Multi-Class Counterfactual Explanations using Support Vector Data Description

Alberto Carlevaro [1], Marta Lenatti [2], Alessia Paglialonga [2], and Maurizio Mongelli [2]

[1]CNR-IEIIT
[2]Affiliation not available

October 30, 2023

Abstract

Explainability is becoming increasingly crucial in machine learning studies and, as the complexity of the model increases, so does the complexity of its explanation. However, the higher the complexity of the problem the higher the amount of information it may provide, and this information can be exploited to generate a more precise explanation of how the model works. One of the most valuable ways to recover such relation between input and output is to extract counterfactual explanations. In binary classification, counterfactuals allow us to find minimal changes from an observation to another one belonging to the opposite class. But how do counterfactuals work in multi-class problems? In this article, we propose a novel methodology to extract multiple counterfactual explanations (MUCH, MUlti Counterfactual via Halton sampling) from an original Multi-Class Support Vector Data Description algorithm (MC-SVDD). To evaluate the performance of the proposed method, we extracted a set of counterfactual explanations from three state-of-the-art datasets achieving satisfactory results that pave the way to a range of real-world applications, for example disease prevention.

Multi-Class Counterfactual Explanations using Support Vector Data Description

A. Carlevaro[†], M. Lenatti[†], A. Paglialonga, and M. Mongelli, *Member, IEEE*

***Abstract*—Explainability is becoming increasingly crucial in machine learning studies and, as the complexity of the model increases, so does the complexity of its explanation. However, the higher the complexity of the problem the higher the amount of information it may provide, and this information can be exploited to generate a more precise explanation of how the model works. One of the most valuable ways to recover such relation between input and output is to extract counterfactual explanations. In binary classification, counterfactuals allow us to find minimal changes from an observation to another one belonging to the opposite class. But how do counterfactuals work in multi-class problems? In this article, we propose a novel methodology to extract multiple counterfactual explanations (MUCH, MUlti Counterfactual via Halton sampling) from an original Multi-Class Support Vector Data Description algorithm (MC-SVDD). To evaluate the performance of the proposed method, we extracted a set of counterfactual explanations from three state-of-the-art datasets achieving satisfactory results that pave the way to a range of real-world applications, for example disease prevention.**

***Impact Statement*—When a system is analyzed by machine learning, the inherent models are posed to the attention of domain experts, thus delegating further possible actions. Counterfactual explanations, on the other hand, directly suggest actuation on the system. Counterfactual control still remains under experts' supervision, but the system improves its level of autonomy. The long-term goal is to make the machine learning model aware of how to affect the environment properly (both in terms of performance and safety). Examples may include: manoeuvering of autonomous cars, clinical diagnosis, decision making in cyber warfare, and applications in finance. The proposed approach generalizes counterfactuals intelligibility and control to the multi-class case. The validation over practical scenarios (e.g., the FIFA dataset) corroborates both control precision and quality of counterfactual explanations, thus increasing the readiness level of the approach.**

***Index Terms*—Counterfactual explanations, Multi-class classification, Support Vector Data Description.**

I. Introduction

A. Background and Rationale

1) eXplainable AI: Over the past decade, Artificial Intelligence (AI) models have achieved astounding levels of accuracy in countless application areas. However, the pervasive presence of opaque or black box architectures can become an obstacle to their application in everyday life. This opacity in decision-making has motivated the investigation of new techniques that provide deeper insights into the inner logic of AI models, i.e. eXplainable AI (XAI) algorithms [1]. The rapid spread of XAI techniques has been mainly driven by the demand to increase the transparency of AI models [2] and the need to allow humans to actively interact with these models. Among the various techniques available, *counterfactual explanations* [3] have recently gained attention thanks to their capability to explain why a model makes a certain decision, given a specific observation. More specifically, counterfactual explanations describes what should be changed in a certain input sample (the *factual*) to obtain a different model decision.

2) Controllability: Counterfactuals can be used to introduce control over the machine learning model in a flexible way [4]–[6]. The process consists of generating counterfactuals around controllable variables, still under non-controllable constraints. Several sets of controllable variables may be considered to look at the problem under different angles and understand reachability over specific conditions. Counterfactual controllability in some ways extends canonical AI understanding, opening the door to increased autonomy.

3) Multi-class: Examples may help understand the importance of counterfactual reasoning in multi-class situations. In healthcare, several diseases present different stages of severity (e.g., cancer, chronic obstructive pulmonary disease...) that can worsen drastically in a short time if not properly treated. In this case, multi-class counterfactuals can be a valuable instrument to monitor the stage of disease progression in order to detect minimal changes in the patient's condition and apply appropriate countermeasures before the disease progresses to the next stage. Another example may involve the study of the transitions of a phenomenon that develops over several stages (e.g., A, B, C, D). The counterfactual analysis can be useful to check for differences between different transitions (e.g., direct paths skipping intermediate transitions or progressive sequential paths). Several practical applications may be mentioned of this type, such as vehicular platooning [7] and predictive maintenance [8].

This work was supported in part by REXASI-PRO H-EU project, call HORIZON-CL4-2021-HUMAN-01-01, Grant agreement ID: 101070028. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

[†] A. Carlevaro and M. Lenatti contributed equally to the development of the article. (*Corresponding author*: A. Carlevaro.)

A.C., M.L., A.P, and M.M. are with Cnr-Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, 00185 Rome, Italy (e-mail: alberto.carlevaro@ieiit.cnr.it, marta.lenatti@ieiit.cnr.it, alessia.paglialonga@ieiit.cnr.it, maurizio.mongelli@ieiit.cnr.it). A.C. is with University of Genoa, Department of Electrical, Electronics and Telecommunications Engineering and Naval Architecture (DITEN), 16145 Genoa, Italy.

B. Contribution

The objective of this paper is to develop a novel method based on Support Vector Data Description under multi-class setting (MC-SVDD) to identify multiple counterfactual explanations from a given observation under varying constraints. The use of SVDD envelopes may provide several advantages, e.g. detection of anomalous points (outside SVDD clusters) and flexible contour of different classes, by including the control of false positives/false negatives rates [6]. To the best of our knowledge, this is the first work aimed at the generation of counterfactuals for multi-class classification problems based on data envelopes extracted via SVDD. The method developed in this study addresses: **1)** explainability, through the use of of counterfactuals, **2)** controllability of counterfactuals via MC-SVDD and **3)** validation of counterfactual quality.

II. Related Works

A. MC-SVDD

Multi-class classification is the task of classifying a new instance into one among at least three classes. As always, when the variability of a problem increases, so does the effort to solve it. There exist different approaches to address the increase of the classes. For example, some algorithms, such as decision trees and Neural Networks, automatically handle multiple outputs. Other algorithms provide exclusively binary outputs (e.g., SVM, logistic regression, perceptron). In these cases, binary classifiers must be adapted to handle multiple outputs. Therefore, we can distinguish two types of multi-class classification techniques [9]: *one-vs-one* and *one-vs-rest*. In one-vs-one techniques the problem is divided into $\frac{m(m-1)}{2}$ binary classifiers, where m is the number of classes and each binary classifier predicts a class label. Then, an instance is assigned to the class with the highest number of counts. In one-vs-rest techniques, instead, the model is trained for m different datasets, where each target class is trained against the rest of the classes. Then, an instance is assigned to the class with the highest probability. Due to the their incremental adaptation to multiple outputs, these approaches lack a comprehensive view of relationship in among the classes. In addition, due to multiple trainings, they are not feasible for large datasets. The approach here proposed *Multi-ClassSVDD* (MC-SVDD)*, solves the problem in one shot, without repetitive adaptations. All uncertainties and data characteristics are handled at the same time, providing a result that best fits the problem [10]. The algorithm generalizes the well-known SVDD by Tax and Duin [11] to the multi-class case, quite naturally as an extension of the original method. Other attempts address multi-class SVDD, but identifying objects belonging to multiple anomalies rather than providing canonical classification. The algorithm proposed by [12] generalizes the unsupervised one class classifier of [13] to multiple outputs. A different approach is proposed by [14], in which the canonical SVDD is merged with binary tree to handle the multi classification problems. Guo et al. [15] proposed a multi kernel learning adaptation to SVDD (MKL-SVDD) to design the kernel weights for multiple kernels and obtain the optimal kernel combination. Hou et al. [16] developed a multi-class SVDD algorithm to classify multiple classes of planetary gear faults based on the method proposed by [17] that minimizes the radius of each hypersphere, while maximizing the distance between them. However, the boundary between couples of classes is optimized for each pair of centers, without including further constraints inherent to the other classes. Recently, a generalization of SVDD to the multi-class case has been proposed [18], but the focus is on anomaly detection.

B. Counterfactual explanations

Following the XAI taxonomies suggested in the literature (e.g., [1]), counterfactual explanations can be defined as local post-hoc XAI techniques, either model-specific or model agnostic, depending on their generation process. Counterfactuals generation methods may be designed to handle different data types like tabular data, images or text and may deliver explanations in different forms including numerical values, regions of pixels and linguistic expressions, as remarked in a recent survey [4]. In a previous work [5], we introduced a method to generate counterfactual explanations for tabular data based on sampled classification regions defined by a Two-Class Support Vector Data Descriptor (TC-SVDD). The method was then extended in [6] and applied to provide clinical recommendations for Type 2 diabetes risk reduction, showing a better counterfactual quality, in terms of availability and similarity, with respect to Diverse Counterfactual Explanations (DiCE) [19]. The present paper extends the analysis with respect to the multi-class problem, as described in Sections III and IV.

III. MC-SVDD: Multi-Class Support Vector Data Description

The training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is composed by m classes of objects of different sizes $n_1, n_2, \dots, n_m$ $(n_1 + n_2 + \dots + n_m = n)$, labelled according to their class

$$\mathbf{y} = \begin{bmatrix} 1 & \dots & 1 & 2 & \dots & 2 & \dots & m & \dots & m \end{bmatrix}^\top.$$

In order to find the m hyperspheres with minimum total volume, we should minimize the total volume of the m hyperspheres with the constraint that, for each object, (i) the distance between the center of one hypersphere and the object is smaller than the radius of that hypersphere (i.e., the object belongs to a specific output class) and (ii) the object should not fall into other hyperspheres (i.e., the object should not belong to other output classes).

Let $\mathbf{a}_k$ and R_k denote the center and radius of the hypersphere k. To allow a flexible description of the hyperspheres we introduce $\varphi : \mathcal{X} \longrightarrow \mathcal{V}$, a *feature map* from the space of the input features $\boldsymbol{x} \in \mathcal{X}$ to an higher dimensional inner product space $\mathcal{V}$.

Searching for hyperspheres of minimum volume that satisfy

*https://github.com/AlbiCarle/MultiClass_SVDD.git the above constraints means finding the solution of the following optimization problem

$$\min\ F(R_k;\mathbf{a}_k)=\sum_{k=1}^{m}R_k^2 \tag{1a}$$

$$\text{s.t.}\quad \left\|\varphi(\boldsymbol{x}_i^k)-\mathbf{a}_k\right\|^2 \le R_k^2,\ i\in[n_k],\forall k \tag{1b}$$

$$\left\|\varphi(\boldsymbol{x}_i^k)-\mathbf{a}_h\right\|^2 \ge R_h^2,\ i\in[n_k],\forall h\neq k \tag{1c}$$

We can follow the classical approach as in [11], which consists in reducing (1) to a quadratic programming problem. To allow for the possibility of outliers in the training set, the distance from an object belonging to class k, $\varphi(\boldsymbol{x}_i^k)$, to its own centre $\mathbf{a}_k$ should not be strictly smaller than R_k^2 , but larger distances should be penalized, and the distance from $\varphi(\boldsymbol{x}_i^k)$ to the other centres $\mathbf{a}_h$, $h{\neq}k$, should not be strictly larger than R_h^2, i.e. smaller distances should be penalized. Therefore we introduce slack variables $\boldsymbol{\xi}^{kk}\ge 0,\boldsymbol{\xi}^{kh}\ge 0$ and the minimization problem changes into

$$\min\ F(R_k;\mathbf{a}_k;\xi^{kh})=\sum_{k=1}^{m}R_k^2+\sum_{k=1}^{m}\sum_{h=1}^{m}C_{kh}\sum_{i=1}^{n_k}\xi_i^{kh} \tag{2a}$$

$$\text{s.t.}\quad \left\|\varphi(\boldsymbol{x}_i^k)-\mathbf{a}_k\right\|^2 \le R_k^2+\xi_i^{kk},\ i\in[n_k],\forall k \tag{2b}$$

$$\left\|\varphi(\boldsymbol{x}_i^k)-\mathbf{a}_h\right\|^2 \ge R_h^2-\xi_i^{kh},\ i\in[n_k],\forall h\neq k \tag{2c}$$

$$\text{and}\quad \xi_i^{kk}\ge 0\quad \forall k,\xi_i^{hk}\ge 0\quad \forall h\neq k \tag{2d}$$

where the parameter C_{kh} controls the misclassification error between the classes.

Now, we consider the dual problem of (2) by incorporating the constraints (2b) and (2c) into (2a) with the introduction of Lagrange multipliers

$$\begin{aligned}
&L(R_k;\mathbf{a}_k;\boldsymbol{\xi}^{kk},\boldsymbol{\xi}^{kh};\boldsymbol{\alpha}^{kk},\boldsymbol{\alpha}^{kh};\boldsymbol{\gamma}^{kk},\boldsymbol{\gamma}^{kh})\\
&=\sum_{k=1}^{m}R_k^2+\sum_{k=1}^{m}\sum_{h=1}^{m}C_{kh}\sum_{i=1}^{n_k}\xi_i^{kh}\\
&-\sum_{k=1}^{m}\sum_{i=1}^{n_k}\alpha_i^{kk}\left(R_k^2+\xi_i^{kk}-\left\|\varphi(\boldsymbol{x}_i^k)-\mathbf{a}_k\right\|^2\right)\\
&-\sum_{h\neq k}\sum_{i=1}^{n_h}\alpha_i^{kh}\left(\left\|\varphi(\boldsymbol{x}_i^k)-\mathbf{a}_h\right\|^2-R_h^2+\xi_i^{kh}\right)\\
&-\sum_{k=1}^{m}\sum_{i=1}^{n_k}\gamma_i^{kk}\xi_i^{kk}-\sum_{h\neq k}\sum_{i=1}^{n_h}\gamma_i^{kh}\xi_i^{kh}
\end{aligned} \tag{3}$$

with the Lagrange multipliers $\boldsymbol{\alpha}^{kk},\boldsymbol{\alpha}^{kh},\boldsymbol{\gamma}^{kk},\boldsymbol{\gamma}^{kh}\ge 0$ (4). In the dual form, L should be maximized with respect to the Lagrange multipliers so setting partial derivatives to zero gives the new constraints

$$\frac{\partial L}{\partial R_k}=0\Rightarrow\sum_{i=1}^{n_k}\alpha_i^{kk}-\sum_{h\neq k}\sum_{i=1}^{n_h}\alpha_i^{kh}=1 \tag{5}$$

$$\frac{\partial L}{\partial \mathbf{a}_k}=0\Rightarrow\mathbf{a}_k=\sum_{i=1}^{n_k}\alpha_i^{kk}\varphi(\boldsymbol{x}_i^k)-\sum_{h\neq k}\sum_{i=1}^{n_h}\alpha_i^{kh}\varphi(\boldsymbol{x}_i^h) \tag{6}$$

$\forall k\in[m]$ and $\forall h\neq k$. And with respect to the slack variables

$$\frac{\partial L}{\partial \xi_i^{ss}}=0\Rightarrow C_{ss}-\alpha_i^{ss}-\gamma_i^{ss}=0\Rightarrow 0\le\alpha_i^{ss}\le C_{ss} \tag{7}$$

$$\frac{\partial L}{\partial \xi_i^{st}}=0\Rightarrow C_{st}-\alpha_i^{st}-\gamma_i^{st}=0\Rightarrow 0\le\alpha_i^{st}\le C_{st} \tag{8}$$

$\forall s\in[m]$ and $\forall t\neq s$ respectively.

Substituting (5) and (6) in (4) the Lagrangian in the dual takes this form

$$\begin{aligned}
L\ =\ &\sum_{k=1}^{m}\sum_{i=1}^{n_k}\alpha_i^{kk}\left(\varphi(\boldsymbol{x}_i^k)\cdot\varphi(\boldsymbol{x}_i^k)\right)\\
-\ &\sum_{h\neq k}\sum_{i=1}^{n_k}\alpha_i^{kh}\left(\varphi(\boldsymbol{x}_i^k)\cdot\varphi(\boldsymbol{x}_i^k)\right)\\
-\ &\sum_{i=1}^{m}\sum_{i,j=1}^{n_k}\alpha_i^{kk}\alpha_j^{kk}\left(\varphi(\boldsymbol{x}_i^k)\cdot\varphi(\boldsymbol{x}_j^k)\right)\\
-\ &\sum_{h\neq k}\sum_{i,j=1}^{n_k}\alpha_i^{kh}\alpha_j^{kh}\left(\varphi(\boldsymbol{x}_i^k)\cdot\varphi(\boldsymbol{x}_j^k)\right)\\
+\ &2\sum_{h\neq k}\sum_{i=1}^{n_k}\sum_{j=1}^{n_h}\alpha_i^{kk}\alpha_j^{kh}\left(\varphi(\boldsymbol{x}_i^k)\cdot\varphi(\boldsymbol{x})_j^h\right)
\end{aligned} \tag{9}$$

The maximization of (10) under the constraints (4)-(5) and (7)-(8) gives the set of $\boldsymbol{\alpha}^{kk},\boldsymbol{\alpha}^{kh}$ $\forall k\in[m]$, $\forall h\neq k$ ($\boldsymbol{\gamma}^{kk}$ and $\boldsymbol{\gamma}^{kh}$ can be eliminated by exploiting their positivity and the first-order conditions on the slack variables).

Depending on the position of the training objects in the feature space, the Lagrange multipliers take on different values in the way the training objects do or do not satisfy the constraints (2b) and (2c)

$$\begin{aligned}
\left\|\varphi(\boldsymbol{x}_i^k)-\mathbf{a}_k\right\|^2<R_k^2\quad&\Rightarrow\quad\alpha_i^{kk}=0\\
\left\|\varphi(\boldsymbol{x}_i^k)-\mathbf{a}_h\right\|^2>R_h^2\quad&\Rightarrow\quad\alpha_i^{kh}=0\\
\left\|\varphi(\boldsymbol{x}_i^k)-\mathbf{a}_k\right\|^2=R_k^2\quad&\Rightarrow\quad 0<\alpha_i^{kk}<C_{kk}\\
\left\|\varphi(\boldsymbol{x}_i^k)-\mathbf{a}_h\right\|^2=R_h^2\quad&\Rightarrow\quad 0<\alpha_i^{kh}<C_{kh}\\
\left\|\varphi(\boldsymbol{x}_i^k)-\mathbf{a}_k\right\|^2>R_k^2\quad&\Rightarrow\quad\alpha_i^{kk}=C_{kk}\\
\left\|\varphi(\boldsymbol{x}_i^k)-\mathbf{a}_h\right\|^2<R_h^2\quad&\Rightarrow\quad\alpha_i^{kh}=C_{kh}
\end{aligned} \tag{10}$$

$\forall k\in[m]$ and $\forall h\neq k$ respectively.

Then, according with the literature around SVDD [11], the objects $\boldsymbol{x}_i^k$ with $\alpha_i^{kk}>0$ and $\alpha_i^{kh}>0$ are called *Support Vectors* (SVs) for the Class k.

By definition, (11), the radius R_k is the distance from the center $\mathbf{a}_k$ of the hypersphere to any of the SVs of Class k with Lagrange multipliers strictly minor than the parameters

$C_{k\{\cdot\}}$. Therefore

$$\begin{aligned}
R_k^2 &= \left\|\varphi(\boldsymbol{x}_s^k)-\mathbf{a}_k\right\|^2 = \big(\varphi(\boldsymbol{x}_s^k)\cdot\varphi(\boldsymbol{x}_s^k)\big)\\
&- 2\sum_{i=1}^{n_k}\alpha_i^{kk}\big(\varphi(\boldsymbol{x}_s^k)\cdot\varphi(\boldsymbol{x}_i^k)\big)\\
&+ 2\sum_{h\neq k}\sum_{i=1}^{n_h}\alpha_i^{kh}\big(\varphi(\boldsymbol{x}_s^k)\cdot\varphi(\boldsymbol{x}_i^h)\big)\\
&+ \sum_{i,j=1}^{n_k}\alpha_i^{kk}\alpha_j^{kk}\big(\varphi(\boldsymbol{x}_i^k)\cdot\varphi(\boldsymbol{x}_j^k)\big) \qquad (11)\\
&- 2\sum_{h\neq k}\sum_{i=1}^{n_k}\sum_{j=1}^{n_h}\alpha_i^{kk}\alpha_j^{kh}\big(\varphi(\boldsymbol{x}_i^k)\cdot\varphi(\boldsymbol{x}_j^h)\big)\\
&+ \sum_{h\neq k}\sum_{i,j}^{n_h}\alpha_i^{kh}\alpha_j^{kh}\big(\varphi(\boldsymbol{x}_i^h)\cdot\varphi(\boldsymbol{x}_j^h)\big)
\end{aligned}$$

for any SVs $\varphi(\boldsymbol{x}_s^k)$ of Class k with $0<\alpha_i^{kk}<C_{kk}$ or $0<\alpha_i^{kh}<C_{kh}$, for $h\neq k$.
To test an object $\mathbf{z}$ it is necessary to calculate its distance from the centre of the hypersphere k, i.e.

$$\begin{aligned}
d_k &\doteq \left\|\mathbf{z}-\mathbf{a}_k\right\|^2\\
&= \big(\varphi(\mathbf{z})\cdot\varphi(\mathbf{z})\big) - 2\sum_{i=1}^{n_k}\alpha_i^{kk}\big(\varphi(\mathbf{z})\cdot\varphi(\boldsymbol{x}_i^k)\big)\\
&+ 2\sum_{h\neq k}\sum_{i=1}^{n_h}\alpha_i^{kh}\big(\varphi(\mathbf{z})\cdot\varphi(\boldsymbol{x}_i^h)\big)\\
&+ \sum_{i,j=1}^{n_k}\alpha_i^{kk}\alpha_j^{kk}\big(\varphi(\boldsymbol{x}_i^k)\cdot\varphi(\boldsymbol{x}_j^k)\big) \qquad (12)\\
&- 2\sum_{h\neq k}\sum_{i=1}^{n_k}\sum_{j=1}^{n_h}\alpha_i^{kk}\alpha_j^{kh}\big(\varphi(\boldsymbol{x}_i^k)\cdot\varphi(\boldsymbol{x}_j^h)\big)\\
&+ \sum_{h\neq k}\sum_{i,j}^{n_h}\alpha_i^{kh}\alpha_j^{kh}\big(\varphi(\boldsymbol{x})_i^h\cdot\varphi(\boldsymbol{x}_j^h)\big)
\end{aligned}$$

a test object $\mathbf{z}$ is accepted by the following criterion:

1) If $d_k\leq R_k^2$ and $d_k>R_h^2$ $\forall h\neq k$, then $\mathbf{z}$ belongs to Class k;
2) If $d_k\leq R_k^2$ and $d_h<d_k$ $\forall h\neq k$, then $\mathbf{z}$ belongs to Class h;
3) If $d_k>R_h^2$ $\forall h$, then $\mathbf{z}$ is unclassified.

That is, the distances between all samples in each class and the center should be smaller than the radius of the corresponding hypersphere and the distances between all samples in each class and the centers of other classes should be larger than the radius of the corresponding hypersphere. And if a new sample belongs to more than a hypersphere, the sample is assigned to the class corresponding to the minimum distance. In any other case the sample is unclassified. Figure 1 clearly shows the behavior of the algorithm for linearly separated data: each sphere encloses the points related to an output class by minimizing its volume and excluding unclassified points.
It is worth to underline the following remarks:

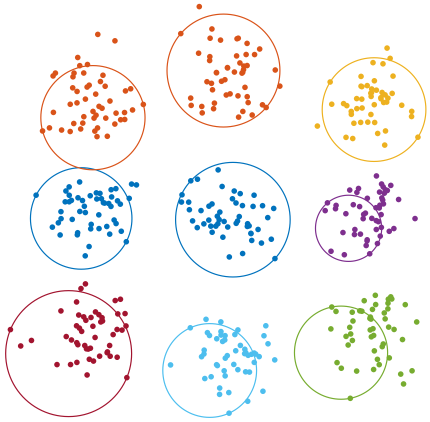

Figure 1. MC-SVDD applied to 9 classes extracted randomly from Gaussian distributions with different means and variances. Here, a linear MC-SVDD has been trained by fixing 9^2 parameters C_{kh} to control the trade-off between class covering and error between the classes.

Remark 3.1: In order to obtain a more compact form of the Lagrangian L, and to make it clear that the problem is quadratic, we define these quantities for all $k\in[m]$

$$\boldsymbol{\alpha}^k \doteq \left[\ \boldsymbol{\alpha}^{k1},\boldsymbol{\alpha}^{k2},\ldots,\boldsymbol{\alpha}^{km}\ \right]^\top,\quad \boldsymbol{\alpha}\doteq\left[\ \boldsymbol{\alpha}^1,\boldsymbol{\alpha}^2,\ldots,\boldsymbol{\alpha}^m\ \right]^\top$$

$$\mathbf{y}^k = \left[\ y_1^k\quad y_2^k\quad\ldots\quad y_n^k\ \right]^\top,$$

where $y_i^k=\begin{cases}+1 & \text{if } y_i=k\\ -1 & \text{if } y_i\neq k\end{cases}\quad\forall i\in[n].$

Defined then, for all $k\in[m]$

$$\begin{aligned}
\Phi_k &\doteq \left[\ \varphi(\boldsymbol{x}_1^k)\quad\varphi(\boldsymbol{x}_2^k)\quad\ldots\quad\varphi(\boldsymbol{x}_n^k)\ \right], && (13)\\
D_k &\doteq \operatorname{diag}\{y_1^k,y_2^k,\ldots,y_n^k\}, && (14)\\
K_k &\doteq \Phi_k^\top\Phi_k, && (15)
\end{aligned}$$

and $K_{i,j}=K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\varphi(\boldsymbol{x}_i)^\top\varphi(\boldsymbol{x}_j),\ i\in[n],j\in[n]$, is the kernel matrix which satisfies the Mercer's theorem [20]. Then let them be

$$\begin{aligned}
H_k &\doteq 2D_kK_kD_k,\\
f_k &\doteq D_k\operatorname{diag}(K_k).
\end{aligned}$$

Finally, defining

$$H \doteq \begin{pmatrix} H_1 & & & \\ & H_2 & & \\ & & \ddots & \\ & & & H_m\end{pmatrix},\ f\doteq\begin{bmatrix} f_1\\ f_2\\ \cdots\\ f_m\end{bmatrix}$$

we obtain that the Lagrangian L, (10), can be rewritten as

$$L=-\frac{1}{2}\boldsymbol{\alpha}^\top H\boldsymbol{\alpha}+f^\top\boldsymbol{\alpha}, \qquad (16)$$

i.e. L is a quadratic form, that can be easily maximized with a quadratic optimizer.

IV. MUCH: MULTI COUNTERFACTUAL VIA HALTON SAMPLING

A dataset $\mathcal{D}$ can be described by a subset of modifiable features $\mathbf{u}$ and a subset of non-modifiable features $\mathbf{z}$. As a consequence, an observation $\boldsymbol{x} \in \mathcal{D}$ can be defined as

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

MC-SVDD is applied to obtain m classification regions defined as follows:

$$S_i \doteq \{\boldsymbol{x} \in \mathbb{R}^N : \|\boldsymbol{x} - \mathbf{a}_i\|^2 \leq R_i^2, \|\boldsymbol{x} - \mathbf{a}_j\|^2 \geq R_j^2; \\ j \in [m]; j \neq i\} \tag{17}$$

where $R_i^2, R_j^2, \mathbf{a}_i, \mathbf{a}_j$ represent the radii and the centers of the spheres, as defined in Section III. Once the m classification regions are defined, the search for a counterfactual explanation of an observation $\boldsymbol{x}_f \in S_i$, called *factual*, consists of determining the minimum joint variation $\Delta\mathbf{u}^*$ of the modifiable variables to obtain the closest observation

$$\boldsymbol{x}^*_{f,j} \doteq (\mathbf{u} + \Delta\mathbf{u}^*, \mathbf{z})_{f,j} \tag{18}$$

that belongs to class S_j different from the original class S_i. Specifically, $\Delta\mathbf{u}^*$ is estimated by solving the following minimization problem:

For all $j \in [m], j \neq i$

$$\begin{aligned} \min_{\Delta\mathbf{u} \in \mathbb{R}^p} \quad & d\big(\boldsymbol{x}_f, (\mathbf{u} + \Delta\mathbf{u}, \mathbf{z})_{f,j}\big) && \text{(19a)} \\ \text{subject to} \quad & \|(\mathbf{u} + \Delta\mathbf{u}, \mathbf{z})_{f,j} - \mathbf{a}_j\|^2 \leq R_j^2 && \text{(19b)} \\ & \|(\mathbf{u} + \Delta\mathbf{u}, \mathbf{z})_{f,j} - \mathbf{a}_k\|^2 \geq R_k^2, && \text{(19c)} \\ & \text{with } k \in [m] \text{ and } k \neq j \end{aligned}$$

where d is the selected distance metrics (e.g., the Euclidean norm), (19b) constraints $\boldsymbol{x}^*$ to lie inside S_i and (19c) constraints $\boldsymbol{x}^*$ to lie outside all the regions $S_k \neq S_j$. It is worth noting that, for each factual $\boldsymbol{x}_f \in S_i$, we can find a set $\mathbf{C}_f = \{\mathbf{x}^*_{f,j} \mid j \in [m]; j \neq i\}$ of $m-1$ counterfactual explanations, that is, one for each class j different from i. In other words, for a set of factuals $\mathbf{F}_i$ we obtain a set of counterfactual explanations $\mathbf{E}$ with maximum size $(|\mathbf{F}_i|, m-1)$.

A. Numerical solution

Since each S_j theoretically includes an infinite set of real points, a numerical approximation is necessarily introduced whereby counterfactual explanations are sought in a sampled region obtained by applying quasi-random Halton sampling [21][*]. Since counterfactual explanations are searched among a finite set of points, the availability and minimality of each explanation depends on the density of the sampling. However, the higher the number of points in the sampled region, the higher the computational cost. As a consequence, a trade-off between accuracy and runtime must be reached. Counterfactual explanations are extracted for each factual observation belonging to each class. Once a factual $\boldsymbol{x}_f \in \mathbf{F}_i$, $i \in [m]$ is defined, the algorithm returns the set of counterfactuals $\mathbf{C}_f$, i.e., each counterfactual $\boldsymbol{x}^*_{f,j}, j \in [m] \setminus \{i\}$.

Algorithm 1 MUCH

Dataset $\mathcal{D}$ is divided in training set $\mathcal{D}_{tr}$ and validation set $\mathcal{D}_{vl}$.

A MC-SVDD is performed on $\mathcal{D}_{tr}$ and validated on $\mathcal{D}_{vl}$, getting $S_1, S_2, \ldots, S_m$. A set of factuals related to the class i, $\mathbf{F}_i$, is chosen.

```
1         $\mathbf{C_{F_i}} = [\,]$
2         for $\mathbf{x}_f = (\mathbf{u}_f, \mathbf{z}_f) \in \mathbf{F_i}$
2.1           $\mathbf{C}_f = [\,]$
2.2           for $j \in [m]$, $j \neq i$
2.2.1             Sample quasi-randomly $\tilde{S}_j$
2.2.2             $d_f = d\big(\boldsymbol{x}_f, \tilde{S}_{j_{|\mathbf{z}=\mathbf{z_f}}}\big)$
2.2.3             $\boldsymbol{x}'_f = \min(d_f)$
2.2.4             if$\big(\boldsymbol{x}_f \in S_i \ \&\ \boldsymbol{x}'_f \in S_j\big)$
2.2.4.1               $\mathbf{C}_f = \mathbf{C}_f \cup \{\mathbf{x}'_f\}$
2.2.5             end
2.2.6             $\mathbf{C_{F_i}} = \mathbf{C_{F_i}} \cup \mathbf{C}_f$
2.3           end
2.4       end
3         return $\mathbf{C}_{\mathbf{F}_i}$
```

The first step of the MUCH algorithm[†] (MUltiCounterfactual via Halton sampling) (**Algorithm 1**) is the classification of data by MC-SVDD, which defines m regions S_i, $i \in [m]$, into which data are classified. The MC-SVDD algorithm is trained on $\mathcal{D}_{tr}$ and validated on $\mathcal{D}_{vl}$, each belonging to the same probability distribution of the data, recovering the best classification after hyperparameter tuning. Then, for each region S_i a randomly sampled region $\tilde{S}_i$ is constructed: this region is the one designated to the numerical search for counterfactuals of class $j \neq i$, i.e., for each factual $\boldsymbol{x}_f$ the respective counterfactual related to the class $j \neq i$, $\boldsymbol{x}^*_{f,j}$, is searched in $\tilde{S}_j$. Among all points in the sampled region $\tilde{S}_j$, the one that minimizes the distance d w.r.t factual $\boldsymbol{x}_f$ is chosen. The distance d plays a key role in the search for counterfactuals as changing the distance may changes the returned counterfactuals. The most natural choice of distance is the distance induced by the classification kernel:

$$d(\boldsymbol{x}, \mathbf{y}) = k(\boldsymbol{x}, \boldsymbol{x}) - 2k(\boldsymbol{x}, \mathbf{y}) + k(\mathbf{y}, \mathbf{y}).$$

The reason for this choice is motivated by the fact that the topology defined by the kernel in the classification affects the relationship between the points in the sampled regions, so keeping the same distance relationship would help the algorithm find the best counterfactual explanation.

The estimation of the computational cost of MUCH can be easily retrieved from the computational cost of the binary counterfactual generator algorithm proposed in [5]. Denoting with n the number of points, with d the number of features and m the number of classes, the

[*] Halton is a low discrepancy sequence generator; other generators of this type, such as Sobol, may be applicable in the sampling step of the algorithm.

[†] https://github.com/AlbiCarle/MUCH.git

computational cost of MC-SVDD, that is, $O(\text{MC-SVDD})$ is estimated in $O\left(m\left(\max(n,d)\min(n,d)^2\right)\right)$. In accordance with [5], the computational cost related to the counterfactuals search, for each set of factuals $\mathbf{F}_i$, is $O\left(\max\left(\sum_{j\neq i} q_j, |\mathbf{F}_i|\max\left(D, \sum_{j\neq i}\tilde{s}_j\right)\right)\right)$, where $O(q_j)$ is the computational cost of the random sampling of $\tilde{S}_j$ [22], $O(D)$ is the computational cost for the computation of the distance d [23] and $O(\tilde{s}_j)$ is the computational cost of the research of the minimum of the vector of distances relative to the $j-$th random sampling ($\tilde{S}_j$) [24]. So, considering all m classes, the computational cost of the counterfactuals search, $O\left(\text{SCF}\right)$, can be estimated in $O\left(m\left(\max\left(\sum_{j\neq i} q_j, |\mathbf{F}_i|\max\left(D, \sum_{j\neq i}\tilde{s}_j\right)\right)\right)\right)$. Finally, the total computational cost of MUCH can be estimated with $O\left(\text{MUCH}\right) = O\left(\max(\text{MC-SVDD, SCF})\right)$. The complete procedure for generation of a set of explanations is summarized in Fig. 2.

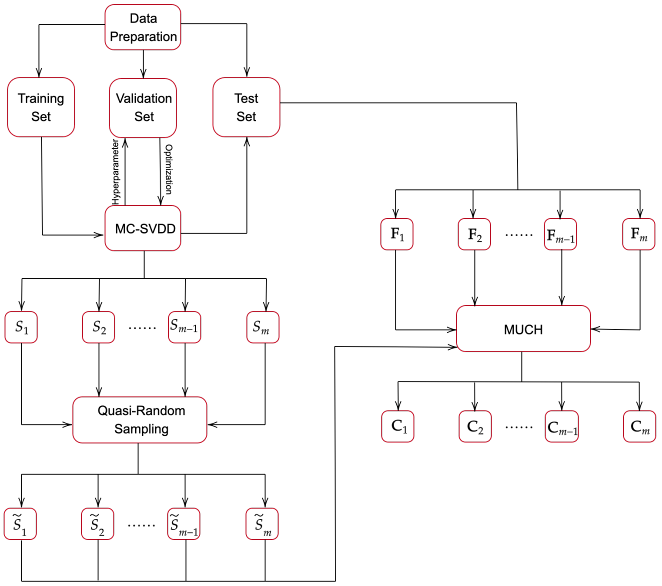

Figure 2. Diagram of the counterfactual extraction procedure.

B. Counterfactual quality

As reported in a recent review by Guidotti [4], counterfactual explanations should fulfill a set of ideal properties and adherence to these properties shall be assessed, for a set of factuals, in terms of appropriate evaluation metrics such as availability, actionability, similarity and discriminative power. *Availability* measures the number of counterfactuals actually returned by the counterfactual explainer for each class and it can be measured as the ratio between the number of counterfactuals of class j i.e., $|\mathbf{E}_j|$ and the total number of factuals of class i, i.e., $|\mathbf{F}_i|$. *Actionability* measures the ability of counterfactual explanations to vary only modifiable features and it is calculated, for each class j, as the ratio of the number of constrained features and the total number of non-modifiable features i.e., $|\mathbf{z}|$. *Similarity* evaluates the average distance (e.g., Euclidean) between each factual in $\mathbf{F}_i$ and the corresponding counterfactual explanations in $\mathbf{E_j}$. In order to be similar, the distance between these two points should be lower than a fixed threshold ε. To evaluate similarity, data points were normalized between 0 and 1 and the computed distance was compared to the maximum theoretical distance in the standardized modifiable-feature space (i.e., $\sqrt{|\mathbf{u}|}$) and represented in terms of average and 95% confidence interval (C.I.). Finally, *discriminative power* measures the ability to distinguish points of the factual class in S_i from counterfactuals in $\mathbf{E}_j$. This metrics was estimated in this study by evaluating the accuracy of a k-Nearest Neighbor (KNN) classifier trained on a dataset including the counterfactuals in $\mathbf{E}_j$ and real data points in S_i. Discriminative power was then computed as the average test accuracy obtained with 5-Fold cross-validation.

In a multi-class classification problem, such as the one considered in this paper, where $|\mathbf{C}_f| > 1$, each evaluation metric can be considered as the average value obtained across the $m-1$ set of counterfactuals.

V. Applicative example: the FIFA dataset

A. Dataset description

FIFA is one of the most famous football videogames in the world. The FIFA dataset[‡] includes latest edition FIFA attributes related to more than 17000 players from different football leagues. In this study, a subset of 50 attributes were selected from the initial set of 89 attributes. Specifically, the attributes related to the player's physical and athletic characteristics were retained, whereas those not relevant (e.g., team, graphical visualization) were discarded. Besides age, height and weight, the selected attributes can be summarized in three main categories: mental, physical and technical skills. These attributes depict different aspects of the player's individual abilities and they are usually represented in terms of rating, on a scale from 1 to 100. Attributes can be grouped in three categories, according to the ability to which they relate: Mental, Physical and Technical Skills. Moreover, the main attributes can be combined in 6 fundamental attributes, namely *Pace* (55% sprint speed, 45% acceleration), *Shooting* (ability to score: 45% finishing, 20% shot power, 20% long shots, 5% penalties, 5% positioning, 5% volleys), *Passing* (capability to successfully pass the ball to other teammates:35% short passing, 20% vision, 20% crossing, 15% long passing, 5% curve, 5% free kick accuracy), *Dribbling* (50% dribbling, 35% ball control, 10% agility, 5% balance), *Defending* (ability to intercept the ball and mark the opponent: 30% marking, 30% sliding tackle, 20% interception, 10% heading accuracy, 10% sliding tackle) and *Physical* (50% strength, 25% stamina, 20% aggression, 5% jumping). These key attributes can be directly derived from the others, and for this reason, only the 44 secondary attributes were considered as input features. Given these input attributes, the classification task consisted in predicting the correct player's position among 4 possible classes: *Midfielder* (MF),

[‡] Retrieved [November 2022] from https://www.kaggle.com/datasets/cashncarry/fifa-23-complete-player-dataset

Defender (DE), *Forward* (FO), or *Goalkeeper* (GK). To obtain a balanced dataset, 2000 records were extracted for each player's position (8000 records in total). The dataset was then splitted in training set (70%, 5600 records) and test set (30%, 2400 records). The parameters of MC-SVDD were optimized by performing a cross validation on the training set, as explained in Section III. MC-SVDD with the best combination of hyperparameters was then tested on the remaining data. Table I shows the MC-SVDD training and test classification performance.

Table I
CLASSIFICATION PERFORMANCE: FIFA DATASET

	%OUT	ACC	F1-SCORE	Cohen's Kappa
Training	0.59%	78.03%	73.08%	0.71
Test	1.25%	77.50%	72.99%	0.70

Specifically, the performance was evaluated in terms of classification accuracy, macro-averaged F1-score (i.e., the mean of F1-scores computed by class), Cohen's Kappa Coefficient [25] (i.e., the level of agreement between ground truth and predicted values) and the percentage of unclassified points (i.e., points lying outside all m SVDD regions). Accuracy and F1-SCORE are satisfactory as they are both are above 72%, moreover there is no presence of overfitting as these values remain stable even when the model is applied to test data. The percentage of unclassified points is really small, meaning that the spherical regions identified by MC-SVDD are able to enclose almost all points and the presence of anomalous points in the selected dataset is limited.

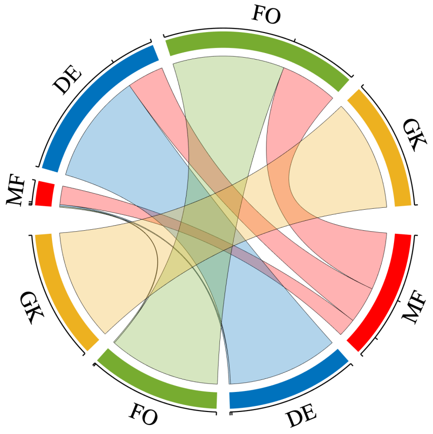

Figure 3. Chord diagram representation of the confusion matrix corresponding to the classification of the FIFA testing dataset.

Table II
AVAILABILITY (%), SIMILARITY(%), AND DISCRIMINATIVE POWER (MEAN% AND C.I.%) OF COUNTERFACTUALS GENERATED FROM FIFA DATASET, FOR DIFFERENT FACTUALS CLASSES

	FIFA			
Factual Class	MF	DE	FO	GK
C1 Class	DE	MF	MF	MF
Availability	100.00%	100.00%	100.00%	100.00%
Similarity (Mean)	21.73%	21.38%	21.39%	40.14%
Similarity (C.I.)	13.49% 29.96%	12.74% 30.02%	13.48% 29.31%	35.80% 44.48%
C2 Class	FO	FO	DE	DE
Availability	100.00%	100.00%	100.00%	100.00%
Similarity (Mean)	23.35%	24.05%	24.34%	38.21%
Similarity (C.I.)	15.80% 30.89%	16.94% 31.17%	16.65% 32.04%	34.11% 42.31%
C3 Class	GK	GK	GK	FO
Availability	100.00%	100.00%	100.00%	100.00%
Similarity (Mean)	40.13%	36.66%	37.60%	41.48%
Similarity (C.I.)	30.65% 49.61%	27.71% 45.62%	28.45% 46.75%	36.95% 46.01%
Discriminative Power	95.58%	98.27%	98.89%	99.84%

Table III
CLASSIFICATION PERFORMANCE: IRIS AND STELLAR DATASETS.

	IRIS	Stellar classification
ACC$_{\mathbf{tr}}$	95.24%	93.83 %
OUT$_{\mathbf{tr}}$	0.00%	0.01%
ACC$_{\mathbf{ts}}$	97.78 %	92.11 %
OUT$_{\mathbf{ts}}$	0.00%	0.02%
Macro F1-SCORE$_{\mathbf{ts}}$	97.78 %	94.18%
Cohen's Kappa$_{\mathbf{ts}}$	0.97	0.88

As it can be noticed from Fig. 3, classes DE, FO, and GK can be accurately classified. On the contrary, class MF is more difficult to discriminate. Indeed, the single class F1-score on the test set is more than acceptable when considering DE, FO and GK (i.e., 84.78%, 79.24%, and 100%, respectively), whereas it is noticeably lower when considering MF (27.96%). This is due to the fact that points in the MF class are easily confused with those in DE and FO classes as the characteristics of MF players are, in practice, intermediate between those of DE and FO players. It can also be observed that GK are perfectly distinguishable from footballers in other game positions, because of the peculiar skills that this kind of player must demonstrate.

B. Multi-counterfactuals generation

1) Setting: To evaluate the MUCH approach, a set of counterfactuals is generated starting from a set of points belonging to the test set. Specifically, given a player belonging to the chosen factual class and the corresponding set of attributes, the algorithm aims to find a counterfactual in each of the other classes, that is, to find the minimal changes in the player's attributes able to change his preferable position. Once $\mathbf{F}_i$ has been defined, a sufficiently large set of candidate counterfactuals is obtained by sampling 10000 points for each of the $m-1$ MC-SVDD regions using Halton sampling (see Section IV-A). As already mentioned, $\mathbf{F}_i$ is a set of test data points, but the corresponding counterfactuals explanations does not necessarily belong to the original dataset. Indeed, counterfactuals explanations as returned by the proposed algorithm are plausible combinations of features sampled inside the classification regions. Thus, the proposed approach is categorized as *exogenous* [4].

Age and *height* were considered as non-modifiable features, hence they were constrained during counterfactual search. Actually, counterfactuals have been accepted within a certain tolerance δ (i.e., $\delta = \pm 2cm$ for *height*) in order to ensure the availability of counterfactuals. Obviously, the smaller the delta, the greater is the probability that the algorithm will not return a counterfactual (i.e., lower availability), especially as the number of non-modifiable variables increases.

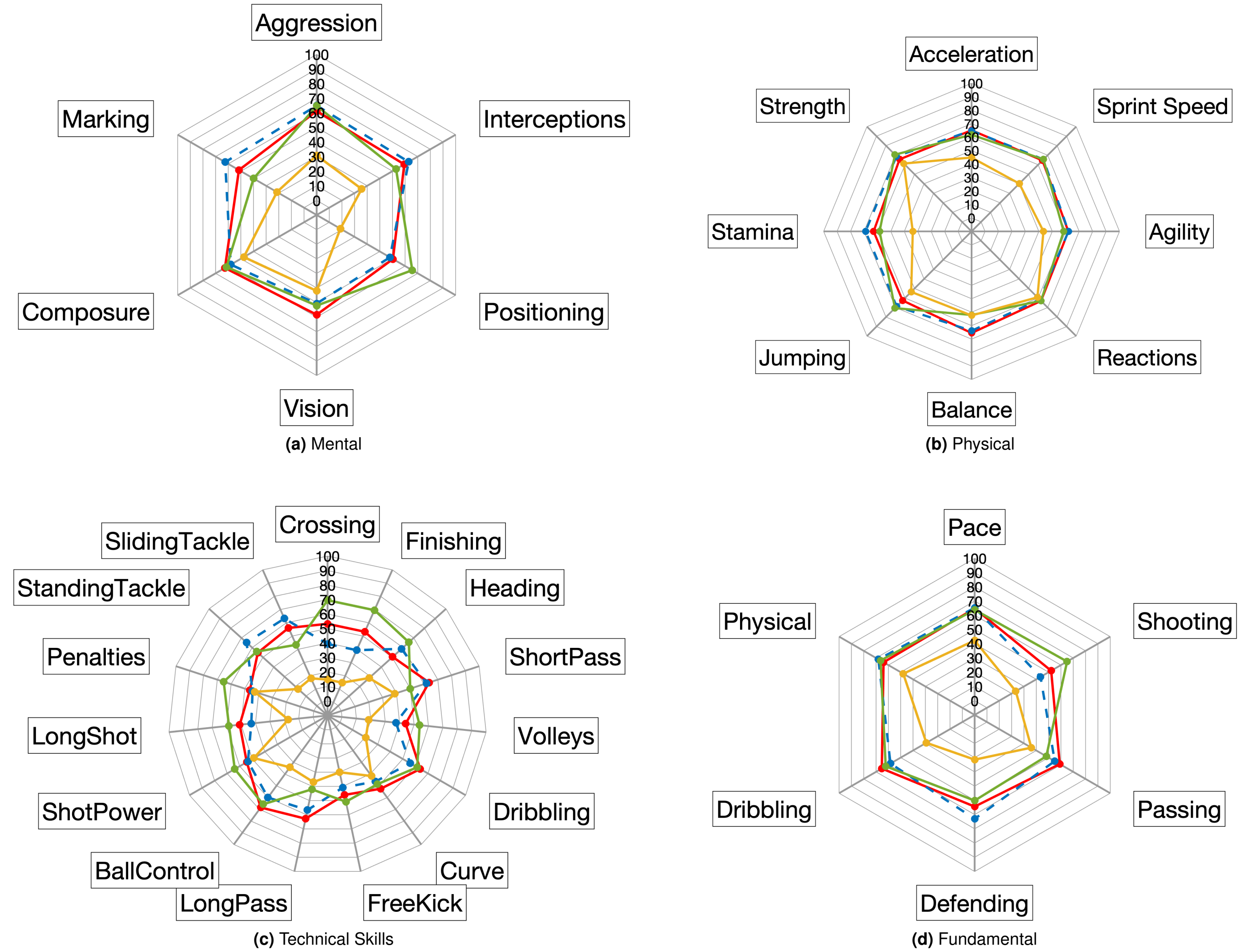

Figure 4. Each spiderplot represents the variation of the average of the factuals (dashed line) and counterfactuals (solid line) for DE class for each attribute category (Mental, Physical, Technical Skills and Fundamental). The value scale ranges from 0 to 100, and the output classes colors are the same as those used in Figure 3 (MF: red, DE: blue, FO: green, and GK: yellow).

Table IV

Availability (%), similarity(%), and discriminative power (mean% and C.I.%) of counterfactuals generated from IRIS and Stellar Classification datasets, for different factuals classes

	IRIS			Stellar classification		
Factual Class	1	2	3	1	2	3
C1 Class	2	1	1	2	1	1
Availability	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Similarity (Mean)	33.93%	28.77%	49.93%	39.14%	16.15%	14.91%
Similarity (C.I.)	27.80%	16.89%	38.72%	18.79%	3.72%	2.50%
	40.07%	40.66%	61.14%	59.49%	28.58%	27.33%
C2 Class	3	3	2	3	3	2
Availability	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Similarity (Mean)	39.93%	11.83%	19.19%	14.78%	17.40%	38.68%
Similarity (C.I.)	33.92%	1.38%	9.13%	3.25%	6.29%	19.61%
	45.95%	22.29%	29.25%	26.31%	28.51%	57.76%
Discriminative Power	100.00%	82.91%	91.99%	95.09%	98.16%	98.10%

2) Results: Table II lists the properties of the sets of counterfactuals (see Section IV-B for the definition) obtained for each different class of factuals $\mathbf{F}_i$. The discriminative power, for the different classes appears to be high, that is, above 95%, as shown in Table II. This indicates that counterfactuals, although searched at a minimum distance, are easily distinguishable from points belonging to the factual class. The highest discriminative power is computed with factuals belonging to the GK class, which, as previously mentioned, has more peculiar characteristics than the others. The algorithm successfully returned all counterfactuals (100% availability), demonstrating a sufficiently dense sampling of the SVDD regions. Lastly, similarity values are also satisfactory, with average values between 21% and 42%, depending on the factual class.

3) Knowledge extraction: The goal of the analysis is to identify which types of players are most characterized in their role and how different training plans can help specialize in a different role. For example, Fig. 4 analyzes the behaviour of the DE role, showing a spiderplot for each attribute category. It should be noted that the GK class differs significantly from the other classes. This is not surprising, since GK role requires different skills compared to the other roles. Concerning mental attributes, DE shows higher marking abilities than MF and FO. Moreover, DE positioning ability is similar to that of MF but remarkably lower than that of FO, whereas interceptions capabilities of DE are slightly higher than those of MF and FO. The remaining mental abilities present comparable values among DE, FO, and MF players. Physical attributes, instead, remain barely unchanged when considering DE, FO, and MF players. The only exception is the fact that DE and MF have on average greater balance than FO. Technical skills present different distributions when focusing on different classes of footballers. For example, DE short passing and long passing abilities are similar to those of MF and significantly higher than those of FO. Moreover, DE has higher values for both standing and sliding tackles than MF and FO. Intuitively, DE possesses worse abilities than FO when considering attributes strictly related to the attack phase including shot power, long shot, penalties, crossing, and finishing. Lastly, regarding the six fundamental attributes, on average DE, FO, and MF present comparable abilities in terms of pace, physical and dribbling abilities. Intuitively, DE players have higher defending abilities w.r.t MF and FO, and passing abilities intermediate between those of FO and MF. Reasonably, shooting capabilities are slightly lower than those of MF and strongly lower than those of a FO. After similar analysis of FO and MF spiderplots[§], the following conclusion arises. Workouts should be common on most abilities and strongly differentiated in target roles. For example, DE should focus on tackles and interceptions, FO on shooting and finishing, MF on passing. Other attributes, such as physical, aggressiveness, and dribbling, does not impact the specialization. Although such a conclusion may appear intuitive, it may be of extreme interest to help experts (tactical and athletic coaches) in the selection of the target variables.

VI. Characterization on Additional Datasets

This section discusses the performance of the proposed approach on a set of frequently referenced multi-class open source datasets, including the IRIS dataset [¶] and the Stellar Classification Dataset - SDSS17 dataset [‖]. These experiments help demonstrate that the approach can potentially scale well to tabular datasets of different size and different nature (i.e., physical measurements in the IRIS and SDSS17 datasets vs simulated play in the FIFA dataset).

The IRIS dataset consists of 150 observations related to peculiar characteristics of three different iris species (i.e., *Setosa*-1, *Versicolor*-2, and *Virginica*-3). Data records are equally balanced in terms of classes and records of the *Setosa* species are linearly separable from the others.

The Stellar Classification dataset includes 100,000 records of 3 type of objects (i.e., *galaxy*-1, *star*-2 or *quasar*-3) described by different spectral characteristics. Every observation consists of 17 input features, however only a subset of 10 features was considered in this experiment. Data records are equally balanced in terms of classes and records of the *Setosa* species are linearly separable from the others. Both datasets were split in training (70%) and test set (30%). Table III shows the training and test classification performance obtained by applying the MC-SVDD model, as presented in Section III. Specifically, the classification performance is summarized in terms of accuracy and percentage of unclassified points on both training and test sets, macro-averaged F1-score and Cohen's Kappa on the test set. Table IV shows the main properties of the set of counterfactuals obtained applying the method presented in Section IV to the two state-of-the art datasets. Since class 1 in the IRIS dataset is linearly separable from the other 2 classes, counterfactuals belonging to classes 2 and 3 are very easily distinguishable from class 1 points. Indeed, the discriminative power for factual class 1 is 100% for both classes of counterfactuals.

VII. Discussion and Conclusion

This work aims to formalize a multi-class generalization of an SVDD (MC-SVDD) and extract a set of counterfactual explanations from the classification results using a multi-class extension (MUCH) of a previously proposed counterfactuals explainer [5]. In order to be considered meaningful, a counterfactual should not only achieve the desired outcome minimizing the variation, but it should also be feasible, actionable, and retrieved fast enough. Experiments on three diverse datasets demonstrate that MC-SVDD is accurate in enclosing different classes of data points, with a negligible percentage of unclassified points. The use of a one-shot approach allows us to directly account for relationships and intersections between classes that would have been disregarded by considering multiple binary classifiers. Moreover, MUCH demonstrated satisfactory performance in terms of availability, similarity and discriminative power of the generated counterfactual explanations. The proposed MUCH approach has been applied starting from classification regions extracted from MC-SVDD, but it is in principle applicable to data regions derived from any machine learning method, like for example K-Nearest-Neighbors or rule based method. Future studies should therefore focus on proving that the MUCH generator is model-agnostic. Counterfactual explanations are minimal variations in input features that change the output class. This technique allows us to investigate the

[§] The corresponding spiderplots are available in: GitHub https://github.com/AlbiCarle/MUCH/blob/main/SpiderImages/FIFA_SpiderPlots.pdf.

[¶] Retrieved [December 2022] from https://www.kaggle.com/datasets/uciml/iris

[‖] fedesoriano. (January 2022). Stellar Classification Dataset - SDSS17. Retrieved [December 2022] from https://www.kaggle.com/fedesoriano/stellar-classification-dataset-sdss17 changes needed to move from the original class to a desired target class, as shown in Section V. Similarly, in cases where it makes no sense to talk about passing between classes, counterfactuals can be used to characterize a dataset through the analysis of the peculiar characteristics that differentiate one class from another, as shown in Section VI. Three datasets have been shown as an example, but obviously the presented approach can be applied in several domains, such as the medical one, for example to study the impact of certain risk factors on the development of one or more diseases and subsequent preventive strategies. Future studies will focus in this direction. Moreover, the presented method should be further extended to handle different kinds of data, for example text, including textual explanations.

Acknowledgment

Marta Lenatti is a PhD student enrolled in the National PhD in Artificial Intelligence, XXXVIII cycle, course on Health and life sciences, organized by Università Campus Bio-Medico di Roma.

References

[1] V. Belle and I. Papantonis, "Principles and practice of explainable machine learning," *Frontiers in Big Data*, vol. 4, 2021.

[2] "General data protection regulation (gdpr)." https://gdpr.eu/tag/gdpr/. [Retrieved December 16, 2022].

[3] S. Wachter, B. D. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Cybersecurity*, 2017.

[4] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, 2022.

[5] A. Carlevaro, M. Lenatti, A. Paglialonga, and M. Mongelli, "Counterfactual building and evaluation via explainable support vector data description," *IEEE Access*, vol. 10, pp. 60849–60861, 2022.

[6] M. Lenatti, A. Carlevaro, A. Guergachi, K. Keshavjee, M. Mongelli, and A. Paglialonga, "A novel method to derive personalized minimum viable recommendations for type 2 diabetes prevention based on counterfactual explanations," *PLOS ONE*, vol. 17, no. 11, 2022.

[7] M. Mirabilio, A. Iovine, E. De Santis, M. D. D. Benedetto, and G. Pola, "String stability of a vehicular platoon with the use of macroscopic information," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 9, pp. 5861–5873, 2021.

[8] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in *2008 International Conference on Prognostics and Health Management*, pp. 1–9, 2008.

[9] P. Mills, "Solving for multi-class: a survey and synthesis," 2018.

[10] P. D. Moral, S. Nowaczyk, and S. Pashami, "Why is multiclass classification hard?," *IEEE Access*, vol. 10, pp. 80448–80462, 2022.

[11] D. M. J. Tax and R. P. W. Duin, "Support vector domain description," *Pattern Recognition Letters*, vol. 20, pp. 1191–1199, 1999.

[12] G. Xie, Y. Jiang, and N. Chen, "A multi-class support vector data description approach for classification of medical image," in *2013 Ninth International Conference on Computational Intelligence and Security*, pp. 115–119, 2013.

[13] S. S. Khan and M. G. Madden, "One-class classification: taxonomy of study and review of techniques," *The Knowledge Engineering Review*, vol. 29, no. 3, pp. 345–374, 2014.

[14] L. Duan, M. Xie, T. Bai, and J. Wang, "A new support vector data description method for machinery fault diagnosis with unbalanced datasets," *Expert Systems with Applications*, vol. 64, pp. 239–246, 2016.

[15] W. Guo, Z. Wang, S. Hong, D. Li, H. Yang, and W. Du, "Multi-kernel support vector data description with boundary information," *Engineering Applications of Artificial Intelligence*, vol. 102, p. 104254, 2021.

[16] H. Hou and H. Ji, "Improved multiclass support vector data description for planetary gearbox fault diagnosis," *Control Engineering Practice*, vol. 114, p. 104867, 2021.

[17] J. Fang, W. Wang, X. Wang, Z. Long, D. Liang, and Q. Zhou, "A svdd method based on maximum distance between two centers of spheres," *Chinese Journal of Electronics*, vol. 21, no. 1, p. 107 – 111, 2012.

[18] M. Turkoz, S. Kim, Y. Son, M. K. Jeong, and E. A. Elsayed, "Generalized support vector data description for anomaly detection," *Pattern Recognition*, vol. 100, p. 107119, 2020.

[19] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, (New York, NY, USA), p. 607–617, Association for Computing Machinery, 2020.

[20] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 209, pp. 415–446, 1909.

[21] C. Cervellera, M. Gaggero, D. Macciò, and R. Marcialis, "Quasi-random sampling for approximate dynamic programming," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2013.

[22] S. Sen, T. Samanta, and A. Reese, "Quasi-versus pseudo-random generators: Discrepancy, complexity and integration-error based comparison," *Int J Innov Comput Info Control*, vol. 2, 2006.

[23] D. Burago, Y. D. Burago, and S. O. Ivanov, "A course in metric geometry," 2001.

[24] 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, p. 448–461, 1973.

[25] J. Cohen, "A coefficient of agreement for nominal scales," 1960.

Alberto Carlevaro received the Master Degree in Applied Mathematics from the University of Genoa with a physics-mathematics thesis. He is now a PhD student in the Department of Electrical, Electronic and Telecommunications Engineering and Naval Architecture (DITEN), in collaboration with CNR-IEIIT and S.M.E. Aitek. His current fields of research are Machine Learning, Deep Learning, Statistical Learning and Explainable AI.

Marta Lenatti received a Master Degree in Biomedical Engineering, from Politecnico di Milano in 2020. She is now a PhD student of the Italian National PhD program on Artificial Intelligence (Health and life sciences area) at University Campus Bio-Medico of Rome in collaboration with CNR-IEIIT, and a Visiting Scientist at Toronto Metropolitan University. Her research interests are related to Explainable AI for the extraction of predictive and descriptive biomarkers in patients with chronic disease.

Alessia Paglialonga obtained her PhD Degree in Biomedical Engineering (2009) from Politecnico di Milano, Italy. She is a researcher at CNR-IEIIT, Adjunct Professor at Politecnico di Milano, and Visiting Scientist at Toronto Metropolitan University. Her research interests include health data analytics, eHealth, audiological technology, machine learning, biosignal processing. She is Associate Editor for BioMedical Engineering OnLine, Heliyon, BMC Digital Health, and the International Journal of Audiology.

Maurizio Mongelli obtained his PhD. Degree in Electronics and Computer Engineering from the University of Genoa in 2004. He worked for Selex and the Italian Telecommunications Consortium (CNIT) from 2001 until 2010. He is now a researcher at CNR-IEIIT, where he deals with machine learning applied to health and cyber-physical systems. He is co-author of over 100 international scientific papers, 2 patents and is participating in the SAE G-34/EUROCAE WG-114 AI in Aviation Committee.