SG-CF: Shapelet-Guided Counterfactual Explanation for Time Series Classification

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

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10020866

***Abstract*—EXplainable Artificial I ntelligence (XAI) methods have gained much momentum lately given their ability to shed the light on the decision function of opaque machine learning models. There are two dominating XAI paradigms: feature attribution and counterfactual explanation methods. While the first family of methods explains why the model made a decision, counterfactual methods aim at answering *what-if* the input is slightly different and results in another classification decision. Most of the research efforts have focused on answering the why question for time series data modality. In this paper, we aim at answering the *what-if* question by finding a g ood b alance b etween a set of desirable counterfactual explanation properties. We propose Shapelet-guided Counterfactual Explanation (SG-CF), a novel optimization-based model that generates interpretable, intuitive post-hoc counterfactual explanations of time series classification models that balance validity, proximity, sparsity, and contiguity. Our experimental results on nine real-world time-series datasets show that our proposed method can generate counterfactual explanations that balance all the desirable counterfactual properties in comparison with other competing baselines.**

***Index Terms*—Explainable Artificial I ntelligence (XAI), counterfactual explanations, shapelet-guided, time series classification**

I. Introduction

In recent years, time series classification m ethods have been widely adopted in various domains of physical and life sciences such as aerospace [1], astronomy [2], [3], medicine [4], [5]. The success of data-driven time series classification methods is enabled by the availability of big data which has provided researchers the unprecedented opportunity to develop highly accurate models and deploy them in real-life applications [6], [7], [8]. However, most of the time series models, especially deep learning models, consider accuracy as their foremost priority without necessarily providing a tracing mechanism of the model's decision-making process which limits their interpretability [9]. Although accurate, non-explainable models are hardly adopted by science communities where domain experts are unlikely to put their judgment aside in favor of a machine [10].

To address the model opacity concern, EXplainable Artificial I ntelligence (XAI) m ethods h ave b een p roposed by experts from different domains [11]. Significant research advancements have been accomplished on explainability in the computer vision and natural language processing (NLP) domains, but there are still many challenges to be addressed to provide interpretability for the time series domain [12].

In this work, we focus on generating counterfactual explanations for the time series classification. In the context of time series classification task, given an input query instance x and the model's class prediction c_i, a counterfactual instance x_{cf} includes the minimal necessary change in the input features that flips the predicted class to the desired label c_j [12]. According to the recent literature on counterfactual explanations for various data modalities, an ideal counterfactual explanation should satisfy the following properties: **validity, proximity, sparsity, and contiguity** [13], [14].

In light of the aforementioned counterfactual method properties, we propose a Shapelet-Guided Counterfactual Explanation (SG-CF), a novel model that generates interpretable, intuitive post-hoc counterfactual explanations for time series classification. Our paper contributions are as follows: (1) We define a new objective function that encapsulates the desirable properties (validity, proximity, sparsity, and contiguity) of a counterfactual explanation for time series data. Our new loss does not require the use of class activation maps to search for the counterfactual explanation, which makes it model-agnostic. (2) We leverage the prior mined shapelets for guiding the perturbations toward an interpretable counterfactual explanation. (3) We conduct experiments on nine real-life datasets from various domains (image, Spectro, ECG, motion, sensor, and simulated) and show the superiority of our methods over other baselines.

To the best of our knowledge, this is the first effort to leverage the prior mined shapelets to produce high-quality counterfactual explanations of the time series classification problems. The rest of this paper is organized as follows: in section II, we lay the ground of our research by introducing the background and 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. Background and Related work

One of the most prominent post-hoc explanation approaches is to determine the feature attributions given a prediction through local approximation. Ribeiro et al. [15] propose a feature-based approach, LIME, that examines the effect of the perturbed input in the vicinity of the decision boundary to generate insight. Guidotti et al. [16] extend this approach by proposing LORE, which is a local black box model-agnostic

Authorized licensed use limited to: Utah State University. Downloaded on April 03,2023 at 09:10:16 UTC from IEEE Xplore. Restrictions apply.

explanation approach based on logic rules. Similarly, Lundberg and Lee [17] present a unified framework that assigns each feature an importance value for a particular prediction, which is known as SHAP. Similar to identifying feature importance, visualizing the decision of a model is a common technique for explaining model predictions [18]. In the computer vision community, visualization techniques are widely applied to different applications, ranging from highlighting the most important parts of images to class activation maps (CAM) in convolutional neural networks [19].

Not until recently, Wachter et al. [20] proposed a counterfactual (CF) explanation method as an alternative to feature-based methods. Wachter counterfactual method (wCF) minimizes a loss function, using adaptive Nelder-Mead optimization, that encourages the counterfactual to change the decision outcome and keep the minimum Manhattan distance from the original input instance. Following the same line of thought, GECO has recently been proposed to tackle the plausibility and feasibility issues of the generated counterfactual explanation. The model achieves the desirable counterfactual properties by introducing a new plausibility-feasibility language (PLAF) [21]. Both GeCo and wCF focus on structured tabular datasets. Given that both methods explore a complete search space, they are not adequate to use in a high-dimensional feature space such as a time-series data modality.

More recently, several studies have been proposed to provide explanations for time series data. Native Guide (NG) in [13] integrates feature attribution information into the counterfactual generation process to generate sparse counterfactual explanations that provide information about discriminative and meaningful areas of the time series. TimeX that developed in [14] is an extension based on wCF. TimeX introduces a new dynamic barycenter average Loss term that encapsulates the desirable properties of time series counterfactual explanations. Study in [22] proposed a Motif-guided method that generates counterfactual explanation by naively substituting the mined motif part from the class of interest for time series classification. CoMTE proposed by Ates et al. [23] is a counterfactual explanation method for multivariate time series data. CoMTE focuses on selecting time series from the training set and substituting variables to obtain the explanations. Study in [24] develops a counterfactual generation algorithm SETS for multivariate time series data. SETS is a shapelet explainer that combines the original class-shapelet removal with target class-shapelet introduction to generate instance-based counterfactual explanations.

In this work, we introduce a new shapelet-guided loss term that enforces each counterfactual instance x_{cf} to be close to the prominent shapelet that is representative of a specific class.

III. Preliminary Concepts

In this section, we formally describe the time series counterfactual explanation problem and the key concepts that are used throughout the paper.

Problem Definition. Assume a time series $x = \{t_1, t_2, ..., t_m\}$ is an ordered set of real values, where m is the length, then we can define a time series dataset $\mathbf{T} = \{x_0, x_1, ..., x_n\}$ as a collection of such time series where each time series has mapped to a mutually exclusive set of classes $C = \{c_0, c_1, ..., c_n\}$. We split the dataset $\mathbf{T}$ into a training set $\mathcal{D}$ and a test set respectively. The training set is used to train a time series classifier f. For each query instance x in the test set associated with a predicted class $f(x) = c_i$, a counterfactual explanation model $\mathcal{M}$ generates a perturbed sample x_{cf} with the minimal perturbation that leads to $f(x_{cf}) = c_j$ such that $c_j \neq c_i$.

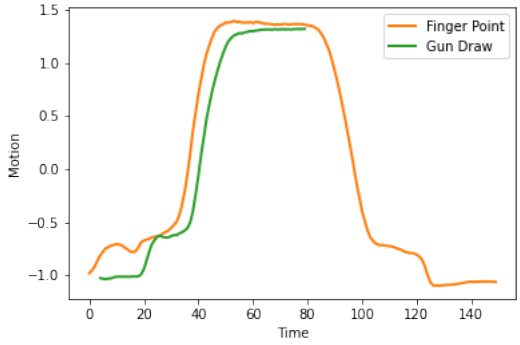


Fig. 1: Classification of Gun-draw/Finger-point movement. We observe that a classifier f predicts the input time series x as class Finger-point. When transforming x to x_{cf} by changing one segment (indicated in green), the prediction result converts to class Gun-draw.

To elaborate on the problem of counterfactual explanation, we present one example from the motion domain:

Gun-draw versus finger-point: Consider the problem of distinguishing whether one specific motion corresponds to a gun-draw or finger-pointing. In Fig. 1, we can see the orange time series which records a regular finger-pointing motion (query instance x). The objective of counterfactual explanation is to generate a perturbed sample (counterfactual instance x_{cf}), such that a time series classifier f predicts it a gun-point motion instead. The green curve in Fig. 1 represents the perturbation result x_{cf} that is generated by a counterfactual explanation model $\mathcal{M}$, which makes the time series classifier f predicts the generated counterfactual instance x_{cf} a gun-point motion.

Shapelets. Shapelets, also known as motifs, are time series sub-sequences that are maximally representative of a class [25]. Shapelets provide interpretable information beneficial for domain experts to better understand their data [25]. Figure 4a and Figure 4b in Figure 4 show the example of the mined shapelets from the Coffee dataset. The magenta lines show two instances from the Robusta and Arabica classes while the highlighted red segments represent their representative shapelets. Shapelets are used in SG-CF to generate shapelet-guided representative samples for assessing the closeness of the produced CF to the desired class outcome as later discussed in Section IV.

Authorized licensed use limited to: Utah State University. Downloaded on April 03,2023 at 09:10:16 UTC from IEEE Xplore. Restrictions apply.

IV. Shapelet-Guided Counterfacutal Explanation

In this section, we describe our proposed SG-CF counterfactual explanation method. SG-CF perturbs time series instances to generate an explanation by explicitly embedding the desired properties of an ideal explanation in the loss function. The original w-CF formulation of the counterfactual discovery problem proposed by Wachter et al. involves minimizing an objective function that encourages the perturbed instance to change the label with minimal perturbed data points [20]. The first objective is measured using the class prediction probabilities of the model f, and the second objective is assessed using the Manhattan distance (L1-norm) between the query instance x and the perturbed instance x_{cf}. Equation 1 defines the loss function with respect to the prediction loss L_{pred} and the distance loss L_{L1}.

$$\arg\min_{x_{cf}} \max_{\lambda} \lambda \underbrace{(c_i - c_j)^2}_{\mathrm{L}_{pred}} + \underbrace{d\,(x, x_{cf})}_{\mathrm{L}_{L1}}, c_j \neq c_i \qquad (1)$$

where c_j is the class of interest that is different than the original class c_i of the query time series x.

A. Shapelet-Guided Loss

Following the original w-CF formulation, we introduce a new shapelet-guided loss term $L_{shapelet}$ to Equation 1 that guides the perturbations performed on x yielding to the counterfactual x_{cf} to fall in the distribution of the counterfactual class. To achieve this purpose, prior to the post-hoc explanation, we mine the most prominent shapelets for each class and introduce them in the objective function to guide the perturbations toward an interpretable counterfactual. The new loss term enforces the counterfactual instance x_{cf} to be close to the prominent shapelet $x_{shapelet}$ extracted from the class of interest c_j. We define the shapelet-representative instance as being the most prominent shapelet of each class. Shapelets' prominence is assessed by their discriminatory power. To mine the most prominent shapelets, we apply the Shapelet Transform (ST) algorithm proposed by [26]. Our new shapelet-guided loss is defined in Equation 2 and the new combined proposed loss is shown in Equation 3.

$$L_{shapelet} = d(x_{cf}, x_{shapelet}) \qquad (2)$$

$$L = L_{pred} + L_{L1} + L_{shapelet} \qquad (3)$$

B. Shapelet-Guided Counterfactual Explanation Algorithm

The general architecture of our proposed counterfactual model is illustrated in Figure 2. For the sake of discussion, in this example, we assume a binary class classification and a deep learning model as the prediction function f. Shapelet-guided Counterfactual Explanation starts by training a classifier f using the training dataset $\mathcal{D}$, then the label of each query instance in the test dataset is predicted using the pre-trained model f. The newly predicted dataset is then split based on the number of classes that exist in $\mathcal{D}$. In this example, the newly predicted dataset results in two splits: predicted positive class (+) and predicted negative class (-). The next step consists

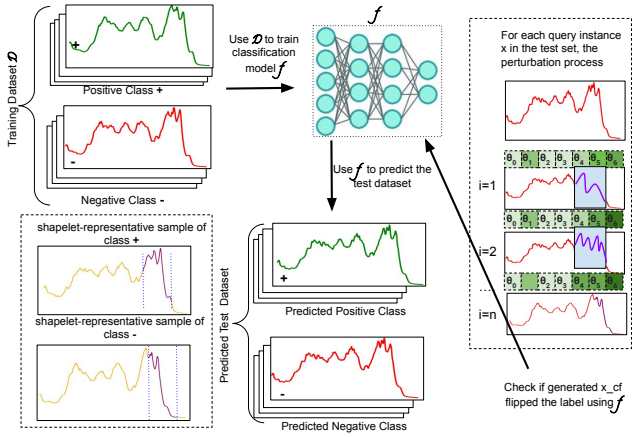


Fig. 2: Shapelet-guided Counterfactual Explanation

of generating the most prominent shapelets from the training dataset $\mathcal{D}$ for the two sets that will be later used during the counterfactual perturbation. Figure 4-(a) and Figure 4-(b) show the two shapelet-representative samples we generated for the Arabica and Robusta classes in the Coffee dataset.

To generate an explanation of a query time series X, the SG-CF optimizer takes the query instance and the most representative shapelet of the opposite predicted class of x. In the example in Figure 2, the predicted input query belongs to the negative class (red class); therefore, the shapelet representative sample of the positive class (green class) is fed to the optimizer. In the first iteration step (i=0), the shapelet existing part (purple part in Fig. 2) is chosen as the best candidate for perturbation. Random noise is then injected into the best candidate segment whose shape is optimized following the desired CF properties encoded in the new proposed loss (refer to Equation 3). At the end of the learning iteration, the optimized counterfactual is evaluated to assess its validity using the pre-trained model f. Fig 4c and Fig 4d in Fig 4 show the generated counterfactual explanations for the Coffee dataset of the Arabica and Robusta classes.

.

V. Experimental Results

In this section, we outline our experimental design and discuss our findings. We conducted our experiments on the publicly-available univariate time series data sets from the University of California at Riverside (UCR) Time Series Classification Archive [27]. We tested our model on nine real-life datasets from various domains (image, Spectro, ECG, motion, sensor, and simulated). We conduct our experiments on both binary class and multiple class datasets. The length of the time series in these datasets varies between 82 and 637. Table I shows the details of the nine datasets we used in our experiments.

A. Baseline Methods

We evaluated our proposed SG-CF model with the other two baselines, Alibi and Native guide counterfactual.

Authorized licensed use limited to: Utah State University. Downloaded on April 03,2023 at 09:10:16 UTC from IEEE Xplore. Restrictions apply.

TABLE I: UCR Datasets Metadata

ID	Dataset Name	Label	TS length	Type
0	Coffee	2	286	SPECTRO
1	BirdChicken	2	512	IMAGE
2	BeetleFly	2	470	IMAGE
3	GunPoint	2	150	MOTION
4	ECG200	2	96	ECG
5	Lightning2	2	637	SENSOR
6	TwoLeadECG	2	82	ECG
7	FaceFour	4	350	SENSOR
8	CBF	3	128	SIMULATED

- **Alibi Counterfactual (Alibi):** Alibi exposes four methods for finding counterfactuals: counterfactual instances (CFI), contrastive explanations (CEM), counterfactuals guided by prototypes (CFP), and counterfactuals with reinforcement learning (CFRL). These four methods are proposed to provide interpretability for the tabular and image data. In our experiment, we fit the first method CFI to our time series datasets. CFI loosely follows the work of Wachter et al. [20], which constructs counterfactual instances from a query instance by running gradient descent on a new instance to minimize the prediction loss L_{pred} and the L1 distance loss L_{L_1}.
- **Native guide counterfactual (NG-CF):** NG-CF is the latest time series counterfactual method that uses the explanation weight vector (from the class activation mapping) and the in-sample counterfactual (the nearest unlike neighbor from another class) to generate a new counterfactual solution. Although the method is model agnostic, the method reaches its full potential when class activation weights are available.

The source code of our model is available on the SG-CF project website [1].

B. Prediction Model Details

For fairness purposes, we evaluated all the aforementioned counterfactual baselines on the same prediction model f. Following the lead of NG-CF, we used a fully convolutional neural network (FCN) originally proposed by Wang et al. [28] for time series classification, closely following the implementation by Fawaz et al. [9].

C. Evaluation metrics

The goal of our experiments is to assess the performance of the baseline methods with respect to all the desired properties of an ideal counterfactual method. To evaluate our proposed method, we compare our method with the other two baselines in terms of several metrics: the target class probability (validity), L1 distance (proximity), sparsity, and the number of independent perturbed segments (contiguity). Fig. 3 shows the results of different datasets. The validity, proximity, sparsity, and contiguity of each dataset that are shown in Fig. 3 is the mean value over the set.

[1]https://sites.google.com/view/sg-couterfactual/

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.

We use L1 distance to demonstrate the proximity, which measures the closeness between the counterfactuals and the original input instance, a smaller L1 distance is preferred.

The third evaluation metric that we used is the sparsity level (Equations 4-5) of the baseline models, which is defined by [14]. The highest sparsity is an indicator that the time series perturbations made in x to achieve x_{cf} are minimal. Therefore, a higher sparsity level is desirable.

$$Sparsity = 1 - \frac{\sum_{i=0}^{len(x)} g({x_{cf}}^i, x^i)}{len(x)} \tag{4}$$

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

Finally, we investigate the number of independent non-contiguous segments that were used by each baseline to perturb to achieve a counterfactual explanation, which is related to the contiguity property. The lower the contiguity metric value the better.

D. Evaluation results

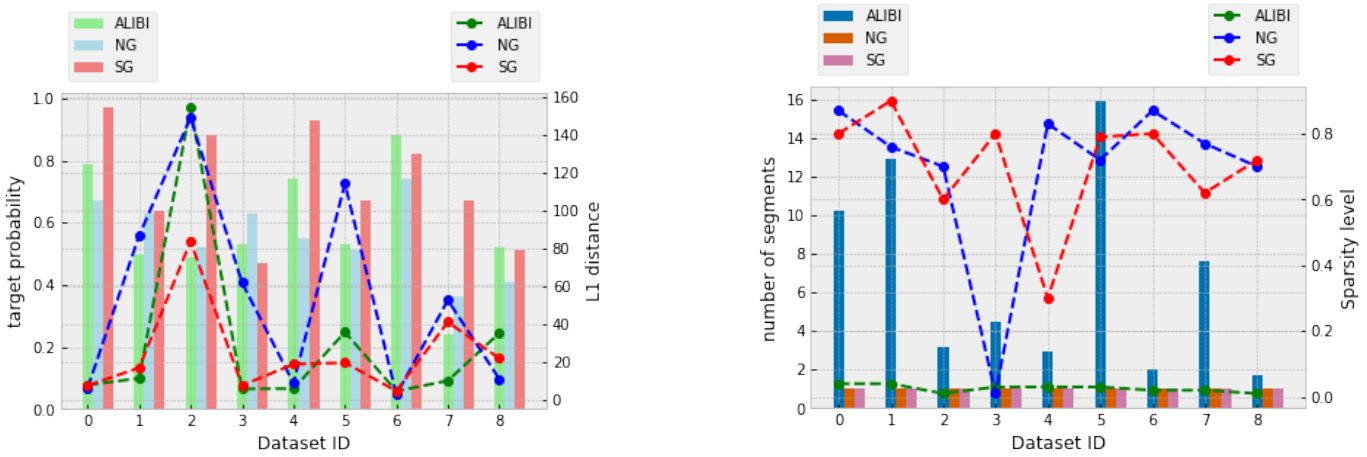
Fig. 3 shows the evaluation results on the CF desired properties assessed using the aforementioned metrics. Fig. 3a shows the comparison of the validity (target probability) and proximity (L1 distance) properties for the generated CF among all counterfactual explanation models. For BirdChicken (ID 1), ECG200 (ID 4), and FaceFour (ID 7) datasets, we note that Alibi achieves the lowest L1 distance (proximity), which is highly desirable. However, Alibi minimizes proximity in the cost of validity. Alibi achieves the lower target class probability (validity), which entails that the method generates CF explanations that are relatively less valid. For the CBF dataset(ID 8), NG-CF achieves the lowest L1 distance but also results in the lowest target probability. Our proposed SG-CF achieves competitive L1 distance while resulting in the highest target probability in general. In sum, compared with ALIBI and NG-CF, SG-CF shows a good balance without maximizing one property and compromising the other in terms of proximity and validity.

Fig. 3b shows the comparison of the contiguity (number of segments) and sparsity (sparsity level) properties for the generated CF among all counterfactual explanation models. Given the constraint on the size of the perturbation range that is used in SG-CF and NG-CF methods, the sparsity levels and the number of independent segments of NG-CF and SG-CF are similar. However, ALIBI results in the lowest sparsity level and the highest number of independent perturbed segments compared to NG-CF and SG-CF.

E. Case Study for Coffee dataset

Food spectrographs are used in chemometrics to classify food types, a task that has obvious applications in food safety and quality assurance [29]. The coffee data set is a two-class problem to distinguish between Robusta and Arabica

Authorized licensed use limited to: Utah State University. Downloaded on April 03,2023 at 09:10:16 UTC from IEEE Xplore. Restrictions apply.



(a) Target probability (the higher the better); L1 distance (the lower the better)

(b) Number of independent segments (the lower the better); sparsity level (the higher the better)

Fig. 3: The evaluation result of the CF explanations generated by ALIBI, NG-CF, and SG-CF. (All of the reported results are the average value over the counterfactual set)

coffee beans. In this section, we consider a coffee beans classification task and demonstrate a user-case example from the Coffee spectrogram dataset. Fig. 4a and 4b show the two representative samples that we extracted from Robusta and Arabica classes respectively. The red lines shown in Fig. 4a and 4b are the representative shapelets that are used to guide the counterfactual explanations. Giving two query instances (see the magenta lines in Fig. 4c and Fig. 4d) from Robusta and Arabica classes respectively, we generate the counterfactual explanations (see the green lines in Fig. 4c and Fig. 4d) for each class by applying the objective function that shows in equation 3. The perturbation parts are shown between two blue parallel dashed lines in Fig. 4c and Fig. 4d, the perturbation parts are limited in shapelet existing range, which contributes to getting contiguous and sparsity counterfactual explanations. The small distances between the query instance and the counterfactual instance that we can see from Fig. 4c and Fig. 4d) also verify the proximity property of our generated counterfactual explanations.

VI. Conclusion

In this paper, we propose a novel model SG-CF that generates interpretable, intuitive post-hoc counterfactual explanations for time series classification. Existing work of providing counterfactual explanations for time series data suffer from the problem of maximizing one desirable property at the cost of others. We address these challenges by proposing SG-CF, an optimization-based model that produces high-quality counterfactual explanations that are close to the original input instance, contiguous, and sparse. The shapelet-guided counterfactual explanation method includes a new shapelet-guided loss that guides the perturbations on the query time series resulting in significantly sparse and more contiguous explanations. As a future step, we would like to extend the scope of this work to multivariate time series data by applying Allen's interval algebra to connect the shapelets from different dimensions to generate high-quality counterfactual explanations.

References

[1] J. Lin, E. Keogh, S. Lonardi, J. P. Lankford, and D. M. Nystrom, "Visually mining and monitoring massive time series," in *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 460–469.

[2] S. F. Boubrahimi, B. Aydin, M. A. Schuh, D. Kempton, R. A. Angryk, and R. Ma, "Spatiotemporal interpolation methods for solar event trajectories," *APJs*, vol. 236, no. 1, p. 23, 2018.

[3] S. F. Boubrahimi, S. M. Hamdi, R. Ma, and R. Angryk, "On the mining of the minimal set of time series data shapelets," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 493–502.

[4] J. Lin, E. Keogh, A. Fu, and H. Van Herle, "Approximations to magic: Finding unusual medical time series," in *IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*. IEEE, 2005, pp. 329–334.

[5] K. Buza, A. Nanopoulos, L. Schmidt-Thieme, and J. Koller, "Fast classification of electrocardiograph signals via instance selection," in *2011 IEEE First International Conference on Healthcare Informatics, Imaging and Systems Biology*. IEEE, 2011, pp. 9–16.

[6] W. A. Chaovalitwongse, O. A. Prokopyev, and P. M. Pardalos, "Electroencephalogram (eeg) time series classification: Applications in epilepsy," *Annals of Operations Research*, vol. 148, no. 1, pp. 227–250, 2006.

[7] Y.-H. Lee, C.-P. Wei, T.-H. Cheng, and C.-T. Yang, "Nearest-neighbor-based approach to time-series classification," *Decision Support Systems*, vol. 53, no. 1, pp. 207–217, 2012.

[8] J. Wiens, E. Horvitz, and J. Guttag, "Patient risk stratification for hospital-associated c. diff as a time-series classification task," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

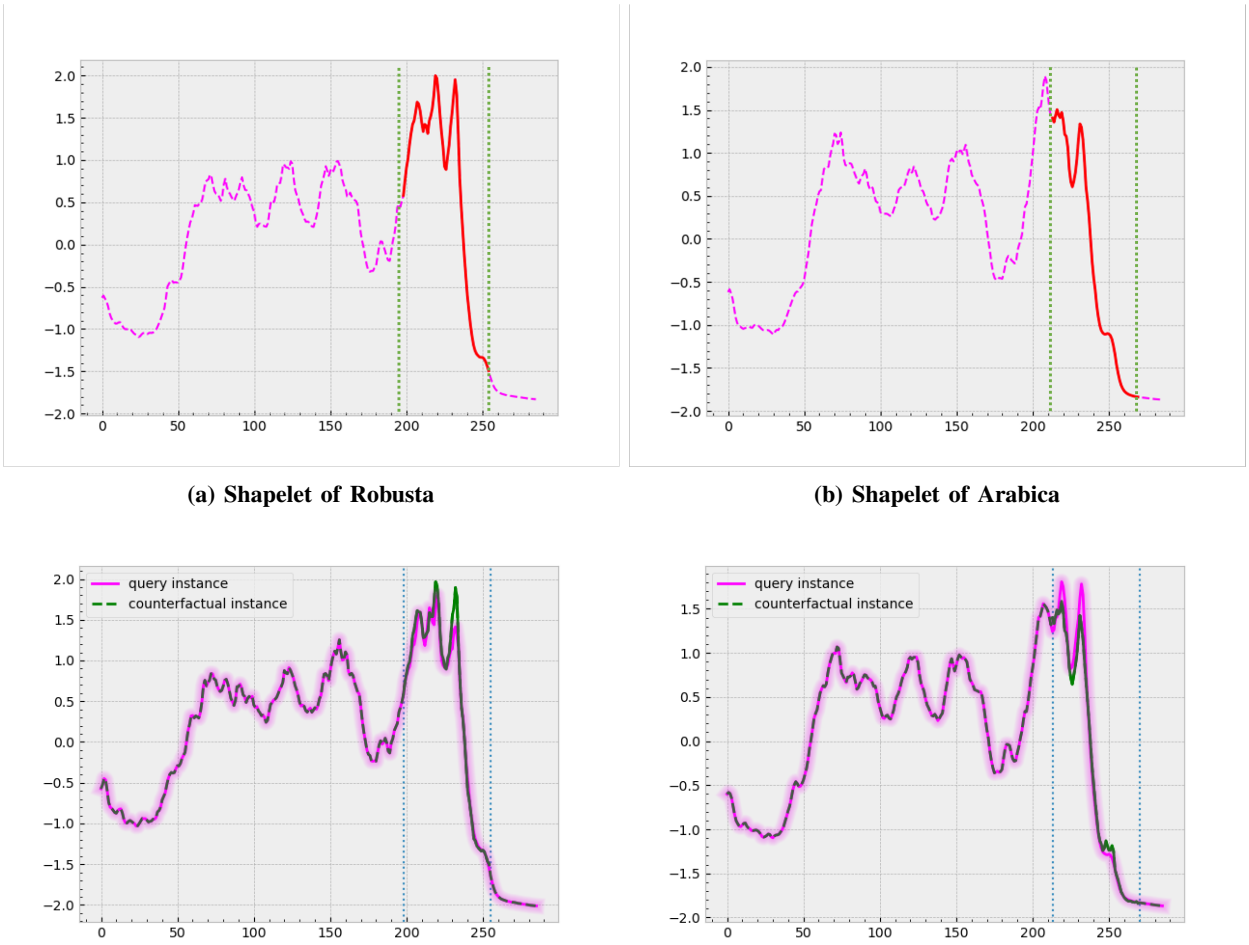
[9] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data mining and knowledge discovery*, vol. 33, no. 4, pp. 917–963, 2019.

[10] S. Kundu, "AI in medicine must be explainable," *Nature Medicine*, vol. 27, no. 8, pp. 1328–1328, 2021.

[11] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[12] T. T. Nguyen, T. Le Nguyen, and G. Ifrim, "A model-agnostic approach to quantifying the informativeness of explanation methods for time series classification," in *International Workshop on Advanced Analytics and Learning on Temporal Data*. Springer, 2020, pp. 77–94.

Authorized licensed use limited to: Utah State University. Downloaded on April 03,2023 at 09:10:16 UTC from IEEE Xplore. Restrictions apply.



(a) Shapelet of Robusta

(b) Shapelet of Arabica

(c) Robusta CF and query

(d) Arabica CF and query

Fig. 4: Case study of SG-CF method for the Coffee dataset

[13] 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.

[14] S. Filali Boubrahimi and S. M. Hamdi, "On the mining of time series data counterfactual explanations using barycenters," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 3943–3947.

[15] 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.

[16] 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.

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Info. Processing Systems*, vol. 30, 2017.

[18] 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*, 2020, pp. 607–617.

[19] 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.

[20] 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.

[21] M. Schleich, Z. Geng, Y. Zhang, and D. Suciu, "Geco: Quality counterfactual explanations in real time," *arXiv preprint arXiv:2101.01292*, 2021.

[22] P. Li, S. F. Boubrahimi, and S. M. Hamdi, "Motif-guided time series counterfactual explanations," *arXiv preprint arXiv:2211.04411*, 2022.

[23] 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.

[24] O. Bahri, S. F. Boubrahimi, and S. M. Hamdi, "Shapelet-based counterfactual explanations for multivariate time series," *arXiv preprint arXiv:2208.10462*, 2022.

[25] L. Ye and E. Keogh, "Time series shapelets: a new primitive for data mining," in *ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 947–956.

[26] J. Lines, L. M. Davis, J. Hills, and A. Bagnall, "A shapelet transform for time series classification," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 289–297.

[27] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, and E. Keogh, "The UCR time series archive," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 6, pp. 1293–1305, 2019.

[28] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *2017 International Joint Conference on Neural Networks*. IEEE, 2017, pp. 1578–1585.

[29] 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, 2014.

Authorized licensed use limited to: Utah State University. Downloaded on April 03,2023 at 09:10:16 UTC from IEEE Xplore. Restrictions apply.