Multiclass Counterfactual Explanations Using Support Vector Data Description

Alberto Carlevaro, Marta Lenatti, Alessia Paglialonga, and Maurizio Mongelli, *Member, IEEE*

***Abstract*—Explainability has become crucial in artificial intelligence 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 input–output relation is to extract counterfactual explanations that allow us to find minimal changes from an observation to another one belonging to a different class. In this article, we propose a novel methodology to extract multiple counterfactual explanations [MUltiCounterfactual via Halton sampling (MUCH)] from an original multiclass support vector data description algorithm. 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.**

***Impact Statement*—When a system is analyzed by artificial intelligence, 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 artificial intelligence (AI) model aware of how to affect the environment properly (both in terms of performance and safety). Examples may include: maneuvering of autonomous cars, clinical diagnosis, and finance. The proposed approach generalizes counterfactuals intelligibility and control to the multiclass 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.**

Manuscript received 3 March 2023; revised 1 August 2023 and 2 October 2023; accepted 22 November 2023. Date of publication 27 November 2023; date of current version 21 June 2024. This work was supported in part by the REXASI-PRO H-EU project, call HORIZON-CL4-2021-HUMAN-01-01, under Grant 101070028 and in part by the Future Artificial Intelligence Research (FAIR) project, Recovery and Resilience Plan ("Piano Nazionale di Ripresa e Resilienza"), Spoke 3—Resilient AI. This article was recommended for publication by Associate Editor Supratik Mukhopadhyay upon evaluation of the reviewers' comments. *(Alberto Carlevaro and Marta Lenatti contributed equally to this work.) (Corresponding author: Alberto Carlevaro.)*

Marta Lenatti, Alessia Paglialonga, and Maurizio Mongelli are with CNR-Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni (CNR-IEIIT), 00129 Turin, Italy (e-mail: marta.lenatti@ieiit.cnr.it; alessia.paglialonga@cnr.it; maurizio.mongelli@cnr.it).

Alberto Carlevaro is with CNR-Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni (CNR-IEIIT), 00129 Turin, Italy and also with the Department of Electrical, Electronics and Telecommunications Engineering and Naval Architecture (DITEN), University of Genoa, 16145 Genoa, Italy (e-mail: alberto.carlevaro@ieiit.cnr.it)

This article has supplementary downloadable material available at https://doi.org/10.1109/TAI.2023.3337053, provided by the authors.

Digital Object Identifier 10.1109/TAI.2023.3337053

***Index Terms*—Counterfactual explanations, multiclass 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 describe what should be changed in a certain input sample (the *factual*) to obtain a different model decision.

2) Controllability: Counterfactual explanations can be used to introduce control over the AI model in a flexible way [4], [5], [6]. The process consists of generating counterfactuals around controllable variables, still under noncontrollable 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) Multiclass: Examples may help understand the importance of counterfactual reasoning in multiclass situations. In healthcare, several diseases present different stages of severity (e.g., cancer) that can worsen drastically in a short time if not properly treated. Multiclass 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, and D). Thus, counterfactual analysis can be useful to check for differences between different transitions (e.g., direct paths skipping intermediate transitions or progressive

sequential paths). Some practical applications of this kind include predictive maintenance and vehicular platooning [7].

B. Contribution

The objective of this article is to develop a novel method based on support vector data description in multiclass frameworks (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 multiclass classification problems based on data envelopes extracted via SVDD. The method developed in this study addresses: 1) explainability, through the use of counterfactuals; 2) controllability of counterfactuals via MC-SVDD; and 3) validation of counterfactuals quality.

II. Related Works

A. MC-SVDD

Multiclass 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 multiclass problems: some algorithms (e.g., decision trees and neural networks) automatically handle multiple outputs, whereas other algorithms (e.g., logistic regression) provide exclusively binary outputs. In the latter case, the classifiers must be adapted to handle multiple outputs. Therefore, we can distinguish two types of multiclass classification techniques [8]: one-vs-one (OvO) and one-vs-rest (OvR). In OvO techniques, the problem is divided into $m(m-1)/2$ 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. Due to its incremental adaptation to multiple outputs, the OvO approach lack a comprehensive view of relationship among the classes. In OvR techniques, instead, m different classifiers are trained, where each target class is classified against the rest of the classes. Then, an instance is assigned to the class with the highest probability. The MC-SVDD[1] approach here proposed, solves the problem in one shot, without repetitive adaptations and providing the weights for classification as an exact solution to an optimization problem. All uncertainties and data characteristics are handled at the same time, providing a result that best fits the problem [9]. The algorithm generalizes the well-known SVDD by Tax and Duin [10] to the multiclass case, quite naturally as an extension of the original method. Other attempts address multiclass SVDD, but focus on identifying anomalous objects rather than providing canonical classification [11]. The algorithm proposed by [12] generalizes the unsupervised one class classifier of [13] to multiple outputs; however, it does not consider the fact that the classification regions (i.e., the hyperspheres) may intersect with each other. A different approach is proposed by [14], in which the canonical SVDD is merged with binary trees to handle the multiclassification problems. Guo et al. [15] proposed a multikernel learning adaptation to SVDD (MKL-SVDD) to design the kernel weights for multiple kernels and obtain the optimal kernel combination. Hou and Ji [16] developed a multiclass 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.

[1]https://github.com/AlbiCarle/MultiClass_SVDD.git

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]. For example, Mothilal et al. [18] introduced a gradient-based method that produces a set of diverse counterfactual explanations (DiCEs) for each input observation in tabular format and proposed a set of quantitative metrics to evaluate the proposed explanations. Vermeire et al. [19], instead, introduced a method for the generation of visual counterfactual explanations for multiclass, model-agnostic image classification and compared the proposed explanations with other state-of-the-art explainability methods, including LIME and SHAP, in terms of stability and computational time. Another method was proposed by Wu et al. [20] to generate grammatically and semantically correct counterfactual explanations starting from text in a more efficient and cost-saving way compared to manual generation from scratch. 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 DiCE [18]. The present article extends the analysis with respect to the multiclass problem, as described in Sections III and IV.

III. Multiclass Support Vector Data Description (MC-SVDD)

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

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

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: 1) 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 2) 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 the above constraints means finding the solution of the following optimization problem:

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

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

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

We can follow the classical approach as in [10], 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 center $\mathbf{a}_k$ 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 centers $\mathbf{a}_h$, $h \neq k$, should not be strictly larger than R_h^2, i.e., smaller distances should be penalized. Therefore, we introduce slack variables $\boldsymbol{\xi}^{kk} \geq 0, \boldsymbol{\xi}^{kh} \geq 0$ and the minimization problem changes into

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

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

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

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

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

$$\boldsymbol{\alpha}^{kk}, \boldsymbol{\alpha}^{kh}, \boldsymbol{\gamma}^{kk}, \boldsymbol{\gamma}^{kh} \geq 0. \tag{4}$$

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

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

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

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

Substituting (5) and (6) in (3), the Lagrangian in the dual takes the following 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 (9) 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):

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

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

Then, according to the literature around SVDD [10], 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, 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,

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

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 $\mathbf{t}$, it is necessary to calculate its distance from the center of the hypersphere k, i.e.,

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

a test object $\mathbf{t}$ 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 $\mathbf{t}$ belongs to class k;
2) If $d_k \leq R_k^2$ and $d_h < d_k\ \forall h\neq k$, then $\mathbf{t}$ belongs to class h;
3) If $d_k > R_h^2\ \forall h$, then $\mathbf{t}$ 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.

Remark III.1: In order to obtain a more compact form of the Lagrangian L and to clarify that the problem is quadratic, we define these quantities for all $k\in[m]$

$$\boldsymbol{\alpha}^k \doteq \left[\boldsymbol{\alpha}^{k1},\boldsymbol{\alpha}^{k2},\ldots,\boldsymbol{\alpha}^{km}\right]^\top,\ \boldsymbol{\alpha}\doteq\left[\boldsymbol{\alpha}^1,\boldsymbol{\alpha}^2,\ldots,\boldsymbol{\alpha}^m\right]^\top$$
$$\mathbf{y}^k = \begin{bmatrix} y_1^k & y_2^k & \ldots & y_n^k\end{bmatrix}^\top,$$

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 $k\in[m]$

$$\Phi_k \doteq \begin{bmatrix}\varphi(\boldsymbol{x}_1^k) & \varphi(\boldsymbol{x}_2^k) & \ldots & \varphi(\boldsymbol{x}_n^k)\end{bmatrix},\tag{13}$$
$$D_k \doteq \text{diag}\{y_1^k, y_2^k,\ldots,y_n^k\},\tag{14}$$
$$K_k \doteq \Phi_k^\top\Phi_k,\tag{15}$$

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 [21]. Then, let them be

$$H_k \doteq 2D_kK_kD_k,$$
$$f_k \doteq D_k\text{diag}(K_k).$$

Finally, defining

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

we obtain that the Lagrangian L (9) can be rewritten as

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

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

IV. MULTICOUNTERFACTUAL VIA HALTON SAMPLING (MUCH)

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

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

Multiclass classification. A multiclass classifier (e.g., MC-SVDD) is applied to obtain m classification regions defined as follows:

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

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.

Counterfactual search. Once the m classification regions are defined, the search for a counterfactual explanation of an observation $\boldsymbol{x}_{f_i}=(\mathbf{u},\mathbf{z})_{f_i}\in S_i$, called *factual*, consists of determining the minimum joint variation $\Delta\mathbf{u}^*$ of the modifiable variables to obtain the closest observation

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

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

$$\min_{\Delta \mathbf{u} \in \mathbb{R}^p} \quad d\big(\boldsymbol{x}_{f_i}, (\mathbf{u}+\Delta\mathbf{u}, \mathbf{z})_{f_i}^{cf_j}\big) \tag{19a}$$

$$\text{subject to} \quad \left\|(\mathbf{u}+\Delta\mathbf{u}, \mathbf{z})_{f_i}^{cf_j} - \mathbf{a}_j\right\|^2 \leq R_j^2 \tag{19b}$$

$$\left\|(\mathbf{u}+\Delta\mathbf{u}, \mathbf{z})_{f_i}^{cf_j} - \mathbf{a}_k\right\|^2 \geq R_k^2, \tag{19c}$$

$$\text{with } k \in [m] \text{ and } k \neq j,$$

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

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 [22] that is a low discrepancy sequence generator; other generators (e.g., Sobol) may be applicable in this sampling step. Since counterfactual explanations are searched among a finite set of points, the availability and minimality of each explanation depend 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 tradeoff between accuracy and runtime must be reached.

Counterfactual explanations are extracted for each factual observation belonging to each class. Once a factual $\boldsymbol{x}_{f_i} \in \mathbf{F}_i$, $i \in [m]$, is defined, the algorithm returns the set of counterfactuals $\mathbf{C}_{F_i}$, i.e., each counterfactual explanation $\boldsymbol{x}_{f_i}^{cf_j}, j \in [m], j \neq i$. The first step of the MUCH algorithm[2] (Algorithm 1) is the classification of data. In this work, data are classified by MC-SVDD, which defines m closed classification regions S_i, $i \in [m]$. 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 $\boldsymbol{x}_{f_i}$, the respective counterfactual related to the class $j \neq i$, $\boldsymbol{x}_{f_i}^{cf_j}$ 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 $\boldsymbol{x}_{f_i}$ is chosen. The distance d plays a key role in the search for counterfactuals as changing the distance may change 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, hence, keeping the same distance relationship would help the algorithm find the best counterfactual explanation.

Denoted with n the number of points, with d the number of features, and with m the number of classes, the computational cost of MC-SVDD, that is, O(MC-SVDD) is estimated considering the two most expensive computations: the solution of the quadratic programming problem to compute the Lagrangian multipliers and the kernelization, i.e., the computation and storage of the kernel matrix. The time complexity for solving a quadratic programming problem is generally in the order of $O(K^3)$ to $O(K^4)$, where K is the number of variables, that is, the number of Lagrange multipliers ($\boldsymbol{\alpha}^{kk}$ and $\boldsymbol{\alpha}^{kh}$) and the variability is due to the type of optimizer that can be chosen. The number K depends on the number of samples and classes, specifically:

1) $\boldsymbol{\alpha}^{kk}$: there are $n_1 + n_2 + \cdots + n_m$ Lagrange multipliers for each class k. These are associated with the data points that belong to class k;
2) $\boldsymbol{\alpha}^{kh}$: there are $n_1n_2 + n_1n_3 + \cdots + n_1n_m + n_2n_3 + \cdots + n_2n_m + \cdots + n_{(m-1)}n_m$ Lagrange multipliers for each pair of classes (k, h).

Algorithm 1 MUCH

1.1 Dataset $\mathcal{D}$ is divided in training set $\mathcal{D}_{tr}$ and validation set $\mathcal{D}_{vl}$.
1.2 A classifier is trained on $\mathcal{D}_{tr}$ and validated on $\mathcal{D}_{vl}$, getting $S_1, S_2, \ldots, S_m$.
1.3 A set of factuals related to the class i, $\mathbf{F}_i$, is chosen.

2 $\mathbf{C}_{\mathbf{F_i}} = [\,]$
3 **for** $\mathbf{x}_{f_i} = (\mathbf{u}, \mathbf{z})_{f_i} \in \mathbf{F_i}$
3.1 $\mathbf{C}_{f_i} = [\,]$
3.2 **for** $j \in [m], j \neq i$
3.2.1 **Sample** quasi-randomly $\tilde{S}_j$
3.2.2 $d_i^j = d\big(\boldsymbol{x}_{f_i}, \tilde{S}_{j_{|\mathbf{z}=\mathbf{z_{f_i}}}}\big)$
3.2.3 $\boldsymbol{x}_{f_i}^{cf_j} = \min(d_i^j)$
3.2.4 **if** $\big(\boldsymbol{x}_{f_i} \in S_i \;\&\; \boldsymbol{x}_{f_i}^{cf_j} \in S_j\big)$
3.2.4.1 $\mathbf{C}_{f_i} = \mathbf{C}_{f_i} \cup \{\mathbf{x}_{f_i}^{cf_j}\}$
3.2.5 **end**
3.2.6 $\mathbf{C}_{\mathbf{F_i}} = \mathbf{C}_{\mathbf{F_i}} \cup \mathbf{C}_{f_i}$
3.3 **end**
3.4 **end**
4 **return** $\mathbf{C}_{\mathbf{F}_i}$

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

Thus, the total number of Lagrange multipliers K can be calculated as follows:

$$K = (n_1 + n_2 + \cdots + n_m) + (n_1 n_2 + n_1 n_3 + \cdots + n_1 n_m + n_2 n_3 + \cdots + n_2 n_m + \cdots + n_{(m-1)} n_m)$$

The computational cost for kernelization, instead, varies from kernel to kernel [23]. Focusing on the linear kernel, the cost for building the Gram matrix can be estimated in $O(n^2)$ but it rises in complexity when a polynomial kernel ($O(n^{2p})$, where p is the degree of the polynomial) or a Gaussian kernel ($O(n^{2g})$, where g takes into account the complexity brought by the exponential function and the Euclidean distance) are considered. In any case, however, the kernelization cost does not overcome the computational cost for the solution of the quadratic optimization problem, making the overall complexity of the algorithm estimable only with the cost to compute the Lagrangian multipliers (i.e., $O(K^3)$ or $O(K^4)$). 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 $\tilde{S}_j$ [24], $O(D)$ is the computational cost for the computation of the distance d [25] and $O(\tilde{s}_j)$ is the computational cost of the research of the minimum of the vector of distances relative to the jth random sampling ($\tilde{S}_j$) [26]. So, considering all m classes, the computational cost of the counterfactuals search, O (SCF), 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 (MUCH) $= O$ (max(MC-SVDD, SCF)). The complete procedure for the generation of a set of explanations is summarized in Fig. 1.

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, discriminative power, and plausibility. *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 i, i.e., $|\mathbf{C}_{\mathbf{F}_i}|$ 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 i, as the ratio of the number of constrained features and the total number of nonmodifiable 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 $\mathbf{C}_{\mathbf{F}_i}$. 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.). *Discriminative power* [4], [27] measures the ability to distinguish points of the factual class in S_i from counterfactuals in $\mathbf{C}_{\mathbf{F}_i}$. It was estimated in this study by evaluating the accuracy of a k-nearest neighbor (KNN) classifier trained on a dataset including the counterfactuals in $\mathbf{C}_{\mathbf{F}_i}$ and real data points in S_i. Discriminative power was then computed as the average test accuracy obtained with fivefold cross validation. Finally, *plausibility* measures the ability of $\mathbf{C}^j_{\mathbf{F}_i}$ to be representative of the reference population (i.e., real data) of class j. Plausibility was computed as the Hellinger distance between counterfactuals of class j generated from class i and the training set distribution for each class $j \in [m]$ (the lower the better). In a multiclass classification problem, such as the one considered in this article, where $|\mathbf{C}_{F_i}| > 1$ for all $i \in [m]$, each evaluation metric can be considered as the average value obtained across the $m - 1$ set of counterfactuals.

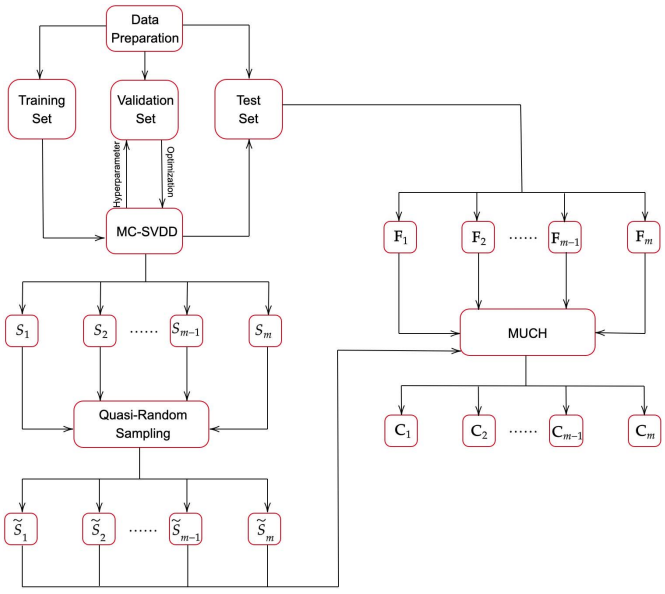

Fig. 1. Diagram of the counterfactual explanations extraction procedure.

V. Applicative Example: The FIFA Dataset

A. Dataset Description and Classification

FIFA is one of the most famous football videogames in the world. The FIFA dataset[3] includes latest edition FIFA attributes related to more than 17 000 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 and 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

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

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

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. Moreover, the main attributes can be combined in six 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. The classification task consisted in predicting the correct player's position among four possible classes: *midfielder* (MF), *defender* (DE), *forward* (FO), and *goalkeeper* (GK). To obtain a balanced dataset, 2000 records were extracted for each player's position (8000 records in total). The dataset was splitted in training set (70%) and test set (30%). The parameters of MC-SVDD were optimized by performing a cross validation on the training set, as explained in Section III.

Table I shows the best MC-SVDD training and test classification performance obtained by selecting a Gaussian kernel. Specifically, the performance is evaluated in terms of classification accuracy, macroaveraged F1-score (i.e., the mean of F1-scores computed by class), Cohen's kappa coefficient [28] (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 macroaveraged F1-score are satisfactory as the 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 regions identified by MC-SVDD are able to enclose almost all points and the presence of anomalous points in the dataset is limited.

As it can be noticed from Fig. 2, 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.

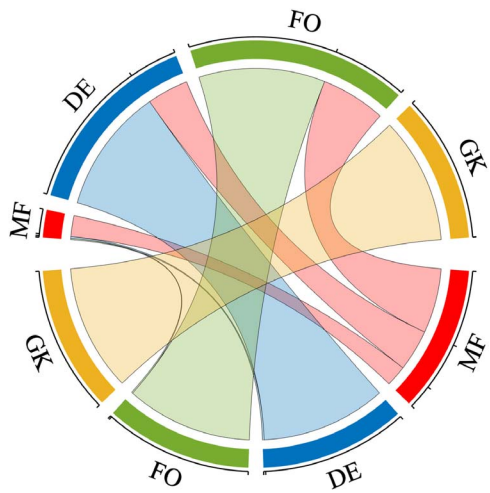

Fig. 2. Chord diagram representation of the confusion matrix corresponding to the classification of the FIFA testing dataset.

For completeness, the MC-SVDD classification performance obtained with different kernels (i.e., linear and cubic polynomial) is presented in Section II of the supplementary material.

B. Multicounterfactual Generation

1) Setting: To evaluate the MUCH approach, a set of counterfactuals are 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 are obtained by sampling 10 000 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 do 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 nonmodifiable features, hence they were constrained during counterfactual search. Actually, counterfactuals have been accepted within a certain tolerance δ (i.e., $\delta = \pm 2\,\mathrm{cm}$ for *height*) in order to ensure their availability. 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 nonmodifiable variables increases.

2) Knowledge Extraction: Table II lists the properties of the sets of counterfactuals (as defined in Section IV-B) 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%. 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,

TABLE II
AVAILABILITY (%), SIMILARITY (MEAN% AND C.I.%), DISCRIMINATIVE POWER (%), AND PLAUSIBILITY 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%	12.74%	13.48%	35.80%
	29.96%	30.02%	29.31%	44.48%
Plausibility	0.068	0.003	0.002	0.002
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%	16.94%	16.65%	34.11%
	30.89%	31.17%	32.04%	42.31%
Plausibility	0.026	0.005	0.058	0.151
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%	27.71%	28.45%	36.95%
	49.61%	45.62%	46.75%	46.01%
Plausibility	0.123	0.088	0.082	0.050
Discriminative Power	95.58%	98.27%	98.89%	99.84%

TABLE III
EXAMPLE OF FACTUALS ($\boldsymbol{x}_{\mathrm{MF}}$, $\boldsymbol{x}_{\mathrm{FO}}$, AND $\boldsymbol{x}_{\mathrm{GK}}$) AND RELATED COUNTERFACTUAL EXPLANATIONS ($\boldsymbol{x}_{\mathrm{MF}}^{\mathrm{DE}}$, $\boldsymbol{x}_{\mathrm{FO}}^{\mathrm{DE}}$, AND $\boldsymbol{x}_{\mathrm{GK}}^{\mathrm{DE}}$). IMPROVEMENTS IN FUNDAMENTAL SKILLS ARE SHOWN IN BOLD

	Example 1		**Example 2**		**Example 3**	
	$\boldsymbol{x}_{\mathrm{MF}}$	$\boldsymbol{x}_{\mathrm{MF}}^{\mathrm{DE}}$	$\boldsymbol{x}_{\mathrm{FO}}$	$\boldsymbol{x}_{\mathrm{FO}}^{\mathrm{DE}}$	$\boldsymbol{x}_{\mathrm{GK}}$	$\boldsymbol{x}_{\mathrm{GK}}^{\mathrm{DE}}$
Pace	89.30	74.66	86.65	71.36	39.35	**53.43**
Shooting	61.3	51.32	72.60	52.37	16.20	**35.79**
Passing	52.15	50.05	62.45	52.75	19.80	**39.51**
Defending	50.40	**62.44**	44.90	**66.44**	10.60	**64.99**
Dribbling	65.75	53.46	84.25	63.26	12.20	**47.68**
Physical	67.20	60.13	63.85	61.18	43.20	**62.73**

which, as previously mentioned, has peculiar characteristics. The algorithm successfully returned all counterfactuals (100% availability), demonstrating a sufficiently dense sampling of the MC-SVDD regions. Similarity values are also satisfactory, with average values between 21% and 42%, depending on the factual class. Lastly, the low plausibility values (i.e., $\ll 1$) indicate that the counterfactuals are close to the real distribution of the class they aim for.

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, Table III lists three examples of factuals belonging to class MF, FO, and GK and the corresponding counterfactual explanations belonging to class DE. These examples quantify the changes that specific players, trained for a specific role, would have to make in terms of fundamental characteristics to transition to the DE class. In particular, in example 1 the transition from the specific MF player to DE is described by higher defending skills, almost unchanged passing and physical abilities, and lower values of the remaining attributes. Likewise, the FO player in example 2 should increase the training of defending attributes and lower the other fundamental skills to become DE. Finally, the GK player in example 3 needs to improve its defending skills to transition to the DE role, but in a significantly greater amount with respect to the previous two examples. Besides, the GK player under consideration should also focus on improving all the other fundamental skills. In all the three examples reported, pace, defending, and physical attributes are the most relevant attributes for the DE class. Each counterfactual explanations provide insights related to a specific input observation, however, it can be extremely useful to consider the overall trend of changes required by the counterfactual explanations of the entire test set. Fig. 3 analyzes the average behavior 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 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 values 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,[4] 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, and MF on passing. Other attributes, such as physical, aggressiveness, and dribbling, do not impact the specialization. Although such a conclusion may appear intuitive, it may be of extreme interest to help experts (e.g., athletic coaches) in the selection of the target variables.

[4]https://github.com/AlbiCarle/MUCH/blob/main/SpiderImages/FIFA_SpiderPlots.pdf.

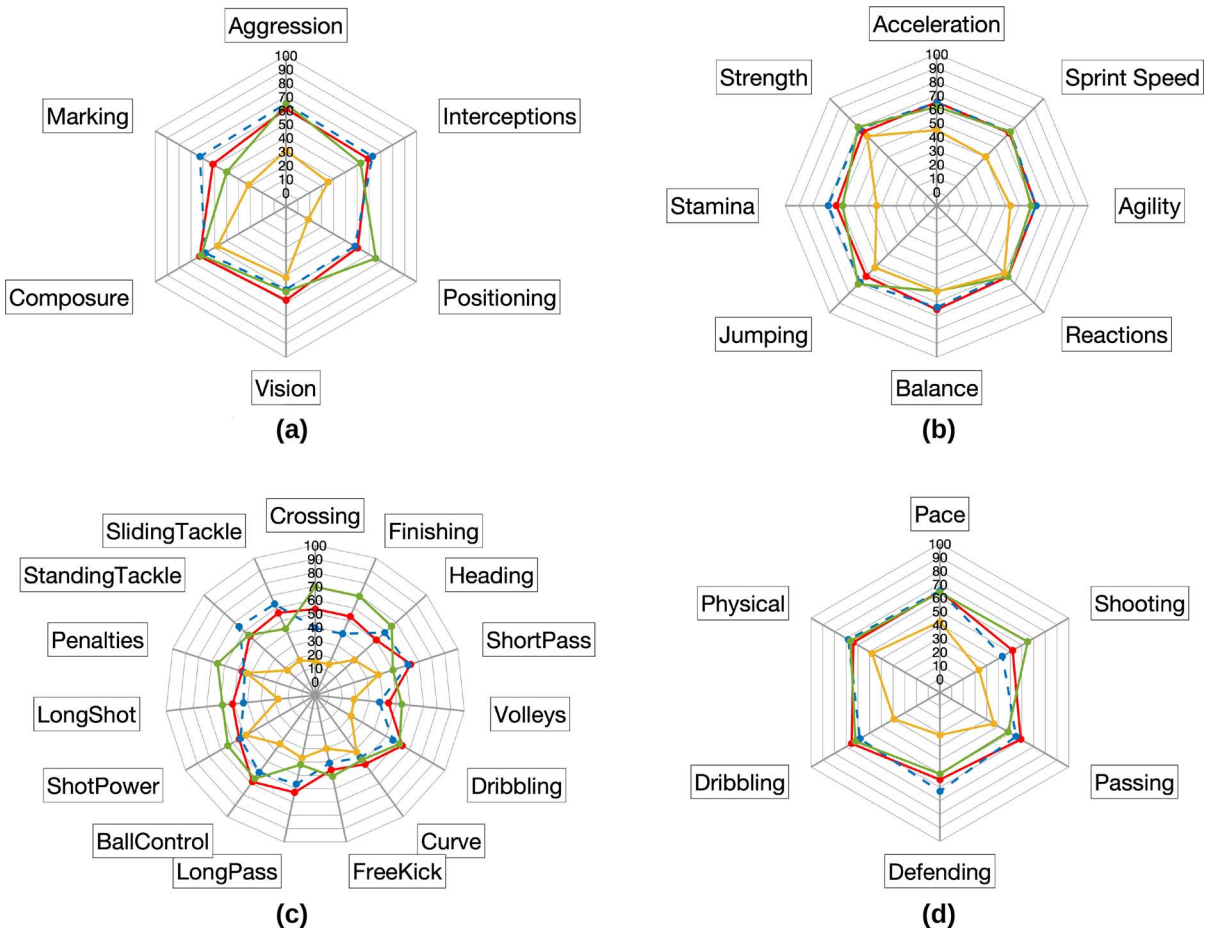

Fig. 3. Each spiderplot represents the variation of the average of the factuals (dashed line) and counterfactuals (solid line) for DE class for each attribute category: (a) mental; (b) physical; (c) technical; and (d) fundamental skills. The value scale ranges from 0 to 100, and the output classes colors are the same as those used in Fig. 2 (MF: red, DE: blue, FO: green, and GK: yellow).

VI. Characterization on Additional Datasets

This section discusses the performance of the proposed approach on a set of frequently referenced multiclass open source datasets, including the IRIS dataset[5] and the Stellar Classification Dataset—SDSS17 dataset.[6] 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 three type of objects (i.e., *galaxy*—1, *star*—2, and *quasar*—3) described by different spectral characteristics. Every observation consists of 17 input features, however only

[5]Retrieved [Dec 2022] from https://www.kaggle.com/datasets/uciml/iris

[6]fedesoriano. Stellar Classification Dataset—SDSS17. Retrieved [Dec 2022] from https://www.kaggle.com/fedesoriano/stellar-classification-dataset-sdss17

TABLE IV
Classification Performance: IRIS and Stellar Datasets

		IRIS	Stellar Classification
ACC_{tr}		95.24%	93.83%
OUT_{tr}		0.00%	0.01%
ACC		97.78%	92.11%
OUT_{ts}		0.00%	0.02%
Macroaveraged F1-$SCORE_{ts}$		97.78%	94.18%
Cohen's Kappas		0.97	0.88
Canonical Machine Learning Models			
Decision Tree	ACC_{ts}	99.54%	94.83%
	Macro F1 $SCORE_{ts}$	99.00%	94.75%
Random Forest	ACC_{ts}	99.32%	96.00%
	Macro F1 $SCORE_{ts}$	99.99%	95.92%
Gradient Boosting	ACCts	99.39%	94.00%
	Macro F1 $SCORE_{ts}$	99.29%	93.75%
Support Vector Machine	ACC_{ts}	98.00%	92.00%
	Macro F1 $SCORE_{ts}$	75.00%	69.00%

TABLE V
AVAILABILITY (%), SIMILARITY (MEAN% AND C.I.%), DISCRIMINATIVE POWER (%), AND PLAUSIBILITY 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%
Plausibility	0.32	0.29	0.17	0.24	0.12	0.04
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%
Plausibility	0.32	0.19	0.38	0.03	0.08	0.21
Discriminative Power	100.00%	82.91%	91.99%	95.09%	98.16%	98.10%

a subset of ten features was considered in this experiment. Both datasets were split in training (70%) and test set (30%). Table IV 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, macroaveraged F1-score, and Cohen's kappa on the test set. Additionally, the MC-SVDD classification performance has been compared with state-of-the-art multiclass classifiers, including decision tree (criterion: "entropy"), random forest (criterion: "gini", bootstrap: true), gradient boosting (criterion: "gini"), and support vector machine (kernel: "Gaussian"). As a result, the MC-SVDD yields comparable, but slightly lower accuracy and macroaveraged F1-score values on the test set (i.e., 1%–2% lower) than the four well-established methods. Table V 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 two 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 multiclass generalization of an SVDD (MC-SVDD) and extract a set of counterfactual explanations from the classification results using a multiclass extension (MUCH) of a previously proposed counterfactuals explainer [5]. In principle, both OvO and OvR methods can be used in multiclass classification problems. As previously stated, the proposed MUCH algorithm is agnostic to the classifier to be used. Therefore, the difference in counterfactual extraction using a OvO or OvR method depends only on the quality of the classification. In a preliminary phase, we have compared the two approaches, and the OvO approach yielded lower classification performance on the FIFA dataset (i.e., accuracy of 57% and macroaveraged F1-score of 50% on the test set), resulting in lower counterfactuals plausibility (details are reported in the supplementary material). However, the difference in classification performance is highly dependent on the selected classification problem. 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. As summarized in Table IV and in Section II of the supplementary material, the classification performance obtained by applying the proposed MC-SVDD to benchmark datasets results comparable to that of well-established classification algorithms (i.e., decision tree, random forest, gradient boosting, and support vector machine). In addition, MC-SVDD presents advantages in terms of capability to detect outliers, error control, and definition of closed regions of data points.

MUCH demonstrated satisfactory performance in terms of availability, similarity, discriminative power, and plausibility of the generated counterfactual explanations. This technique allows us to investigate the 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, counterfactual explanations 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, such as text. As a further development, sets of counterfactual explanations obtained with different AI methods or different sets of modifiable features can be compared to better understand the inner logic of various models and exploit local explainability to address potential model-induced biases.

ACKNOWLEDGMENT

Marta Lenatti is a Ph.D. student enrolled in the National Ph.D. in artificial intelligence, XXXVIII cycle, course on health and life sciences, organized by Università Campus Bio-Medico di Roma. 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.

REFERENCES

[1] V. Belle and I. Papantonis, "Principles and practice of explainable machine learning," *Frontiers Big Data*, vol. 4, 2021.
[2] "General data protection regulation (GDPR)." GDPR.EU. Accessed: Dec. 16, 2022. [Online]. Available: https://gdpr.eu/tag/gdpr/
[3] S. Wachter, B. D. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard J. Law Technol.*, 2017.
[4] R. Guidotti, "Counterfactual explanations and how to find them: Literature review and benchmarking," in *Data Min. Knowl. Discov.*, 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 Trans. Intell. Transp. Syst.*, vol. 22, no. 9, pp. 5861–5873, Sep. 2021.
[8] P. Mills, "Solving for multi-class: A survey and synthesis," 2018.
[9] P. D. Moral, S. Nowaczyk, and S. Pashami, "Why is multiclass classification hard?" *IEEE Access*, vol. 10, pp. 80448–80462, 2022.
[10] D. M. J. Tax and R. P. W. Duin, "Support vector domain description," *Pattern Recognit. Lett.*, vol. 20, pp. 1191–1199, 1999.
[11] M. Turkoz, S. Kim, Y. Son, M. K. Jeong, and E. A. Elsayed, "Generalized support vector data description for anomaly detection," *Pattern Recognit.*, vol. 100, 2020, Art. no. 107119.
[12] G. Xie, Y. Jiang, and N. Chen, "A multi-class support vector data description approach for classification of medical image," in *Proc. 9th Int. Conf. Comput. Intell. Secur.*, 2013, pp. 115–119.
[13] S. S. Khan and M. G. Madden, "One-class classification: Taxonomy of study and review of techniques," *Knowl. Eng. Rev.*, 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 Syst. Appl.*, 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," *Eng. Appl. Artif. Intell.*, vol. 102, 2021, Art. no. 104254.
[16] H. Hou and H. Ji, "Improved multiclass support vector data description for planetary gearbox fault diagnosis," *Control Eng. Pract.*, vol. 114, 2021, Art. no. 104867.
[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," *Chin. J. Electron.*, vol. 21, no. 1, pp. 107–111, Jan. 2012.
[18] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," Barcelona, Spain. New York, NY, USA: ACM, 2020.
[19] T. Vermeire, D. Brughmans, S. Goethals, R. M. B. de Oliveira, and D. Martens, "Explainable image classification with evidence counterfactual," Jan. 2022.
[20] T. Wu, M. T. Ribeiro, J. Heer, and D. Weld, "Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.,* vol. 1: Long Papers, Association for Computational Linguistics, 2021, pp. 6707–6723.
[21] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," *Philos. Trans. R. Soc. London, Ser. A*, vol. 209, pp. 415–446, 1909.
[22] C. Cervellera, M. Gaggero, D. Macciò, and R. Marcialis, "Quasi-random sampling for approximate dynamic programming," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2013, pp. 1–8.
[23] N. Cesa-Bianchi, Y. Mansour, and O. Shamir, "On the complexity of learning with kernels," in *Proc. 28th Conf. Learn. Theory*, Paris, France, vol. 40, PMLR, 2015, pp. 297–325.
[24] S. Sen, T. Samanta, and A. Reese, "Quasi-versus pseudo-random generators: Discrepancy, complexity and integration-error based comparison," *Int. J. Innovative Comput. Inf. Control*, vol. 2, 2006.
[25] Y. Burago and D. Shoenthal, "Metric geometry," in *New Analytic and Geometric Methods in Inverse Problems*, K. Bingham, Y. V. Kurylev, and E. Somersalo, Eds., Berlin, Germany: Springer-Verlag, 2004, pp. 3–50.
[26] 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, 1973.
[27] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.
[28] J. Cohen, "A coefficient of agreement for nominal scales," 1960.

Alberto Carlevaro received the master's degree in applied mathematics from the University of Genoa, Genoa, Italy. He is currently working toward the Ph.D. degree with the Department of Electrical, Electronic and Telecommunications Engineering and Naval Architecture (DITEN), in collaboration with CNR-IEIIT, Rome, Italy, and S.M.E. Aitek.

His current fields of research are in machine learning, deep learning, statistical learning, and explainable AI.

Marta Lenatti received the master's degree in biomedical engineering from Politecnico di Milano, in 2020. She is currently working toward the Ph.D. degree with the Italian National Ph.D. program on artificial intelligence (health and life sciences area) in collaboration with CNR-IEIIT, Rome, Italy, and a Visiting Scientist with Toronto Metropolitan University.

Her research interests are related to explainable AI for the prediction and management of chronic diseases.

Alessia Paglialonga received the Ph.D. degree in biomedical engineering, in 2009, from Politecnico di Milano.

She is a Senior Researcher with CNR-IEIIT, Rome, Italy, an Adjunct Professor with Politecnico di Milano, and a Visiting Scientist with Toronto Metropolitan University. Her research interests include health data analytics, eHealth, audiological technology, and machine learning.

Dr. Paglialonga is an Associate Editor for *BioMedical Engineering Online*, *BMC Digital Health*, 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, Genoa, Italy, in 2004.

He worked for Selex and the Italian Telecommunications Consortium (CNIT) from 2001 to 2010. He is now a Senior Researcher with CNR-IEIIT, where he deals with machine learning applied to health and cyber–physical systems.

Dr. Mongelli is a Co-Author of over 100 international scientific papers and two patents and is participating in the SAE G-34/EUROCAE WG-114 AI at Aviation Committee.