IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 4, 2022, accepted May 29, 2022, date of publication June 3, 2022, date of current version June 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3180026

Counterfactual Building and Evaluation via eXplainable Support Vector Data Description

ALBERTO CARLEVARO[1,2], MARTA LENATTI[2], ALESSIA PAGLIALONGA[2], AND MAURIZIO MONGELLI[2], (Member, IEEE)

[1]Department of Electrical, Electronics and Telecommunications Engineering and Naval Architecture (DITEN), University of Genoa, 16145 Genoa, Italy
[2]Institute of Electronics, Information Engineering and Telecommunications (IEIIT), National Research Council of Italy (CNR), 10129 Turin, Italy

Corresponding author: Alberto Carlevaro (alberto.carlevaro@edu.unige.it)

ABSTRACT Increasingly in recent times, the mere prediction of a machine learning algorithm is considered insufficient to gain complete control over the event being predicted. A machine learning algorithm should be considered reliable in the way it allows to extract more knowledge and information than just having a prediction at hand. In this perspective, the counterfactual theory plays a central role. By definition, a counterfactual is the smallest variation of the input such that it changes the predicted behaviour. The paper addresses counterfactuals through Support Vector Data Description (SVDD), empowered by explainability and metric for assessing the counterfactual quality. After showing the specific case in which an analytical solution may be found (under Euclidean distance and linear kernel), an optimisation problem is posed for any type of distances and kernels. The vehicle platooning application is the use case considered to demonstrate how the outlined methodology may offer support to safety-critical applications as well as how explanation may shed new light into the control of the system at hand.

INDEX TERMS Counterfactuals, support vector data description, eXplainable machine learning.

I. INTRODUCTION

A. BACKGROUND

Counterfactual explanations (CEs), a concept borrowed from philosophy of language and logic, has been first declined in the context of machine learning by Wachter *et al.* [1] as the minimal change that is required in the input features of a certain observation in order for the prediction of that observation to fall into the opposite class, in a binary classification problem. Specifically, a change of a certain delta in the features describing the observation $\boldsymbol{x}$, belonging to class $\mathbf{C}$, leads to the generation of an observation $\boldsymbol{x}'$ (i.e., the counterfactual of $\boldsymbol{x}$) that will be classified as belonging to class $\mathbf{C}'$. These kind of local explanations are assuming a certain importance, especially in machine learning models dealing with images [2], as they allow to add a certain degree of interpretability to the underlying behavior of complex models like neural networks, in line with the demand of the European General Data Protection Regulation (GDPR)[1] for greater transparency when handling decisions made by a model.

Different approaches have been recently proposed to produce realistic and feasible counterfactuals to provide local explanations for automated decision making processes. Table 1 provides an overview of related literature with regards to the methods for CEs generation, the use cases, the validation approach and open issues. For example, White *et al.* [3] determined counterfactuals by applying minimum perturbations for each feature separately and use them to generate local regression models, then evaluating the fidelity of these regressions, in five different case studies. Poyiadzi *et al.* [4], instead, proposed a method for generating CEs by considering a trade-off between the length of the path from the point to its corresponding counterfactual and the data density along this path. Finally, Mochaourab *et al.* [7] considered the design of robust CEs for privacy preserving mechanisms based on binary Support Vector Machines (SVM), by applying the bisection method between two points belonging to different classes and evaluating the trade off between accuracy, privacy and explainability.

CEs are a rather versatile solution that can be applied to different contexts, with various purposes. For example they

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

[1]https://gdpr.eu/tag/gdpr/

TABLE 1. Overview of related literature.

Authors	Method for generation of CEs	Use cases	Validation	Open issues
White et al [3]	CLEAR: minimization of the fidelity error, obtained by iteratively comparing progressive b-perturbations of each single feature with estimates of b-perturbations calculated using a local regression equation built around the initial point.	Numerical: Pima Indians Diabetes, Iris, Default of Credit Card Clients, Adult and Breast Cancer Wisconsin data sets	Fidelity of regression, against LIME	CEs quality depends on the neighbourhood dataset used for step-wise regression
Poyiadzi et al [4]	FACE: minimization of the f-distance describing the trade-off between path length and data density along the path, through the Shortest Path First Algorithm applied to a graph constructed over data points by using KDE, KNN or ε-graph.	Numerical: synthetic data; Images: MNIST data set	Comparison with CEs generated with a baseline method [1]	Limited validation of generated CEs
Van Looveren et al [5]	Addition of a prototype loss term in the objective function, to guide and fasten the search process. Encoders or K-d trees may be used to define class prototypes.	Numerical: Breast Cancer Wisconsin data set; Images: MNIST	Quantitative and visual interpretability, sparsity, and speed.	The interpretability measures depend on AE trained on the original and counterfactual classes, hence are associated with prediction uncertainty.
Nemirovsky et al [6]	CounteRGAN: fixed target classifier (e.g., NN) coupled with a RGAN trained to produce residuals that are added to the input, to produce CEs.	Numerical: PIMA Indian Diabetes, COMPAS recidivism data set; Images: MNIST data set	Prediction gain, realism, latency and actionability against others related works (e.g., [5])	The application of this method to large images data sets would require need more complex architectures and finer hyper-parameters tuning.
Mochaourab et al [7]	Bisection method: starting from two prototypes with opposite class, according to Privacy preserving SVM with RBF kernel.	Numerical: Breast Cancer Wisconsin data set	Trade-off between accuracy, privacy and explainability	Privacy requirement degrades the quality of generated CEs. CEs assessment is based on SVM prediction confidence.
Dhurandhar et al [8]	CEM: optimization of the perturbation variable using the fast iterative shrinkage-thresholding algorithm (FISTA) coupled with the use of a CAE to evaluate the distance from the data manifold.	Numerical: procurement fraud; Images: MNIST, Autism Brain Imaging Data Exchange	Comparison with LIME and LRP	Fully deep learning-based
Albini et al [9]	Mapping the variables that influence the assignment of observations to classifications in Bayesian Network classifiers (single or multi-label, binary or multidimensional)	Numerical: Voting records, Parole Violation, and Child Bayesian Network	Comparison with a baseline method ([10])	Model specific method for Bayesian Network Classifiers. Numerical features are mapped into categorical features

can be generated in order to understand what are the changes in the characteristics of a medical image that lead to a certain diagnosis of pathology (e.g., [8]and [5]). Another possible use of counterfactuals recently proposed in the literature [11] concerns their application to generate actionable feedback (e.g., realistic changes in expected salary or increase in work experience word count) to candidates in a hiring marketplace in order to improve their profile.

Whether an observation belongs to a certain class may depend on two categories of features: *controllable features*, which can be manipulated through internal/external intervention (e.g., therapies or lifestyle changes in clinical classification problems or control algorithms in systems modelling and control problems) and *non-controllable features*, which by their nature are not manipulable (e.g., the age of a subject in health prediction algorithms). Therefore, the search for realistic counterfactuals should be performed by perturbing only controllable variables. To our knowledge, only a limited number of attempts to force the generated CEs to have no change in terms of non-controllable characteristics have been carried out. For example, Nemirovsy et al [11] developed a method to produce counterfactuals able to provide actionable feedbacks in real-time using Generative Adversarial Networks (GANs). However, in that case, feature immutability was imposed after the application of the counterfactuals search algorithm by setting the values of non-controllable features to the original values rather than to the values suggested by the counterfactuals search algorithm. By contrast, in this study, the search for counterfactuals is guided by directly perturbing only controllable features.

Previous related works validated the proposed CEs with respect to explanations obtained with other local explainability methods, like Local Interpretable Model-agnostic Explanations (LIME) or Layer-Wise Relevance Propagation (LRP) [3], [8] or with respect to other state-of-the-art method for generation of CEs [4], [9], [11]. Often, the validation measure relies on verifying that the CE is correctly associated with its target outcome, based on the prediction of a classifier. However, this measure is characterized by a degree of uncertainty, since it is not guaranteed that the real class matches the predicted class. To our knowledge, none of the approaches presented in the literature is supported by a validation of the generated CEs with computational simulations, capable of verifying that the CE belongs to a certain class, and rule-based models that explain the reason for this belonging.

The aim of this paper is to introduce a novel methodology for counterfactual generation and validation. The counterfactuals generation method uses regions defined by Two Class-Support Vector Data Descriptors (TC-SVDDs) and is here introduced in both analytical (II-A) and numerical (II-B)

form. The validation method combines computational simulations and eXplainable AI (XAI), specifically in the form of rule-based classification of counterfactuals. An example of application to collision detection in vehicle platooning is introduced to demonstrate the method (III).

B. CONTRIBUTION

The main contributions of this paper include:

- the introduction of constrained counterfactuals, whose search is based on perturbations of controllable features;
- the analytical and numerical formulations for the generation of counterfactual that include:
  - the introduction of the minimum distance problem between SVDD classes and its analytical solution in the linear case;
  - an SVDD-based counterfactuals generation algorithm which is simpler than deep learning-based solutions;
  - the assessment of Counterfactual Distance (i.e., whether it is over- or under- dimensioned);
- the use of an XAI global method to extract knowledge from counterfactuals;
- the application of the newly introduced method to an example of cyberphysical system and the validation by means of simulations, together with the identification of the rules that characterise the decisions.

C. STRUCTURE OF THE PAPER

The paper is structured as follows: section II introduces the concept of counterfactual SVDD, its analytical (II-A) and numerical solution (II-B), and the natively explainable method used to define the rules that characterize both factuals and counterfactuals (II-C), section III describes an example of application of counterfactual SVDD to a case of truck platooning, and section IV discusses our findings with respect to related literature.

II. METHODOLOGY

Suppose we have a dataset $\mathcal{X} \times \mathcal{Y} \subset R^N \times \{-1, +1\}, N \geq 2$, consisting of a subset of controllable features $\mathbf{u}$ and a subset of non-controllable features $\mathbf{z}$, so that an observation $\boldsymbol{x} \in \mathcal{X}$ can be described as

$$\boldsymbol{x} = \left(u^1, u^2, \ldots, u^n, z^1, z^2, \ldots, z^m\right) \in R^{n+m=N}$$

We perform a TC-SVDD classification as in [12], obtaining two regions

$$S_1 \doteq \{\boldsymbol{x} \in R^N : \|\boldsymbol{x} - \mathbf{a}_1\|^2 \leq R_1^2, \|\boldsymbol{x} - \mathbf{a}_2\|^2 \geq R_2^2\}$$

and

$$S_2 \doteq \{\boldsymbol{x} \in R^N : \|\boldsymbol{x} - \mathbf{a}_2\|^2 \leq R_2^2, \|\boldsymbol{x} - \mathbf{a}_1\|^2 \geq R_1^2\},$$

where $R_1^2, R_2^2, \mathbf{a}_1, \mathbf{a}_2$ are, respectively, the radii and the centers of the spheres of the computed TC-SVDD.

Given an object $\boldsymbol{x} = (\mathbf{u}, \mathbf{z}) \in S_1$, our goal is to determine the minimum variation $\Delta\mathbf{u}^*$ of the controllable variables so that the point

$$\boldsymbol{x}^* = (\mathbf{u} + \Delta\mathbf{u}^*, \mathbf{z}) \tag{1}$$

belongs to the class S_2. To determine $\Delta\mathbf{u}^*$, we define the following minimization problem

$$\min_{\Delta\mathbf{u} \in \mathbb{R}^n} \quad d(\boldsymbol{x}, (\mathbf{u} + \Delta\mathbf{u}, \mathbf{z})) \tag{2a}$$

$$\text{subject to} \quad \|(\mathbf{u} + \Delta\mathbf{u}, \mathbf{z}) - \mathbf{a}_2\|^2 \leq R_2^2 \tag{2b}$$

$$\|(\mathbf{u} + \Delta\mathbf{u}, \mathbf{z}) - \mathbf{a}_1\|^2 \geq R_1^2 \tag{2c}$$

where d is a distance and (2b), (2c) are the constraints that require $\boldsymbol{x}^*$ to belong to S_2 and not to S_1, respectively. In other words, the *counterfactual* $\boldsymbol{x}^*$ is the nearest point, with respect to distance d, that belongs to the class opposite to the original class of a given point $\boldsymbol{x}$, taking into account that *only* controllable features $\mathbf{u}$ can be modified.

1) OPTIMALITY

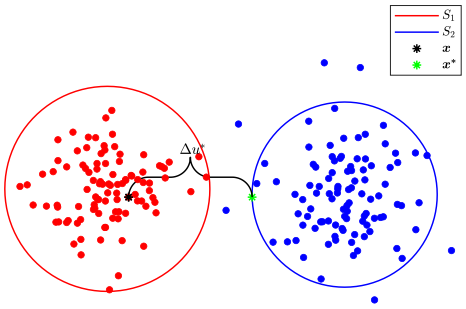
The *optimality* of a counterfactual refers to the identification, in the target output class (i.e., the class opposite to the original class the point belongs to), of the point that exhibits the *joint* minimum variation of the input features with respect to the starting point (i.e., the factual, that is by definition a point outside the target class), as shown in (2). Typically, it is possible to have variations of several combinations of features although only one of these joint variations would be at minimal distance. The proposed algorithm searches for the minimal joint variation (of all the control variables) through the minimum distance from the factual.

2) CLOSED-FORM VERSUS NUMERICAL SOLUTION

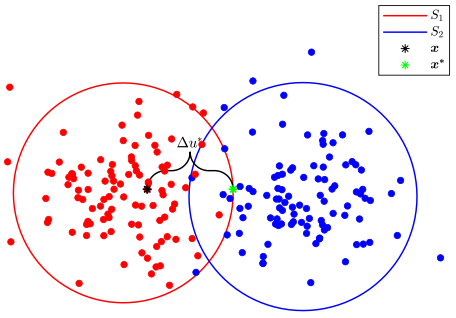
Finding an analytical solution of (2) is not an easy task and might be impossible since the space of constraints is not convex (i.e., the constraint (2c) is not convex), also it is necessary to take into account the choice of distance d. However, there are some cases where it is possible to analytically explicate the solution of (2), for example choosing as distance the Euclidean norm, performing a linear TC-SVDD and assuming to be only in two dimensions, with one feature controllable and the other non-controllable. In other cases, the solution of (2) will be performed numerically by sampling the classification regions with quasi-random methods and searching for the closest point of a given observation with respect to a fixed distance.

A. $\mathbb{R}^2$ ANALYTICAL SOLUTION

Let be $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^2 \times \{-1, 1\}$ a labelled two-dimensional dataset, in which each object $\boldsymbol{x} \in \mathcal{X}$ consists of a controllable component u and a non-controllable one z, i.e. $\boldsymbol{x} = (u, z) \in \mathbb{R}^2$. After performing a linear TC-SVDD [12] and determining two regions $S_1, S_2 \subset \mathbb{R}^2$, our goal is, given an object $\boldsymbol{x} = (u, z) \in S_1$, to find the minimum change in the



(a) Counterfactual solution for $S_1 \cap S_2 = \emptyset$. The solution in this case is obtained by simply posing $\lambda_2 = 0$, i.e., imposing nullity on the constraint (3c).



(b) Counterfactual solution for $S_1 \cap S_2 \neq \emptyset$. In this case the optimal solution is not on the edge of the region S_2 but it is inside it.

FIGURE 1. Counterfactual solutions for a 2-dimensional linear TC-SVDD with Euclidean distance. Points were sampled from a Gaussian distribution of variance 0.5 and mean 0 and 5 for red and blue points, respectively. The controllable variables lie on the abscissas while those not controllable on the ordinates, i.e. *u*Oz plane.

controllable variable Δu^* so that the object $\boldsymbol{x}^* = (u+\Delta u^*, z)$ is the closest point to $\boldsymbol{x}$ belonging to S_2 and not belonging to S_1.

In $\mathbb{R}^2$, the problem to be solved is the following:

$$\begin{aligned} \min_{\Delta u \in \mathbb{R}} \quad & \|(u,z)-(u+\Delta u,z)\|^2 && \text{(3a)} \\ \text{subject to} \quad & \|(u+\Delta u,z)-\mathbf{a}_2\|^2 \leq R_2^2 && \text{(3b)} \\ & \|(u+\Delta u,z)-\mathbf{a}_1\|^2 \geq R_1^2 && \text{(3c)} \end{aligned}$$

Two slack variables ξ_1, ξ_2 are introduced and the above problem changes in:

$$\begin{aligned} \min_{\Delta u \in \mathbb{R}} \quad & \Delta u^2 + D_1\xi_1 + D_2\xi_2 && \text{(4a)} \\ \text{subject to} \quad & \|(u+\Delta u,z)-\mathbf{a}_2\|^2 \leq R_2^2+\xi_1,\ \xi_1 \geq 0 && \text{(4b)} \\ & \|(u+\Delta u,z)-\mathbf{a}_1\|^2 \geq R_1^2-\xi_2,\ \xi_2 \geq 0 && \text{(4c)} \end{aligned}$$

where the parameters D_1, D_2 control the trade-off between the distance and the error.

Introducing the Lagrange multipliers $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0$ we get the Lagrangian function

$$\begin{aligned} \mathcal{L}(\Delta u, \xi_1, \xi_2) = {} & \Delta u^2 + D_1\xi_1 + D_2\xi_2 \\ & -\lambda_1\Big(R_2^2+\xi_1-\|(u+\Delta u,z)-\mathbf{a}_2\|^2\Big) \\ & -\lambda_2\Big(\|(u+\Delta u,z)-\mathbf{a}_1\|^2-R_1^2+\xi_2\Big) \\ & -\lambda_3\xi_1-\lambda_4\xi_2 \end{aligned} \tag{5}$$

Setting partial derivatives to zero gives the following constraints:

$$\frac{\partial\mathcal{L}}{\partial\Delta u} = 0 \Rightarrow \Delta u = \frac{\Big(\lambda_2\big(u-a_1^u\big)-\lambda_1\big(u-a_2^u\big)\Big)}{1+\lambda_1-\lambda_2} \tag{6}$$

$$\frac{\partial\mathcal{L}}{\partial\xi_1} = 0 \Rightarrow D_1-\lambda_1-\lambda_3 = 0 \Rightarrow 0 \leq \lambda_1 \leq D_1 \tag{7}$$

$$\frac{\partial\mathcal{L}}{\partial\xi_2} = 0 \Rightarrow D_2-\lambda_2-\lambda_4 = 0 \Rightarrow 0 \leq \lambda_2 \leq D_2 \tag{8}$$

TABLE 2. Algorithm 1 legend.

Line	Symbol	Description
1.	$\mathcal{C}$	Set of counterfactuals
3.	$\triangle$	Symmetric difference, $G_1 = G \cap (S_1 \backslash S_2)$, $G_2 = G \cup (S_2 \backslash S_1)$
4.1	$\boldsymbol{x}_i$	Factual point
4.2	d	Distance function
4.2	$G_{2_{\mathbf{z}=\mathbf{z}_i}}$	G_2 points with component $\mathbf{z}$ equal to $\mathbf{z}_i$
4.4.1	$\boldsymbol{x}_i'$	Counterfactual point

where a_1^u, a_2^u are the projections of $\mathbf{a}_1, \mathbf{a}_2$ onto the controllable variable u.

By substituting (6) into the expression of $\mathcal{L}$ we get:

$$\begin{aligned} & \mathcal{L}(\lambda_1,\lambda_2) \\ & = -\frac{\Big(\lambda_2\big(u-a_1^u\big)-\lambda_1\big(u-a_2^u\big)\Big)^2}{1+\lambda_1-\lambda_2} \\ & \quad -\lambda_1\Big(R_2^2-\|(u,z)-\mathbf{a}_2\|^2\Big)-\lambda_2\Big(\|(u,z)-\mathbf{a}_1\|^2-R_1^2\Big) \end{aligned} \tag{9}$$

which must be maximized under the constraints (7) and (8) to get λ_1^* and λ_2^* to be substituted into (6) to obtain the minimum variation Δu^*.

B. NUMERICAL SOLUTION

As the size of the feature space increases and for more complicated distances d or kernels, the solution of (2) may be analytically unfeasible. Thus, a discreet search algorithm has been developed.

1) `CounterfactualSVDD` ALGORITHM

Algorithm 1 returns the set $\mathcal{C}$ of counterfactuals of points belonging to S_1. Of course, the same procedure can be applied to find the counterfactuals of the points belonging to S_2 simply by reversing the roles of S_1 and S_2. For better

understanding, Table 2-B1 shows the meaning of the symbols and variables used in **Algorithm 1**.

Algorithm 1 `CounterfactualSVDD`
Dataset $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^N \times \{-1, +1\}$ is divided in training set $\mathcal{X}_{tr} \times \mathcal{Y}_{tr}$ and validation set $\mathcal{X}_{vl} \times \mathcal{Y}_{vl}$.
A TC-SVDD [12] is performed on $\mathcal{X}_{tr} \times \mathcal{Y}_{tr}$ and validated on $\mathcal{X}_{vl} \times \mathcal{Y}_{vl}$ in order to derive S_1 and S_2.
$N_C > 0$ is fixed.

---

1. $\mathcal{C} = [\,]$
2. **Sample** quasi-randomly a new dataset G
3. $G_1 \cup G_2 \doteq G \cap \left(S_1 \triangle S_2\right)$
4. **for** $i = 1 : N_C$
4.1 $\boldsymbol{x}_i = (\mathbf{u}_i, \mathbf{z}_i) \in S_1$
4.2 $d_i = d\left(\boldsymbol{x}_i, G_{2_{|\mathbf{z}=\mathbf{z}_i}}\right)$
4.3 $\boldsymbol{x}'_i = \min(d_i)$
4.4 **if**$\left(\boldsymbol{x}_i \in S_1 \;\&\; \boldsymbol{x}'_i \in S_2\right)$
4.4.1 $\mathcal{C} = \mathcal{C} \cup \{\boldsymbol{x}'_i\}$
4.5 **end**
5. **end**
6. **return** $\mathcal{C}$

The points for which a counterfactual is desired are randomly or directly sampled in S_1, while their counterfactual is sought in the set G_2, obtained from the intersection of S_2 with the set G, sampled in feature space using quasi-random sampling techniques [13], with the non-controllable features fixed. Thus, the accuracy of the counterfactual is related to the granularity of the sampling: the denser the sampling, the more accurate the counterfactual will be (bounds on the best number of random sampling points can be found in the literature [14]). Moreover, since the concept of counterfactual is closely related to explainability, a set of rules for each TC-SVDD class, $\mathcal{R}(S_i)$, is defined according to `ExplainableSVDD` algorithm [15], [16]. This is a further validation that will then also be used as a basis for extracting knowledge from the rules that characterise counterfactuals (see Section III).

2) CONVERGENCE

The counterfactual generation method can, in principle, converge to the optimal counterfactual based on the information available. According to statistical learning theory, this information corresponds to the set of points available for the method to choose the candidate optimal one. More specifically, this depends on to the size (L) of the set of candidate counterfactuals, taken within the randomly sampled SVDD target region, on which the distance from the starting point (the factual) is computed to find the point at minimum distance. In this respect, [14] gives convergence assurance, whose rate is linear with respect to L. It is also worth noting that the gap between the solution and the optimum grows exponentially in the dimension of the feature space.

3) COMPUTATIONAL COST

The estimation of the computational cost of **Algorithm 1** takes into account several aspects and considerations that need to be thoroughly investigated. First, there are two complexities involved: the SVDD and the research of the counterfactuals. Then, the counterfactual search itself involves other methods with their own complexities.

Since the SVDD is closely related to the SVM, we can assume that the computational cost is similar without losing any information, and denoting with n the number of points and with d the number of features, its computational cost is estimated in $O(\max(n, d)\min(n, d)^2)$ [17]. Let us indicate this computational time with $O(SVDD)$.

Regarding instead the research of the counterfactuals, we have to take into account

- the complexity of the quasi-random sampling,
- the number of the counterfactuals N_C,
- the computation of the distance,
- the search of the minimum of a vector.

The complexity of the quasi-random sampling depends on the method used for the sampling and references for its estimation can be found in [18]. Let us denote with $O(q)$ the complexity of the quasi-random sampling. The number of counterfactuals N_C affects the computational time of the for-loop, that is $O(N_C)$. Inside this loop, we have to compute the distance d which, in principle, can be based on any kind of distance definition. Let us indicate with $O(D)$ its computational cost. Finally, the cost of the research of the minimum of a vector can be estimated to be linear in the order of the number of the elements composing the vector [19]. So its computational cost, denoting with $g = \#G_{2_{|\mathbf{z}=\mathbf{z}_i}}$, is $O(g)$. Therefore, putting together all the components computed so far, the total complexity of the search of the counterfactuals, $O(SC)$), can be estimated with $O(\max(q, N_C \cdot (\max(D, g))))$. And then, the total computational cost of the **Algorithm 1** can be estimated with $O(\max(SVDD, SC))$.

4) COUNTERFACTUAL DISTANCE

Since the counterfactual determined by the algorithm is an approximation of the real counterfactual, a metric of the quality of the extracted counterfactual is needed. Given a point, its counterfactual is, by definition, the nearest point belonging to the opposite class. Thus, a straightforward metric for evaluating the quality q of the counterfactual $\boldsymbol{x}'$ of a point $x \in S_1$ is to evaluate its distance from S_1:

$$q = d(\mathbf{a}_1, \boldsymbol{x}') - R_1 \tag{10}$$

where $\mathbf{a}_1$ and R_1 are respectively the center and the radius of S_1. We define this new metric as *Counterfactual Distance* (CD).

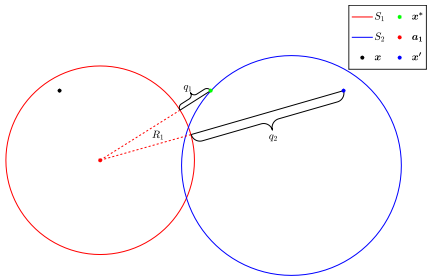


FIGURE 2. 2D-linear example of CD: this metric evaluates the goodness of the counterfactual, the closer q is to zero the more the counterfactual is optimal in terms of minimum distance. In the figure, $q_2 > q_1$ and the blue counterfactual x' is worst than the green (optimal) one x^*.

From Figure 2 it is easy to see that the lower the q, the better the counterfactual and if $q < 0$ then the counterfactual determined is incorrect.

C. EXPLAINABLE AI

XAI has gained a lot of importance in recent years. The already mentioned European GDPR, in 2018, stated that "*the existence of automated decision-making should carry meaningful information about the logic involved*". XAI is therefore a concept related to all those methods which can guarantee trustworthiness and understanding to humans. Hence, they often come in the form of intelligible rules. XAI drives the SVDD counterfactual characterization and knowledge extraction. The Logic Learning Machine (LLM) is used to this aim. The LLM algoritm is based on a four-step process: *discretization* and *latticization*, *shadow clustering*, and *rule generation* as defined in [20], [21]. First, each variable is transformed into a binary string in a proper Boolean lattice, using the inverse only-one code binarization. All strings are eventually concatenated in one unique large string per each sample. Then, a set of binary values, called implicants, which allow the identification of groups of points associated with a specific class, is generated. Finally, implicants are transformed into a collection of simple conditions and combined into a set of intelligible rules. Therefore, the decision process of an LLM algorithm can be summarized as a set of m intelligible rules in the form *IF (premise) THEN (consequence)*, with the premise being the logical product of n_k conditions and the consequence being the output class. The relevance of a rule r_k is associated with two measures, namely:

$$Covering: C(r_k) = \frac{TP(r_k)}{TP(r_k) + FN(r_k)}$$
$$Error: E(r_k) = \frac{FP(r_k)}{FP(r_k) + TN(r_k)}$$

where $TP(r_k)$, $FP(r_k)$, $TN(r_k)$, and $FN(r_k)$ are the true positives, false positives, true negatives, and false negatives associated with the rule r_k. The covering is the percentage of points for which a rule is true and maps the points on a target class. The error is the percentage of points for which the rule is true on classes other than the target one. Like decision trees, the LLM is explainable by design and it is a global method as it discovers rules which map clusters of points into classes. Other XAI methods, such as Anchors and their optimised variations [22], are "local" as they specialise rules locally for each separate sample. More specifically, Anchors explains the results of any black-box classifier, by approximating it locally through linearization as in LIME [23] and an interpretable model.[2] Extending the validity (covering) of a local rule over neighbour points is not a straightforward matter [22]; for this reason, the LLM is preferred to facilitate the knowledge extraction from the SVDD counterfactuals, by following the approach in [15], [16]. This approach applies the LLM around the boundary of the SVDD, thus maintaining the global structure of the rule-based clustering, still limiting the number of involved points and the inherent computational burden.

1) FEATURE RANKING

Feature ranking helps rule interpretation and knowledge discovery. It gives the importance of each feature in inferring the right classification (e.g., distance and speed of vehicles as outlined later on). It is also used for feature reduction in order to synthesize the model (just using the most relevant features). Whatever the XAI solver is, feature ranking may be easily derived from the ruleset, by applying sensitivity analysis on model accuracy, with and without the feature to be ranked. The interested reader is referred to [24] for further details on that subject. Feature ranking is later used to synthetize the knowledge extracted from the factual and counterfactual rulesets at hand.

III. EXPERIMENT: VEHICLE PLATOONING

The following safety-critical application is considered. Vehicle platooning is one of the most challenging problems in smart mobility scenarios. It consists of a group of vehicles interconnected via wireless that travel autonomously; the aim is to find a compromise between performance (e.g., maximize speed and minimize reciprocal distance, thus minimizing air drag resistance and fuel consumption, too) and safety (avoid collisions, even in the presence of anomalous events, such as sudden brakes or cyberattacks, [25]). The aim here is to determine what is the minimum variation in terms of controllable factors (i.e, the initial mutual distance and speed between two consecutive vehicles in the platoon, respectively d_0 and v_0) that allows for a change in system safety (collision/non-collision or vice versa). A point of the dataset is labelled as collision if the distance between any couples of vehicles, during the simulation run, becomes lower than 2 meters.

A. DATA SET DESCRIPTION

The data set concerning collision prediction in vehicle platooning is taken from [25], [26].[3] The machine learning solution is based on a supervised classification task that maps the

[2] C. Molnar, Interpretable machine learning, https://christophm.github.io/interpretable-ml-book/ (2019).

[3] https://github.com/mopamopa/Cyberplatooning and https://github.com/mopamopa/Platooning

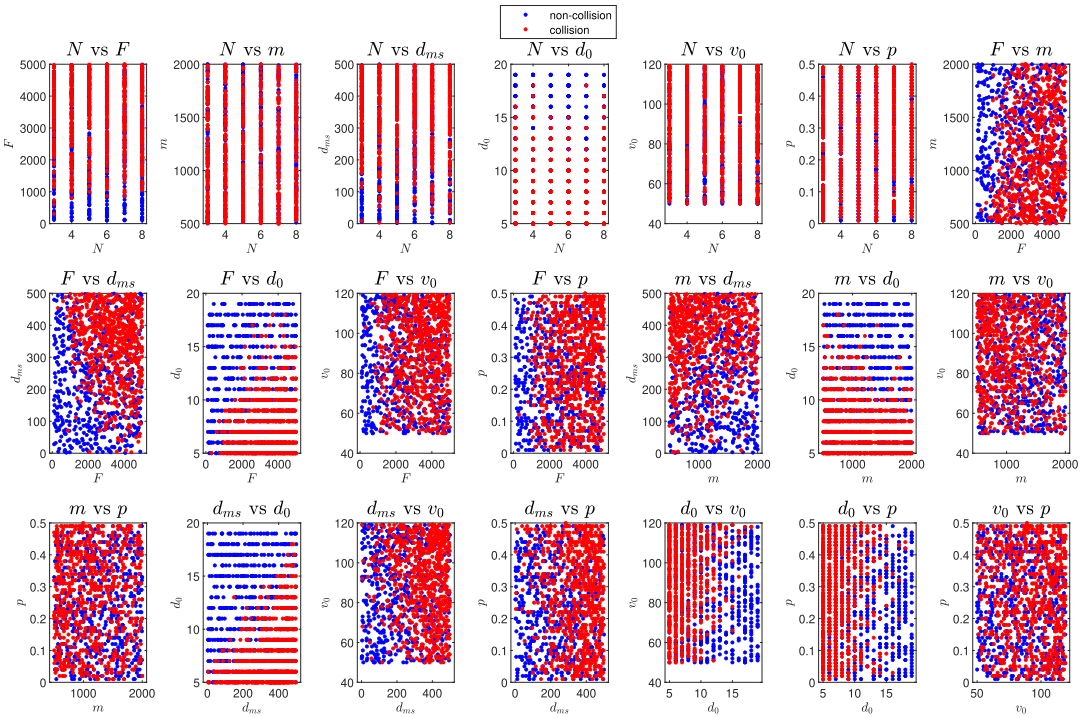


FIGURE 3. Scatter plot of each pair of variables in the platooning data set. Red dots indicate collision and blue dots indicate non-collision.

features into a potential collision in the near future. Features are: braking force of lead vehicle (at the top of the platoon), current speed, distance and acceleration, number and weight of vehicles, as well as quality of service of the communication channel (loss probability and delay). Controllable variables are speed and distance only, thus making the restrictions on counterfactual generation (with respect to the other variables), as well as the search in the grid of the destination SVDD, very tight.

In this scenario, the counterfactual explanation can play an effective role in improving the safety of the platooning system: given a combination of the platoon input parameters that brings the system into collision, the counterfactual finds the minimal change in the controllable features such that the platoon no longer collides. Finding such a minimal change simplifies the recovery operation (from collision).

The behaviour of the platooning system is synthesised by the following vector of features:

$$\mathbf{I} = [N, F, m, d_{ms}, p, d_0, v_0]$$

where N is the total number of vehicles of the platoon, F is the braking force applied by the leader, m is the weight of the vehicles, d_{ms} is the communication delay in milliseconds, p is the probability of packet loss, and d_0 and v_0 are the mutual distance and speed between each pair of vehicles in the initial condition.

Data points are sampled by implementing the CACC simulator as in [25] in the following ranges:

$$N \in [3, 8],\ F \in [1000, 5000]\ \text{N},\ m \in [500, 2000]\ \text{Kg},$$
$$d_{ms} \in [0, 1000]\ ms,\ d_0 \in [4, 20]\ \text{m},$$
$$v_0 \in [30, 130]\ \text{Km/h},\ \mathbf{p} \in [0, 1].$$

The considered ranges are very challenging as they cover a very large set of working conditions. As already said, since the control of the dynamical system reacts by changing the initial distance and speed, we consider the variables d_0 and v_0 as the only controllable ones and the others as non-controllable, therefore, named $\mathcal{X}_{PL}$ the platooning dataset, an observation $\boldsymbol{x} \in \mathcal{X}_{PL}$ can be written as

$$\boldsymbol{x} = (\mathbf{u}, \mathbf{z})$$

where $\mathbf{u} = (d_0, v_0)$ and $\mathbf{z} = (N, F, m, d_{ms}, p)$.

The analysed platooning data set includes 20000 records with equally distributed samples for the collision $(+1)$ and non-collision (-1) classes. A TC-SVDD with Gaussian Kernel [27] has been trained ($\sigma = 1.87$, $C_1 = C_2 = 1$, $C_3 = 1/(\nu N_1)$, $C_4 = 1/(\nu N_{-1})$, where N_1 and N_{-1} are the sizes of the collision and non collision class, respectively, and $\nu = 0.05$ as in [12]) on 60% of the data and evaluated on the remaining 40%. A set of 10000 CEs has been generated through the implementation of **Algorithm 1** and validated both with rule-analysis and simulations.

Figure 3 presents the scatterplots of all the possible pairs of features in the platooning data set, grouped by target class, and reveals how the separation between safety and collision may be hardly found without complex combinations of more than two features.

TABLE 3. Explainable rules extracted from SVDD through the algorithm `ExplainableSVDD` as in [15], [16].

#	output	cond1	cond2	cond3	cond4	cond5	covering	error
1	-1	$F \leq 4690$	$d_{ms} \leq 478$	$d_0 > 13$			0.50	0.04
2	-1	$F \leq 4238$	$m \leq 1963$	$10 < d_{ms} \leq 335$	$v_0 \leq 111$		0.49	0.04
3	-1	$106 < F \leq 2287$	$17 < d_{ms} \leq 496$	$7 < d_0 \leq 17$	$v_0 \leq 118$	$p < 0.5$	0.43	0.04
4	-1	$N \leq 6$	$d_{ms} \leq 493$	$d_0 > 13$			0.37	0.04
5	-1	$m > 806$	$d_0 > 14$				0.37	0.04
6	-1	$F \leq 4941$	$566 < m \leq 1990$	$2 < d_{ms} \leq 399$	$v_0 \leq 80$		0.34	0.04
7	-1	$F \leq 1432$	$m > 801$	$d_{ms} \leq 482$			0.27	0.04
8	-1	$2 < d_{ms} \leq 97$	$d_0 > 8$				0.20	0.03
9	-1	$m > 753$	$2 < d_{ms} \leq 72$	$d_0 > 5$			0.17	0.04
10	-1	$N \leq 6$	$1365 < F \leq 4964$	$m > 1586$	$d_{ms} > 77$	$v_0 > 66$	0.09	0.04
11	+1	$F > 1116$	$d_{ms} > 29$	$d_0 \leq 8$			0.52	0.04
12	+1	$d_0 \leq 7$	$v_0 > 79$				0.37	0.03
13	+1	$d_{ms} > 45$	$d_0 \leq 15$	$p \geq 0.58$			0.29	0.04
14	+1	$N \leq 7$	$3714 < F \leq 4606$	$d_0 \leq 16$	$54 < v_0 \leq 118$		0.26	0.05

TABLE 4. Counterfactual explanation table of ten points randomly sampled from the set of 10000 extracted collision points. The last column contains the minimum change Δu^* of the controllable features d_0, initial distance, and v_0, initial velocity, of the platoon.

Factuals										Counterfactuals									
d_0	v_0	N	F	m	d_{ms}	p	SVDD	LLM	Rule	d_0	v_0	N	F	m	d_{ms}	p	SVDD	LLM	Rule	Δu^*
5	117	8	2976	952	81	0.44	+1	+1	11	18	111	8	2976	952	81	0.44	-1	-1	1	$(13, -6)$
6	97	8	4898	1215	92	0.3	+1	+1	11	15	51	8	4898	1215	92	0.3	-1	-1	5	$(6, -49)$
6	82	4	966	1271	398	0.5	+1	+1	12	6	51	4	966	1271	398	0.5	-1	-1	6	$(0, -31)$
7	73	8	1290	807	338	0.43	+1	+1	11	15	50	8	1290	807	338	0.43	-1	-1	1	$(8, -23)$
5	65	3	1117	535	329	0.48	+1	+1	12	16	116	3	1117	535	329	0.48	-1	-1	1	$(11, 51)$
7	91	6	973	708	458	0.13	+1	+1	12	16	55	6	973	708	458	0.13	-1	-1	1	$(9, -36)$
7	108	5	3451	1895	478	0.19	+1	+1	11	18	84	5	3451	1895	478	0.19	-1	-1	4	$(11, -24)$
8	99	8	3993	634	380	0.01	+1	+1	11	17	99	8	3993	634	380	0.01	-1	-1	5	$(9, 0)$
6	76	6	1785	744	370	0.11	+1	+1	11	18	56	6	1785	744	370	0.11	-1	-1	1	$(12, -20)$
5	119	3	2333	554	272	0.31	+1	+1	11	18	50	3	2333	554	272	0.31	-1	-1	1	$(13, -69)$

B. RESULTS

The TC-SVDD trained on the platooning data achieved the following classification performance: training accuracy of 0.88, test accuracy of 0.88, sensitivity of 1.00, specificity of 0.75. LLM decision rules describing the two SVDD regions are extracted as in [15], [16] and presented in Table 3. Specifically, the collision region is described by four rules (average number of conditions = 2.75), whereas the non collision region is described by ten rules (average number of conditions = 3.3).

The feature ranking in Figure 4 helps understand the most relevant features for classes separation. Distance, braking force and delay are the most meaningful ones; surprisingly, speed and number of vehicles have less importance than expected. The left and right directions of the bars indicate the relevance in decreasing and increasing values, respectively, of the feature. The directions of distance and speed are coherent with intuition, e.g., decreasing distance increases the frequency of collision. The direction of the bar associated with the delay feature in the safety class (no collision) is however counter-intuitive as it states that safety is achieved by increasing delay. This is not uncommon in machine learning analysis as it should give unexpected insights into the problem. In this case, the delay effect is superseded by the ones of the other variables; the delay subplots in Figure 3 show the spread of red (collision) points over almost all the delay ranges (except for very low delays). Together with Table 3, the ranking figures help understand how much global XAI drives a more synthetic knowledge extraction than local XAI (such as through LIME, as often used in counterfactual explanations [28]), which gives rules that are built around the point of interest and have a limited covering over the rest of the dataset. Global XAI still has local explanation property (as outlined in Table 4), but it may give global insight, too (as outlined later in Figure 6c).

C. EXPLANATION

To determine a counterfactual explanation of $\mathcal{X}_{PL}$, 10000 points were randomly sampled from the collision class (+1) and a counterfactual was determined for each of them through **Algorithm 1**, using the Gaussian kernel-induced distance d as the distance [29]

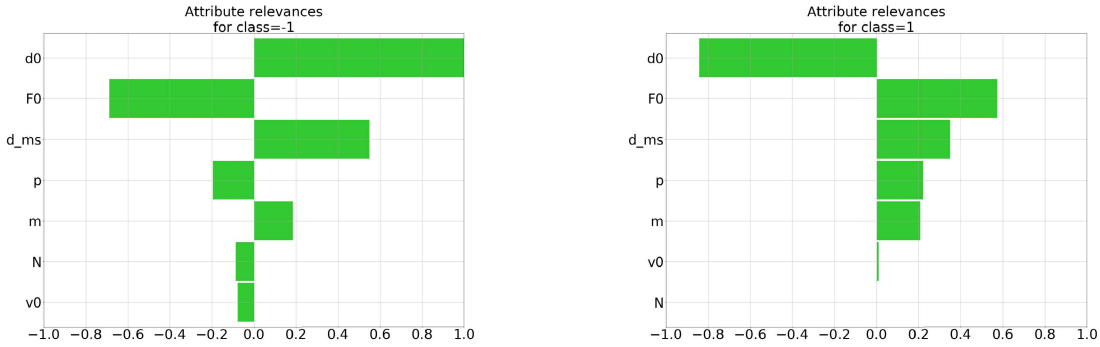
$$d(x, y) = 2 - 2k(x, y)$$

where $k(x, y) = e^{-\frac{||x-y||^2}{2\sigma^2}}$ is the Gaussian kernel. Ten examples are shown in Table 4.

Each row of Table 4 shows the point belonging to the collision class, classified with the SVDD and LLM and the rule, with largest covering, it satisfies; the corresponding CE, also classified with the SVDD and LLM, and the rule it satisfies are reported. The last column reports the minimum change $\Delta\mathbf{u}$ in distance and speed that allowed to move from the collision class to the non-collision class.

D. VALIDATION

The validation of the counterfactuals safety is as follows: the 10000 CEs determined by **Algorithm 1** were tested by the CACC simulator [25], obtaining 7.82% error (i.e., that the determined counterfactual still brings the system



(a) Feature ranking for class "non-collision".

(b) Feature ranking for class "collision".

FIGURE 4. Graph of the most relevant features for the determination of the class.

into collision) and 92.18% actual counterfactuals, of which only 2.07% are found to be overestimated. Overestimation is defined with respect to a final distance larger than 10 meters,[4] such a distance is found at the end of the simulation run, which is driven by the counterfactual. Figure 5 deals with the temporal behaviour of three significant cases; the first two (from top to bottom subplots) are optimal counterfactuals (the first with change in speed and the second one with change in distance), as they lead to a final condition which is very close to collision. The last subplot (at the bottom of the figure) highlights an over-dimensioned counterfactual as the final distance is much larger than the boundary one (between collision and non-collision).

E. ON THE MINIMUM DISTANCE

The analysis would suggest more insightful thinking on the concept of ''minimum'' counterfactual distance, which is ubiquitous in the literature. In the platooning application, that concept would imply ''almost collision'' because the counterfactual, by construction, should lie in the safety SVDD (under the constraint of non-controllable variables), but still closest to the collision one. On the one hand, this corroborates the flexibility of counterfactual construction through the SVDD with respect to deep learning, in which the positioning of the (constrained and with minimum distance) counterfactual should be mapped into a very complex training cost. On the other hand, it would lead to other, more restricted, forms of counterfactual construction, when safety plays a crucial role. This topic is left open for future research.

F. QUALITY

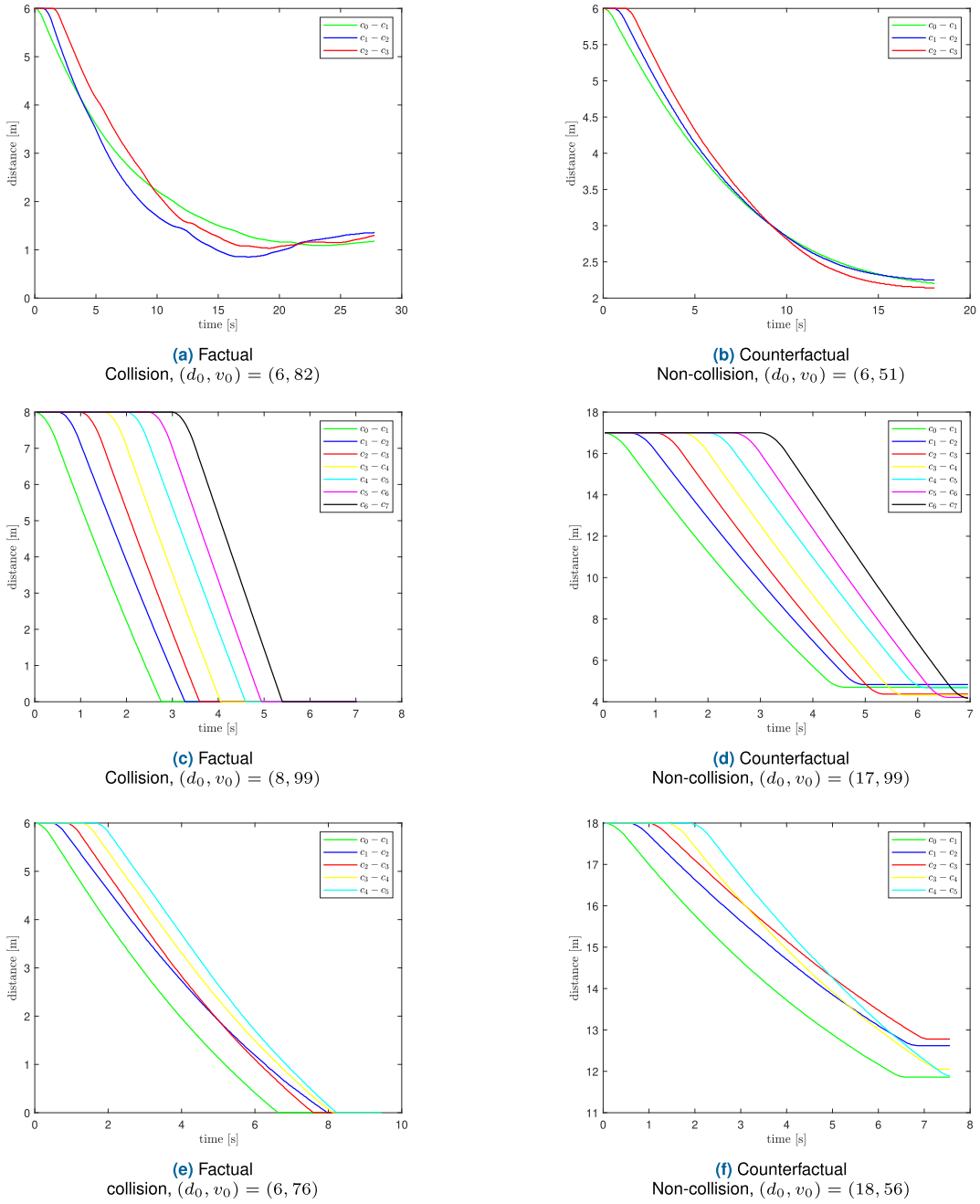
The validation of the counterfactuals quality is as follows. The CD of each CE is calculated (see Section II-B), thus evidencing satisfactory statistics, as shown in Figure 6a, in line with simulation evidence (Figure 6b). The CD metric well synthesises the overestimation issue. Recall that high CD means low quality in counterfactuals. In order to derive further knowledge extraction from the CD analysis, the following supervised problem is defined over the CD values and solved via the LLM. The factuals (i.e., points of the collision class, which are mapped into the corresponding counterfactuals) are mapped into two classes; the classes label CD values under and above the 0.03 threshold. Values larger than the threshold represent overdimensioned and almost overdimensioned points, as evidenced in Figure 6a.

The resulting feature ranking in Figure 6c (for CD>threshold) shows that high CD samples are associated with critical factuals, namely, with increasing delay, leader acceleration (force divided by the mass), loss, speed and number of vehicles as well as decreasing distance. The rationale of the conditions relies on the fact that critical factuals need to go deeper inside the destination class (thus leading to larger CD) to replace the original conditions of collision into new safety ones. Moreover, the rules identifying high CD may drive further optimisation of the respective counterfactuals, e.g., through a finer granularity of the grid in a reduced search space, identified by the ruleset itself [30]. This is left open for future research as well.

IV. DISCUSSION

This study aims to define a new method for generating local explanations by defining counterfactuals from observations characterized by controllable and non-controllable features. Nemirovsky *et al.* [6] first introduced the concept of CEs with controllable and non-controllable features in a diabetes prediction algorithm, however they first applied counterfactual search to all the features and then they removed the perturbations related to non-controllable features like age and the number of pregnancies. In this study, controllable and non-controllable features are handled in a more straightforward way, since the search for counterfactuals is instead done by perturbing only the controllable features (i.e., d_0 and v_0) in the kernel space, keeping the non-controllable variables fixed. Most of the recently proposed methods are deep learning based [6], [8], thus requiring more complex architectures and higher computational cost for training. The use

[4] A collision is considered, in the original dataset, when the distance is below the threshold of 2 meters.



(a) Factual
Collision, $(d_0, v_0) = (6, 82)$

(b) Counterfactual
Non-collision, $(d_0, v_0) = (6, 51)$

(c) Factual
Collision, $(d_0, v_0) = (8, 99)$

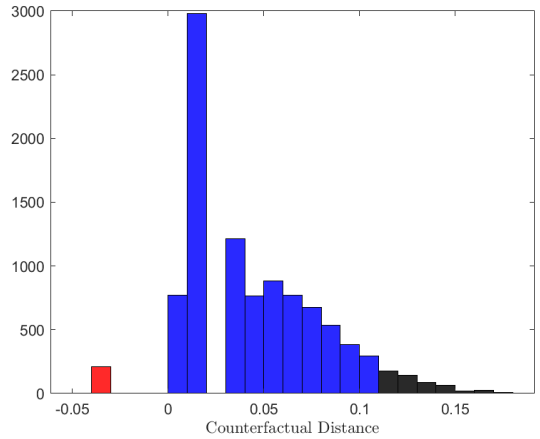
(d) Counterfactual
Non-collision, $(d_0, v_0) = (17, 99)$

(e) Factual
collision, $(d_0, v_0) = (6, 76)$

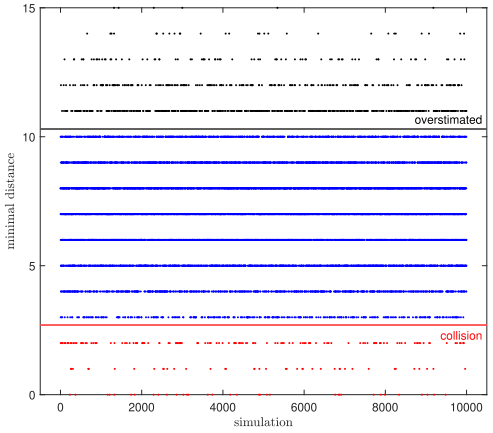
(f) Counterfactual
Non-collision, $(d_0, v_0) = (18, 56)$

FIGURE 5. Table 4, row 3, 8, 9: examples of platoon distance trend of the original features and their counterfactual.

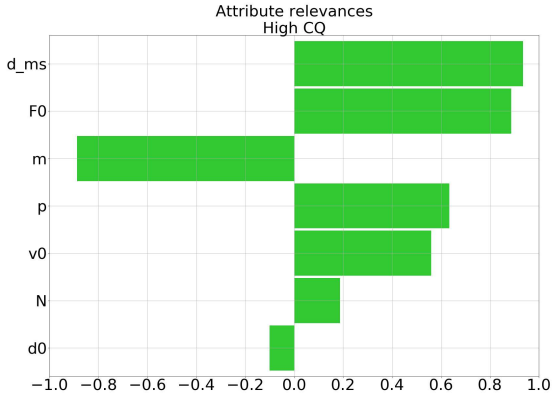
of TC-SVDD allows us to define the two regions with a reduced computational cost, yet still achieving more than satisfactory accuracy (e.g., >85%). Furthermore, the additional rule-based description of the SVDD regions provides transparency to the point classification process, allowing for a robust validation of correctness and consistency of the generated CEs. Specifically, as shown in Table 4, in the platooning example, CEs are generally associated with greater initial distance and reduced initial velocity of the platoon. Moreover, the quality of explanations have been evaluated in terms of distance from the region associated with the opposite outcome. The optimal CE of $\boldsymbol{x}$ is the point, with opposite class, located at minimum distance from $\boldsymbol{x}$. The introduction of a quality metric (CD) allows us to verify the correctness of CEs, generated with the proposed numerical approximation, since a distance greater than zero ensures the non-intersection between the two SVDD regions, thus the belonging of the CE to the correct class, with a certain level of confidence defined by the TC-SVDD (i.e., 88% in the platooning example) and a distance close to zero ensures the minimum distance requirement. Figure 6b shows CD values for the generated platooning CEs, demonstrating the



(a) CD of extracted counterfactuals. The red bin refers to counterfactuals that are incorrect, i.e. $q < 0$. Black bins refer to counterfactuals that overestimate corrections ($q > 0.1$).



(b) Behaviour of simulations with counterfactuals extracted via **Algorithm 1**. The platoon collides when the minimum distance in the simulation is less than or equal to 2 (red dots). Black dots refer to counterfactuals that overestimate the correction (minimum distance greater than 10).



(c) Feature ranking which describes the relevance of the features in classify high value of CQ.

FIGURE 6. Metrics for validating Algorithm 1: (6a) shows the CD of the extracted counterfactuals, (6b) represents the behaviour of the 10000 counterfactual simulations and (6c) shows the feature ranking for the class "High CD".

effectiveness of the proposed method, as most of the points are associated to a low but positive CD value. Indeed, almost 40% of the points are associated to CD lower than 0.02 and about 92% of the points present CD lower than 0.1.

Unlike previous works in this area, the validation of the generated counterfactuals is not only based on class prediction via SVDD, but further supported by validation via simulations. In fact, the attribution of the point to the correct class according to the prediction of the previously trained model does not guarantee its real belonging to that class, because of the existence of a certain number of false positives and false negatives that, even if minimized, should not be neglected. The validation process through the CACC simulator (see 6a) has proven that the generated CEs are descriptive of the non-collision class with a more than satisfactory accuracy, and that only a small part of the generated points overestimates the minimum distance. Hence, the use of CE in platooning results applicable to the generation of control algorithms, based on the correction of the system dynamics, to prevent collisions.

A. OTHER APPLICATIONS

The considered approach is applicable to cyberphysical systems, empowered by simulated digital twins. However, the method is applicable to a wider range of applications. Examples may lie in the following sectors: health sector (e.g, disease prediction and prevention), human behavioral analysis (fraud detection) and social networks (guidance of public opinion [31][5]). The health sector is currently our next step as it introduces some conceptual differences in the validation process. As already pointed out for cyberphysical systems, testing tools (via simulation, emulation, or replicable experiments) may offer support to validation through additional

[5]See also, e.g., https://www.journals.uchicago.edu/doi/pdfplus/10.1086/210513

counterfactual-driven ground truth (i.e., testing the exact counterfactual collision avoidance). Clinical analysis, on the other hand, cannot exploit controllable ground truth in a straigthforward manner (i.e., applying a medical treatment just in accordance of the counterfactual). The health scenarios would claim for additional human interaction between AI and the clinician who interprets the (explained) artificial reasoning (i.e., the suggested counterfactual) and maps it into current clinical practice. In this case, the testing environment would consist of dedicated medical trial campaigns.

B. DIABETES CHARACTERIZATION AND PREVENTION

In [32], CEs were used to characterize the smallest changes in biomarker values that distinguish diabetic patients from non-diabetic ones. Preliminary results have shown that non-diabetics patients have on average lower values in terms of fasting blood sugar ($-0.88\ mmol/L$) and body mass index ($-0.14\ kg/m^2$) and higher values of high-density lipoprotein ($0.26\ mmol/L$) with respect to diabetic ones. Particularly, the changes in biomarkers tend to increase with age. These variations, albeit small, reflect the literature on risk factors for Type 2 diabetes and suggest the importance, in biomedical applications, of integrating AI-generated recommendations with medical knowledge and clinical guidelines. Possible next developments could head in this direction as CEs generated through the application of variable distance perturbations could be useful to provide an estimate of risk in the case of chronic diseases, such as diabetes, and contribute to the formulation of preventive strategies. In fact, CEs generated at minimum distance are associated to an higher risk of developing the disease, whereas CEs generated at a progressively increasing distance are associated with a lower risk.

V. CONCLUSION AND FUTURE WORKS

The proposed counterfactual methodology proves to be trustworthy, thanks to the use of the eXplainable AI, which allows to characterize the extracted counterfactuals through readily interpretable rules that can be easily understood and validated by application domain experts, even if they have no prior knowledge in the field of artificial intelligence. Future research will need to focus on further optimization of the method as anticipated in the results section, as well as on modifying the proposed method to handle categorical variables and images. Moreover, the aforementioned approach shall be compared with other state-of-the-art solutions and investigated with respect to different domains of application, like the field of disease prevention, for example using observations derived from electronic medical records, from longitudinal population studies, or from individual monitoring devices.

REFERENCES

[1] S. Wachter, and B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR (October 6, 2017)," Harvard J. Law Technol., vol. 31, no. 2, 2017. [Online]. Available: https://ssrn.com/abstract=3063289 and http://dx.doi.org/10.2139/ssrn.3063289, doi: 10.2139/ssrn.3063289.

[2] M. Suzuki, Y. Kamcya, T. Kutsuna, and N. Mitsumoto, "Understanding the reason for misclassification by generating counterfactual images," in *Proc. 17th Int. Conf. Mach. Vis. Appl. (MVA)*, Jul. 2021, pp. 1–5.

[3] A. White and Artur S. d'Avila Garcez, "Measurable counterfactual local explanations for any clas-sifier," 2020, *arXiv:1908.03020*.

[4] R. Poyiadzi, K. Sokol, R. Santos-Rodríguez, T. D. Bie, and P. Flach, "Face: Feasible and actionable counterfactual explanations," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2020, pp. 344–350, doi: 10.1145/3375627.3375850.

[5] A. Van Looveren and J. Klaise, "Interpretable counterfactual explanations guided by prototypes," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2021, pp. 650–665, doi: 10.1007/978-3-030-86520-7_40.

[6] D. Nemirovsky, N. Thiebaut, Y. Xu, and A. Gupta, "CounteRGAN: Generating realistic counterfactuals with residual generative adversarial nets," 2020, *arXiv:2009.05199*.

[7] R. Mochaourab, S. Sinha, S. Greenstein, and P. Papapetrou, "Robust counterfactual explanations for privacy-preserving SVM," 2021, *10.48550/ARXIV.2102.03785*. [Online]. Available: https://arxiv.org/abs/2102.03785, doi: 10.48550/arxiv.2102.03785.

[8] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P.-S. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," 2018, *10.48550/ARXIV.1802.0762*. [Online]. Available: https://arxiv.org/abs/1802.07623, doi: 10.48550/ARXIV.1802.07623.

[9] E. Albini, A. Rago, P. Baroni, and F. Toni, "Relation-based counterfactual explanations for Bayesian network classifiers," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 451–457, doi: 10.24963/ijcai.2020/63.

[10] A. Shih, A. Choi, and A. Darwiche, "A symbolic approach to explaining Bayesian network classifiers," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 5103-5111, doi: 10.24963/ijcai.2018/708.

[11] D. Nemirovsky, N. Thiebaut, Y. Xu, and A. Gupta, "Providing actionable feedback in hiring marketplaces using generative adversarial networks," in *Proc. WSDM*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 1089–1092, doi: 10.1145/3437963.3441705.

[12] G. Huang, H. Chen, Z. Zhou, F. Yin, and K. Guo, "Two-class support vector data description," *Pattern Recognit.*, vol. 44, no. 2, pp. 320–329, Feb. 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320310004115

[13] C. Cervellera, M. Gaggero, D. Maccio, and R. Marcialis, "Quasi-random sampling for approximate dynamic programming," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Aug. 2013, pp. 1–8.

[14] C. Cervellera and M. Muselli, "Deterministic design for neural network learning: An approach based on discrepancy," *IEEE Trans. neural Netw.*, vol. 15, no. 3, pp. 533–544, Jun. 2004.

[15] A. Carlevaro and M. Mongelli, "A new SVDD approach to reliable and explainable AI," *IEEE Intell. Syst.*, vol. 37, no. 2, pp. 55–68, Mar. 2022.

[16] A. Carlevaro and M. Mongelli, "Reliable AI trough SVDD and rule extraction," in *Proc. Int. IFIP Cross Domain (CD) Conf. Mach. Learn. Knowl. Extraction (MAKE) (CD-MAKE)*, pp. 153–171, 2021.

[17] O. Chapelle, "Training a support vector machine in the primal," *Neural Comput.*, vol. 19, no. 5, pp. 1155–1178, 2007.

[18] S. Sen, T. Samanta, and A. Reese, "Quasi-versus pseudo-random generators: Discrepancy, complexity and integration-error based comparison," *Int. J. Innov. Comput. Inf. Control*, vol. 2, pp. 621–651, Jan. 2006.

[19] M. Blum, R. W. Floyd, V. Pratt, R. L. Rivest, and R. E. Tarjan, "Time bounds for selection," *J. Comput. Syst. Sci.*, vol. 7, no. 4, pp. 448–461, Aug. 1973, doi: 10.1016/S0022-0000(73)80033-9.

[20] M. Muselli, "Switching neural networks: A new connectionist model for classification," in *Neural Nets*, B. Apolloni, M. Marinaro, G. Nicosia, and R. Tagliaferri, Eds. Berlin, Germany: Springer, 2006, pp. 23–30.

[21] M. Muselli and E. Ferrari, "Coupling logical analysis of data and shadow clustering for partially defined positive Boolean function reconstruction," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 1, pp. 37–50, Jan. 2011, doi: 10.1109/tkde.2009.206.

[22] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini, "Factual and counterfactual explanations for black box decision making," *IEEE Intell. Syst.*, vol. 34, no. 6, pp. 14–23, Nov. 2019.

[23] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Sep. 2019, doi: 10.1145/3236009.

[24] D. Cangelosi, F. Blengio, R. Versteeg, A. Eggert, A. Garaventa, C. Gambini, M. Conte, A. Eva, M. Muselli, and L. Varesio, "Logic learning machine creates explicit and stable rules stratifying neuroblastoma patients," *BMC Bioinf.*, vol. 14, no. S7, p. S12, Apr. 2013, doi: 10.1186/1471-2105-14-S7-S12.

[25] M. Mongelli, "Design of countermeasure to packet falsification in vehicle platooning by explainable artificial intelligence," *Comput. Commun.*, vol. 179, pp. 166–174, Nov. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366421002504

[26] M. Mongelli, E. Ferrari, M. Muselli, and A. Fermi, "Performance validation of vehicle platooning through intelligible analytics," *IET Cyber-Phys. Syst., Theory Appl.*, vol. 4, no. 2, pp. 120–127, Jun. 2019, doi: 10.1049/iet-cps.2018.5055.

[27] A. Carlevaro. (2022). NCLASS SVDD. MATLAB Central File Exchange. Accessed: Jan. 8, 2022. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/104660-nclass-svdd

[28] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*. New York, NY, USA: Association for Computing Machinery, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[29] B. Schölkopf, "The kernel trick for distances," in *Proc. 13th Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Cambridge, MA, USA: MIT Press, 2000, pp. 283–289.

[30] S. Narteni, M. Ferretti, V. Orani, I. Vaccari, E. Cambiaso, and M. Mongelli, "From explainable to reliable artificial intelligence," in *Proc. Int. IFIP Cross Domain (CD) Conf. Mach. Learn. Knowl. Extraction (MAKE) (CD-MAKE)*, 2021, pp. 255–273.

[31] P. Sun and L. Zhang, "Public opinion guidance under the background of big data technology," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Jul. 2019, p. 226.

[32] M. Lenatti, A. Carlevaro, K. Keshavjee, A. Guergachi, A. Paglialonga, and M. Mongelli, "Characterization of type 2 diabetes using counterfactuals and explainable AI," in *Proc. 32nd Med. Informat. Eur. (EFMI MIE) Conf.*, Nice, France, May 2022, pp. 98–103, doi: 10.3233/SHTI220404.

ALBERTO CARLEVARO received the master's degree *(cum laude)* in applied mathematics from the University of Genoa, in May 2020, with a physics–mathematics thesis on the behavior of liquid crystals under electromagnetic fields, where he is currently pursuing the Ph.D. degree with the Department of Electrical, Electronic and Telecommunications Engineering and Naval Architecture (DITEN), in the research topic "Traffic Analysis in the Smart City," in collaboration with CNR and S.M.E. Aitek. He was a Research Fellow at the Institute of Electronics, Informatics and Telecommunications Engineering (IEIIT), National Research Council (CNR), where he worked on machine learning and explainable AI in collaboration with Rulex Inc. His current research interests include machine learning, deep learning, statistical learning, and explainable AI.

MARTA LENATTI received the master's degree *(cum laude)* in biomedical engineering from the Politecnico di Milano, in December 2020. She is currently a Graduate Research Fellow at the Institute of Electronics, Information Engineering and Telecommunications (IEIIT), Italian National Research Council of Italy (CNR), Milan, Italy, and a Visiting Scientist at Ryerson University, Toronto. Her current research interests include explainable AI, eHealth in audiology, and the application of machine learning methods for the extraction of predictive and descriptive biomarkers in patients with chronic pathologies.

ALESSIA PAGLIALONGA received the M.Sc. and Ph.D. degrees in biomedical engineering from the Politecnico di Milano, Italy, in 2005 and 2009, respectively. She is currently a Researcher at the Institute of Electronics, Information Engineering and Telecommunications (IEIIT), National Research Council of Italy (CNR), Milan, Italy, an Adjunct Professor at the Politecnico di Milano, and a Visiting Scientist at Ryerson University, Toronto. Her research interests include data analytics and predictive modeling for health, eHealth and mHealth, audiological technology, machine learning, and biosignal processing. She is a member of the European Society of Cardiology and a Senior Member of the European Alliance for Innovation (EAI). She is serving as an Associate Editor for *BioMedical Engineering Online* and the *International Journal of Audiology*.

MAURIZIO MONGELLI (Member, IEEE) received the Ph.D. degree in electronics and computer engineering from the University of Genoa (UNIGE), in 2004. The Ph.D. was funded by Selex Communications S.p.A. (Selex). He worked with Selex and the Italian Telecommunications Consortium (CNIT), from 2001 to 2010. During his Ph.D. and in the following years, he worked on the quality of service for military networks with Selex. He was the CNIT Technical Co-ordinator of a research project concerning satellite emulation systems, funded by the European Space Agency; and he spent three months working on the project at the German Aerospace Center in Munich. He is currently a Researcher at the Institute of Electronics, Computer and Telecommunication Engineering (IEIIT), National Research Council (CNR), where he deals with machine learning applied to bioinformatics and cyber-physical systems. He is the coauthor of over 100 international scientific articles, two patents, and is participating in the SAE G-34/EUROCAE WG-114 AI in Aviation Committee.