension data and then apply the evaluation metrics on the flattened data. In addition, for each evaluation metric, the result we show is the average

value among the whole dataset. The details of each evaluation metric are shown below.

L1 distance, which measures the distance between the counterfactual instance and the query instance, a smaller L1 distance is desired. Table I shows that our proposed FAST-CF method achieves the minimum L1 distance when compared with the other two baselines.

Sparsity level, which indicates the level of time series perturbations. A high sparsity level that is approaching 100% is desirable, which means the time series perturbations made in X to achieve X' is minimal. We computed the sparsity level using the Equations 4-5. From the blue dash line plot in Figure 3, we can notice that our proposed FAST-CF performs best in terms of sparsity level compared with the other two baselines.

$$sparsity = 1 - \frac{\sum_{i=0}^{len(X)} g(X'_i, X_i)}{len(X)} \quad (4)$$

$$g(x, y) = \begin{cases} 1, & \text{if } x \neq y \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

The number of independent non-contiguous segments is also investigated to show the contiguity. The lower the number of independent non-contiguous segments the better. From the red bar plot in Figure 3, we can see that our proposed FAST-CF method results in the minimum number of independent non-contiguous segments.

In addition, we define the validity metric by comparing the target class probability for the prediction of the counterfactual explanation result. The closer the target class probability is to 1, the better. From Table I, we can see that FAST-CF achieves 1.0 target class probability, which is much larger than the target class probability generated by ALIBI.

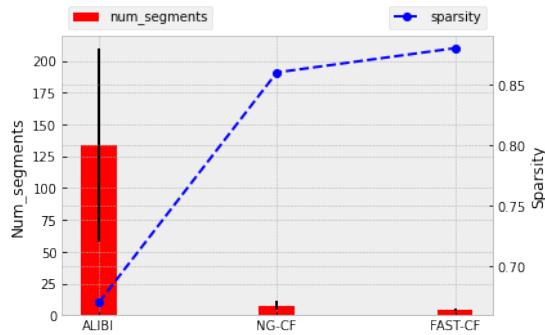


Fig. 3. Comparing the performances of NG-CF, Alibi, and FAST-CF models in terms of Sparsity and the number of independent segments

We also apply an embedding model to visualize the solar flare prediction training data and their associated counterfactual explanations in 2D space which can demonstrate the within distribution property of our generated counterfactual instances. To achieve this purpose, we use the first two eigenvectors of the Principal component analysis (PCA) [21] and visualize the transformed time series vector representations. For better comparison, we show the transformed time series

TABLE I
COMPARING THE PERFORMANCES OF NG-CF, ALIBI, AND FAST-CF MODELS IN TERMS OF L1 DISTANCE AND THE TARGET PROBABILITY (THE WINNER IS BOLDED).

Method	L1 distance		Target probability	
	Mean	Std	Mean	Std
NG-CF	1.68e+12	3.22e+13	1.0	0
ALIBI	495.16	376.69	0.56	0.11
FAST-CF	69.03	69.79	1.0	0

vector representations of the original query set in Figure 4 and show the transformed time series vector representations of the original query set and the generated counterfactual explanations using FAST-CF in Figure 5. In Figure 5, the red stars show the generated counterfactual instances' two-dimensional embedding, while the circle markers show the original query set data points. By comparing the two figures, we can see that the original query set data points and the generated counterfactual instances are highly overlapped, which means that our generated counterfactual instances from FAST-CF are within the distribution of the original query set.

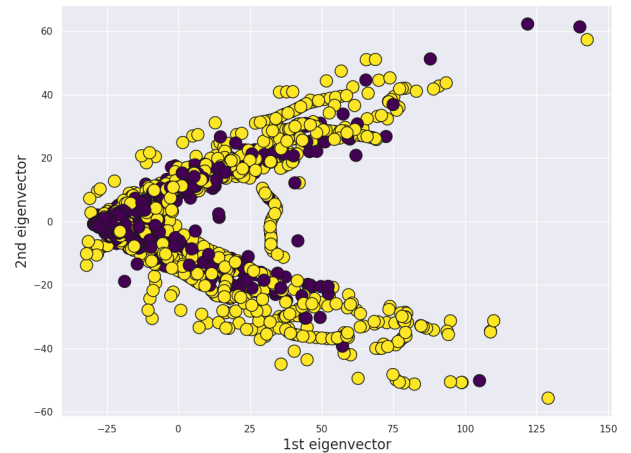


Fig. 4. PCA for solar flare training data

Finally, we visualize an example of the generated counterfactual instance using FAST-CF and compared it with the original query instance in Figure 6. For better visualization, we flattened the 33-dimensions query instance and the generated counterfactual instance into 1-dimension. From Figure 6, we can see that there are 4 top important dimensions (4 contiguous segments) that have been substituted from the original query instance (original instance) to make sure the counterfactual instance (CF) is classified to the class of desired.

VI. CONCLUSION

In this paper, we propose a novel model that generates intuitive, interpretable post-hoc counterfactual explanations for solar flare prediction. In the case of solar flare prediction, we are dealing with multivariate time series data. Due to the high-dimensional nature of multivariate time series data, existing

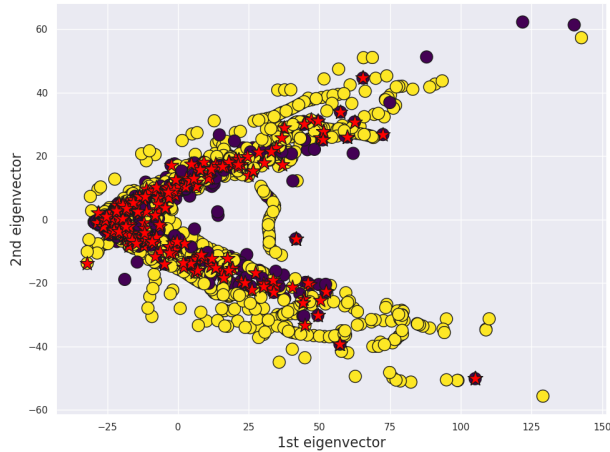


Fig. 5. 2-D PCA projection of the original query set (training data) and the counterfactual explanations. The red stars represent the counterfactual instances generated by FAST-CF.

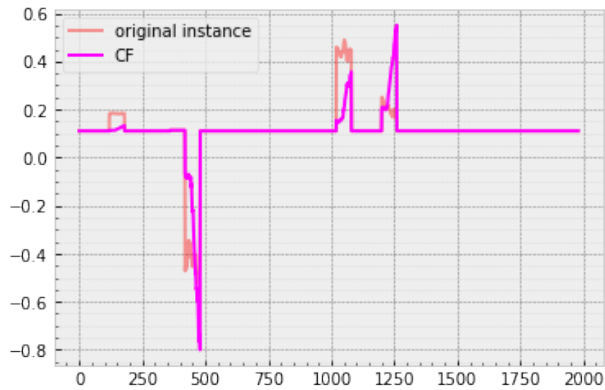


Fig. 6. An example of original query instance and its counterfactual instance (flattened to one dimension)

work that focuses on generating counterfactual explanations for tabular data can not be applied to multivariate time series data directly. We address the high-dimension challenge by proposing FAST-CF, which incorporates the nearest unlike neighbor for guiding the counterfactual search. By focusing on a small set of important dimension substitutions, the FAST-CF method guides the perturbations on the query solar flare prediction data resulting in significantly sparse and more contiguous explanations than other baseline methods. To the best of our knowledge, this is the first effort to focus on a small set of dimension substitutions while generating counterfactual explanations for multivariate TSC. There are spaces for extensions of our work with counterfactual explanations for solar flare prediction with high dimensions. As a future direction of this work, we would like to leverage our method to fit into multivariate time series data in other domains with different dimension complexity.

REFERENCES

- [1] 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 *et al.*, “Multivariate time series dataset for space weather data analytics,” *Scientific data*, vol. 7, no. 1, pp. 1–13, 2020.
- [2] S. M. Hamdi, D. Kempton, R. Ma, S. F. Boubrahimi, and R. A. Angryk, “A time series classification-based approach for solar flare prediction,” in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 2543–2551.
- [3] X. Huang, H. Wang, L. Xu, J. Liu, R. Li, and X. Dai, “Deep learning based solar flare forecasting model. i. results for line-of-sight magnetograms,” *The Astrophysical Journal*, vol. 856, no. 1, p. 7, 2018.
- [4] N. Nishizuka, K. Sugiura, Y. Kubo, M. Den, and M. Ishii, “Deep flare net (defn) model for solar flare prediction,” *The Astrophysical Journal*, vol. 858, no. 2, p. 113, 2018.
- [5] E. Park, Y.-J. Moon, S. Shin, K. Yi, D. Lim, H. Lee, and G. Shin, “Application of the deep convolutional neural network to the forecast of solar flare occurrence using full-disk solar magnetograms,” *The Astrophysical Journal*, vol. 869, no. 2, p. 91, 2018.
- [6] Y. Chen, W. B. Manchester, A. O. Hero, G. Toth, B. DuFumier, T. Zhou, X. Wang, H. Zhu, Z. Sun, and T. I. Gombosi, “Identifying solar flare precursors using time series of sdo/hmi images and sharp parameters,” *Space Weather*, vol. 17, no. 10, pp. 1404–1426, 2019.
- [7] K. Domijan, D. S. Bloomfield, and F. Pitić, “Solar flare forecasting from magnetic feature properties generated by the solar monitor active region tracker,” *Solar Physics*, vol. 294, no. 1, pp. 1–19, 2019.
- [8] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, “Xai—explainable artificial intelligence,” *Science Robotics*, vol. 4, no. 37, p. eaay7120, 2019.
- [9] A. Van Looveren, J. Klaise, G. Vacanti, and O. Cobb, “Conditional generative models for counterfactual explanations,” *arXiv preprint arXiv:2101.10123*, 2021.
- [10] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘‘ why should i trust you?’’ explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.
- [11] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, “Local rule-based explanations of black box decision systems,” *arXiv preprint arXiv:1805.10820*, 2018.
- [12] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.
- [13] M. Schleich, Z. Geng, Y. Zhang, and D. Suciu, “Geco: Quality counterfactual explanations in real time,” *arXiv preprint arXiv:2101.01292*, 2021.
- [14] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the gdpr,” *Harv. JL & Tech.*, vol. 31, p. 841, 2017.
- [15] E. Delaney, D. Greene, and M. T. Keane, “Instance-based counterfactual explanations for time series classification,” in *International Conference on Case-Based Reasoning*. Springer, 2021, pp. 32–47.
- [16] A. Mahendran and A. Vedaldi, “Understanding deep image representations by inverting them,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5188–5196.
- [17] K. Yi, Y.-J. Moon, D. Lim, E. Park, and H. Lee, “Visual explanation of a deep learning solar flare forecast model and its relationship to physical parameters,” *The Astrophysical Journal*, vol. 910, no. 1, p. 8, 2021.
- [18] E. Ates, B. Aksar, V. J. Leung, and A. K. Coskun, “Counterfactual explanations for multivariate time series,” in *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*. IEEE, 2021, pp. 1–8.
- [19] A. V. Looveren and J. Klaise, “Interpretable counterfactual explanations guided by prototypes,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2021, pp. 650–665.
- [20] J. Klaise, A. V. Looveren, G. Vacanti, and A. Coca, “Alibi explain: Algorithms for explaining machine learning models,” *Journal of Machine Learning Research*, vol. 22, no. 181, pp. 1–7, 2021. [Online]. Available: <http://jmlr.org/papers/v22/21-0017.html>
- [21] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang, “Two-dimensional pca: a new approach to appearance-based face representation and recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 1, pp. 131–137, 2004.