This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3492730

Estimation and Conformity Evaluation of Multi-Class Counterfactual Explanations for Chronic Disease Prevention

Marta Lenatti†, Alberto Carlevaro† *Student Member, IEEE*, Aziz Guergachi, Karim Keshavjee, Maurizio Mongelli *Member, IEEE*, and Alessia Paglialonga

***Abstract*— Recent advances in Artificial Intelligence (AI) in healthcare are driving research into solutions that can provide personalized guidance. For these solutions to be used as clinical decision support tools, the results provided must be interpretable and consistent with medical knowledge. To this end, this study explores the use of explainable AI to characterize the risk of developing cardiovascular disease in patients diagnosed with chronic obstructive pulmonary disease. A dataset of 9613 records from patients diagnosed with chronic obstructive pulmonary disease was classified into three categories of cardiovascular risk (low, moderate, and high), as estimated by the Framingham Risk Score. Counterfactual explanations were generated with two different methods, MUlti Counterfactuals via Halton sampling (MUCH) and Diverse Counterfactual Explanation (DiCE). An error control mechanism is introduced in the preliminary classification phase to reduce classification errors and obtain meaningful and representative explanations. Furthermore, the concept of *counterfactual conformity* is introduced as a new way to validate single counterfactual explanations in terms of their conformity, based on proximity with respect to the factual observation and plausibility. The results indicate that explanations generated with MUCH are generally more plausible (lower implausibility) and more distinguishable (higher discriminative power) from the original class than those generated with DiCE, whereas DiCE shows better availability, proximity and sparsity. Furthermore, filtering the counterfactual explanations by eliminating the non-conformal ones results in an additional improvement in quality. The results of this study suggest that combining counterfactual explanations generation with conformity evaluation is worth further validation and expert assessment to enable future development of support tools that provide personalized recommendations for reducing individual risk by targeting specific subsets of biomarkers.**

***Index Terms*— explainable AI (XAI), counterfactual explanations, conformal predictions, chronic disease prevention**

I. Introduction

THE extensive development of eXplainable Artificial Intelligence (XAI) techniques has paved the way for data-driven clinical decision support systems that aim at incorporating transparency in automated decision pathways, beyond ordinary levels of performance [1]. Nevertheless, there are several challenges to the practical implementation of these tools, particularly in relation to concerns around privacy, scalability, fairness and accountability, which have the potential to erode the trustworthiness of the system [2], [3]. Moreover, several of the currently available XAI techniques struggle to produce explanations that are interpretable in human terms and that can be used to provide readily applicable and actionable interventions [4], [5].

Counterfactual explanations [6], falling under the umbrella of local post-hoc XAI techniques, can help bridge this gap. Specifically, counterfactual explanations aim to clarify why a particular decision was made by an AI model by showing how changing the input data would lead to a different result, i.e., a different output. Counterfactual explanations from tabular data are typically recovered through the minimization of a loss function, incorporating measures of the distance between the original instance and the candidate counterfactual explanation, as well as the distance between the candidate explanation and its target class [6]. More complex terms may be incorporated into the process, for example, to promote diversity of the

This work was funded in part by the European Union - NextGenerationEU and by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.5, project RAISE - Robotics and AI for Socio-economic Empowerment (ECS00000035) and in part by the European Union - NextGenerationEU and by the Italian Ministry for Universities and Research (MUR), National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1 program “Progetti di Ricerca di Rilevante Interesse Nazionale” (PRIN), through project QUASAR-AI (QUAntification of SARcopenia in patients with Chronic Obstructive Pulmonary Disease using Artificial Intelligence for CT analysis of respiratory muscles: development of new biomarkers and outcome measures), project code 2022FA4RN2, CUP D53D23013480006.

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

M.Lenatti, A.Carlevaro, M.Mongelli, and A.Paglialonga are with Cnr-Istituto di Elettronica e di Ingegneria dell’Informazione e delle Telecomunicazioni (CNR-IEIIT), 00185 Rome, Italy (e-mail: martalenatti@cnr.it, albertocarlevaro@cnr.it, maurizio.mongelli@cnr.it, alessia.paglialonga@cnr.it).

A.Carlevaro is also with University of Genoa, Department of Electrical, Electronics and Telecommunications Engineering and Naval Architecture (DITEN), 16145 Genoa, Italy.

A.Guergachi is with Ted Rogers School of Management and Ted Rogers School of Information Technology Management, Toronto Metropolitan University, ON M5G 2C3 Toronto, Canada, and with Department of Mathematics and Statistics, York University, ON M3J 1P3 Toronto, Canada (e-mail: a2guerga@torontomu.ca).

K.Keshavjee is with Institute of Health Policy, Management and Evaluation, University of Toronto, ON M5T 3M6 Toronto, Canada (e-mail: karim.keshavjee@utoronto.ca).

retrieved explanations [7], to foster causal consistency [8] or to maintain correlations between features [9]. As an alternative approach, gradient-free optimization based on heuristic search strategies (e.g., genetic algorithms [10]) or reinforcement learning [11] have been proposed in the context of non-differentiable models. Despite the search typically occurring in the original feature space, recent efforts have been made to recover counterfactual explanations in a transformed latent space, with optimized dimensions, in binary classification problems [12]. Application of this technique in healthcare has shown promising results [13], [14], [15], [16], yet its practical use remains limited. An important issue is the lack of common benchmark criteria for assessing the quality of counterfactual explanations as the definition of quality criteria may be highly dependent on the clinical goal [17]. Generally, counterfactual explanations are designed to satisfy different properties [18]. For example, counterfactual explanations must be classified into a different class than the original one while remaining as close as possible to it. Additionally, each explanation should be representative of the target destination class. All the relevant and distinct characteristics of this XAI technique should be considered in comprehensive quality metrics. Furthermore, particularly relevant is the concept of reliability, i.e., the amount of confidence an XAI-based model is capable of providing, ensuring an output that is not just interpretable but also trustworthy.

In this regard, the present study proposes a novel approach in the field of counterfactual explanations for personalized disease prevention by introducing substantial methodological and application-oriented advancements, that can be summarized as:

- the introduction of an algorithm to control the classification error of multi-class Support Vector Data Descriptors (MC-SVDD, [19]);
- the introduction of an original “counterfactual conformity” measure, leveraging conformal predictions (CPs, [20]) guarantees to filter counterfactual explanations that do not reach the desired level of confidence;
- the application of the proposed methods and state-of-the-art methods to real-world data in an original health-related application, i.e., creating personalized risk reduction strategies to reduce the risk of developing cardiovascular diseases (CVDs) in patients diagnosed with Chronic Obstructive Pulmonary Disease (COPD).

II. Materials and methods

This section is structured as follows. Section II-A describes the extraction of a dataset of routinely collected biomarkers from patients diagnosed with COPD receiving primary care services, with the aim of estimating the individual 10-year CVD risk. Sections II-B-II-E thoroughly describe the methodological pipeline, according to the workflow summarized in Figure 1. First, a multi-class classifier is optimized and trained on the training set, as detailed in Section II-B. Then, a set of counterfactual explanations is generated based on observations from the test set that have been predicted as high risk of developing CVDs. Two distinct generators of explanations are compared, namely, MUlti Counterfactuals via Halton sampling (MUCH, [19]) and Diverse Counterfactual Explanations (DiCE, [7]), as described in detail in Section II-C. At last, a novel measure of the conformity of counterfactual explanations, as formulated in Section II-E, is computed to evaluate the quality of the explanations and to provide a way to discard explanations that do not reach the desired level of confidence. Related codes are available at [21].

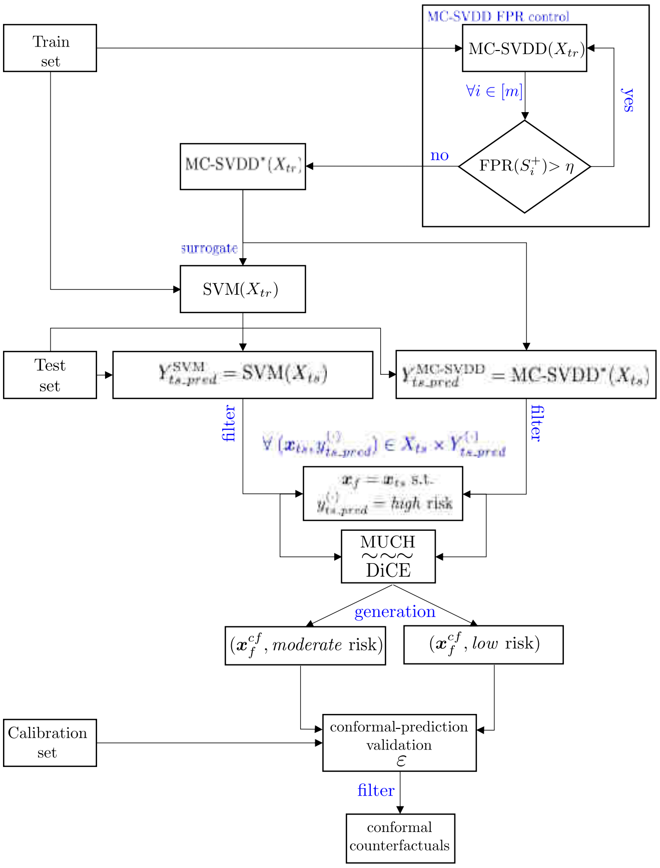

MC-SVDD FPR control
Train set
MC-SVDD(X_{tr})
$\forall i \in [m]$
yes
no
FPR(S_i^+)> η
MC-SVDD*(X_{tr})
surrogate
SVM(X_{tr})
Test set
$Y_{ts\_pred}^{SVM} = SVM(X_{ts})$
$Y_{ts\_pred}^{MC\text{-}SVDD} = MC\text{-}SVDD^*(X_{ts})$
filter
filter
$\forall (x_{ts}, y_{ts\_pred}^{(\cdot)}) \in X_{ts} \times Y_{ts\_pred}^{(\cdot)}$
$x_f = x_{ts}$ s.t. $y_{ts\_pred}^{(\cdot)} = high$ risk
MUCH ∼∼∼ DiCE
generation
(x_f^{cf}, *moderate* risk)
(x_f^{cf}, *low* risk)
Calibration set
conformal-prediction validation ε
filter
conformal counterfactuals

Fig. 1: Methodological workflow: multi-class classification, generation and evaluation of counterfactual explanations.

A. Dataset extraction

The study dataset was ad hoc extracted from part of the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) [22] that includes de-identified electronic health records collected by primary care providers between 2000 and 2015. A waiver of ethics review (reference number: REB 2013-261) was granted by the Review Ethics Board of Toronto Metropolitan University (formerly Ryerson University) as this portion of CPCSSN database includes de-identified and anonymized records. A sample of patients older than 20 years and diagnosed with COPD was extracted and the following features were considered: age at COPD onset, sex assigned at birth, body mass index (BMI), systolic and diastolic blood pressure (sBP and dBP, respectively), fasting blood sugar (FBS), low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglycerides (TG), total cholesterol (totChol), smoking (yes, no, ex), presence of hypertension and/or diabetes, if diagnosed up to 6 months before the onset of COPD. The extracted biomarkers refer to the medical encounter closest to COPD date. Values collected up to 6 months before COPD diagnosis were considered to account for possible uncertainty in the diagnosis date. Using these extraction criteria, a sample of 9613 records with no missing values (one record for each subject) was extracted from an initial set of 37504 subjects diagnosed with COPD. The output variable considered in this study is the Framingham Risk Score (FRS), a sex-specific multivariable indicator that can be used to estimate the 10-year-risk of developing CVDs. The rationale was to identify an output variable that could be used to generate counterfactual explanations with the final aim of guiding personalized preventive actions, e.g., by reducing the individual 10-year-risk of CVDs in patients with COPD. The importance of such measures has been suggested, for example, in [23]. The FRS for each subject was calculated, converted into a percentage risk value, and then grouped into three classes using the Framingham Risk Score Worksheet provided by the Canadian Cardiovascular Society [24]: *low* risk (10-year CVD Risk$<$ 10%, 3944 records), *moderate* risk (10% $<$10-year CVD Risk$<$ 19%, 3274 records), and *high* risk (10-year CVD Risk$\geq$ 20%, 2395 records). A summary of the distribution of the dataset features as a function of the output risk class is shown in Table I. Each feature is marked as modifiable, partially modifiable or not modifiable depending on its ability to be manipulated for the sake of risk reduction (e.g., through lifestyle interventions). For each modifiable feature, the maximum acceptable value shown in Table I is the value used as an upper bound when generating counterfactual explanations, as described in Section II-C. Partially modifiable features, like smoking habits, are permitted to vary only in certain directions. For example, an individual who smokes (Smoke="y") may be able to cease smoking (Smoke = "ex") but cannot transition to the category Smoke = "n" which represents those who have never been smokers.

TABLE I: Features distribution as a function of the output class, degree of modifiability, and maximum acceptable value. Numerical features: median (inter-quartile range); categorical features: number of samples for each category.

Feature	Low risk (N=3944)	Moderate risk (N=3274)	High risk (N=2395)	Modif	Max acceptable value
Age [years]	56 (49-64)	66 (59-73)	73 (67-78)	No	/
Sex at birth	f: 3010 m: 934	f: 1673 m: 1601	f: 360 m: 2035	No	/
HTN	n:3185 y:759	n:1704 y:1570	n:1635 y:760	No	/
Diabetes	n:3320 y:624	n:2487 y:787	n:1630 y:765	No	/
Smoke	n: 859 ex: 1082 y: 2003	n: 709 ex: 1106 y: 1459	n: 563 ex: 961 y: 871	Partial	/
sBP [mmHg]	120 (110-128)	130 (122-140)	140 (130-150)	Yes	140
dBP [mmHg]	73 (68-80)	77 (70-82)	78 (70-84)	Yes	90
BMI [kg/m^2]	27.0 (23.0-32.5)	28.2 (24.3-32.7)	28.0 (25.0-32.0)	Yes	35
FBS [mmol/L]	5.2 (4.8-5.8)	5.5 (5.0-6.1)	5.6 (5.2-6.4)	Yes	7
LDL [mmol/L]	2.65 (2.00-3.33)	2.61 (1.93-3.35)	2.42 (1.78-3.20)	Yes	5
HDL [mmol/L]	1.44 (1.19-1.75)	1.30 (1.06-1.60)	1.34 (1.13-3.66)	Yes	2.5
TG [mmol/L]	1.21 (0.87-1.72)	1.35 (1.92-7.67)	1.39 (1.00-1.99)	Yes	5.7
totChol [mmol/L]	4.76 (4.02-5.50)	4.69 (3.88-5.56)	4.40 (3.56-5.33)	Yes	6.2

B. Multi-class classification

The process of counterfactual explanations generation begins with the classification of data samples. Two state-of-the-art generators of counterfactual explanations were used in this study (i.e., MUCH and DiCE). MC-SVDD was used as the underlying classifier because, as shown in [19], it is flexible, reliable and easily controllable, thus making generation of explanations fast and accurate. Moreover, as shown in Section II-B.2, a new method to control the percentage of unclassified points in the MC-SVDD prediction is introduced to help derive counterfactual explanations that are highly representative of the class they are meant to target. DiCE demonstrated to work well with both non-differentiable and differentiable models [7] implemented with common Python frameworks such as sklearn, TensorFlow, or PyTorch. However, the MC-SVDD algorithm is not yet compatible with these frameworks and cannot be directly applied to DiCE in its current form. To facilitate a direct comparison of the reported results, we employed a surrogate Support Vector Machine (SVM) model that emulates the input/output behavior of the MC-SVDD, as indicated in the workflow in Figure 1. The scikit-learn implementation of the SVM classifier was chosen to surrogate the MC-SVDD, given their inherent similarities [25]. A 70:30 ratio was used to separate train and test data. Max scaling was applied to normalize data between 0 and 1. Tuning of the MC-SVDD hyperparameters was performed by 3-fold cross-validation on 50% of the training data as in [19]. A grid search with 3-fold cross-validation was performed to tune the regularization parameter and the kernel of the SVM (best model parameters: C=7, gamma="auto", kernel="rbf"). Besides accuracy, the capacity of the SVM model to surrogate the original MC-SVDD model was assessed using the Cohen's Kappa coefficient [26]. This coefficient assumes values between -1 and 1, with -1 indicating total disagreement, 0 indicating random chance agreement and 1 indicating total agreement between the two models.

1) Multi-class Support Vector Data Description (MC-SVDD): The SVDD [27] is a state-of-the-art algorithm for outlier detection able to enclose labeled target points within a hypersphere with center and radius computed on the training data points. Its generalization to the multi-class case (MC-SVDD) is briefly described below; a comprehensive description of the algorithm can be found in [19].

Given a training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \in \mathcal{X}_{tr} \times \mathcal{Y}_{tr}$ composed by m classes of objects of different sizes $n_1, n_2, \ldots, n_m$ ($n_1+n_2+\ldots+n_m = n$), labeled and ordered according to their class $\mathbf{y} = \begin{bmatrix} 1 & \ldots & 1 & 2 & \ldots & 2 & \ldots & m & \ldots & m \end{bmatrix}^\top$, MC-SVDD allows to search for the smallest hyperspheres that separate data, i.e.

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

$$\text{s.t.} \quad \left\|\varphi(\boldsymbol{x}_i^k) - \mathbf{a}_k\right\|_2^2 \leq R_k^2, \; i \in [n_k], \forall k, \tag{1b}$$

$$\left\|\varphi(\boldsymbol{x}_i^k) - \mathbf{a}_h\right\|_2^2 \geq R_h^2, \; i \in [n_k], \forall h \neq k, \tag{1c}$$

where $\varphi : \mathcal{X} \longrightarrow \mathcal{V}$ is a feature map from the space of the input features $\boldsymbol{x} \in \mathcal{X}$ to an higher dimensional inner product space $\mathcal{V}$ that allows to identify more flexible descriptions exploiting kernels $K_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi(\boldsymbol{x}_i)^\top \varphi(\boldsymbol{x}_j), \; i \in [n], j \in [n]$ that satisfy the Mercer's theorem [28].

Once a relaxed version of (1) has been solved (i.e. introducing slack variables, as detailed in [19]) and the centers $\mathbf{a}_k$ and radii R_k have been computed for all $k \in [m]$, the classification regions S_i, $i \in [m]$ are defined. A new test sample $\tilde{\boldsymbol{x}}$ is classified based on its distance from each center $d_k \doteq \|\tilde{\boldsymbol{x}} - \mathbf{a}_k\|$:

1) If $d_k \leq R_k$ and $d_h > R_h$ $\forall h \neq k$, then $\tilde{\boldsymbol{x}}$ belongs to Class k;
2) If $d_k \leq R_k$ for several $k \in [m]$ then $\boldsymbol{x}$ belongs to class $k' = argmin_{k \in K} d_k$, where $K = \{k \in [m] \mid d_k \leq R_k\}$;
3) If $d_k > R_k$ $\forall k$ or $\#\{k' \,|\, k' = argmin_{k \in K} d_k\} > 1$, then $\tilde{\boldsymbol{x}}$ is unclassified.

In simple words, the distance between all samples in each class and the center of that class should be smaller than the radius of the corresponding hypersphere, and the distances between all samples in each class and the centers of all the other classes should be larger than the radius of the corresponding hyperspheres. If a given sample belongs to more than one hypersphere, the sample is assigned to the class that lies at minimum distance. In any other case, the sample is unclassified.

2) False Positive Rate control for SVDD: In this study, we extended the algorithm to control the classification error of the binary SVDD (originally proposed in [29]) to the multi-class case with the aim of obtaining well-defined and reliable classification regions, highly representative of the target class (**Algorithm 1**). The algorithm is based on a one-vs-all approach. First, a given class is selected and its false-positive rate (FPR) is optimized by training successive negative SVDDs (i.e., SVDDs with a single target class [27]) until the number of misclassified points for that class is below a predefined threshold η (here set equal to 0.1) or until a maximum iteration limit k_{maxIter} (e.g., 1000) is achieved. We indicate with SVDD$(\cdot)$ the operator that executes the (trained) negative SVDD algorithm, i.e.

$$\begin{aligned} \text{SVDD}: \quad & \mathcal{X} \longrightarrow \mathcal{X} \times \mathcal{Y} \\ & \boldsymbol{x} \longmapsto (\boldsymbol{x}, y) \end{aligned} .$$

Given a dataset $D \subseteq \mathcal{X}$, we indicate with SVDD(D) the application of the negative SVDD algorithm to the dataset D. The procedure is repeated for all the remaining classes.

Algorithm 1 `MC-SVDD FPR control`

Input $S_1, S_2, \ldots, S_m$ regions from multi-class SVDD, threshold on FPR η, maximum number of iterations k_{maxIter}.

Output FPR reduced regions $S_1^*, S_2^*, \ldots, S_m^*$.

```
1:      For all $i \in [m]$:
1.2:        For all $\boldsymbol{x} \in \mathcal{X}$, assign
                $\boldsymbol{x} \longrightarrow y \doteq \begin{cases} +1 & \text{if} \quad \boldsymbol{x} \in \mathcal{S}_i \\ -1 & \text{otherwise} \end{cases}$ ,
            and build a dataset
                $S_{i_0} = \{(\boldsymbol{x}, y) \mid \boldsymbol{x} \in \mathcal{X}, y \in \{-1, +1\}\}$ .
1.3:        Compute $S_{i_0}^* \doteq$ SVDD$(S_{i_0})$.
1.4:        Set $k = 1$.
1.4.1:          While FPR$(S_{i_k}^*) > \eta$ AND $k \leq k_{\text{maxIter}}$
                    $S_{i_k}^*$=SVDD$(S_{i_{k-1}}^*)$
                    $k = k + 1$.
1.5:        $S_i^*$=$S_{i_k}^*$
```

In disease risk prediction applications, a large number of misclassifications could lead a potential clinical decision support system to miss patients in need of a treatment or to unnecessarily treat healthy patients. In this perspective, it would be advisable to not assign a sample to any class if its classification is uncertain and handle it as an outlier, thus abstaining from providing an automated decision in case of doubtful samples. Therefore, FPR control increases the reliability of MC-SVDD, by distinguishing points classified with high confidence from points whose classification is uncertain.

C. Generation of counterfactual explanations

Counterfactual explanations aim to find a "what-if" scenario in the target class while manipulating only the subset of input features that can be changed through internal or external interventions (modifiable and partially modifiable features, denoted as $\mathbf{u}$, for example clinical biomarkers) and constraining those that are immutable (non-modifiable features, denoted as $\mathbf{z}$, for example, age and diagnosed chronic diseases). More specifically, given an observation $\boldsymbol{x}_{f_i} = (\mathbf{u}, \mathbf{z})_{f_i}$ (the *factual*) predicted as belonging to a certain class i, the search for its counterfactual explanation $\boldsymbol{x}_{f_i}^{cf_j}$ for a class $j \neq i$ consists in determining the minimum joint variation $\Delta\mathbf{u}^*$ of the set of modifiable and partially modifiable features necessary to obtain the closest observation that belongs to class j:

$$\boldsymbol{x}_{f_i}^{cf_j} \doteq (\mathbf{u} + \Delta\mathbf{u}^*, \mathbf{z})_{f_i}^{cf_j} \tag{2}$$

The generators of counterfactual explanations used in this study, namely MUCH and DiCE, share some common features, for example: the capability to handle tabular datasets with mixed data (either continuous or categorical), the possibility to specify a set of modifiable and non-modifiable features, and the capability to provide constraints on during generation, that is, to provide a range of admissible values for each feature. To generate explanations that aim to improve or otherwise non-worsen the patient health status, a maximum acceptable value for each modifiable feature (Table I) is specified as an upper limit during counterfactual explanations generation. As it can be observed from Table I, these values indicate cut-off values that normally determine a clinically relevant worsening of the patient's health status (e.g., class 2 obesity or worse for BMI over 35 kg/m^2, hyperlipidemia for total cholesterol above 6.2 mmol/L).

1) MUCH: MUlti Counterfactuals via Halton sampling [19] is an algorithm to generate counterfactual explanations from Halton sampling of the output class distributions of any machine learning classifier. Specifically, for a given factual $\boldsymbol{x}_{f_i}$, $\Delta\mathbf{u}^*$ is estimated by solving the following minimization problem for all $j \in [m], j \neq i$:

$$\begin{aligned} \min_{\Delta\mathbf{u}\in\mathbb{R}^p} \quad & \mathrm{dist}\big(\boldsymbol{x}_{f_i}, (\mathbf{u} + \Delta\mathbf{u}, \mathbf{z})_{f_i}^{cf_j}\big) && \text{(3a)} \\ \text{subject to} \quad & \left\|(\mathbf{u} + \Delta\mathbf{u}, \mathbf{z})_{f_i}^{cf_j} - \mathbf{a}_j\right\|_2^2 \leq R_j^2 && \text{(3b)} \\ & \left\|(\mathbf{u} + \Delta\mathbf{u}, \mathbf{z})_{f_i}^{cf_j} - \mathbf{a}_k\right\|_2^2 \geq R_k^2, && \text{(3c)} \\ & \text{with } k \in [m] \text{ and } k \neq j, \end{aligned}$$

where dist$(\cdot,\cdot)$ is the selected distance metrics (e.g., the Euclidean norm), (3b) constraints $\boldsymbol{x}_{f_i}^{cf_j}$ to lie inside S_j and (3c) 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 $\boldsymbol{x}_{f_i} \in S_i$, a set $\mathbf{C}_{f_i} = \{\mathbf{x}_{f_i}^{cf_j} \mid j \in [m]; j \neq i\}$ of $m-1$ counterfactual explanations is found, 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 maximum size equal to $(m-1)|\mathbf{F}_i|$. Since the solution of (3) is not always feasible, a quasi-random sampling (Halton sequence) research in the output-class space is implemented, obtaining an approximate numerical solution of the optimization problem. Therefore, each $\boldsymbol{x}_{f_i}^{cf_j}$ is searched within the finite set of points in the sampled region $\tilde{S}_j$, aiming to minimize the distance with respect to $\boldsymbol{x}_{f_i}$.

2) DiCE: Diverse Counterfactual Explanation [7] is a state-of-the-art algorithm for the generation of counterfactual explanations. It is based on an optimization procedure such that, given a machine learning classifier, it is possible to find the minimal variation of the input features to change the output of the classification by minimizing a suitable cost function that allows for the possibility of returning more than one counterfactual explanation $\boldsymbol{x}_{f_i}^{cf_j}$ for each class $j \neq i$. An additional regularization term ensures diversity in the set of returned explanations, while keeping proximity with the factual. To allow for a fair comparison with MUCH, the total number of returned counterfactual explanations for each factual was set to 1. The optimization problem in the model agnostic framework can be solved using three different search methods, namely an independent random sampling of features, a genetic algorithm, or a K-dimensional tree. In this study, the heuristic genetic algorithm was selected as search method due to its faster convergence. However, using approximate search algorithms, there is no guarantee that DiCE will converge to a solution. Additionally, the obtained solution is, in general, sub-optimal as the algorithm is likely to get stuck in a local optimum.

D. Evaluation of counterfactual explanations

Counterfactual explanations were evaluated in this study based on several criteria: the percentage of returned counterfactual explanations (*availability*), the ability to distinguish the set of counterfactual explanations from the factual class (*discriminative power*, as computed in [19]), the average percentage of features changed (*sparsity*, as computed in [7]), the average distance from the factual observation (*proximity*), the average distance from real target observations (*implausibility*), and the average distance among the generated set of explanations (*diversity*). The last three metrics were computed as described in [12]. For sparsity, a tolerance of 0.1 was used. For implausibility, the barycenter of the target class, calculated from the training set, served as the reference. Higher values of availability, discriminative power, sparsity, and diversity indicate better counterfactual explanations, while lower values are preferable for proximity and implausibility.

The Wilcoxon Signed-Rank Test for paired samples was applied to assess possible statistical differences between counterfactual explanations and the corresponding factuals, whereas the Mann-Whitney U test was used to assess possible differences in counterfactual explanations generated with the two methods. The same test was used to compare counterfactual explanations generated from subpopulations of patients with/without comorbidities (hypertension, diabetes). A significance level $\alpha = 0.05$ was considered for statistical comparisons and Bonferroni correction was applied to correct for multiple comparisons.

E. Counterfactual conformity

Providing predictions that can guarantee a sufficiently high level of reliability is crucial in safety-critical areas like healthcare [30]. Likewise, offering reliable explanations is also fundamental to improve trustworthiness. With this aim, we introduce a measure of *counterfactual conformity* by adapting and expanding the concept of Conformal Prediction (CP) [20]. The rationale is to quantify the uncertainty of counterfactual explanations with respect to their ideal properties. According to CP theory [31], once defined:

- a *calibration set* $\mathcal{X}_c \times \mathcal{Y}_c$ with size n_c (typically, the size is greater than 500 observations [31]),

- a desired *error level* $\varepsilon \in (0,1)$,
- a real-valued *score function* s: $\mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$ measuring how much a label y is *conformal* to the sample $\boldsymbol{x}$,

for any $\boldsymbol{x} \in \mathcal{X}$, we can determine the following *prediction set* at *level of confidence* $1-\varepsilon$

$$\mathcal{C}_\varepsilon(\boldsymbol{x}) = \{\hat{y} \mid s(\boldsymbol{x}, \hat{y}) \leq s_\varepsilon\} \in 2^{\mathcal{Y}}, \tag{4}$$

where s_ε is the $\lceil (n_c+1)(1-\varepsilon)\rceil / n_c$ *quantile* of the score values computed on the calibration set. Hence, CP measures the uncertainty of predictions of machine learning models with a certain confidence level.

In this section, we draw inspiration from the CP framework to develop a new metric to evaluate the goodness of counterfactual explanations. This preliminary approach will not adhere strictly to the formalism of CPs since we are interested in quantifying the uncertainty related to the *generation* of counterfactual explanations rather than the prediction of a model, as CP do. Specifically, given an instance $\boldsymbol{x}$, we "substitute" the conformal label with the conformal counterfactual explanation, relying on the idea that a counterfactual uniquely belongs to its target class. The definition of counterfactual conformity here introduced assumes that the quality of a counterfactual explanation can be measured considering [18]:

1) the distance between the counterfactual explanation and its factual (i.e., the smaller the distance, the better the counterfactual, considering that the optimal counterfactual explanation should be, by definition, the minimal variation of the input parameters that realizes a change in the prediction label); and
2) the distance between the counterfactual explanation and the center of the corresponding counterfactual class (i.e., the smaller the distance, the more representative the counterfactual explanation is for the class).

The combination of these two requirements leads to a trade-off between the properties of *proximity* (i.e., the counterfactual explanation should be close to the classification boundary) and *implausibility* (i.e., the counterfactual explanation should be representative of the new class).

As a measure to assess, concurrently, the two properties, we defined a new score function as the weighted combination of the distances between the counterfactual explanation $\boldsymbol{x}_{f_i}^{cf_j}$ and its factual $\boldsymbol{x}_{f_i}$ and between the counterfactual explanation $\boldsymbol{x}_{f_i}^{cf_j}$ and the barycenter of the counterfactual class $\boldsymbol{x}_j^O$ (computed on the training set), respectively:

$$\begin{aligned} s(\boldsymbol{x}_{f_i}, \boldsymbol{x}_{f_i}^{cf_j}) = &\tau \cdot \text{mix\_dist}(\boldsymbol{x}_{f_i}, \boldsymbol{x}_{f_i}^{cf_j}) + \\ &(1-\tau) \cdot \text{mix\_dist}(\boldsymbol{x}_{f_i}^{cf_j}, \boldsymbol{x}_j^O) \end{aligned} \tag{5}$$

where $\tau \in (0,1)$ is a real valued weight and

$$\begin{aligned} \text{mix\_dist}(x,y) = &\left(\frac{\alpha}{\alpha+\beta}\right) \cdot \text{Hamming}(x,y) + \\ &\left(\frac{\beta}{\alpha+\beta}\right) \cdot \text{Cosine}(x,y) \end{aligned} \tag{6}$$

is a mixed distance borrowed from [12], with α being the number of categorical input features and β being the number of numerical input features. In this study, τ was set to 0.5 to give equal importance to the two contributions. Then, given a factual $\boldsymbol{x}_f$, the *conformal counterfactual set* of $\boldsymbol{x}_f$ is the set of all the counterfactual explanations $\boldsymbol{x}_{f_i}^{cf_j}$ such that the score value $s(\boldsymbol{x}_{f_i}, \boldsymbol{x}_{f_i}^{cf_j})$ is less than or equal to the almost $(1-\varepsilon)-$quantile s_ε computed on the calibration set, i.e.

$$\mathcal{C}_\varepsilon(\boldsymbol{x}_{f_i}) = \{\boldsymbol{x}_{f_i}^{cf_j} \mid s(\boldsymbol{x}_{f_i}, \boldsymbol{x}_{f_i}^{cf_j}) \leq s_\varepsilon\}. \tag{7}$$

With this interpretation of conformity, an empirical error is made whenever the conformal set of a factual does not contain the counterfactual explanation related to a certain class, as defined in the followings:

$$\text{err}_j = \Pr\{\boldsymbol{x}_{f_i}^{cf_j} \notin \mathcal{C}_\varepsilon(\boldsymbol{x}_{f_i}) | \varepsilon\} = \frac{\#\{\boldsymbol{x}_{f_i}^{cf_j} \notin \mathcal{C}_\varepsilon(\boldsymbol{x}_{f_i}) | \varepsilon\}}{\#\{\boldsymbol{x}_{f_i}^{cf_j}\}} \tag{8}$$

The term "fully conformal counterfactual" applies when the computed counterfactual explanations for a certain factual $\boldsymbol{x}_{f_i}$ (i.e., $\boldsymbol{x}_{f_i}^{cf_j}$) for all the classes $j \neq i$, adhere to the aforementioned conformity criterion. Conversely, the term "non-conformal counterfactual" is used when none of the computed counterfactual explanations for a certain observation meet the aforementioned criterion. The term "partially conformal counterfactual" is used in any other case.

The score function for counterfactual conformity was calibrated on 80% of the test set and the error was computed on the remaining 20%. Calibration was performed separately for MUCH and DiCE.

III. Results

The MC-SVDD classifier achieved an accuracy of 76.0% on the training set with 0.07% of unclassified points. The accuracy on the training set increased to 85.6% when FPR control (**Algorithm 1**) was applied, bringing the amount of unclassified points up to 10%. The approach taken by FPR control proves to be more reliable because it prefers to not classify the data points rather than misclassify them (before control: $\text{FP}_{\text{low risk}}$ = 49, $\text{FP}_{\text{moderate risk}}$ = 212, $\text{FP}_{\text{high risk}}$ = 1342; after FPR control: $\text{FP}_{\text{low risk}}$ = 41, $\text{FP}_{\text{moderate risk}}$ = 125, $\text{FP}_{\text{high risk}}$ = 103). The classification performance on the test set was slightly lower yet satisfactory. The accuracy was 70.2% (4.2% of unclassified points) before FPR control, which increased to 78.6% (11.1% of unclassified points) after FPR control. The class-specific sensitivity after FPR control was equal to 88.2% for low risk, 75.0% for moderate risk, and 95.9% for high risk on the training set; 83.3% for low risk, 69.0% for moderate risk, and 83.4% for high risk on the test set. The surrogate SVM model achieved high accuracy in predicting the output of the MC-SVDD (96.9% accuracy on the training set and 92.6% accuracy on the test set). The Cohen's Kappa coefficient was equal to 0.89, suggesting a satisfactory level of agreement between the MC-SVDD and the surrogate SVM models.

The set of factuals here considered included only those elements of the test set that were predicted as belonging to the high-risk class by the underlying classifier. This resulted in a factual set with 682 test records for MUCH and 690 for

DiCE. For each factual $\boldsymbol{x}_{f_{high}}$, two counterfactual explanations were generated, i.e. one from high to moderate risk class ($\boldsymbol{x}^{cf_{moderate}}_{f_{high}}$) and one from high to low risk class ($\boldsymbol{x}^{cf_{low}}_{f_{high}}$). Table II shows the performance of MUCH and DiCE in terms of availability, discriminative power, proximity, sparsity, implausibility and diversity. The two methods yielded a high percentage of explanations despite constraints in the generation process, with MUCH having an average availability of 84.6% and DiCE reaching 98.2%. Both methods produced counterfactual explanations that could be discriminated from points of the factual class with a satisfactory level of accuracy (i.e., discriminative power >77%, with MUCH performing better than DiCE). MUCH is slightly superior than DiCE in terms of implausibility, and diversity. Conversely, DiCE exhibits better proximity and sparsity compared to MUCH. Counterfactual explanations in the moderate risk class have worse discriminative power and slightly worse diversity but better proximity, sparsity and slightly better implausibility than those in the low risk class, for both methods.

Table III summarizes the error and size of the non-conformal, partially conformal and fully conformal counterfactual explanations sets of the two methods as a function of ε. From the first column of the table, we can notice that both the algorithms are well calibrated since the average error in the evaluation set (i.e., 20% of the test set) is close to the desired error level ε, hence representing a quasi-linear relationship. According to our definition of "counterfactual conformity", the higher the number of fully conformal counterfactual explanations, the more reliable the counterfactual extraction procedure is. Both methods here considered, for small values of ε, output a sufficiently high number of fully conformal counterfactuals, meaning that both counterfactual explanations $\boldsymbol{x}^{cf_{low}}_{f_{high}}$ and $\boldsymbol{x}^{cf_{moderate}}_{f_{high}}$ are representative of the target class while maintaining, by definition, also a minimal distance from the factual. In the following analysis, $\varepsilon = 0.1$ was selected as a compromise between the severity of the conformal check and the number of retained counterfactual explanations. Furthermore, in healthcare applications such as the one presented here, the use of a higher ε (i.e., a more selective filtering process with respect to counterfactual explanations) may assist in identifying more realistic explanations with regard to the necessary changes in features to determine a change in output class. In Table IV, conformal and non-conformal counterfactual explanations are compared in terms of desired properties. Conformal explanations exhibit superior quality in comparison to non-conformal ones (i.e., lower proximity and implausibility, higher diversity and sparsity). Non-conformal explanations demonstrate higher discriminative power, which can be attributed to their greater distance from the factual points (i.e., poorer proximity), thereby making them more readily distinguishable from the factual points. The comparison between the entire set of retrieved counterfactual explanations (Table II) and the conformal explanations (Table IV) shows improved quality after discarding non-conformal explanations, as suggested by the values of proximity, sparsity and implausibility observed, while diversity and discriminative power remain similar.

Regarding the use of the generated counterfactual explanations for the reduction of CVD risk in patients with COPD, a closer inspection revealed that non-conformal explanations are primarily associated with the generation of explanations with unrealistically high changes in feature values compared to the observed factual. As an example, Table V presents one conformal and one non-conformal factual-counterfactual pair (high to low risk transition) generated using MUCH. The two factuals shown in Table V describe male patients who are overweight, are aged between 60-65 years, and are diagnosed with diabetes and chronic hypertension. Notably, the non-conformal explanation (E2) is associated with higher changes in feature values compared to the conformal one (E1), some of which are unrealistic. For example, a lift in BMI from Class 1 obesity to Class 2 obesity and an increase in triglycerides are usually associated with increased CVD risk, and a decrease of about 40 mmHg in systolic blood pressure can be difficult to achieve from a clinical point of view.

Figure 2 shows the distributions (median, 25% and 75% percentiles) of the average changes requested by MUCH and DiCE to pass from the high risk class to the moderate risk class (panel 2a) and to the low risk class (panel 2b). To ensure a fair comparison between the two methods, only common factuals and only fully conformal counterfactuals were considered

TABLE II: Quality measures computed on counterfactual explanations generated with MUCH and DiCE methods: full set of explanations. ↑: Higher values indicate better quality; ↓: Lower values indicate better quality.

	Availability↑	**Discr Power**↑	**Proximity**↓	**Sparsity**↑	**Implausibility**↓	**Diversity**↑
			MUCH			
$\boldsymbol{x}^{cf_{moderate}}_{f_{high}}$	100.0%	94.6%	0.080	0.590	0.629	0.551
$\boldsymbol{x}^{cf_{low}}_{f_{high}}$	69.1%	98.8%	0.092	0.454	0.643	0.557
			DiCE			
$\boldsymbol{x}^{cf_{moderate}}_{f_{high}}$	98.0 %	77.0 %	0.002	0.792	0.749	0.545
$\boldsymbol{x}^{cf_{low}}_{f_{high}}$	98.4 %	92.3%	0.009	0.658	0.757	0.549

TABLE III: Error and size of the non-conformal, partially-conformal, and fully conformal sets at varying desired error levels (ε).

	Error			**Size**		
	Average error	High → Moderate error	High → Low error	Non conformal	Partially conformal	Fully conformal
			MUCH			
$\varepsilon = 0.01$	0.006	0.000	0.011	0.000	0.011	0.989
$\varepsilon = 0.05$	0.050	0.044	0.056	0.022	0.056	0.922
$\varepsilon = 0.10$	0.117	0.122	0.111	0.067	0.100	0.833
			DiCE			
$\varepsilon = 0.01$	0.011	0.015	0.008	0.008	0.008	0.985
$\varepsilon = 0.05$	0.050	0.053	0.046	0.046	0.008	0.947
$\varepsilon = 0.10$	0.141	0.160	0.122	0.084	0.114	0.801

TABLE IV: Quality of counterfactual explanations generated with MUCH and DiCE methods: conformal vs non-conformal explanations ($\varepsilon = 0.1$). ↑: Higher values indicate better quality; ↓: Lower values indicate better quality.

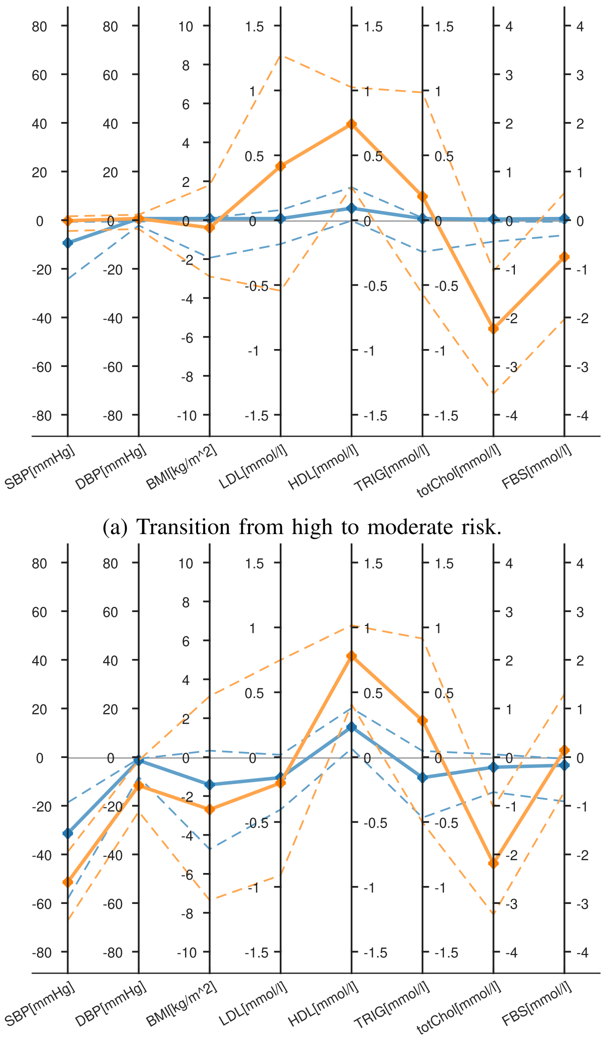
	Type	Discr Power↑	Proximity↓	Sparsity↑	Implausibility↓	Diversity↑
		MUCH				
$x^{cf_{moderate}}_{f_{high}}$	Non Conformal	99.49%	0.133	0.559	0.925	0.004
	Fully Conformal	94.34%	0.077	0.591	0.576	0.545
$x^{cf_{low}}_{f_{high}}$	Non Conformal	99.85%	0.143	0.461	0.932	0.002
	Fully Conformal	98.43%	0.080	0.454	0.599	0.558
		DiCE				
$x^{cf_{moderate}}_{f_{high}}$	Non Conformal	97.31%	0.003	0.736	1.151	0.227
	Fully Conformal	77.27%	0.002	0.801	0.692	0.526
$x^{cf_{low}}_{f_{high}}$	Non Conformal	98.47%	0.012	0.613	1.162	0.229
	Fully Conformal	92.65%	0.009	0.664	0.700	0.529

TABLE V: Examples of conformal (E1) and non-conformal (E2) counterfactual explanations generated using MUCH and setting $\varepsilon = 0.1$.

		sBP [mmHg]	dBP [mmHg]	BMI [kg/m^2]	FBS [mmol/L]	LDL [mmol/L]	HDL [mmol/L]	TG [mmol/L]	totChol [mmol/L]
E1	$x_{f_{high}}$	150	78	31.9	8.0	4.7	0.9	2.7	6.8
	$x^{cf_{low}}_{f_{high}}$	124	72	29.7	6.9	4.5	2.0	2.5	3.2
E2	$x_{f_{high}}$	138	72	32.1	8.5	3.0	1.1	1.3	4.8
	$x^{cf_{low}}_{f_{high}}$	97	81	37.7	6.2	1.37	2.1	2.8	1.3

(N=337, $\varepsilon = 0.1$). Counterfactual explanations generated with MUCH and DiCE differ in terms of changes in modifiable characteristics requested for moving from high to moderate risk and from high to low risk, with MUCH suggesting larger variations than DiCE. Statistically significant changes (i.e., variations in feature values statistically different from 0) were observed in transitions from high to moderate risk in terms of sBP, BMI, LDL, HDL, TRIG, totChol, and FBS for MUCH, and in terms of sBP, BMI, HDL, TRIG, and totChol for DiCE. In transitions from high to low risk, statistically significant changes in terms of all modifiable features, except FBS and LDL were observed using MUCH, and statistically significant changes in terms of all modifiable features were observed using DiCE.

The counterfactual explanations obtained with MUCH are statistically different from those obtained with DiCE for sBP, LDL, HDL, TRIG, totChol, and FBS when analysing high to moderate risk transitions (Figure 2a) whereas all features distributions except for BMI and LDL are statistically different when comparing MUCH and DiCE in terms of high to low transitions (Figure 2b). The changes suggested by MUCH to reduce the risk class are associated, on average, with a reduction in sBP and dBP, BMI, and totChol and an increase in HDL. These trends are coherent with a general improvement in the patients' health status and a reduction in cardiovascular risk, i.e., decreased blood pressure, lower weight, and better lipidic profile. The counterfactual explanations generated by the two methods were also consistent with the individual characteristics, for example in relation to the comorbidities here considered (diabetes and hypertension). Specifically, by comparing conformal counterfactual explanations with their factuals, we observed a greater median decrease in sBP for patients with stage 2 hypertension compared to non-hypertensive ones considering both high to moderate risk transitions (21.00 mmHg higher with DiCE, p = 3.48×10^{-38}; 1.91 mmHg higher with MUCH, p = 1.83×10^{-9}) and high to low risk transitions (39.00 mmHg higher with DiCE, p = 2.44×10^{-33}; 25.77 mmHg higher with MUCH, p = 1.43×10^{-29}).



(a) Transition from high to moderate risk.

(b) Transition from high to low risk.

Fig. 2: Distributions of conformal counterfactual explanations ($\varepsilon = 0.1$) simulating transitions from high to moderate (2a) and from high to low (2b) CVD risk, obtained using MUCH (in orange) and DiCE (in blue), respectively. Solid lines: medians of the distributions; dashed lines: 25% and 75% percentiles.

Similarly, a significantly higher median decrease in FBS is observed for diabetic patients compared to non diabetic ones considering both high to moderate (0.90 mmol/L higher with DiCE, p = 7.70×10^{-12}; 1.17 mmol/l higher with MUCH, p = 2.50×10^{-7}) and high to low transitions (1.40 mmol/L higher with DiCE, p = 1.48×10^{-10} ; 1.09 mmol/l higher with MUCH, p = 8.89×10^{-11}).

IV. Discussion

This study investigated multi-class counterfactual explanations as an original, interpretable data-driven method to support the design of tailored disease prevention strategies. Furthermore, a new metric named counterfactual conformity was introduced to ensure reliability for the end user by providing a confidence value for each explanation produced and enabling rejection of non-conformal explanations.

The proposed approach has been applied to estimate personalized recommendation for reducing the 10-year CVD risk in COPD patients. The importance of reducing CVD risk in patients with COPD is well documented in the literature (e.g., [34]), and COPD patients may present a two to five times higher likelihood of CVD occurrence with respect to non-COPD subjects [35]. Cardiovascular conditions in patients with COPD can lead to further complications and more difficult disease management and CVD prevention is key to reduce the individual risk. Currently, the presence of CVDs in patients with COPD is mainly treated following general CVDs guidelines [36], [23]. However, the use of personalized strategies derived specifically on patients with COPD could lead to more effective disease prevention and patient management.

Generation of counterfactual explanations. Counterfactual explanations retrieved from a set of common factual observations using MUCH and DiCE showed differences in terms of suggested changes for most of the features. The higher changes and higher discriminative power observed using MUCH (Figure 2, Table II) are largely attributable to the reliance of MUCH on the shape of the classification regions provided by MC-SVDD, which are further refined and narrowed by the FPR control algorithm. In general, DiCE has better proximity (i.e., the generated counterfactual explanations are closer to the factual) compared to MUCH, and this metric is further improved once the non-conformal counterfactuals are filtered out, keeping only the conformal ones (Table IV). The better proximity is reflected in the less pronounced variation trend, especially in transitions from high to moderate risk (Figure 2a). DiCE provides a higher availability of explanations in the low risk class whereas the two methods have similar availability of explanations in the moderate class.

In both methods, the number of retrieved explanations is reduced compared to a theoretical value of 100% due to two types of constraints related to the design of the generation algorithms, specifically: (i) non-modifiability of a subset of features and (ii) rejection of candidate explanations with one or more features exceeding the maximum acceptable values in Table I. In general, the performance of the algorithms depends on the design criteria and the more constraints are imposed (e.g., the higher the number of non-modifiable features and/or the lower the maximum acceptable values), the more challenging it is to find a solution during the optimization phase.

The use of counterfactual explanations in clinical applications holds potential as a data-driven method for identifying personalized minimum viable changes to decrease the individual risk [14]. However, there is not a one-size-fits-all solution as there may be differences in the explanations generated using different algorithms that should be evaluated by the physician on a case-by-case basis, based on patient characteristics and clinical feasibility. For example, some counterfactual explanations might suggest a significant change in biomarkers (e.g., blood pressure, BMI, triglycerides, as shown in the Example E2 in Table V) that could be deemed unrealistic or unfeasible to achieve in practice, even with the help of medications and intensive lifestyle interventions.

Counterfactual conformity. To facilitate a semi-automatic approach for selecting counterfactual explanations, we introduced the concept of counterfactual conformity, a novel quality metric for filtering out explanations that are not compliant with the desired properties. The definition of a score function as in (5) combines the measurement of two key properties: proximity and plausibility. Analyzing how the produced explanations are statistically distributed with respect to these properties can help understand the *global* quality of the generated counterfactual explanations. Furthermore, by using counterfactual conformity, each explanation is accompanied by a *local* reliability value. Table IV effectively shows that counterfactual explanations deemed conformal exhibit better quality compared to non-conformal ones in terms of desired properties, demonstrating how this value provides the physician with additional information to determine whether to consider or discard the specific output.

Limitations and future research. This study evaluates counterfactual explanations from a computational perspective using a range of quality measures and the newly introduced conformity metric. However, it is subject to certain limitations. Despite the high agreement between MC-SVDD and its surrogate SVM (Cohen's Kappa coefficient equal to 0.89), it is important to acknowledge that the comparison between MUCH and DICE may be slightly influenced by the differences between the two underlying classifiers. Additionally, to ensure the effective application of this approach in practice, it is essential to guarantee the clinical feasibility of the proposed interventions. Preliminary findings suggest that conformal counterfactual explanations may be more realistically applicable than non-conformal ones. To fully achieve the goal of estimating viable recommendations for disease prevention, further research should focus on incorporating medical knowledge into the counterfactual generation process and into the definition of counterfactual conformity, for example by defining expert-driven dynamic bounds that indicate a plausible range of acceptable changes for each subject.

Although preliminary, the counterfactual conformity measure here introduced is a step towards a more precise methodology for assessing the quality of counterfactual explanations. However, an optimal value of ε has not yet been defined. In the future, it will be necessary to establish criteria for selecting ε by balancing the trade-off between the number of discarded points and desired characteristics, defined by a combination of quality metrics and expert knowledge. Moreover, further research should include a deeper investigation of counterfactual conformity on a wider range of datasets and applications and in relation to various measures of counterfactual quality.

V. Conclusion

This study presented an XAI based methodology for the extraction of target risk-reduction strategies from electronic medical data in the form of counterfactual explanations. The parallel use of MUCH and DiCE demonstrated the potential of creating high quality explanations with respect to standard criteria such as discriminative power, proximity, plausibility, and sparsity using different classifiers and generators. Moreover, by introducing the counterfactual conformity measure, we ensured the possibility to discard all those counterfactual explanations that did not guarantee the desired level of compliance in terms of target properties. The results of this study are promising and may have implications in the field of personalized medicine, specifically offering clinicians actionable suggestions to reduce the risk of developing cardiovascular complications in patients already diagnosed with COPD. These explanations, being accompanied by global and local quality metrics, can provide an additional tool to choose among potential alternative strategies based on the desired requirements. However, further validation is necessary to confirm these findings. In the future, the proposed procedure may potentially generalize to other chronic diseases and it may be integrated in clinical decision support tools.

Acknowledgment

Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the Italian Ministry for Universities and Research (MUR). Neither the European Union nor the Italian Ministry for Universities and Research (MUR) can be held responsible for them. 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. Maurizio Mongelli and Alessia Paglialonga are part of RAISE Innovation Ecosystem.

References

[1] H. Hakkoum, I. Abnane, and A. Idri, “Interpretability in the medical field: A systematic mapping and review study,” *Applied Soft Computing* vol. 117, pp. 108391, 2022. doi: 10.1016/j.asoc.2021.108391.

[2] M.A. Alsalem, A.H. Alamoodi, O.S. Albahri, A.S. Albahri, Luis Martínez, R. Yera, A.M. Duhaim and I. M. Sharaf, “Evaluation of trustworthy artificial intelligent healthcare applications using multi-criteria decision-making approach”, *Expert Systems with Applications*, vol. 246, pp. 123066, 2024. doi: 10.1016/j.eswa.2023.123066.

[3] M. Frasca, D. La Torre, G. Pravettoni, and I. Cutica, “Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review,” *Discover Artificial Intelligence*, vol. 4, no. 15, 2024. doi: 10.1007/s44163-024-00114-7.

[4] W. Yang, Y. Wei, H. Wei, Y. Chen, G. Huang, X. Li et al., “Survey on Explainable AI: From Approaches, Limitations and Applications Aspects,” *Human-Centric Intelligent Systems*, vol. 3, pp. 161188, 2023. doi: 10.1007/s44230-023-00038-y.

[5] C. Combi, B. Amico, R. Bellazzi, A. Holzinger, J.H. Moore, M. Zitnik, and J.H. Holmes, “A manifesto on explainability for artificial intelligence in medicine”, *Artificial Intelligence in Medicine*, vol. 133, pp.102423,2022. doi: 10.1016/j.artmed.2022.102423.

[6] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: automated decisions and the GDPR,” *Harvard Journal of Law and Technology*, vol. 31, no. 2, pp. 841887, 2018.

[7] R.K. Mothilal, A. Sharma, and C. Tan, “Explaining machine learning classifiers through diverse counterfactual explanations," Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020.

[8] D. Mahajan, C. Tan, and A. Sharma, “Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers”, *arXiv*, 2019. doi: 10.48550/ARXIV.1912.03277.

[9] K. Kanamori, T. Takagi, K. Kobayashi, and H. Arimura, “DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization,” Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, pp. 2855–2862, 2020. doi: 10.24963/ijcai.2020/395.

[10] T. D. Duong, Q. Li, and G. Xu,“Causality-based counterfactual explanation for classification models,” *Knowledge-Based Systems*, vol. 300, pp. 112200,2024. doi: 10.1016/j.knosys.2024.112200.

[11] W. Yang, J. Li, C. Xiong, and S.C.H. Hoi, “MACE: An Efficient Model-Agnostic Framework for Counterfactual Explanation”, 2022, textitarXiv:2205.15540.

[12] F. Bodria, R. Guidotti, F. Giannotti, and D. Pedreschi, “Transparent Latent Space Counterfactual Explanations for Tabular Data”, IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), pp.1-10, 2022. doi:10.1109/DSAA54385.2022.10032407.

[13] S. Mertes, T. Huber, K. Weitz, A. Heimerl, and E. André, “GANterfactual-Counterfactual Explanations for Medical Non-experts Using Generative Adversarial Learning,” *Frontiers in Artificial Intelligence*, vol. 8, no. 5, pp. 825565, 2022. doi: 10.3389/frai.2022.825565.

[14] 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. 7, no. 11, pp. e0272825. doi: 10.1371/journal.pone.0272825.

[15] H. Benkirane, M. Vakalopoulou, S. Michiels, P.H. Cournède, and W. Lotter, “Counterfactual Analysis for Digital Histopathology Slides Using Human Interpretable Features”, *Medical Imaging with Deep Learning*, 2024. https://openreview.net/forum?id=0Pod0c2Av6

[16] D. Console, M. Lenatti, D. Simeone, K. Keshavjee, A. Guergachi, M. Mongelli, and A. Paglialonga, “Exploring Prediabetes Pathways Using Explainable AI on Data from Electronic Medical Records,” *Studies in health technology and informatics*, vol. 316, pp. 736-740, 2024. doi: 10.3233/SHTI240519

[17] W. Jin, X. Li, M. Fatehi, and G. Hamarneh, “Guidelines and evaluation of clinical explainable AI in medical image analysis,” *Medical Image Analysis*, vol. 84, pp. 102684, 2023. doi: 10.1016/j.media.2022.102684.

[18] R. Guidotti, “Counterfactual explanations and how to find them: literature review and benchmarking,” *Data Mining and Knowledge Discovery*, vol. 38, pp. 2770-2824, 2024. doi: 10.1007/s10618-022-00831-6.

[19] A. Carlevaro, M. Lenatti, A. Paglialonga, and M. Mongelli, “Multi-Class Counterfactual Explanations using Support Vector Data Description," in *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 6, pp. 3046-3056, 2024. doi: 10.1109/TAI.2023.3337053

[20] V. Vovk, A. Gammerman and G. Shafer, “Algorithmic Learning in a Random World”. Ed. New York, NY, USA: Springer-Verlag, 2005.

[21] A. Carlevaro et al., (2024) MUlti Counterfactual Halton sampling [Source code]. https://github.com/AlbiCarle/MUCH/tree/main/CVD_risk.

[22] Canadian Primary Care Sentinel Surveillance Network (CPCSSN). [Online] Available: http://cpcssn.ca/. Accessed on: Jan 28, 2024.

[23] K. Brassington, S. Selemidis, S. Bozinovski, and R. Vlahos, “Chronic obstructive pulmonary disease and atherosclerosis: common mechanisms and novel therapeutics,” *Clinical Science,* vol. 136, no. 6, pp. 405423, 2022. doi: https: 10.1042/CS20210835

[24] “Framingham Risk Score Worksheet. Canadian Cardiovascular Society.” [Online] Available: https://ccs.ca/app/uploads/2020/12/FRS_eng_2017_fnl_greyscale.pdf. Accessed on: Jan 16, 2024.

[25] Y. L. Chen, Y.F. Zheng, and Y. Liu, “Margin and Domain Integrated Classification for Images,” *International Journal of Information Acquisition*, vol. 08, no. 1, pp. 1-16, 2011. doi: 10.1142/S0219878911002343.

[26] H. C. Kraemer, “Kappa Coefficient,” *Wiley StatsRef: Statistics Reference Online*. Wiley, pp. 14, 2014. doi: 10.1002/9781118445112.stat00365.pub2.

[27] D.M. Tax and R.P. Duin, “Support Vector Data Description,” *Machine Learning,* vol. 54, pp. 4566, 2004. doi: 10.1023/B:MACH.0000008084.60811.49.

[28] 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. 41546, 1909. Available: `http://www.jstor.org/stable/91043`. Accessed on: Jan. 14, 2024.

[29] A. Carlevaro and M. Mongelli, “A New SVDD Approach to Reliable and Explainable AI," in *IEEE Intelligent Systems*, vol. 37, no. 2, pp. 55-68, 1 March-April 2022, doi: 10.1109/MIS.2021.3123669.

[30] X. Huang, G. Jin, and W. Ruan, “Machine Learning Safety”. Springer Nature Singapore, 2023. doi: 10.1007/978-981-19-6814-3.

[31] A.N. Angelopoulos and S. Bates, “Conformal prediction: A gentle introduction." *Foundations and Trends in Machine Learning*, vol. 16, no. 4, pp. 494591, 2023. doi: 10.1561/2200000101.

[32] G. Shafer, and V. Vovk, “ A tutorial on conformal prediction”, *The Journal of Machine Learning Research*, vol. 9, pp. 371-421, 2008. doi: 10.5555/1390681.1390693.

[33] G. Zeni, M. Fontana and S. Vantini, “Conformal Prediction: a Unified Review of Theory and New Challenges,” *Bernoulli*, vol. 29, no. 1, pp. 1-23, 2023. doi: 10.3150/21-BEJ1447.

[34] A. D. Morgan, R. Zakeri, and J. K. Quint, “Defining the relationship between COPD and CVD: what are the implications for clinical practice?,” In *Therapeutic Advances in Respiratory Disease,* SAGE Publications, vol. 12, 2018. doi: 10.1177/1753465817750524.

[35] W. Chen, J. Thomas, M. Sadatsafavi, and J. M. FitzGerald, “Risk of cardiovascular comorbidity in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis,” *The Lancet Respiratory Medicine,* vol. 3, no. 8, pp. 631-639, 2015. doi: 10.1016/S2213-2600(15)00241-6.

[36] K. Brassington, S. Selemidis, S. Bozinovski, and R. Vlahos, “New frontiers in the treatment of comorbid cardiovascular disease in chronic obstructive pulmonary disease,” *Clinical Science,* vol. 133, no. 7, pp. 885904, 2019. doi: 10.1042/CS20180316.

Marta Lenatti received a Master Degree in Biomedical Engineering, from Politecnico di Milano, Italy. She is now a PhD student of the Italian National PhD program on Artificial Intelligence (Health and life sciences area) in collaboration with CNR-IEIIT, and a Visiting Scientist at Toronto Metropolitan University. Her research interests are related to Explainable AI for prediction and management of chronic diseases.

Alberto Carlevaro received a Master Degree in Applied Mathematics from the University of Genoa, Italy. He is now a PhD student in the Department of Electrical, Electronic and Telecommunications Engineering and Naval Architecture, in collaboration with CNR-IEIIT and Aitek. He has been visiting research scholar at the department of Electrical engineering and Computer Science of UC Berkeley, California USA. His fields of research are Machine Learning, Conformal Prediction and Explainable AI.

Aziz Guergachi is a tenured full professor in Toronto Metropolitan University, Canada. He holds a Bachelor of Science in Pure Mathematics from Aix-Marseille Université, and a Bachelors degree in Engineering from Ecole Centrale Marseille, France. He holds a PhD in Engineering from the University of Ottawa. Over the last decade, he has been an instructor for the MBA course on Product Development and Commercialization. He is an adjunct professor in the Department of Mathematics and Statistics at York University.

Karim Keshavjee is a family physician by training with over 25 years of experience in designing and implementing technology for clinical research. His current research is focused on how to use AI in the service of diabetes prevention through the PREVENT program. He is an Assistant Professor, Teaching Stream and Program Director for the Health Informatics program at the University of Toronto and a Visiting Researcher at Toronto Metropolitan University.

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

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