When to Explain? Model Agnostic Explanation Using a Case-based Approach and Counterfactuals

Betül Bayrak[0000−0002−0554−9823] and Kerstin Bach[0000−0002−4256−7676]

Department of Computer Science
Norwegian University of Science and Technology (NTNU),
Trondheim, Norway
{betul.bayrak,kerstin.bach}@ntnu.no

Abstract. Explainable Artificial Intelligence (XAI) systems have gained importance with the increasing demand for understanding why and how an artificial intelligence system makes decisions. Counterfactual explanations, one of the rising trends of XAI, benefit from human counterfactual thinking mechanisms and aim to follow a similar way of reasoning. In this paper, we create an eXplainable Case-Based Reasoning system using counterfactual samples with a model-agnostic approach. While CBR methodology allows us to use past experiences to create new explanations, using counterfactuals helps to increase understandability. The main idea of this paper is to generate an explanation when necessary. The proposed method is sample-centric. Thus, an adaptive explanation area is calculated for each data point in the dataset. We detect if there is any existing counterfactual of the samples to increase the coverage of the system, and we create explanation cases from detected sample-counterfactual pairs. If a query case is in the explanation area, at least one explanation case will be triggered, and a two-phase explanation will be created using a text template and a bi-directional bar graph. In this work, we will show (1) how explanation cases are created, (2) how the nature of a dataset influences the explanation area, (3) how understandable explanations are created, and (4) how the proposed method works on open datasets.

Keywords: Explainable AI · XCBR · Counterfactual · Model-Agnostic Explanation Generation

1 Introduction

With the popularization of *artificial intelligence* (AI), AI models have become a significant factor in many areas of life, like e-commerce, health, or banking. It becomes more substantial to understand how and why decisions affect our lives. For this reason, many studies have been conducted in the literature on the understandability of models over time [9, 3, 14]. Several studies focused on interpretable models [20, 4], while others concentrated on explainable models [18, 19, 17], especially in the more recent periods.

In this paper, we develop an eXplainable-CBR system using a counterfactual approach and a model-agnostic Case-Based Reasoning (CBR) based methodology. The counterfactuals improve comprehension, while the CBR methodology enables us to draw on the past to develop new explanations. Our primary goal is to create an explanation when one is required. The proposed method is sample-centric and therefore an adaptive explanation area is calculated. We detect if there is any existing counterfactual of the samples to increase the coverage of the system, and we create explanation cases from detected sample-counterfactual pairs. In the case where an explanation is required for a classification result a two-phase explanation will be created using a text template and a bi-directional bar graph.

Our work will demonstrate how to build explanation cases, show how characteristics of datasets affect the explanation area, describe how to create understandable explanations, and present how the suggested method performs on open datasets.

This paper is structured as follows: in Section 2 we provide a background for paper and discuss relevant work in Section 3. In Section 4 we explain the processes of the proposed method. Section 5 shows how open datasets perform with the proposed method, and we discuss significant points. The last section concludes the paper and gives directions for future works.

2 Background

Interpretability and explainability terms are frequently mentioned and discussed in the literature [24, 2, 18]. However, the terms are often used interchangeably [2]. Briefly, *interpretability* is a feature of the model and represents the understandability of the cause-effect relationship for the model. *Explainability* is an external feature that is constituted by a process. In contrast to interpretability, parameters such as target audience or disclosure scope are controllable and configurable in the explainability feature. Earlier, the interpretability was higher due to the lower complexity of the models used [2]. However, over time, the models developed and became more complex, and as models have become more complex, interpretability decreased and opacity increased. Therefore, the interpretability-performance trade-off discussion started [20, 15, 2].

As black-box models with high opacity are widely used in decision-making systems in every division in life, explainability becomes a more important concept, and studies on explainability begin to be conducted in the literature. Thus, the concept of *eXplainable Artificial Intelligence* (XAI) came into existence, and nowadays, it is known as a significant sub-topic of AI [5] which arose from interpretable machine learning [27].

In the applications of XAI, creating explanations may hold many different purposes and ways. Explanation types may be categorized based on many different views, *Nunes and Jannach*'s paper has one of the broadest views in the literature [16]. The authors listed 17 different explanation types under four different categories. Also, they pointed out the importance of considering the aim of the explanation in the explanation generation process. From another view, which is similar to the paper of *Arrieta et al.* [2], ways of creating explanations can be discussed under two different concepts, model-dependent, and model-agnostic explanation systems. In the first category, model-dependent explanation systems, explanation mechanisms are designed depending on a specific model and its capability; the explanation mechanism can also be performed only on the model. (*e.g.* [12, 29, 10]) In the second category, model-agnostic explanation systems, explanation mechanisms are not designed depending on a specific model and its capability; the explanation mechanism can perform on any model, notwithstanding the model structure or complexity. (*e.g.* [17, 6, 7])

Another challenge in the applications of XAI is generating or selecting the best explanation for a case. While creating explanations, there are many challenges to overcome. Some of the quality criteria for an explanation can be listed as follows:

- **Trustworthy:** Explanation trustworthiness is about how accurately reflected why the decision was made.
- **Understandable:** Explanation medium and content should be created according to the target audience.
- **Informative:** Explanations should convey the proper arguments to the target audience.
- **Sufficient:** Explanations might not be complete but should be sufficient to convey the main reasons.
- **Unbiased:** Explanations should not contain data that is discriminatory, biased or emphasizing existential characteristics for humans and animals.

Many approaches from different fields meet the quality criteria mentioned above. One of which is the *Case-Based Reasoning* (CBR) methodology which is a problem-solving methodology with high interpretability, which has four different steps and benefits from past experiences [1]. In the CBR approach, previous cases are stored in the case base, and to solve a new problem, the following four steps are applied:

- **Retrieve:** Retrieving similar cases to new problems from the case base.
- **Reuse:** Reuse a solution from retrieved similar cases.
- **Revise:** Adapt the solution according to the new problem if needed.
- **Retain:** Retain the new problem and solution (case) to solve upcoming problems.

Sormø et al. [24] mentioned that CBR concentrates on open-ended, often changing, uncertain and incomplete problems and underscored that the CBR approach is flexible and can be applied to a large variety of problems through its simplified problem-solving strategy.

Characteristics of the problems that CBR focused on are very similar to the challenges of XAI, so the CBR methodology is often used to create explanations for AI models; in this way, the concept of *eXplainable Case-Based Reasoning* (XCBR) emerged. XCBR is categorized as a sub-field of XAI [22] and gains ground in XAI. Since the CBR methodology allows the reuse of experiences to generate explanations for both model-agnostic and model-dependent explanations in every kind of AI model and application, XCBR systems are flexible, interpretable, sustainable, and evolved over time. These advantages enable creating local and understandable explanations, explanation systems adaptable to data distribution changes, and trustworthy explanations with a small amount of data.

Counterfactual explanations (CE) have been studied in different papers [8, 11, 28], and it is a rising trend of XAI. A counterfactual explanation provides a causal situation with a contrastive argument, like "If you were younger than 59 years old, you would get the loan". CE can be in different formats like text, image, or graph, but the important thing is to follow the counterfactual thinking method to increase the understandability. CE are not only for explaining the decisions of the models but also for providing insights on how to change the outcome. [28]

3 Related Work

An *explanation* is a concept dating back to the end of the 1940s [24]. Sormo et al. [24] give a detailed recap of the explanation concept from the view of philosophy and cognitive science societies. Using this perspective, the authors described the transition of explanation concepts, and they discussed explanations in expert systems and explanations in CBR systems. Furthermore, they clearly stated common and diverse points of these concepts, challenging points, and gaps in the fields.

XAI has developed into a popular topic over the past years, and XCBR is also a pioneer sub-field of XAI that has its roots in the late 1980s with SWALE [21]. Arrieta et al. [2] Schoenborn et al. [22] provide one of the most complete perspectives of concepts, taxonomies, open-ended questions, and challenges, respectively, in XAI and XCBR fields. While categorizing the studies in the literature, they both used a very similar approach to our work.

Explanations on Smart Stores paper [9], provides a view on how to apply XCBR in daily life. The authors proposed an explanation system for un-staffed smart stores by taking advantage of the CBR methodology. This paper is a good example of creating understandable explanations with template tables and representing cases from real-life data for the target audience. The explanation template is divided into header, body, and footer: these parts are used for parameters to define the case, text explanation, and similarity score, respectively.

Ribeiro et al. published a paper in 2016 [19], one of the famed XAI studies, proposed a model-agnostic explanation technique named LIME (Local Interpretable Model-agnostic Explanations) for classifiers. It claims that the LIME technique is interpretable, faithful, and flexible. This technique derives an explanation for a new case by using random local cases and weighting them by distance. Afterward, many studies in the literature have been built on this paper. (*i.e. [23, 30, 17, 26]*) The CBR-LIME work [17] pointed out a problem in

LIME: the configuration of parameters. The paper proposes a CBR solution to the problem for an image classifier. In the presented system, the case base reflects the human perception of the explanation qualities for different LIME configurations. Using CBR methodology, they use human knowledge to create more proper explanations.

In XAI applications, the understandability of explanations plays an essential role in the explanation quality. There shall be various ways to create understandable explanations; as an example, in the paper, we discussed above Huertos et al. [9] use a table template. Furthermore, Lamy et al. proposed a visual explanation system that does not require domain knowledge to understand explanations [13]. They provide a model-agnostic visual explanation system using the simplicity of CBR methodology; in the explanation part, they show both quantitative (a radar plot - displays similarity) and qualitative (rainbow boxes - displays common features) approaches. Another real-life concentrated paper is about predicting and explaining running-related injuries from Strava [1] marathon training histories [7]. The paper described the case representation, prediction, and explanation processes; the explanation creation technique benefits from both cases and their counterfactual cases. Besides feature analysis and visualizations, the authors stated that using different types of explanations, like text templates with statistical information, will increase the understandability of the explanation.

Keane and Smyth published a paper about exploring and generating counterfactuals for explanation systems [11]. According to the authors, a good counterfactual is the nearest unlike neighbor to the case with at most two feature differences. In the first part, the authors discussed the exploration of good counterfactuals, while in the second part they conducted experiments on 20 different UCI datasets, and mentioned that good counterfactuals are not frequently encountered in datasets. To observe the salience points of the model for explanations, the second part of the paper proposes synthetic counterfactual creation from existing cases. Although successful inferences are made with the CBR approach, using created explanations with synthetic data on fields that might affect humans, animals, and nature may be open to dispute.

4 Development of an XCBR System

The creation of an XCBR system includes many steps; in this part, we describe our proposed method in the following three subsections. Also, in Section 4.4, we show the details of how the proposed method works with a dataset.

4.1 When Do We Need Explanation?

In real-life applications, there is usually no need to explain ordinary situations, but an explanation is needed for an unexpected result or in uncertain situations.

[1] https://www.strava.com/

For example, in a system where obesity diagnosis is made from people's height, weight, sex, age, and activity level, if the patient data is not close to obesity, there is no need to explain. However, if the patient is pretty close to having obesity, explaining why there is a risk is crucial to avoid obesity and warn the patient. Therefore, we follow the idea of explaining AI model decisions when necessary by detecting samples around decision boundaries. When there is a data distribution, as in Figure 1a, selecting the samples in the explanation area is fairly easy because the distributions of the classes are separate. An imaginary decision boundary can be drawn, and uncertain situations exist around this boundary. However, almost no dataset has such clear decision boundaries.

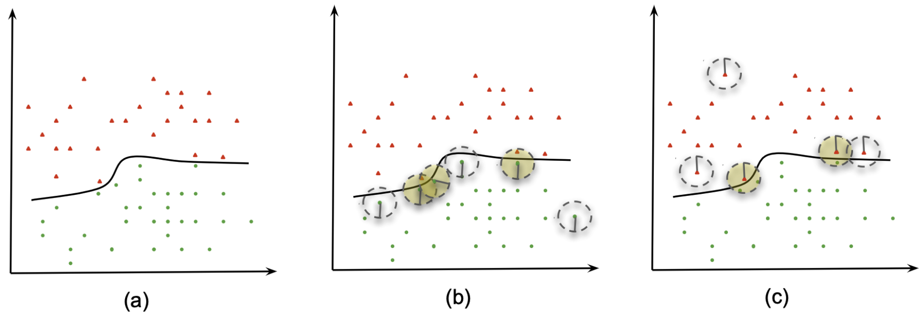


Fig. 1. **(a)** Imaginary decision boundary in a two dimensional dataset, **(b,c)** Sample-centric pair detection, highlighted areas are detected as explanation area.

One of the critical points in this paper is detecting the explanation area to decide in which areas samples need explanations. As mentioned above, we consider uncertain cases to explain. The cases with at least one counterfactual in a determined distance are identified in the explanation area and added to the case base as explained in the following section.

4.2 Case Base Elicitation

A primary case base must be provided for a CBR system to solve problems. The case base elicitation process is adapting and representing existing previous knowledge as cases. *Algorithm-1* shows the flow of case base elicitation process.

In *Algorithm-1*, the case base elicitation process, X and y are the requirements for the first step. X is the list of data points and x_i is a data point with m features.

$$X = [x_1, x_2, ..., x_n],\ x_i = \langle f_1, f_2, ..., f_m\rangle\ where\ i \leq n \tag{1}$$

and y is the list of target values for the data points:

$$y = [y_1, y_2, ..., y_n],\ y_i \in 0, 1\ where\ i \leq n. \tag{2}$$

The pre-processing step applies one-hot encoding for categorical features. Standard scaling is applied for numerical features of X ensuring unbiased measurements of the distance between cases.

After the pre-processing step, the global variables k and avg_r are initialized. k is the best performing neighbour when training the kNN classifier for the dataset. The classifier is optimized using the accuracy running a cross-validation using grid search. Every sample has a r value, where r is the mean distance between the sample and the k nearest neighbours. avg_r is the average of r values of all samples; therefore, it is named as global r.

To detect explanation areas, a sample-centric approach is used. Namely, a radius (r) is calculated for a sample and if there is at least one counterfactual in the sample's circle, this area will be considered as explanation area. Also, detected pairs will be added to the case base.

For each sample (s), we calculate r as the mean of the k nearest neighbours distance to s. k_nn is a list of k nearest neighbours of s.

$$s = \langle f_1, f_2, ..., f_m, y \rangle \tag{3}$$

$$r = \sum_{i=1}^{n} \frac{dist(s, k\_nn_i)}{n} \tag{4}$$

The r value can become large for samples that are outliers. Such outliers may cause the detection of irrelevant counterfactual samples. To avoid detecting irrelevant counterfactuals, we update r with avg_r if r is larger than avg_r.

$$r \leftarrow min(find\_r(), avg_r) \tag{5}$$

cf_s is the list of counterfactual samples that it is empty when the algorithm is initialized. The counterfactual detection process detects all existing counterfactuals within maximum r distance from s *(See Figure 1b,c)* and adds them to the cf_s list.

$$cf_{si} = \langle f_1, f_2, ..., f_m, \neg\ y \rangle \ where\ i \in 1, 2, ..., n\ ,\ dist(s, cf_{si}) < r \tag{6}$$

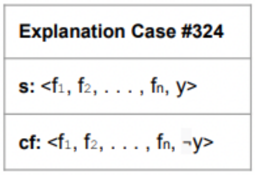
Algorithm 1 Case base elicitation

```
Require: : $X, y, cf$
$X$ = preprocessing($X$)
$k \leftarrow find\_k()$
$avg_r \leftarrow find\_avg\_r()$
for all $s$ in $X$ do
    $r \leftarrow min(find\_r(), avg_r)$
    $cf_s \leftarrow k$ nearest counterfactual sample
    Filter(dist($s$, $cf_s$) $< r$)
    if $cf_s$ is not empty then
        $cf['sample']$.append($s$)
        $cf['cf\_list']$.append($cf_s$)
    end if
end for
```

To create explanation cases, explanation pairs can be derived from the sample and its counterfactuals. An explanation case (exp_{si}) is defined as:

$$exp_{si} = \langle s, cf_{si} \rangle \ where \ i \in 1, 2, ..., n \tag{7}$$

All possible pairs are created and used to generate cases. Figure 2 shows the explanation case format in which s represents a*Sample* and exp_{si} represents one *Counterfactual* in an explanation case.



Explanation Case #324
s: <f1, f2, . . . , fn, y>
cf: <f1, f2, . . . , fn, ¬y>

Fig. 2. Case representation

After the explanation cases are created and added to the case base, the similarity functions are modelled using the method described in Verma et al. [25]. Hereby, our CBR system is ready to create explanations for new cases.

4.3 Generating Explanations

This section describes how the information provided by an explanation case is used to create an explanation for a user. A new sample (s_{new}) and the prediction result (s_{new_y}) from the black-box model *bb_model* of the sample are needed as input for the explanation process.

$$s_{new} = \langle f_1, f_2, ..., f_n \rangle \tag{8}$$

$$s_{new\_y} = bb\_model.predict(s_{new}) \tag{9}$$

In the next step, we retrieve the four most similar explanation cases with a higher global similarity score than the threshold for which an explanation should be generated. The threshold is determined according to data distribution. As part of the explanation method, we show which classes the neighboring cases belong to and compare those classes with the prediction of the incoming query using the black-box model. Given the information from the existing and query cases, we can explore how decisions change depending on the feature values in similar cases.

Showing the retrieved explanation cases can be enough to explain a new case. However, one of the most critical requirements of an explanation is understandability. Therefore, we create a two-phase explanation instead of showing the pure explanation cases. Firstly, a textual explanation that describes how many similar cases are in the same class and situations in which the prediction result might change. Secondly, a visual explanation that describes how the counterfactuals' features differ from the samples' features using a bi-directional bar graph.

4.4 Application on Artificially Generated Dataset

We use a generated dataset with two features f_1, f_2 and 1000 samples (*see Figure 4a*) to explain the steps of our proposed method. The generated dataset has mixed and separated areas, as shown in Figure 3. The distribution allows us to test the proposed method in different situations.

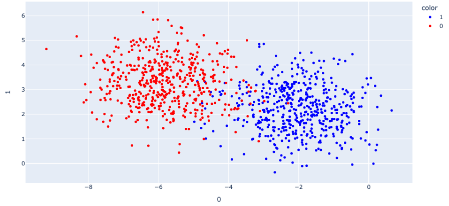
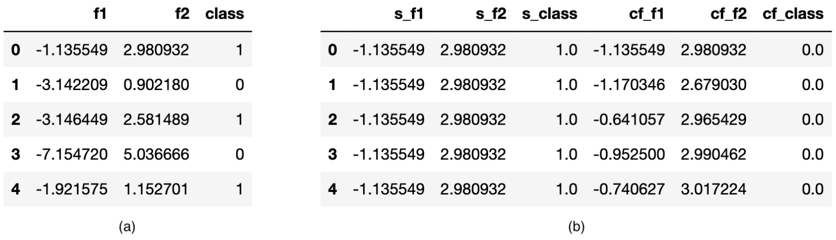


Fig. 3. Generated two feature dataset

Firstly, we apply all steps described in Section 4.2. For the given 1000 samples, the best k value is calculated as four, and the explanation area is detected using the r values. 502 samples were detected in the explanation area, and 8350 explanation pairs were created from those 502 samples. The instances of the created explanation sample-counterfactual pairs are shown in Figure 4b. After adding the generated explanation pairs to the CBR system and modeling their similarities, the CBR system is ready to query new cases and propose explanations. As described in Section 4.3, we need a new sample and its classification result as input to create an explanation. To test our explanation system, we trained a Gradient Boosting Classifier, which has accuracy of 0.975.



	f1	f2	class
0	-1.135549	2.980932	1
1	-3.142209	0.902180	0
2	-3.146449	2.581489	1
3	-7.154720	5.036666	0
4	-1.921575	1.152701	1

(a)

	s_f1	s_f2	s_class	cf_f1	cf_f2	cf_class
0	-1.135549	2.980932	1.0	-1.135549	2.980932	0.0
1	-1.135549	2.980932	1.0	-1.170346	2.679030	0.0
2	-1.135549	2.980932	1.0	-0.641057	2.965429	0.0
3	-1.135549	2.980932	1.0	-0.952500	2.990462	0.0
4	-1.135549	2.980932	1.0	-0.740627	3.017224	0.0

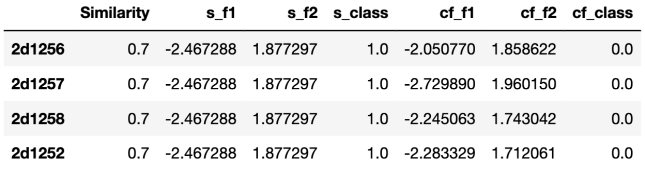
(b)

Fig. 4. (a) Samples of randomly generated dataset (b) Created explanation pairs

As an example, for a query sample s the predicted class is '1'.

$$s = \{f1 : -2.547,\ f2 : 1.854,\ class : 1\} \tag{10}$$

s queried in the case base and four most similar samples, with at a global similarity of at least 0.6 is retrieved with their counterfactuals (*see Figure 5*).



	Similarity	s_f1	s_f2	s_class	cf_f1	cf_f2	cf_class
2d1256	0.7	-2.467288	1.877297	1.0	-2.050770	1.858622	0.0
2d1257	0.7	-2.467288	1.877297	1.0	-2.729890	1.960150	0.0
2d1258	0.7	-2.467288	1.877297	1.0	-2.245063	1.743042	0.0
2d1252	0.7	-2.467288	1.877297	1.0	-2.283329	1.712061	0.0

Fig. 5. Retrieved four explanation cases for *s*

To meet the *understandable and informative explanation requirement*, the textual explanation template created as follows:

> ” The prediction result is the same with **4** out of **4** closest samples. However, the sample is at risk; in similar cases, when the **f1** feature increases by **0.14** and **f2** decreases by **0.05**, decisions change. ”

This explanation format showed that the black-box classifier made the same prediction as the closest samples, but it may flip the class with slight differences, which means the sample is close to a decision boundary. This is an informative and quantitative way to warn the user about the decision boundaries and risks.

A visual explanation method, which includes qualitative features, is generated using a bi-directional bar graph to reinforce the textual explanation. In Figure 6, every bar group implies a counterfactual, and every colour implies a feature. A difference (y-axis) for a counterfactual indicates a deviation between features. No difference means that the query and the counterfactual match. Thereby the audience can easily understand which features can affect the classification results.

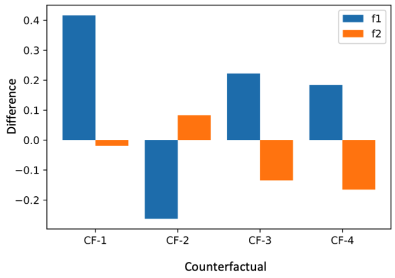


Fig. 6. Visual explanation of s

Figure 6 shows that $f1$ is more dominant than $f2$ in terms of changing classification result for s. Also, while increasing $f1$ and decreasing $f2$ the decision of the classifier can be changed.

5 Discussion

We proposed a model-agnostic XCBR system to explain the decisions of black-box models. The proposed system uses sample-counterfactual pairs to create a two-phase explanation, textual and visual. As mentioned in Section 1, explanations must be created for the target audience and should be trustworthy, understandable, informative, sufficient, and unbiased. In this paper, the textual explanation provides a quantitative approach, while the visual explanation provides a qualitative approach. Therefore, generated explanations appeal to the audience from all aspects, whether the audience is a domain expert or has no idea about the domain. Also, the proposed method can be applied to every domain.

Detection of the sample-counterfactual (explanation) pairs complies with the dataset distribution. For example, in Algerian Forest Fires dataset (Figure 8a, Table 1), there are 243 samples and k value calculated as 3. 137 pieces of samples selected for creating the explanation pairs. The number of the explanation pairs is 141. Meanwhile, in Cesarean dataset (Figure 7a, Table 1), from 80 samples 39 samples selected for creating the explanation pairs and 41 explanation cases generated.

Table 1. Information about datasets

	# samples	# features	k	# samples has explanation	# exp. pairs
Cesarean [2]	80	5	2	29	41
Algerian Forest Fires [3]	243	11	3	137	141

In visual explanations, a color dominance based approach is used. The proportion of the colors implies the features' importance to flip classes. For instance, in Figure 7b, only blue color exists, and blue implies 'age' feature. There are no other differences between samples and counterfactuals. It means, for queried sample, 'age' is the most crucial feature to flip the decision.

In some cases, overlapping points can exist and have different labels. These points will be detected as explanation pairs, and when the visual explanations are created, there will not be any bars because there is no difference between a sample and a counterfactual. A prominent example of this can be shown in Figure 8b, $CF-2$ and $CF-4$ are empty; to increase understandability, the

[2] https://archive.ics.uci.edu/ml/datasets/Caesarian+Section+Classification+Dataset

[3] https://archive.ics.uci.edu/ml/datasets/Algerian+Forest+Fires+Dataset++

audience should be informed. Using textual and visual explanations together helps in this situation.

Diversity is an essential requirement while detecting counterfactuals. The proposed method considers diverse sample-counterfactual pairs. As evidence of meeting the diversity requirement, in Figure 8b, $CF-1$ and $CF-3$ have the same distance values from the sample. However, the difference vectors are not the same because of counterfactuals located at different points, and our pair detection method detected both.

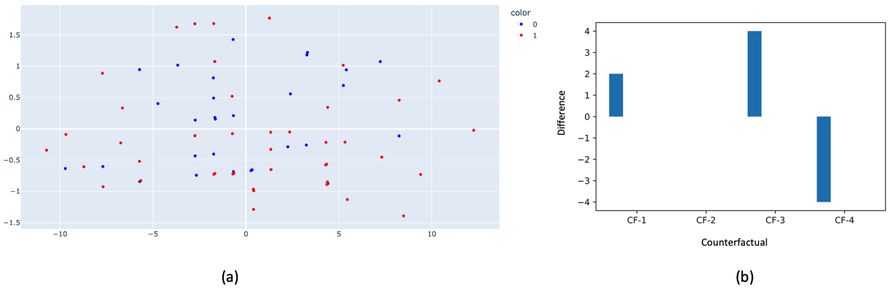


Fig. 7. (a) Distribution of Cesarean dataset in two dimension (PCA), **(b)** Visual explanation of a sample from Cesarean dataset

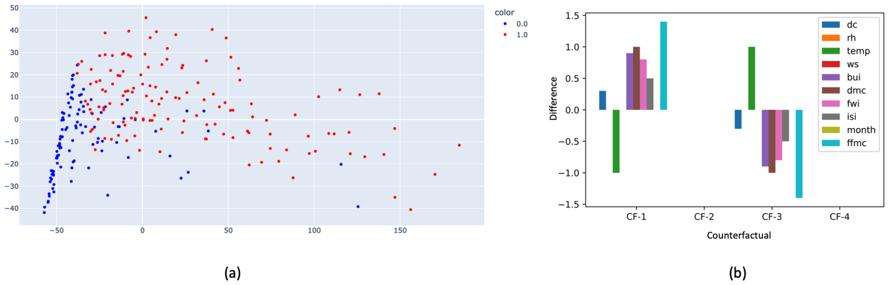


Fig. 8. (a) Distribution of Algerian Forest Fires dataset in two dimension (PCA), **(b)** Visual explanation of a sample from Algerian Forest Fires dataset

This research, however, is subject to several limitations. One of these limitations is the shortcomings introduced by the kNN algorithm. The problem of sensitivity to dataset scale and irrelevant features has been partially overcome by establishing a dynamic structure. Nevertheless, the shortcoming of dependence on data quality is open to improvement. Also, the proposed approach is applied to binary-classification models on tabular data. However, there is room to improve and generalize the presented method for future work, making it applicable to different data types and tasks.

6 Conclusion

This paper aimed to create a model-agnostic XCBR system and present qualitative and quantitative explanations when necessary. We proposed methods for creating explanation cases from counterfactuals that benefit from the human counterfactual reasoning mechanism and presenting explanations in qualitative and quantitative ways using text templates and bi-directional bar graphs. We showed that the explanation system can be applied to various datasets and models through the interpretability and flexibility of the methods. We also discussed how the nature of a dataset influences the explanation area and how the proposed method works on open datasets.

References

1. Aamodt, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI communications **7**(1), 39–59 (1994)
2. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)
3. Castelvecchi, D.: Can we open the black box of ai? Nature News **538**(7623), 20 (2016)
4. Che, Z., Purushotham, S., Khemani, R., Liu, Y.: Interpretable deep models for icu outcome prediction. In: AMIA annual symposium proceedings. vol. 2016, p. 371. American Medical Informatics Association (2016)
5. Darias, J.M., Dıaz-Agudo, B., Recio-Garcia, J.A.: A systematic review on model-agnostic xai libraries (2021)
6. Datta, A., Sen, S., Zick, Y.: Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In: 2016 IEEE symposium on security and privacy (SP). pp. 598–617. IEEE (2016)
7. Feely, C., Caulfield, B., Lawlor, A., Smyth, B.: A case-based reasoning approach to predicting and explaining running related injuries. In: International Conference on Case-Based Reasoning. pp. 79–93. Springer (2021)
8. Grath, R.M., Costabello, L., Van, C.L., Sweeney, P., Kamiab, F., Shen, Z., Lecue, F.: Interpretable credit application predictions with counterfactual explanations. arXiv preprint arXiv:1811.05245 (2018)
9. Huertos, A.A., Garbın, I.P., Dıaz-Agudo, B., Sánchez-Ruiz, A.A.: Explanations on smart stores (2021)
10. Karpathy, A., Johnson, J., Fei-Fei, L.: Visualizing and understanding recurrent networks. arXiv preprint arXiv:1506.02078 (2015)
11. Keane, M.T., Smyth, B.: Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai). In: International Conference on Case-Based Reasoning. pp. 163–178. Springer (2020)

12. Krishnan, S., Wu, E.: Palm: Machine learning explanations for iterative debugging. In: Proceedings of the 2Nd workshop on human-in-the-loop data analytics. pp. 1–6 (2017)
13. Lamy, J.B., Sekar, B., Guezennec, G., Bouaud, J., Séroussi, B.: Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. Artificial intelligence in medicine **94**, 42–53 (2019)
14. Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)
15. Marchese Robinson, R.L., Palczewska, A., Palczewski, J., Kidley, N.: Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets. Journal of Chemical Information and Modeling **57**(8), 1773–1792 (2017)
16. Nunes, I., Jannach, D.: A systematic review and taxonomy of explanations in decision support and recommender systems. User Modeling and User-Adapted Interaction **27**(3), 393–444 (2017)
17. Recio-García, J.A., Díaz-Agudo, B., Pino-Castilla, V.: Cbr-lime: a case-based reasoning approach to provide specific local interpretable model-agnostic explanations. In: International Conference on Case-Based Reasoning. pp. 179–194. Springer (2020)
18. Recio-García, J.A., Parejas-Llanovarced, H., Orozco-del Castillo, M.G., Brito-Borges, E.E.: A case-based approach for the selection of explanation algorithms in image classification. In: International Conference on Case-Based Reasoning. pp. 186–200. Springer (2021)
19. Ribeiro, M.T., Singh, S., Guestrin, C.: "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. pp. 1135–1144 (2016)
20. Rudin, C.: Please stop explaining black box models for high stakes decisions. Stat **1050**, 26 (2018)
21. Schank, R.C., Leake, D.B.: Creativity and learning in a case-based explainer. Artificial intelligence **40**(1-3), 353–385 (1989)
22. Schoenborn, J.M., Weber, R.O., Aha, D.W., Cassens, J., Althoff, K.D.: Explainable case-based reasoning: A survey. In: AAAI-21 Workshop Proceedings (2021)
23. Sokol, K., Hepburn, A., Santos-Rodriguez, R., Flach, P.: blimey: surrogate prediction explanations beyond lime. arXiv preprint arXiv:1910.13016 (2019)
24. Sørmo, F., Cassens, J., Aamodt, A.: Explanation in case-based reasoning–perspectives and goals. Artificial Intelligence Review **24**(2), 109–143 (2005)
25. Verma, D., Bach, K., Mork, P.J.: Modelling similarity for comparing physical activity profiles-a data-driven approach. In: International Conference on Case-Based Reasoning. pp. 415–430. Springer (2018)
26. Visani, G., Bagli, E., Chesani, F.: Optilime: Optimized lime explanations for diagnostic computer algorithms. arXiv preprint arXiv:2006.05714 (2020)
27. Weber, R., Shrestha, M., Johs, A.J.: Knowledge-based xai through cbr: There is more to explanations than models can tell. arXiv preprint arXiv:2108.10363 (2021)
28. Yang, W., Li, J., Xiong, C., Hoi, S.C.: Mace: An efficient model-agnostic framework for counterfactual explanation. arXiv preprint arXiv:2205.15540 (2022)
29. Zhang, Y., Wallace, B.: A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820 (2015)
30. Zhang, Y., Song, K., Sun, Y., Tan, S., Udell, M.: " why should you trust my explanation?" understanding uncertainty in lime explanations. arXiv preprint arXiv:1904.12991 (2019)