Application of Multi-Instance Counterfactual Explanation in Road Safety Analysis

André Artelt[1,2], Andreas Gregoriades[3]

[1]Bielefeld University, Germany
[2]University of Cyprus, Cyprus
[3]Cyprus University of Technology, Cyprus

Abstract

Road accidents cause millions of fatalities worldwide and impose a significant economic burden on societies. To address this problem, road safety researchers have mainly applied statistical or machine learning methods to predict accident occurrences and identify the causes of such crashes, but to the best of our knowledge, no work has explored how to optimally address the causes of such crashes to minimise accidents' severity. Recently, eXplainable AI (XAI) techniques have been applied in transportation to evaluate the effect of accidents' contributing factors. Limited work however delved into optimum ways to reduce accident severity by minimising changes needed to road network's infrastructure using XAI. In this work, we apply counterfactual explanations, a popular XAI technique, to road accident data, to identify optimal changes to infrastructure to improve road safety, by converting severe accidents into minor accidents. Traditionally, counterfactual explanations are used for single instances(accidents), which is not appropriate to this problem since the goal is to find actionable changes to the road infrastructure to convert as many severe accidents to non-severe accidents. Thus, our proposed methodology is based on multi-instance explanations. The proposed methodology is evaluated in a case study with real accident data from Cyprus.

Keywords

Road Safety, XAI, Counterfactual Explanations

1. Introduction

According to the World Health Organisation (WHO), approximately 1.4 million accidents occur each year worldwide leaving millions of people injured. Road accidents constitute the eighth leading cause of death worldwide and this number is likely to increase if this problem is not addressed effectively.

Predominantly research on road safety utilizes classical statistical techniques such as Logistic, Poisson, and Negative binomial regression [1, 2]. These methods undoubtedly provided insights; however, the fundamental characteristics of accident data often result in methodological limitations that cannot be accounted for. Recently, research has indicated that machine learning (ML) techniques outperformed conventional statistical methods by offering superior prediction and greater accuracy due to their ability to work with massive amounts of multidimensional and noisy data, while also being able to address generalizability, as reported by Iranitalab et al. (2017). With the increased availability of data from various sources such as Internet of Things devices, installed in the road network, connected vehicles, and naturalistic driving studies, machine learning is becoming a key methodology in transportation.

During road safety analysis, traffic accident data is used to develop models to predict and understand the causes of accidents through the identification of relationships among contributing factors and the outcome of the models. Contributing factors can be combinations of properties relating to the road infrastructure, driving behavior, and environmental conditions. These are considered independent variables and the outcome (accident occurrence or severity) is the dependent variable. Predicting an accident before it happens gives the chance to policymakers to take precautionary measures to minimise or prevent accidents from happening. In the machine learning domain, researchers use techniques such

ATT'24: Workshop Agents in Traffic and Transportation

aartelt@techfak.uni-bielefeld.de (A. Artelt); andreas.gregoriades@cut.ac.cy (A. Gregoriades)
https://andreartelt.github.io/ (A. Artelt)
0000-0002-2426-3126 (A. Artelt); 0000-0002-7422-1514 (A. Gregoriades)

as eXplainable AI (XAI) [3], and/or causality analysis to investigate the contributing factors that lead to accidents. The latter approaches use multivariate statistical models to evaluate the effects of the contributing factors. However, statistical relationships are not always causal and could be the result of chance, bias, confounding variables, or others. A cause can be an action or event that changes the outcome that would not have changed otherwise. A cause is therefore a necessary precondition for an event such as an accident. Causal relationships are usually characterised by regularity, meaning the same cause always produces the same result [4]. Causal relationships also are linked to the notion of counterfactual, which explains what would have happened if the cause was not present. Thus, to infer a causal relationship, it is always necessary to establish the counterfactual. In scientific analysis, the most popular approach to do so is to conduct a controlled experiment where treatment is applied at random, and the control group shows what would have happened if the action/treatment was not introduced (counterfactual). Road safety experiments are difficult to perform, with most studies being observational, thus the task of establishing the counterfactual is similar to controlling the confounding factors.

In transportation, one of the main approaches to improve safety is by addressing one or more causes or risk factors that are associated with accident occurrence. The traffic safety literature highlights different modifications to either the policy or road infrastructure to improve safety. For instance, infrastructural modifications could be, changes in horizontal curvature of roads, shoulder widths, the width of the median that separate lanes, etc. Such modifications, however, are usually introduced after analysing the problem, designing a solution, simulating it, and then applying it, without having any guarantees of the effect that these changes can make. Therefore, most of such projects fail to meet their goals. Additionally, such modifications are usually introduced without knowing the optimum degree of change (for example speed limit) to archive the desired effect. Optimisation has been applied in transportation to minimise the budget required to improve safety as reported in work by [5]. Limited work however addressed the problem of optimising infrastructural changes to reduce fatalities or accident severity.

Explainable AI (XAI) [3] is an approach aiming to explain black-box machine learning models and the reasons they come up with decisions through intuitive and human-understandable explanations. The need for explainability is not new since it addresses the question "why" a system behaves the way it does. The term XAI however has been recently coined by DARPA and it is used in a variety of domains where machine learning is applied. For instance, in transportation, two of the main XAI techniques used to extract knowledge from prediction models are Shapley Additive exPlanation (SHAP) and Lime [6]. These approaches, however, do not offer recommendations on how to achieve the desired results. Counterfactual explanations [7] ("counterfactuals" for short) on the other hand are designed for this purpose and thus became a popular technique for explaining black-box models. However, their application in the traffic safety domain is missing, making this work one of the first works that apply counterfactual explanations to road safety.

The goal of this work is the maximisation of road safety (reducing severe accidents) by minimising the infrastructural changes to a road network to improve safety. This is achieved by explaining using counterfactual explanations a machine learning model trained to predict accident severity using historical accident data of the specific road network. Counterfactuals, however, usually provide explanations for single instances (accidents) rather than a group of cases. Moreover, counterfactual explanations are applied on numerical variables (e.g. road width, speed limit, etc.), even though categorical variables (e.g. type or intersection) are a key type of features in different domains, including transportation. The approach proposed herein addresses these two problems by finding counterfactuals that satisfy multiple instances(accidents) simultaneously, characterised by both numerical and categorical features. Therefore, the method provides policymakers with recommendations that not only indicate the factors that contribute to the problem but also provide them with information regarding the degree of change to these factors to archive the desired effect. Therefore, the method can potentially minimise the costs of satisfying stakeholders' goals.

The remainder of this work is organised as follows: Section 2 elaborates on the background of counterfactual explanations, the related literature, and the motivation behind the proposed counterfactual methodology. The next section introduces the methodology (Section 3) and elaborates on its application in a road safety case study using accident data (Section 4). Finally, we summarize and discuss future directions (Section 5).

2. Counterfactual Explanations

A counterfactual explanation [7] ("counterfactual" for short) states how to change a given instance such that the output of the model for this instance changes in a specific way (towards a desired outcome). The popularity of counterfactual explanations comes from the fact that they are very similar to the way humans explain situations [8] and that they provide precise and actionable recommendations that can be directly applied in the real world [7].

To be useful in practice, a counterfactual must not only be feasible (i.e. valid) but also as simple as possible - e.g. not too many recommendations or big changes [7]. Considering these two aspects, the computation of a counterfactual $\vec{\delta}_{\text{cf}}$ for a given case $\vec{x}_{\text{orig}}$ can be formally phrased as an optimization problem that minimises the modifications to the attributes of $\vec{x}_{\text{orig}}$ so that the classifier $h(\cdot)$ changes its prediction to the desired output y_{cf} [7, 9]:

Definition 1 (Counterfactual Explanation)**.** *Assume a prediction function $h : \mathcal{X} \to \mathcal{Y}$ is given. Computing a counterfactual explanation $\vec{\delta}_{cf} \in \mathcal{X}$ for a given instance $\vec{x}_{orig} \in \mathcal{X}$ is phrased as the following optimization problem:*

$$\underset{\vec{\delta}_{cf} \in \mathcal{X}}{\arg\min}\ \ell\left(h(\vec{x}_{orig} \oplus \vec{\delta}_{cf}), y_{cf}\right) + C \cdot \theta(\vec{\delta}_{cf}) \tag{1}$$

where $\ell(\cdot)$ denotes a loss function that penalizes deviation of the output $h(\vec{x}_{orig} \oplus \vec{\delta}_{cf})$ from the requested output y_{cf}, $\theta(\cdot)$ implements the cost of $\vec{\delta}_{cf}$ – i.e. prefer “simple, cheap & easy to execute” explanations –, and $C > 0$ denotes the regularization strength.

In order to not make any assumptions on the data domain, we use the symbol $\oplus$ to denote the application/execution of the counterfactual $\vec{\delta}_{\text{cf}}$ to the original instance $\vec{x}_{\text{orig}}$. While in the case of real and integer numbers (e.g. $\mathcal{X} = \mathbb{R}^d$) this reduces to the translation (i.e. $(\vec{x}_{\text{cf}})_i = (\vec{x}_{\text{orig}})_i + (\vec{\delta}_{\text{cf}})_i$, in the case of categorical features it denotes a substitution – i.e. $(\vec{x}_{\text{cf}})_i = (\vec{\delta}_{\text{cf}})_i$.

Note that Definition 1 constitutes a non-causal approach - i.e. no causal model of the world is included. There exists an entirely different line of research on counterfactuals utilizing structural causal models to incorporate causal knowledge [10]. However, in practice, such causal models are usually not known and have to be estimated from data or carefully specified with the help of domain experts. However, experts are not easily available, thus in this work, we only consider a non-causal approach.

There exists a wide variety of methods for computing counterfactual explanations – i.e. solving the optimization problem. The model-agnostic methods utilize gradient-based optimization methods that can be applied on any black box model [11, 12, 13], while the model-specific methods use details of the specific model's architecture to calculate recommendations [9]. An important limitation of counterfactual explanations is the fact that they are missing uniqueness. This means that there usually exists more than one possible explanation which raises the question of which one to pick - usually, the "simplest" explanation is picked.

2.1. Multi-instance Counterfactual Explanations

In many real-world applications of counterfactuals, one is interested in gaining knowledge about a set or group of instances instead of a single instance, which is the usual case. For instance, in any organization, the human resource department is interested in minimizing employees' attrition, since this causes several problems. To understand the cause of attrition and deploy appropriate (global) countermeasures, an organization needs to consider all cases (employees intending to leave) and find a change (maybe increase employee salary) that will guarantee the retention of as many employees as possible [14, 15].

To find a common feasible recommendation for a group of instances (i.e. employees), counterfactual explanations have been recently extended towards multi-instance counterfactual explanations (also called group-counterfactuals) [14, 15, 16, 17]. A multi-instance counterfactual states what to change on a group level (e.g. increasing some specific attribute for all instances in the group by the same amount) such that the outcome for this group of instances changes simultaneously in some desired way. Like counterfactual explanations, the computation of a multi-instance counterfactual explanation $\vec{\delta}_{\text{cf}}$ for a set of cases $\mathcal{D}$ can be formalized as an optimization problem:

Definition 2 (Multi-instance Counterfactual Explanation)**.** *Let $h : \mathcal{X} \to \mathcal{Y}$ denote a prediction function, and let $\mathcal{D}$ be a set of labeled instances with the same prediction $y \in \mathcal{Y}$ under $h(\cdot)$ – i.e. $h(\vec{x}_i) = y \quad \forall \vec{x}_i \in \mathcal{D}$. We are looking for a single change $\vec{\delta}_{cf} \in \mathbb{R}^d$ that, if applied to the instances in $\mathcal{D}$, changes as many of their predictions to some requested output $y_{cf} \in \mathcal{Y}$.*

We call all pareto-optimal solutions $\vec{\delta}_{cf}$ to the following multi-objective optimization problem multi-instance counterfactuals:

$$\min_{\vec{\delta}_{cf} \in \mathcal{X}} \; \left(\theta(\vec{\delta}_{cf}) \;,\; \ell(h(\vec{x}_1 \oplus \vec{\delta}_{cf}), y_{cf}), \dots, \ell(h(\vec{x}_{|\mathcal{D}|} \oplus \vec{\delta}_{cf}), y_{cf}) \right) \tag{2}$$

where $\theta(\cdot)$ denotes the cost of the counterfactual, and $\ell(\cdot)$ denotes a suitable loss function penalizing deviations from the requested outcome y_{cf} – suitable loss functions might be the mean-squared error or cross-entropy loss, while the cost $\theta(\cdot)$ might be implemented by a p-norm.

Note that in contrast to normal counterfactuals (Definition 1), multi-instance counterfactuals (Definition 2) require multiple constraints – one constraint for each case in the set $\mathcal{D}$. Similar to normal counterfactuals, one could merge all constraints into a single objective which would then enable the use of general gradient-based or black-box methods for solving the optimization problem.

A major challenge in the computation of multi-instance counterfactuals is that there might be no feasible solution – i.e. it might be impossible to find a single change $\vec{\delta}_{\text{cf}}$ that is feasible for all instances in $\mathcal{D}$. Therefore, one might either want to relax the constraints and compute a change $\vec{\delta}_{\text{cf}}$ that is feasible for as many as possible instances in $\vec{\delta}_{\text{cf}}$, or find a grouping/clustering of $\mathcal{D}$ such that there exists a change $\vec{\delta}_{\text{cf}}$ for each of the groups that is feasible for all instances within this sub-group – here the additional challenges of finding such sub-groups arise. We denote the percentage of instances for which the explanation is feasible as accuracy.

Unlike counterfactual explanations (Definition 1), multi-instance counterfactuals (Definition 2) is a novel concept and consequently, existing work on this is rather limited. For instance, the work by [15] applies multi-instance counterfactuals to the employee attrition problem but only considers a linear classifier. One of the latest works such as [14] proposes a counterfactual explanation tree, which assigns counterfactual explanations to a learned decision tree that assigns samples to groups – this method is only applicable if an automatic clustering into sub-groups is needed. Besides that, most existing work for multi-instance counterfactuals can be interpreted as summarizing or aggregating local counterfactuals. In the work of [16], multi-instance counterfactuals are generated by first computing individual counterfactuals and then selecting those that maximizes the cover of a given set of instances. Similarly, [18] tries to obtain a global explanation by simply aggregating local explanations. However, these methods cannot guarantee high accuracy because they consider all instances separately.

3. Multi-instance Counterfactuals for Improving Road Safety

Herein we propose a methodology that utilizes multi-instance counterfactuals to analyze road accident records and compute suggestions on how to improve road safety – the only assumption we make is that the collected road accident records are labeled with accident severity. The proposed methodology consists of three main steps as illustrated in Figure 1:

1. Train a binary classifier using road accident data to predict the severity of accidents.

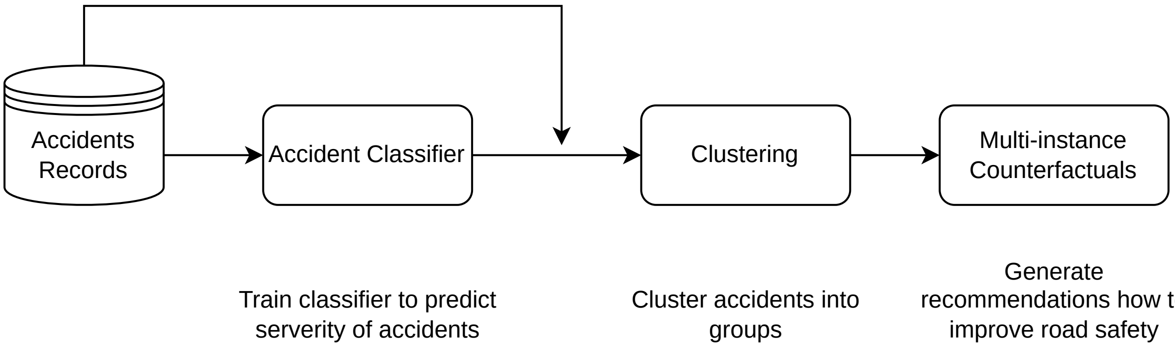
Accidents Records
Accident Classifier
Clustering
Multi-instance Counterfactuals
Train classifier to predict serverity of accidents
Cluster accidents into groups
Generate recommendations how to improve road safety

Figure 1: Proposed methodology for reducing accident severity.

2. (Optional:) Cluster all severe accidents into groups of similar characteristics.
3. Compute a multi-instance counterfactual explanation for each group - i.e. what to change to turn a severe accident into a non-severe (slight).

We interpret the computed multi-instance counterfactuals as potential suggestions on how road safety can be improved. Note that the grouping in the second step of the methodology is optional because one could simply consider all severe accidents as one large group. However, depending on the use case a more refined grouping might allow the computation of suggestions for more specific questions: For instance, depending on the nature of the collected data, it might be possible to group road accidents based on their location (e.g. rural area vs. urban) or timing of the accident (day, night). In this way, one could generate more specific recommendations (e.g. area/time specific) for certain types of accidents.

4. Empirical Case Study

We empirically evaluate our proposed methodology (see Section 3) in a case study using real-world accident data.

4.1. Data

The original data consists of accidents that occurred in Nicosia, Cyprus during 2007 - 2012. After merging and cleaning the data, 9829 cases were left each consisting of 58 attributes. To ensure that the recommendations to be made are actionable, we only consider 12 attributes that could be changed in practice, in contrast to attributes that cannot be changed, such as the age of drivers. Examples of mutable attributes are the road width, speed limit, traffic control, pedestrians crossing, the existence of a median in the road, etc. The accident type attribute is converted into a binary variable by combining fatal with severe accidents and labelling them as severe, and slight(minor) with property damage accidents and labelling these as non-severe accidents.

4.2. Implementation

Because the data mainly consists of numerical (e.g. integer) and categorical attributes, most standard machine learning models are not suitable. Thus, in this work, we decided to use a tree-based classifier (i.e. Xgboost [19]) that can handle such categorical attributes without transforming them. The classifier is trained to predict the severity of a given accident using different hyperparameters that have been tuned to improve the model's performance such as depth of trees, number of estimators, learning rate, and scale-pos-weight to address data imbalance. The classifier achieves an average F1-score of approx. 80% on the test data – i.e. we split the data into train (70%) and test data (30%).

We consider three groups of accidents (among the severe accidents only) in our experiments: All Severe accidents; Severe accidents in rural areas; and Severe accidents in urban areas. These groupings can be changed depending on the case study. For instance, a transportation engineer might use only accidents at a specific location in the road network (black spot). Table 1 shows different multi-instance recommendations for each of these groups. The column multi-instance counterfactuals show the changes that need to be made so that the severe accidents are converted to non-severe. The variables in parenthesis refer to the property of the infrastructure and the number next to each one defines the type and degree of change. Therefore for the case (Traffic-control, 3) which is a categorical variable, it denotes that the type of traffic control needs to change to traffic-light (this refers to the number 3) for all of these accidents so that these change to non-severe. Similarly for continuous variables such as 'speed' the recommended change is defined as a number with a sign indicating positive or negative change. In the case of speed limit, most recommendations indicate a reduction in speed limit which abides to road safety literature [20, 21].

Because in this case study the data also include non-continuous attributes, existing methods for computing multi-instance counterfactuals cannot be applied out of the box. To address this issue, we built an evolutionary algorithm for computing multi-instance counterfactuals that are guaranteed to adhere to the specific attribute ranges and yield feasible solutions. The evolutionary (i.e. genetic) algorithm treats all variables as discrete and iteratively mutates and merges (cross-over) candidate solutions until convergence. To guarantee the feasibility of the final multi-instance counterfactual $\vec{\delta}_{\text{cf}}$, we construct the set of feasible changes for each feature of numerical variables as follows – assuming non-negativity which can be achieved by adding a constant:

$$l_i = \alpha_i - \min_j\{(\vec{x}_j)_i\} \text{ and } u_i = \beta_i - \max_j\{(\vec{x}_j)_i\} \tag{3}$$

where α_i and β_i denote the maximum and minimum feasible value of the i-th feature, and the final set of feasible changes is then given as $[l_i, u_i]$. These sets are used when computing mutations in our evolutionary algorithm of existing individuals during the optimization. As an objective, we use the zero-norm (i.e. setting p=0 in the p-norm) - by this, we aim to minimize the number of suggested changes. Together with the constraints, this yields the following optimization problem:

$$\arg\min_{\vec{\delta}_{\text{cf}}} \left(\|\vec{\delta}_{\text{cf}}\|_0 \;,\; \ell(h(\vec{x}_1 + \vec{\delta}_{\text{cf}}), y_{\text{cf}}), \ldots, \ell(h(\vec{x}_{|\mathcal{D}|} + \vec{\delta}_{\text{cf}}), y_{\text{cf}})\right) \tag{4}$$

The final algorithm is given in Algorithm 1.

4.3. Results

To generate reliable results and avoid recommendations that are based on a bad train-test data split, the experiments are conducted using a 3-fold cross-validation. Thus, we generate three multi-instance explanations for each accident groups (i.e. all severe accidents, severe accidents in urban areas, and rural areas) as shown in Table 1.

Besides listing the explanations in Table 1 (tuple of attribute and recommended change), we also added a column for the accuracy of the explanation which refers to the percentage of instances for which the explanation correctly changes their prediction (Table 1). From the results, we observe that the generated multi-instance counterfactuals are almost always feasible (the accuracy is close to one). This demonstrates that our implemented evolutionary algorithm is able to compute feasible solutions with high reliability.

The results from the multi-instance counterfactuals show the required changes to infrastructure to convert severe accidents into non-severe. The variables that are considered important and thus are part of the recommended changes are the Traffic Control, which takes the values traffic signals, roundabout, police, stop sign and none; Road Width, stating the width in meters; Speed Limit in Km/h; Pedestrian Crossing type, that takes values: zebra crossing, pedestrian traffic signal crossing, and pelican crossing; Constriction variable that takes the values: one-way, two-way bridge, none; and Brake

Algorithm 1 Multi-instance Counterfactuals for Road-safety Analysis

```
Input: Set of labeled (severe vs. non-severe) accidents $\mathcal{D} = \{(\vec{x}_i, y_i)\}$, hyper-parameter $N$ denoting the number of evolutionary steps
Output: Recommendations (i.e. multi-instance counterfactual explanation) $\vec{\delta}_{\text{cf}}$
Split data $\mathcal{D}$ into train $\mathcal{D}_{\text{train}}$ and test $\mathcal{D}_{\text{test}}$ set          ▷ Repeat in k-fold cross validation
Fit XGBoost classifier $h(\cdot)$ to $\mathcal{D}_{\text{train}}$
Consider all accidents $\vec{x}_i \in \mathcal{D}_{\text{test}}$ with $h(\vec{x}_i) =$ "severe" – create $\mathcal{D}_{\text{severe}}$     ▷ Severe accidents only
(Group accidents and pick group of interest)          ▷ Optional
Compute feature bounds $\{[l_i, u_i]\}$ Eq. (3) on $\mathcal{D}_{\text{severe}}$
                    ▷ Evolutionary algorithm for computing a multi-instance counterfactual
$\{\vec{\delta}_{\text{cf}j}\} = \text{random\_init}(\{[l_i, u_i]\})$     ▷ Random initial population of solutions – respect feature bounds!
for N iterations do
    Evaluate fitness of each $\vec{\delta}_{\text{cf}j}$ using Eq. (4)
    Select best $\vec{\delta}_{\text{cf}j}$ for next generation
    Apply random mutations considering feature bounds $\{[l_i, u_i]\}$
    Apply cross-over to create next generation of solutions $\{\vec{\delta}_{\text{cf}j}\}$
end for
                    ▷ Select best solution as the final recommendation $\vec{\delta}_{\text{cf}}$
$\vec{\delta}_{\text{cf}} = \underset{\{\vec{\delta}_{\text{cf}j}\}}{\arg\min} \Big( \|\vec{\delta}_{\text{cf}}\|_0 \;,\; \ell(h(\vec{x}_1 + \vec{\delta}_{\text{cf}}), y_{\text{cf}}), \dots, \ell(h(\vec{x}_{|\mathcal{D}_{\text{severe}}|} + \vec{\delta}_{\text{cf}}), y_{\text{cf}}) \Big)$
```

Grouping	Accuracy ↑	Multi-instance counterfactual
All	1.0	(’TRAFFIC_CONTROL’, ’3’), (’SPEED_LIMIT’, ’-83’), (’PEDESTRIAN_CROSSING’, ’2’)
All	0.968	(’TRAFFIC_CONTROL’, ’3’), (’CONJUNCTION_TYPE’, ’1’), (’SPEED_LIMIT’, ’-15’), (’PEDESTRIAN_CROSSING’, ’2’)
All	1.0	(’ROAD_WIDTH’, ’+60’), (’BREAK_LANE_WIDTH’, ’+8’), (’SPEED_LIMIT’, ’-23’), (’PEDESTRIAN_CROSSING’, ’1’)
Urban	1.0	(’TRAFFIC_CONTROL’, ’1’), (’SPEED_LIMIT’, ’-6’), (’PEDESTRIAN_CROSSING’, ’2’)
Urban	1.0	(’TRAFFIC_CONTROL’, ’2’), (’ROAD_WIDTH’, ’+41’), (’BREAK_LANE_WIDTH’, ’+11’), (’ROAD_DESCR’, ’1’)
Urban	0.96	(’TRAFFIC_CONTROL’, ’2’), (’SPEED_LIMIT’, ’-39’), (’PEDESTRIAN_CROSSING’, ’2’), (’LIGHTING’, ’1’)
Rural	1.0	(’TRAFFIC_CONTROL’, ’1’), (’SPEED_LIMIT’, ’-78’), (’PEDESTRIAN_CROSSING’, ’2’)
Rural	1.0	(’TRAFFIC_CONTROL’, ’3’), (’CONSTRUCTION’, ’2’), (’SPEED_LIMIT’, ’-90’), (’BUS_STOP’, ’-1’), (’LIGHTING’, ’1’)
Rural	1.0	(’TRAFFIC_CONTROL’, ’2’), (’ROAD_WIDTH’, ’+48’), (’SPEED_LIMIT’, ’-18’)

Table 1
Multi-instance counterfactual explanations from the conducted case study.

Lane, that describes the width of the footway/shoulder in meters. The recommended changes mainly highlight the importance of traffic control type and speed limit, factors that are known in the literature to affect road safety [20, 21]. Additional changes include an increase in road width and break lane width, recommendations which are also consistent with the literature [22].

To validate our approach we used SHAP (SHapley Additive exPlanations) [23], an XAI model-agnostic method, to identify which features are influencing mostly accident type. The SHAP summary plot in Figure 2 shows that the road width, break lane width, and traffic control are key features along with

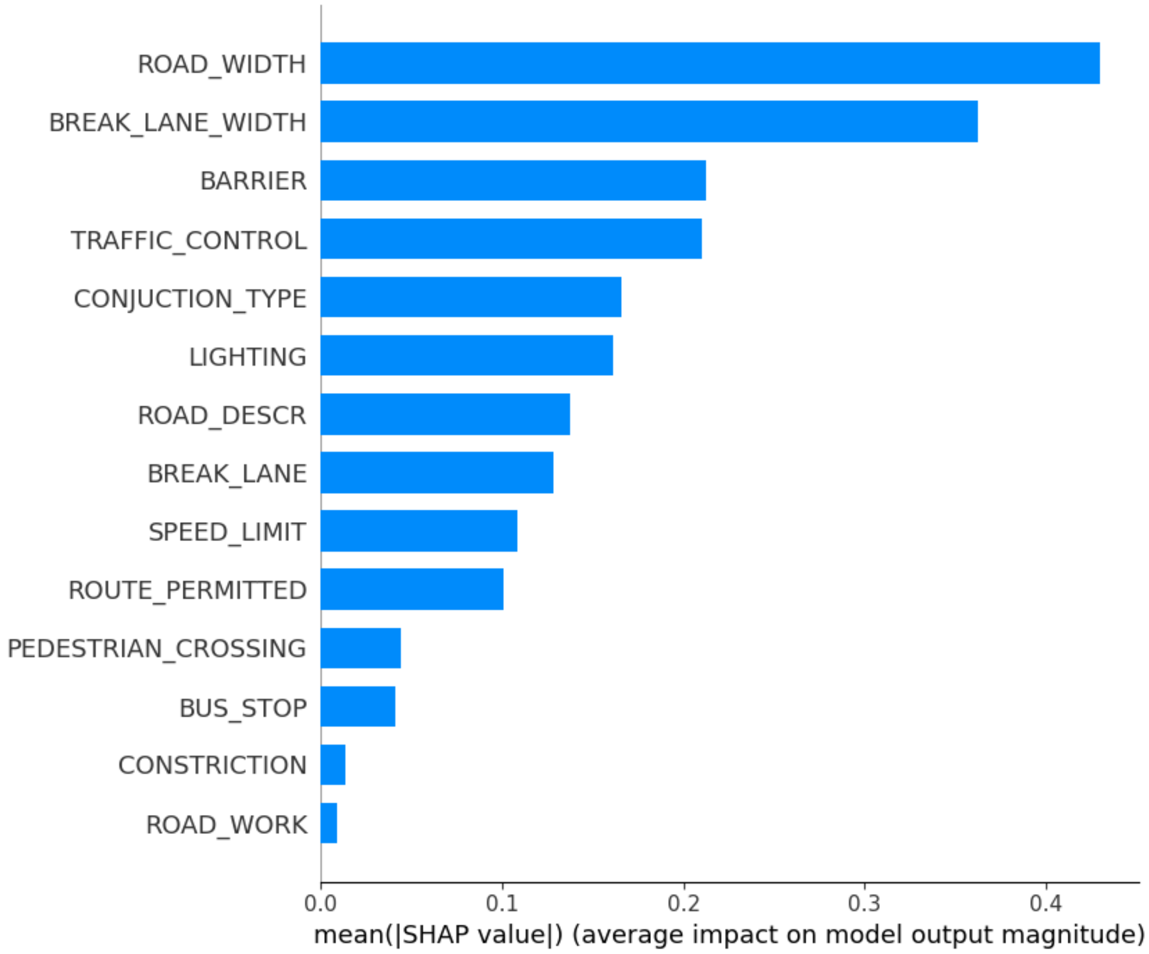

Figure 2: SHAP results show similar features contributing to accident type.

conjunction type and speed limit. These results verify that our method uses these key features to make counterfactual recommendations. The limitation of SHAP however is that it does not indicate which combinations of features must change and by how much so that the accidents are converted from severe to non-severe.

5. Summary & Future Research

In this work, we proposed a methodology for analyzing road accident data and using XAI to generate suggestions on how to improve road safety. Our methodology requires training a classifier to predict the severity of accidents (severe, non-severe) and then recommending changes to the input of a set of severe accidents using multi-instance counterfactual explanations, so that the predictions of the model change to non-severe accidents. The approach enable us to compute a global recommendation on how to reduce the severity of accidents by converting all severe accidents into non-severe using minimum alterations to chargeable features (referring to infrastructure), and by doing this improve road safety. We also conducted an empirical case study on real-world data, which showed that our proposed methodology computes reasonable recommendations that abide with the literature.

Despite the promising results, some aspects need further investigation:

- We observed that the suggested changes are often large (for example the speed limit), and it is not clear how plausible those changes would be in practice. We think this is mainly due to the fact that the designed evolutionary algorithm does not have any distributional knowledge about the data - i.e. which variable values or combinations are more often observed in the real world. Such distributional information might be used as part of the objective function, thus, automatically punishing large changes that would be costly to implement. Currently, we are working on an extension where the evolutionary algorithm is given distributional information which it utilizes when generating random mutations, new individuals, and cross-overs. By this, we hope to improve the quality of the recommended infrastructural changes significantly.

- In the presented case study, we either considered all severe accidents as one large group or manually split them into two sub-groups based on a location attribute from the dataset. While the computation of multi-instance counterfactuals worked well for both groups (large group, and two smaller sub-groups), it remains unclear if other interesting or beneficial (for the generated suggestions) groupings exist. An automatic clustering might not only reveal interesting clusters within the accident data set but also give rise to better and more specific suggestions on how to improve road safety. Currently, we are investigating how to cluster cases into groups such that the resulting multi-instance counterfactuals are as simple as possible.

Acknowledgments

This research was supported by the Ministry of Culture and Science NRW (Germany) as part of the Lamarr Fellow Network. This publication reflects the views of the authors only.

References

[1] A. Kassu, M. Hasan, Factors associated with traffic crashes on urban freeways, Transportation Engineering 2 (2020) 100014. URL: https://www.sciencedirect.com/science/article/pii/S2666691X20300154. doi:https://doi.org/10.1016/j.treng.2020.100014.

[2] J. S. Madushani, R. K. Sandamal, D. Meddage, H. Pasindu, P. A. Gomes, Evaluating expressway traffic crash severity by using logistic regression and explainable & supervised machine learning classifiers, Transportation Engineering 13 (2023) 100190. URL: https://www.sciencedirect.com/science/article/pii/S2666691X23000301. doi:https://doi.org/10.1016/j.treng.2023.100190.

[3] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (xai), IEEE access 6 (2018) 52138–52160.

[4] R. Elvik, Assessing causality in multivariate accident models, Accident Analysis & Prevention 43 (2011) 253–264. URL: https://www.sciencedirect.com/science/article/pii/S0001457510002514. doi:https://doi.org/10.1016/j.aap.2010.08.018.

[5] C. B. Byaruhanga, H. Evdorides, A budget optimisation model for road safety infrastructure countermeasures, Cogent Engineering 9 (2022) 2129363. doi:10.1080/23311916.2022.2129363.

[6] M. Gregurić, F. Vrbanić, E. Ivanjko, Towards the spatial analysis of motorway safety in the connected environment by using explainable deep learning, Knowledge-based systems 269 (2023) 110523.

[7] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.

[8] R. M. J. Byrne, Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 6276–6282. URL: https://doi.org/10.24963/ijcai.2019/876. doi:10.24963/ijcai.2019/876.

[9] A. Artelt, B. Hammer, On the computation of counterfactual explanations–a survey, arXiv preprint arXiv:1911.07749 (2019).

[10] A.-H. Karimi, G. Barthe, B. Balle, I. Valera, Model-agnostic counterfactual explanations for consequential decisions, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 895–905.

[11] S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson, C. Shah, Counterfactual explanations and algorithmic recourses for machine learning: A review, arXiv preprint arXiv:2010.10596 (2020).

[12] I. Stepin, J. M. Alonso, A. Catala, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, IEEE Access 9 (2021) 11974–12001.

[13] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery (2022) 1–55.

[14] K. Kanamori, T. Takagi, K. Kobayashi, Y. Ike, Counterfactual explanation trees: Transparent and consistent actionable recourse with decision trees, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 1846–1870.

[15] A. G. Andre Artelt, "how to make them stay?": Diverse counterfactual explanations of employee attrition, in: Proceedings of the 25th International Conference on Enterprise Information Systems, ICEIS 2023, Volume 1, Prague, Czech Republic, SCITEPRESS, 2023, pp. 532–538. URL: https://doi.org/10.5220/0011961300003467. doi:10.5220/0011961300003467.

[16] G. Warren, M. T. Keane, C. Gueret, E. Delaney, Explaining groups of instances counterfactually for xai: A use case, algorithm and user study for group-counterfactuals, arXiv preprint arXiv:2303.09297 (2023).

[17] Z. Huang, M. Kosan, S. Medya, S. Ranu, A. Singh, Global counterfactual explainer for graph neural networks, in: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, 2023, pp. 141–149.

[18] D. Ley, S. Mishra, D. Magazzeni, Global counterfactual explanations are reliable or efficient, but not both (2023).

[19] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[20] A. Vadeby, Åsa Forsman, Traffic safety effects of new speed limits in sweden, Accident Analysis & Prevention 114 (2018) 34–39. URL: https://www.sciencedirect.com/science/article/pii/S0001457517300532. doi:https://doi.org/10.1016/j.aap.2017.02.003, road Safety on Five Continents 2016 - Conference in Rio de Janeiro, Brazil.

[21] M. Papageorgiou, C. Diakaki, V. Dinopoulou, A. Kotsialos, Y. Wang, Review of road traffic control strategies, Proceedings of the IEEE 91 (2003) 2043–2067. doi:10.1109/JPROC.2003.819610.

[22] P. Pokorny, J. K. Jensen, F. Gross, K. Pitera, Safety effects of traffic lane and shoulder widths on two-lane undivided rural roads: A matched case-control study from norway, Accident Analysis & Prevention 144 (2020) 105614. URL: https://www.sciencedirect.com/science/article/pii/S0001457520303006. doi:https://doi.org/10.1016/j.aap.2020.105614.

[23] A. B. Owen, C. Prieur, On shapley value for measuring importance of dependent inputs, SIAM/ASA Journal on Uncertainty Quantification 5 (2017) 986–1002. doi:10.1137/16M1097717.