Interpretable Deep Learning for Biological Age Prediction: A Counterfactual Approach to Personalized Health Insights

Bhavana Nare

Abstract—**The accurate estimation of biological age from physical activity data has the potential to revolutionize personalized health monitoring and early disease detection. However, existing deep learning models often lack interpretability, limiting their practical application in real-world healthcare settings. In this study, we propose an Explainable Time Series Regression (XTSR) framework that integrates deep learning with counterfactual reasoning to enhance model transparency and user trust. Our approach employs a hybrid Time Series Extrinsic Regression (TSER) model, trained on large-scale wearable sensor data, to predict biological age while simultaneously generating counter- factual explanations. By identifying the most influential activity patterns contributing to aging predictions, our system offers actionable recommendations for personalized health optimization. Experimental results demonstrate that our model outperforms traditional regression methods, achieving higher accuracy and interpretability. This research bridges the gap between predictive analytics and human-centered AI, paving the way for intelligent and user-friendly health monitoring systems that provide action- able insights based on individual behavior patterns.**

Index Terms—**Deep Learning, Time Series, Extrinsic Regression, Counterfactuals, Explanations.**

I. INTRODUCTION

Healthcare costs are globally increasing due to an aging population, technological advancements, medication errors, and a rise in annual spending on medicines (1; 2). The aging population, poor diet, physical inactivity, and tobacco use (including secondhand smoke) (3) contribute to the prevalence of chronic diseases, which are a major cause of deaths among the populations (4). Digital health (5) can fulfill the need for healthcare accessible to everyone, regardless of location or time, while also improving the quality of care and reducing costs. The use of portable edge devices with sensing capabilities allows for the remote monitoring of patient health data, which can be particularly helpful for those with chronic conditions (6). With wearable sensors, mobile phones, or other edge devices, patients can easily record physiological and behavioral data, which can be aggregated to create digital biomarkers that explain, influence, or predict health-related outcomes. Passively measured data such as vital signs, physical activity, and other health-related data allows patients to monitor their health condition without visiting a healthcare provider (7). These real-time and remote monitoring capabilities not only improve patient outcomes but can also reduce healthcare costs by minimizing the need for frequent visits with clinicians.

Processing huge amounts of sequence data typically requires versatile, high-performing, and highly generalizing *DL* mod- els. However, most existing studies have concentrated on tertiary prevention (8), which aims to prevent disease recurrence or complications. Tertiary prevention only deals with diseases that have already occurred and does not proactively reduce the burden on healthcare systems. Therefore, it is crucial to shift the focus towards the early detection of diseases (secondary prevention) or even preventing diseases from occurring in the first place (primary prevention) (9). Both primary and secondary prevention can be very beneficial in preventing the onset of serious health concerns. Research in this field has been limited due to uncertainty about which factors to examine. In general, it is difficult to evaluate the overall health condition of a healthy individual in the absence of disease symptoms.

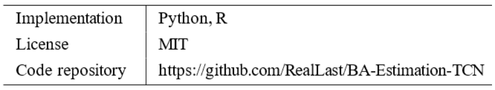
One potential method to determine the general health state of a person is by using the concept of biological age. Prior work has shown that it is possible to use deep learning models to predict a person's biological age non-invasively using physical activity data (10; 11). Nevertheless, existing methods of predicting biological age lack an explanation about what a person can do to improve their health state in general. It is widely known that general recommendations such as taking a minimum number of steps each day (12), maintain- ing good sleeping habits (13), and engaging in recreational activities (14) have a significant impact on health outcomes. However, specific recommendations tailored to the individual's needs are currently lacking, making it difficult to identify what changes to make at a personal level. In order to provide these recommendations, the individual needs to understand how the model assesses their health status.

Despite achieving great performance, *ai* models are limited due to being seen as a black box, resulting in low practical use, especially in healthcare. *xai* helps developers, domain experts, and users understand how *DL* models work and how they make predictions (15). Many state-of-the-art tools for explaining *DL* models rely on visually highlighting important input data areas, which is useful for developers or domain experts but hard for patients to understand. Counterfactual ex- planation systems (16) aim to support counterfactual reasoning by modifying the input data to lead to a different prediction by the model. That way, the users of counterfactual explanation systems are provided with a fully diverse type of illustrative information that complies with the *gdpr*(17) and are easy for humans to understand (18). There is a tremendous potential for counterfactual explanations in the mobile health setting (19). Yet, many of the *xai* techniques predominantly deal with images or texts; time series data has attracted less interest, and the few techniques developed for time series are focused on tasks such as classification or forecasting (20). Especially in medical contexts, where relevant information often consists of time-dependent information, high-quality time series counterfactuals have the potential to give meaningful insights into

decision processes.

With our work, we make the following contributions:

- We present a novel approach for generating counterfactual explanations for time series extrinsic regression.
- We use our approach to adjust four counterfactual methods for time series classification to time series extrinsic regression.
- We compare both qualitatively and quantitatively generated counterfactual explanations in a mobile health setting to estimate biological age from physical activity data.
- We illustrate how counterfactual explanations can be used to generate expressive text recommendations and provide continuous health supervision, thus reducing the need for external supervision and, consequently, healthcare costs.



Implementation	Python, R
License	MIT
Code repository	https://github.com/RealLast/BA-Estimation-TCN

II. RELATED WORK

This section discusses concepts related to Explainable Time Series Extrinsic Regression. First, we examine previous work done in this field. Next, we explore the available options to explain regression models. Finally, we take a closer look at Counterfactuals, a user-oriented explainable method that is extremely useful in mobile health monitoring.

A. Time-Series Extrinsic Regression

tser describes the task of predicting a continuous external variable from a time series. The term *Extrinsic* refers to the variable not being inherently part of the time series distribution. Instead, the time series serves as input to a model, which then infers an additional variable, such as a score. For example, *tser* can be used for *hr* estimation using *ppg* sensors (21). *tser* is closely related to *tsc* and *tsf*. The goal of *tsc* is to understand the relationship between a time series and a categorical variable. For instance, *tsc* learns how the shape of an *ecg* signal changes during diseases such as myocardial infarction or atrial fibrillation (22). *tsf* consists of analyzing a signal and predicting the future values of the same signal. For example, *tsf* is useful in finance when forecasting the closing price of a stock each day (23). Tan et al. (24) formalized the definition of *tser* and assessed popular regression techniques such as Support Vector Machine (25), Linear Regression, and Residual Networks (26) on a new archive consisting of nineteen *tser* datasets (27). More recently, Guijo et al. (28) extended the dataset archive (27) of *tser* problems and implemented new *tser* algorithms based on *tsc* methods, FreshPRINCE and DrCIF. The first is a robust pipeline algorithm that performs regression using two key components - the TSFresh feature extraction algorithm and the Rotation Forest (RotF) (29) estimator. The TSFresh algorithm transforms the input time series data into a feature vector fed to the RotF estimator for model training and label prediction. DrCIF is a type of tree ensemble that generates features by analyzing summary statistics over random intervals.

B. Prior work on explainable time-series analysis

Using findings of prior work by Rojat et al. (30) as a starting point, we investigated numerous explainable methods for time series data published in recent literature. Many methods have been developed for this purpose, and to choose the most suitable method for a specific use case, a few criteria need to be considered. First, it is crucial to understand *what* the method aims to explain. Explainable methods usually have one or multiple goals; they showcase a model's robustness to adversarial attacks, stability to data noise, the trustworthiness of the model's outputs, interactivity with users, explainability to a particular audience, or interpretability by the developer. Second, depending on what the explanation should achieve, explainability *scopes* vary, encompassing local and global perspectives. Local explanations provide insights into individual behaviors, while global explanations discern broader population trends. Another important consideration is the *target audience* for the explanation. Some explanations are designed for developers, while others target experts in specific fields, such as healthcare. Finally, some explanations are intended for end-users themselves. Lastly, the DL *model* they explain has an impact. For example, post-hoc methods wrap an explainability module around the model to generate explanations. They can be model-specific, only usable for a certain model type, or model-agnostic (31), versatile across different models. On the other hand, there are ante-hoc techniques that integrate the explanation module inside the model architecture and provide explanations after the model's training phase (30). For that reason, they only work with specific model architectures.

Multiple sets of methods have been presented in prior literature to accomplish these various objectives (robustness, trustworthiness, interpretability, target audience, scope, etc):

- *Backpropagation-based* methods (26; 32; 33; 34; 35; 36; 37; 38; 39; 40) allow for network explanations through a single forward and backward pass. They are post-hoc model-specific methods, meaning they depend on the mo-del architecture. They use the *cam* (41), a post-hoc method that shows which part of the input is responsible for the classifier output.

TABLE I: Mean and variance of the scores obtained by each counterfactual generation technique on five evaluation metrics for CNN.

Method	Validity ↑	Proximity ↓	Sparsity ↓	Plausibility ↑	Time ↓
TSEvoR	0.37±0.48	159.99±274.71	0.04±0.09	0.39±0.25	8m01s±2m52s
NUNR	0.99±0.08	4740.66±1149.04	0.98±0.02	0.47±0.20	0m01s±0m01s
DBAR	0.10±0.31	2499.74±702.91	0.99±0.03	0.02±0.11	9m28s±4m58s
Wachter	0.00±0.00	0.00±0.00	1.00±0.00	0.00±0.00	1m01s±0m28s

TABLE II: Mean and variance of the scores obtained by each counterfactual generation technique on five evaluation metrics for ConvLSTM.

Method	Validity ↑	Proximity ↓	Sparsity ↓	Plausibility ↑	Time ↓
TSEvoR	0.08±0.27	450.75±445.8	0.07±0.10	0.46±0.29	47m36s±11m21s
NUNR	0.99±0.08	3944.03±796.31	0.98±0.02	0.59±0.25	0m01s±0m01s
DBAR	0.09±0.29	1898.02±328.64	1.00±0.02	0.04±0.15	11m33s±4m32s
Wachter	0.02±0.12	0.00±0.00	0.99±0.02	0.14±0.27	6m32s±1m48s

TABLE III: Mean and variance of the scores obtained by each counterfactual generation technique on five evaluation metrics for TCN.

Method	Validity ↑	Proximity ↓	Sparsity ↓	Plausibility ↑	Time ↓
TSEvoR	0.39±0.49	147.37±242.41	0.04±0.09	0.41±0.26	10m01s±1m14s
NUNR	0.99±0.08	4740.66±1149.04	0.98±0.02	0.51±0.23	0m01s±0m01s
DBAR	0.10±0.30	2344.37±639.02	1.00±0.02	0.01±0.07	12m03s±5m54s
Wachter	0.00±0.00	0.00±0.00	1.00±0.00	0.00±0.00	1m09s±0m03s

- *Perturbation* methods (42; 43; 44; 45; 46; 47; 48) make direct changes to the input by either masking, transform- ing, or mutating certain parts of it. After modifying a part of the input, a forward pass is performed to calculate the difference with the initial input. If the difference is high, it indicates that the modified part significantly impacts the model's decision. These methods are useful because they are model-agnostic. They treat the model as a black box and offer flexibility across arbitrary *DL* architectures.
- *Attention-based* methods (49; 50; 51; 52; 53; 54; 27; 55; 56; 57) use the weights of the attention layer that represents the importance that the mechanism assigns to different parts of the input. For instance, Gao et al. (57) used the attention weights to visualize the feature contribution to the model output as a line plot and the temporal contribution as a heatmap. Attention-based methods are an example of ante-hoc model-specific meth- ods. It is worth noting that Attention-based explanations are currently a matter of debate in the research field. In the context of *nlp* tasks, Jain et al. (58) have claimed that attention weights do not explain predictions clearly. However, Wiegreffe et al. (59) have disagreed with this claim. It is important to mention that this debate is not limited to *nlp* tasks alone. Bibal et al. (60) have analyzed the debate for various data modalities.
- *Fuzzy-logic &* sax methods (61; 62; 63; 64) are specific to time series data. *sax* (61; 62) transforms the time series into informative segments, which are then assigned to a symbol, allowing the detection of recurrent patterns in the data. Fuzzy logic (63; 64) is a type of logic that deals with approximate rather than precise reasoning. It allows for the inclusion of uncertainty in decision-making processes. By assigning degrees of membership to different features or classes, fuzzy membership functions can help explain why a certain decision was made
- *Shapelets* methods (65; 66; 67) identify discriminative subsequences, called shapelets, that can be used to classify the time series. Shapelets are patterns derived from a group of time series or learned to minimize a specific objective function. There are various methods available to discover the shapelets. One method (68) involves training a classifier first and then extracting the shapelets to explain it. However, this approach can be computationally expensive but offers the flexibility of a model-agnostic method. Other approaches (69) involve learning the shapelets representations while simultaneously training the model. This results in an effective, ante-hoc, explainable approach.
- *Prototypes* methods, such as Gee et al. (70) and Li et al. (71), use the latent space created by deep learning models to understand the impact of meaningful represen- tations on the decision-making process. These methods treat prototypes as representative individuals of a class, where a prototype represents a concept learned by the model, such as how the model represents a cat or a dog. According to Obermair et al., (72), a concept is defined as explanatory data containing all the relevant properties humans require to make the same decisions as the black box model.
- *Counterfactual* methods (17; 73; 74; 75; 76) and pertur- bation methods are two techniques that involve changing the input data to study the behavior of machine learning models. However, they differ in their goals. Perturbation methods identify the input features that contribute to the model's decision. In contrast, counterfactuals aim to produce a modified input that the model classifies differently by changing these important input features. To achieve this objective, counterfactuals search for the smallest possible alteration in the input data that could result in a different model output.

In *tsxai*, to understand a model's decision-making process, the methods based on backpropagation or attention rely on the classifier's model architecture, and the same goes for most methods using *sax*, Fuzzy Logic, and prototypes. Many machine learning methods can be adopted from *tsc* and be applied to *tser* as well (77). However, adapting the model- specific explanation method for *tser* is not always possible, as the method was mostly designed for classification tasks only. On the other hand, model-agnostic methods can be more easily adopted to explain *tser* tasks, as they do not depend on the model itself. This argument makes model- agnostic methods advantageous over model-specific methods. Model-agnostic approaches, such as perturbation-based meth- ods, generate explanations intended for the model developer but not the user. The user here refers to someone who would act or make decisions based on the models' output, e.g., a doctor giving treatment recommendations. Perturbation-based methods, like DynaMask (46), give explanations in the form of a heatmap, which can typically be understood by a model developer but cannot be converted into *recommendations* or explanations for the user. In the case of a univariate time series, the heatmap highlights the important segments of the time series for the classifier. Still, it does not provide information on the actionability of these segments, i.e., it does not explain how changing timestamps affects the model's output. *cfe* typically explain the latter, demonstrating how the model's output changes if discriminative timestamps are modified. Recent research indicates (18) that counterfactuals are easy to understand, making them an advantageous model-agnostic approach that targets the user for their explanations.

C. Counterfactual Explanations

Wachter et al. (17) introduced counterfactual theory in 2018 and established key definitions and methodologies, such as the Wachter equation, which is given by

$$\arg\min_{x'}\max_{\lambda} \lambda\,(f_w(x') - c')^2 + d(x\,, x_i'), \qquad (1)$$

where f_w is a black-box classifier, x' a counterfactual, c' the desired class and $d(\cdot,\cdot)$ a distance function that measures how far the counterfactual x' and the original data point x_i are from one another. This equation is used to compute sparse counterfactuals. Wachter et al. (17) suggest using the Man- hattan distance weighted feature-wise with the inverse *mad* to generate sparse and outlier-robust solutions to the equation 1. In practice, maximization over λ is done by iteratively solving for x' and increasing λ until a sufficiently close solution is found. Wachter's method, however, imposes no constraint on the plausibility of the obtained counterfactual and does not necessarily find an optimal λ. This lack of constraint leads to counterfactuals that might be out of distribution, i.e., not representative of the underlying data distribution or simply unrealistic in real-world scenarios.

To resolve the plausibility issue, Delaney et al. (75) introduced Native Guide in 2021. Similar to some *cfe* for other data modalities, such as images, tabular or text data (78; 79; 76; 80; 81), Native Guide leans on existing instances in the training data to generate in-distribution counterfactual explanations. The method works in two steps. First, it retrieves the *NUN s* from the dataset. The *NUNs* are the closest instances in the dataset that are classified differently than the original data point. Then, using the weights of the last layer of the classification model, the algorithm perturbs one of the *NUNs* to move it closer to the decision boundary of the model. More recently, in 2022, *tsevo* (76) used various properties of time series transformations as introduced by Guilleme´ et al.
(43) and Mujkanovic et al. (48). Guilleme´ et al. (43) tailored *lime* (82) and *shap* (83) for time series data. Mujkanovic et al. (48) extended Guilleme´ et al.'s work by creating mappings that utilize the time and frequency domains and the statistical properties of time series. By employing these integrated time series transformations, Ho¨llig et al. (76) could generate different types of counterfactuals, outperforming other time series counterfactual approaches in both uni-and multivariate settings.

III. METHODS

Despite progress in counterfactual explanations for *tsc* or *tsf* (20; 30), Table ?? indicates a notable gap in addressing *tser* tasks. In fact, to the best of our knowledge, there currently does not exist any method for (deep learning) based explainable extrinsic regression. This work introduces novel methods for explainable *tser*. We begin by precisely defining *tser* (cf. Section III-A). Afterward, we showcase our reasoning behind the choice of counterfactuals to explain *tser*. Furthermore, we outline a framework for the transformation of explainable counterfactual-based methods from classification to extrinsic regression (cf. Section III-B) and introduce definitions for desired properties for counterfactuals in *tser* (cf. Section III-A) adopted from prior work (75). We apply this framework to adopt four *tsc* methods for *tser* (Section III-C), namely

1) *wachter* (Section III-C1)
2) *NUNr* (Section III-C2)
3) *dbar* (Section III-C3)
4) *tsevor* (Section III-C4)

Lastly, we evaluate the four methods on the task of biological age estimation to derive recommendations for individuals to improve their health (cf. Section III-D).

A. tser *and User Explainability*

tser is a regression task that learns the mapping from time series data to a scalar value (24). We formally define *tser* by Definition 3.1.

Definition 3.1: Let $x = [x_1, . . . , x_T] \in \mathrm{R}^{N\times T}$ be a uni- or multivariate time series, where T is the number of time steps, and N is the number of features. Let $x_{i,t}$ represent input feature i at time t, and y denote the output. Then, the regression model $f : x \rightarrow y$ returns an extrinsic continuous variable, with f considered a "black box" — i.e., no access to the inner workings of the model is available, and only the result y is observable.

In Section II, we discovered numerous techniques for ex- plaining time series data. Perturbation-based, backpropagation- based, and attention-based methods have one thing in com- mon: They show which part of the time series influences the model's output. While this is useful for the developer or a field expert, a user may not know how to interpret this explanation. When aiming for continuous health assessment, the explainability method should focus on *actionability*, i.e., it should be able to show the user how to change his behavior to improve the model's output, meaning that the technique should have a local scope and target the user. Another criterion to consider is the model-specificity or model-agnosticism of an explainable technique. Indeed, as most model-specific methods focus on models used for classification, and as we explore different black-box model architectures to perform biological age estimation, we considered only post-hoc, model-agnostic methods. Model-agnostic methods allow us to explain existing models without modifying them for transparency. However, this approach does not come without drawbacks: Post-hoc methods can result in explanations based on misconceptions learned by the model rather than actual knowledge from the data (84). From the different already implemented techniques available for *tsc* (see Table ??), the explanation method that would provide local, user-targetted explanations while being model agnostic is called cf. The following subsections define how to adapt existing counterfactual methods for *tsc* to *tser*.

B. *Counterfactual in* tser *via thresholding and its Desired Properties*

The common goal of counterfactual approaches is to provide an explanation via counter-examples given a time series x, called a query, and a model f. In classification scenarios, counter-examples allow users to understand why a classifier- model f predicts a label y for data point x instead of a counterfactual class y^{cf} (17). However, in the case of extrinsic regression, we do not have classes as we are predicting a continuous value and not a categorical value. Using only an inequation such as $y \neq y^{cf}$ would not work since the difference between the query and the counterfactual labels could be infinitely small, providing insufficient information. Yet, we can enforce a minimal change required for a data point to be counterfactual, and this is done via thresholding: we assume that for each x, a counterfactual sample x^{cf} can be computed that is close to x, but with a minimum prediction difference larger than a certain threshold $|y - y^{cf}| > \varepsilon^1$. A counterfactual should meet a few desired properties to be considered a relevant explanation for a user. When dealing with time-series data, we typically consider the following four properties (75):

1) Validity (Def. 3.2)
2) Proximity (Def. 3.3)
3) Sparsity (Def. 3.4)
4) Plausibility (Def. 3.5)

Let $x = [x_1, . . . , x_T] \in \mathrm{R}^{N\times T}$ be a uni- or multivariate time series or so-called query, where T is the number of time steps, N is the number of features, x^{cf} a counterfactual, and X the input space. Then, for a fixed ε, the set of valid counterfactuals denoted as S is defined by Definition 3.2.

Definition 3.2 (Validity of the counterfactual): We define the validity property for TSER as:

$S= \{x \in X: f(x) - \varepsilon \geq f(x^{cf}) \geq f(x) - 2\varepsilon\}$

This equation defines a set of accepted counterfactual labels for each query. It requires that the distance between the query label and the counterfactual label is larger than the threshold ε but does not exceed twice the ε value. The upper limit is because on the label axis, the threshold defines an area around the query where samples are not considered counterfactuals as they are too close to the query. It mimics the behavior of a class; if we move on the label axis from the query label by a distance ε, we are in the counterfactual area, i.e., in another class. If we move again by the same ε distance, we are no longer in this counterfactual area. We are too far from the query. The latter limit is set to ensure that the *cfe* does not differ too much from the original query (85). For example, applied to biological age estimation, if we let the patient query be 56 years old and $\varepsilon = 3$, then a valid counterfactual has a label between 50 and 53 years old. A 55- year-old counterfactual is considered too close to the query to be relevant, and the lower limit ensures that a 20-year-old counterfactual is not suggested, as he would be too distant from the query.

Definition 3.3 (Proximity of the counterfactual): We characterize the proximity property for TSER as follows:

$$\min_{x^{cf}} \; d(x, x^{cf}) \quad \text{s.t.} \quad f(x) - \varepsilon \geq f(x^{cf}) \geq f(x) - 2\varepsilon \tag{2}$$

This property ensures that the resulting x^{cf} is a proximate instance to the query (86). Proximity refers to the distance between the query instance x and the counterfactual instance x^{cf}, calculated as a distance measure d between x and x^{cf} . A commonly used metric for the distance between two time series is the *dtw* distance (87).

Definition 3.4 (Sparsity of the counterfactual): We define the sparsity property for TSER as follows:

$$\min_{x^{cf}} \sum_{i=1}^{N} \sum_{t=1}^{T} 1_{|x_{i,t} - x^{cf}_{i,t}| \neq 0} \quad \text{s.t.} \quad f(x) - \varepsilon \geq f(x^{cf}) \geq f(x) - 2\varepsilon$$

Sparsity refers to the number of changes in data points between x and x^{cf} (86). This key property forces a *cfe* method to make human-interpretable changes. When enforcing the sparsity property, *cfe* methods strive to alter the fewest variables necessary to achieve user-interpretable solutions. Another constraint specific to time series data is that not only the fewest number of variables should change, but the changed variables should be in continuous subsequences of the original time series. Each counterfactual method implements a different technique to overcome this constraint.

Definition 3.5 (Plausibility of the counterfactual): We define the plausibility property of counterfactuals in the context of TSER as follows:

$$x^{cf} \sim D \quad \text{s.t.} \quad f(x) - \varepsilon \geq f(x^{cf}) \geq f(x) - 2\varepsilon \tag{4}$$

A counterfactual x^{cf} is plausible if it could have been drawn from the data D (84). The plausibility property ensures that the post-hoc explanation method produces justified explanations. This property is verified by looking at the neighbors' expla- nation labels and analyzing whether they are close to the explanation's label. A counterfactual with a label far away from his neighbor's label is considered unjustified. In biological age estimation, a justified counterfactual has neighbors in the same age range, i.e., the distance between the counterfactual's label and the neighbor's label is below the threshold ε.

C. Adoption of methods for time-series extrinsic regression

Given our prior definitions of desired properties for counterfactuals in *tser*, we describe our adoption of four methods of TSC to TSER. Based on its results, we chose *tsevo* first, as it seemed to be the more promising approach. Then we adapted Wachter, *NUN*, and *dba* to compare and put in perspective *tsevo*'s results. In the following sections, we present their adoption chronologically.

1) wachter-cf*:* The first candidate for the adoption of *cfe* for *tser* is Wachter et al. (17) (2018). Wachter's approach involves minimizing an equation through gradient descent that combines validity (cf. Definition 3.2) and proximity (cf. Definition 3.3) properties. To adapt Wachter's method for *tser*, we replace the classifier model and introduce the threshold criterion (cf. Section III-B) while the distance function remains unchanged. Adapting Eq. 1 (cf. Section II-C) leads to the following :

$$\arg\min_{x'} \max_{\lambda} \lambda (f_w(x_i) - \varepsilon - f_w(x'))^2 + d(x_i, x') \tag{5}$$

The main adaptations reside in that f_w now denotes a black- box extrinsic regression model, and a threshold ε is introduced.

2) NUNr-cf*:* The second candidate for *cfe* for *tser* is the *NUN* as counterfactuals (80) (2009), and later adapted to time series data by Delaney et al. (75) (2021). In a classification setup, the *NUN* method aims to find the closest instance in the dataset that is classified differently than the query. It works by creating a reference set containing all instances in the dataset with the target classification or a different classification than the query. Then, the *NUN-cf* algorithm computes the *nns* of the query that are in the reference set, obtaining the *NUNs*. The *nns* are computed using *KNeighborsTimeSeries* (88). We must only modify how the reference set is defined to adapt the *NUN-cf* algorithm to the regression setup.

Definition 3.6 (Reference set): The reference set is a

[1] Important note: In the specific case of biological age estimation, as we want to find a healthier patient, we are looking to decrease the *ba* of the patient. Therefore, we are only interested in counterfactuals whose predicted values are lower than the patients' *ba* by a certain margin (e.g., at least three years younger).

subset of all known data D with a valid prediction under def. 3.2.

$$R = \{z \in D : f(x) - \varepsilon \geq f(z) \geq f(x) - 2\varepsilon\}$$

With *NUNr-cf*, if a *NUN* exists, it is guaranteed that the found counterfactual is in the data distribution, as it is an existing sample (cf. Def. 3.5) and valid (cf. Def. 3.2). The proxim- ity (cf. Def. 3.3) of the counterfactuals will be minimized among the existing valid samples, but as shown in previous work (75; 89), the *NUN* is not necessarily close to the model decision boundary, and it is possible to find more proximate counterfactuals by mutating the *NUN* towards the decision boundary. These mutations are typically done by perturbing the query time series on areas where the query and the *NUN* disagree. Another issue is that *NUNs* are not sparse; in the context of biological age estimation, the *NUNr-cf* algorithm locates a physical activity that closely resembles the query's physical activity but is performed by a different individual and slightly varies at each timestamp.

3) dbar-cf*:* The third method we consider is called *dba- cf* and was proposed by Delaney et al. (75) in 2021. The main idea is to bring the found *NUN* closer to the decision boundary by averaging between the query and the *NUN*. Originally, Forestier et al. (90) (2017) proposed *dba* to aug- ment time-series datasets. *dba* is used to compute the average between time series. The concept behind *dba-cf* is to achieve a weighted average between the query and the *NUN*, starting with all the weight on the query and then iteratively moving the weight towards the *NUN* until the decision boundary is reached.

4) tsevor-cf*:* *tsevo* is a technique that combines time series perturbation approaches from the recent work(43) and (48) with a genetic algorithm for multi-objective optimization (?
). This technique allows the creation of model-agnostic coun- terfactual explanations for uni- and multivariate classification problems.

tsevo tackles the challenge of finding a counterfactual that meets the four key properties validity 3.2, proximity 3.3, sparsity 3.4 and plausibility 3.5. To achieve this, *tsevo* treats each property as an objective to optimize, forming a multi-objective optimization problem. We define below how the properties are transformed into objectives in the setting of *tser*.

Definition 3.7 (Multi-Objective Problem): O_1 is derived from def. 3.3 by applying *mae* as distance function d (86), (17).

$$O_1(x, x^{cf}) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} |x_{i,t} - x^{cf}_{i,t}|$$

O_2 is consistent with def. 3.4.

$$O_2(x, x^{cf}) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} 1_{|x_{i,t} - x^{cf}_{i,t}| = 0}$$

O_3 denotes the normalized output distance.

$$O_3(x, x^{cf}) = (f(x) - f(x^{cf}) - \varepsilon)/\varepsilon$$

Combining the desired properties leads to the following multi-objective problem O:

$$\begin{aligned} \min_{x^{cf}} \quad & O(x, x^{cf}) := (O_1(x, x^{cf}), O_2(x, x^{cf}), O_3(x^{cf}))^T \\ \text{s.t.} \quad & f(x) - \varepsilon \geq f(x^{cf}) \geq f(x) - 2\varepsilon \end{aligned} \tag{6}$$

The multi-objective optimization follows the steps described in the original *tsevo* publication (76). In summary, a population of n individuals is initialized, where each individual repre- sents a potential counterfactual. The individuals are evaluated with respect to their objectives score. For g generations, the evolution algorithm selects the best individuals in the popu- lation according to how they fulfill the different objectives. Depending on a certain probability, it performs crossover and/or mutates them. For the mutations, we used the authentic opposing information mutation, first introduced by Guilleme et al. (43), which is based on the assumption that interpretable values of time series can exhibit shapes (e.g., peaks) that are easily understandable to humans. To use those shapes included in a reference set R (cf. Def. 3.6), we draw a random sample $r \in R$. Both r and the selected individual λ_i are segmented with window size w_i, resulting in $S(r)$ and $S(\lambda_i)$. The mutation then draws a random segment index $s \in [0, |S(r)| - 1]$ and replaces the drawn slice $S(\lambda_i)[s]$ with the slice $S(r)[s]$ from the replacement time series. The concept of crossover in genetic algorithms is utilizing the search space by merging the genetic material of high-performing individuals
(91). The reference set is used in the evolution algorithm to mutate the individuals. It ensures that the mutated individuals stay in the data manifold. Note that this is meant to achieve the plausibility property (cf. Def. 3.5) by design, as this property was not expressed as an objective.

D. Experimental Evaluation

Our work was motivated by the findings of Pyrkov et al. (10) and Rahman et al. (11), who showed that deep learning models could estimate the biological age of a patient from his physical activity. Once we predict the biological age, we can use counterfactual explanations to give feedback to the patient. The following Sections describe our efforts to reproduce prior work to train models for biological age estimation, enabling us to test our CFE methods. For the training and data generation, we followed the same steps as described in from data generation to providing recommendations.

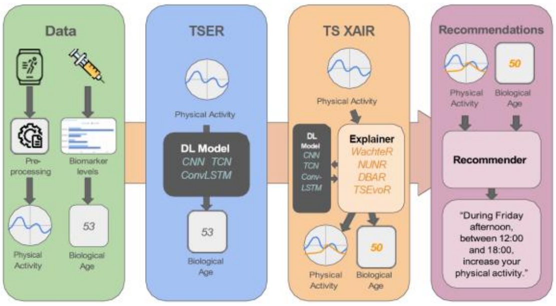


Fig. 1: Pipeline: From data preprocessing to physical activity recommendations.

prior work in Pyrkov et al. (10), Rahman et al. (11), and Shim et al. (13). Figure 1 shows an overview of the whole pipeline, from data generation to providing recommendations.

1) Physical Activity: Description We used physical activity data recorded during the *nhanes* from 2003 to 2004 (92) and from 2005 to 2006 (93). *nhanes* uses a complex sampling design to survey non-institutionalized members of the US population. A subset of *nhanes* participants recorded their activity data. During the survey, participants' physical activity is tracked for seven continuous days using a physical activity monitor[2] to record *activity counts* sampled every minute. The participants wear the monitors on the right hip using an elastic belt. The datasets included 14'631 participants, with 7'176 in 2003-2004 and 7'455 in 2005-2006. Filtering The filtering steps are based on prior work done on biological age estima- tion (10; 11; 13). First, we removed outliers with abnormally low (< 50) or high (> 5000) mean activity count. Then, we removed patients with less than 10'080 (= 7×24×60) activity counts, corresponding to one measurement every minute for seven days. Also, we only considered days where the par- ticipant was active for more than 200 minutes. Therefore, we filtered out participants who had less than four days of meeting this criteria. This filter resulted in a total of 9,591 individuals. Transformation To handle the noisy and outlier-filled nature of the time-series human locomotor data, we first needed to apply some basic data transformation operations, such as smoothing. The physical activity intensity ($pa_{intensity}$) values range over a large magnitude and are always positive, so we applied log transformations to the data, as suggested by Rahman et al. (11). However, since the original data contains some 0 values, we added a negligible value (1) before applying log transformations. The first transformation is a Box-Cox (94) transformation with $\lambda = 1$, which is equivalent to a simple log transformation (other values of λ were investigated by Rahman et al. (11)). Then, a second log transformation is computed. Since the data is a sequential time series of seven days, we applied moving averages on the data with different window sizes and an *ema*.

$$pa_{log} = \log(BoxCox(pa_{intensity} + 1) + 1) \qquad (7)$$

$$pa^{i}_{ema} = \begin{cases} pa^{i}_{log}, & \text{if } i = 1 \\ \alpha pa_{log_i} + (1 - \alpha) pa_{log_{i-1}}, & \text{otherwise} \end{cases} \qquad (8)$$

2) Biological Age: We used the *nhanes* 2003-2006 anthro- pometric and bio-marker datasets to compute each patient's biological age (95). The biomarker dataset contained informa- tion on albumin, alkaline phosphatase, blood urea nitrogen, uric acid, cholesterol, creatinine (96), C-reactive protein (97), body mass index (98), glycohemoglobin (99), systolic and diastolic blood pressure (100), lymphocyte percentage, mean cell volume and white blood cell count (101). Then, we used the Klemera-Doubal method (102) to compute a mapping between the biomarkers information and the biological age. We retained only patients aged between 18 and 85. As a result, we obtained a dataset of 10,184 patients along with their cor- responding biological age. Lastly, we combined the physical activity dataset with the biological age datasets, resulting in 7'222 matched patients. Not all patients in the physical activity dataset were included in the biomarker dataset, hence the lower total number of patients. We split the combined dataset into training (65%), validation (25%), and testing (10%) sets, yielding 4'694, 1'805 and 723 patients, respectively.

3) Deep Learning models for Biological Age Estimation: Using physical activity data to predict biological age is an example of a *tser* task. In our scenario, we trained and tested three different models to predict biological age from physical activity data:

1) A cnn suggested by Pyrkov et al. for

[2] The monitor used was the ActiGraph AM-7164; ActiGraph, Pensacola, FL, USA

biological age estimation (10)

2) A convlstm proposed by Rahman et al. for biological age estimation (11).

3) A tcn (103; 104) network.

To the best of our knowledge, the *tcn* model has not yet been tested for biological age estimation in prior work. It was included in the evaluation in this work because recent literature indicates that *tcn* models can achieve state-of-the-art results on time series tasks while often being less complex than *lstm* models (104). Each model requires a different data representation: The *cnn* takes as input a flat vector of 10’080 values, *convlstm* uses a complex 3D representation (60*,* 24*,* 7), and *tcn* falls in between the two, taking vectors of shape (7*,* 1440), where the days are treated as features. We trained all models on the training data for *convlstm* and *cnn*; we used the steps and hyperparameters described by Rahman et al. (11) and Pyrkov et al. (10) respectively. The *tcn* has ten layers of 128 channels each; we used a kernel size of 4 and a dropout of 0.2. We trained the model for 100 epochs, with a learning rate 0.0001, and we chose the best model according to the *mse* loss. We recorded two other metrics, which are the *mae*, which gives a more intuitive comprehension of the error of the model than the *mse*, and the Pearson correlation, which indicates the strength of the linear association between the physical activity and the biological age.

4) Explainable Biological Age Estimation: Utilizing our counterfactual adaptations *wachter*, *NUNr*, *dbar* and *tsevor*, we generated counterfactuals with threshold ε = 3 (cf. Section III-B) for each sample of the 723 samples in the test set using the three different models. For *wachter*, we defined a loss function from eq. 5:

$$LOSS = \lambda\,(f_{t'}(x) - \varepsilon - f_{t'}(x'))\ + (1 - \lambda)\,d(x, x') \quad (9)$$

and x' was initialised to x_i (we also attempted random initial- ization). Using the L1-Norm as distance function *d*, we then performed a gradient descent using the ADAM (105) optimizer for n = 100 iterations for each lambda λ and increased lambda by 0.05 if no valid (Def. 3.2) counterfactual is found. The valid counterfactual with the smallest loss (Eq. 9) is returned. If no valid counterfactual is found after trying out all lambdas, we return the counterfactual with the smallest corresponding loss. For *NUNr*, once the reference set is adapted (see 3.6), the remainder of the algorithm is the same as in a classification setting. For *dbar*, we retrieved the *NUN* and performed a weighted sum:

$$x^{cf} = DBA(\beta x_{query}, (1 - \beta)x_{NUN}) \qquad (10)$$

We chose to run 10 iterations, increasing β by 0.01 at each iteration and returning x^{cf} as soon as it is valid. For *tsevor*, we used the genetic algorithm over 50 generations and ap- plied mutation to the individuals using the authentic opposing information transformer, as implemented by Guilleme´ et al. (43).

5.) Habits recommendations: The generated counterfactuals depict time series plots of recommended physical activity data. Presenting this data to potential patients allows them to interpret what actions they should take to improve their biological age. However, different patients might interpret the counterfactuals differently. We designed a system that provides the patient with highly interpretable text feedback to give a more concrete, text-based recommendation. This feedback contains a few sentences or *recommendations* on how to adjust the activity based on the generated counterfactuals. Each day is separated into four parts of six hours each to generate a recommendation *R*:

Night, Morning, Afternoon, and *Evening*.

For each part *p*, we compute the mean value of the query’s physical activity $mean_q$ and the mean of the counterfactual’s physical activity $mean_{cf}$. We assume that the obtained means represent a percentage activity on a scale from *No Intensity* (0%) to *Very High Intensity* (100%). This representation allows us to interpret the difference between $mean_q$ and $mean_{cf}$ as a percentage change. Concretely, this percent- age change ($Percentage_{change}$) is defined as follows, where MAX ACTIVITY INTENSITY denotes the maximum value of the intensity of the counterfactual and the query.

$$Percentage_{change} = \frac{mean_{cf} - mean_q}{\text{MAX ACTIVITY INTENSITY}} \times 100 \qquad (11)$$

Using our approach, we generate several maximum *n* rec- ommendations to the user and only include recommendations suggesting a minimal percentage change of at least *p*%, with *n* and *p* being configurable parameters.

IV. RESULTS

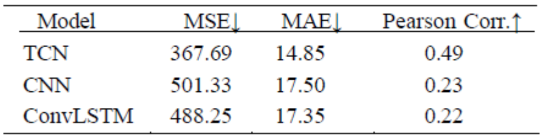
In the results section, we first report the performance of the *DL* models to estimate Biological Age data (cf. Section ??). Then, we evaluate the generated *cfe* using qualitative and quantitative criteria (cf. Section

IV-B).

A. Biological Age Estimation with Deep Learning

We set out to reproduce results presented in prior work (10; 11) that use *DL* models to estimate biological age from physical activity data and achieved the following re- sults (cf. Table IV). In Rahman et al. work (11), the authors reported slightly different results. For the *convlstm*, they re- ported a *mae* of 13.21 years, an *mse* of 282*,* 58 with a Pearson correlation of 0.62. For the *cnn*, they reported an *mae* of 15.49, an *mse* of 353.82, and a Pearson correlation of 0.45.

TABLE IV: Performance Comparison of Models



Model	MSE↓	MAE↓	Pearson Corr.↑
TCN	367.69	14.85	0.49
CNN	501.33	17.50	0.23
ConvLSTM	488.25	17.35	0.22

B. Counterfactuals for Biological Age Estimation

In this section, we evaluate the generated counterfactuals in two ways. First, qualitatively, we plotted one patient evaluated with the four different counterfactual techniques. The plots al- low us to evaluate the user interpretability of the explanations. Looking at the explanations, a user should understand what he does well, what he could improve, and what impact it will have on his health. Then, we evaluate quantitatively using metrics based on the properties defined in Section III.

1) Qualitative evaluation: Fig. 2 shows counterfactuals obtained for a specific patient using the four counterfactual methods. We highlight findings for each method in the fol- lowing subsections.

2) wachter*:* Figure 2a shows the counterfactual result for patient 6306. It was obtained using the *tcn* model and the *wachter* technique. The model's predicted biological labels are at the plot's top. According to the *tcn* model, the patient's biological age is 46.46. The counterfactual physical activity level corresponds to that of someone who is 34.1 years old, which is not a valid counterfactual, as valid counterfactuals need to have a biological age between 40.46 and 43.46 years

3) NUNr*:* Figure 2b displays the counterfactual outcome for patient 6306 using the *tcn* model explained with the *NUNr* technique. The *NUNr* biological age is 42.39, which is a valid result. Based on the explanation, four recommendations were made to improve the patient's health; only the top three are shown in the plot. These include reducing activity levels on Wednesday and Sunday mornings and Saturday afternoons and increasing activity on Friday afternoons.

4) dbar*:* Figure 2c shows the counterfactual for patient 6306 obtained through the *tcn* model with the *dbar* technique. The *dbar* biological age is 43.46, which is an optimal and valid result. It is optimal because 43.46 is the closest accepted label possible. By comparing with the *NUNr* plot, we can observe that the *dbar* technique averages between the *NUNr*'s physical activity and the patient's physical activity to produce a similar counterfactual, with less important changes (percentage changes are lower) but still slight changes at each time stamp. Based on the *dbar*'s output, the recommender suggests only one recommendation to improve the patient's biological age: increasing physical activity on Friday afternoon. It is important to note that this was already a recommendation from the *NUNr* technique.

5) tsevor*:* Figure 2d shows a counterfactual for patient 6306, which was obtained using the *tcn* model and explained using the *tsevor* technique. The *tsevor* biological age is 43.46, which is an optimal and valid result. *tsevor* only had to make a few changes to the timestamps to arrive at a valid and optimal result. Based on this analysis, it then recommended that the patient increase their physical activity on Friday morning and afternoon, which is the same as *dbar* and *NUNr*.

C. Quantitative evaluation

We generated counterfactuals on the 723 samples from the test set. To generate counterfactuals, we first feed each sample from the test set to each of the three deep-learning models (*cnn*, *tcn*, and *convlstm*). We then generated counterfactuals using the four different counterfactual techniques (*wachter*, *NUNr*, *dbar*, and *tsevor*). Finally, we evaluate the 723×3×4 = 8676 generated counterfactuals with the following metrics:

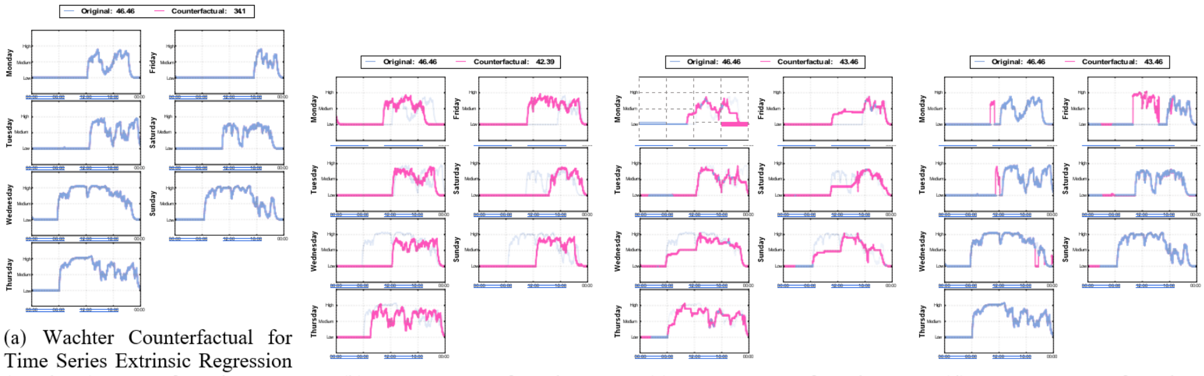
- Validity: Proportion of the 723 generated counterfactuals that are valid (cf. Def. 3.2).
- Proximity: L_1 norm (Manhattan distance) between the query x and the counterfactual x^{cf}.

$$||x - x^{cf}||_1 = \sum_i |x_i - x^{cf}_i| \qquad (12)$$

- Sparsity: Proportion of changed data points to obtain the counterfactual.
- Plausibility: Proportion of Nearest Neighbours with a label close to the counterfactual label. We used $k = 5$ nn, and $\varepsilon = 3$ as the threshold value to differentiate between close and far neighbors.

$$\text{Plausibility} = \frac{1}{k} \sum_{i \in kNN(x^{cf})} 1_{|y_i - y^{cf}| \leq \varepsilon} \qquad (13)$$

- Time: Time needed to explain a sample.



(a) Wachter Counterfactual for Time Series Extrinsic Regression (WachteR) Counterfactual

(b) NUNR Counterfactual

(c) DBAR Counterfactual

(d) TSEvoR Counterfactual

Method was unable to generate any counterfactuals

During Sunday morning, between 06:00 and 12:00, decrease the physical activity by 100.00%

During Wednesday morning, between 06:00 and 12:00, decrease the physical activity by 100.00%

During Friday afternoon, between 12:00 and 18:00, increase the physical activity by 89.48%

...

During Friday afternoon, between 12:00 and 18:00, increase the physical activity by 81.39%

During Friday morning, between 06:00 and 12:00, increase the physical activity by 100.00%

During Friday afternoon, between 12:00 and 18:00, increase the physical activity by 84.09%

Fig. 2: Example of time series counterfactuals (pink line) of an input time series (blue line) for the biological estimation problem, where, given a threshold $\varepsilon = 3$, the counterfactual modifies the original activity so the model predicts a smaller biological age. Each plot represents a different counterfactual technique, and each corresponding text recommendation is listed below the plot. Common recommendations are in bold.

TABLE V: Mean and variance of the scores obtained by each counterfactual generation technique on five evaluation metrics, using three different deep learning models, tested on 723 samples.

Model	Method	Validity ↑	Proximity ↓	Sparsity ↓	Plausibility ↑	Time ↓
CNN	TSEvoR	0.37±0.48	159.99±274.71	0.04±0.09	0.39±0.25	8m01s±2m52s
	NUNR	0.99±0.08	4740.66±1149.04	0.98±0.02	0.47±0.20	0m01s±0m01s
	DBAR	0.10±0.31	2499.74±702.91	0.99±0.03	0.02±0.11	9m28s±4m58s
	Wachter	0.00±0.00	0.00±0.00	1.00±0.00	0.00±0.00	1m01s±0m28s
ConvLSTM	TSEvoR	0.08±0.27	450.75±445.8	0.07±0.10	0.46±0.29	47m36s±11m21s
	NUNR	0.99±0.08	3944.03±796.31	0.98±0.02	0.59±0.25	0m01s±0m01s
	DBAR	0.09±0.29	1898.02±328.64	1.00±0.02	0.04±0.15	11m33s±4m32s
	Wachter	0.02±0.12	0.00±0.00	0.99±0.02	0.14±0.27	6m32s±1m48s
TCN	TSEvoR	0.39±0.49	147.37±242.41	0.04±0.09	0.41±0.26	10m01s±1m14s
	NUNR	0.99±0.08	4740.66±1149.04	0.98±0.02	0.51±0.23	0m01s±0m01s
	DBAR	0.10±0.30	2344.37±639.02	1.00±0.02	0.01±0.07	12m03s±5m54s
	Wachter	0.00±0.00	0.00±0.00	1.00±0.00	0.00±0.00	1m09s±0m03s

Upon analysis, we can first observe that the best-performing explainable technique for each metric remains constant across the models, which only impacts the time aspect. Secondly, we note that the *NUNr* technique performs best across three out of five metrics but falls short on the proximity and sparsity metrics. Compared to *NUNr*, *dbar* improves the proximity metric at the cost of a large drop in validity and plausibility. Thirdly, it is evident that all techniques, except for *tsevor*, cannot produce sparse counterfactuals. It is also important to note that *tsevor* ranks first or second in every metric except for time, where it secures third place. Moreover, when using *tsevor*, the model plays a role in the validity and time metrics. Indeed, the validity drops to 0.08 when explaining the *convlstm* model, compared to 0.37 for *cnn* and 0.39 for *tcn*, while being five times slower.

V. DISCUSSION

Our work introduces novel model-agnostic, user-targeted explanation methods for *tsxair* applications. To achieve this, we adapted four counterfactual techniques from the do- main of time series classification. We outline principal find- ings (cf. Sec. V-A), practical implications (cf. Sec. V-C), and limitations (cf. Sec. V-D) below and also provide a comparison with prior work (cf. Sec. V-B).

A. Principal findings

With our work, we applied counterfactual theory to time se- ries extrinsic regression. We successfully adapted four existing counterfactual methods for *tsc* tasks for *tser*. We evaluated these four adapted methods, the digital health use case of biological age estimation. Our experiment on biological esti- mation demonstrates that we can generate meaningful counter- factual explanations for the univariate *tser* task. Specifically, in biological age estimation, prior works show how to collect the data, preprocess it, and use it to predict the biological age, but they could not provide recommendations. Our work provides the final piece of the puzzle for continuous health assessments using wearable devices. Using counterfactual applications, we can provide objective recommendations to participants on how to improve their health, which could be delivered, for example, via a smartphone app.

B. Comparison with prior work

Previous research tackled similar tasks in the field. For example, Perturbation techniques such as DynaMask (46) can highlight essential subsets of time series data, such as the most discriminative areas used in the model decision process. While it is useful for explaining *tser*, it does not focus on the user-interpretability. Counterfactual techniques have been developed for *tsc* (76; 75) and *tsf* (106) tasks. However, these techniques can not be used for *tser* tasks. To the best of our knowledge, no explainable technique specifically targeting users for *tser* models exists.

C. Practical implications

Our experiment of *tsxair* on biological age estimation is not limited to that particular area, as our approach can be applied to any *DL* technique that aims to learn a score from time series data. For instance, we could use Diaz-Lozano et al.'s work (107). They show that it is possible to use the evolution in the number of COVID-19 contagions to predict the mortality rate of people affected by this particularly contagious disease. With the help of our technique, we could understand how the number of contagions should vary to reduce the mortality rate. This could be used to take the rightful political measures to reduce the contamination number. For further usage of *tsxair*, it is important to note that the ability of *NUNr*, *dbar*, and *tsevor* to identify valid counterfactuals (as defined in Definition 3.2) depends on the size of the reference set (as defined in Section 3.6). As a matter of fact, the number of mutations that *tsevor* can attempt is determined by the size of the reference set. If the reference set is empty, it cannot identify any counterfactuals. In our experiment, we filtered out people younger than 18 years old, meaning it was impossible to find counterfactuals for people between 18 and 21, as their reference set was empty.

D. Limitations

We consider the following three points to be limitations of our work. First, the long computation times for *tsevor*, 8 minutes when using the *cnn* or 47 minutes with the *convl- stm*, suggest that we could benefit from using parallelization strategies to speed up the process. Second, our reliance on deep learning models like *tcn*, *cnn*, and *convlstm* can lead to variable results, which raises questions about the fairness of comparisons and the need for more robust models (108) with lower *mae*. The plausibility (cf. Sec. 3.5) of the coun- terfactuals may vary depending on the accuracy of the *DL* model. Figure 2a shows that the *wachter* technique utilizes *DL* models' noise to produce a counterfactual. *NUNr* has only around 50% of plausibility which is concerning. Indeed, the *NUNr* counterfactual, an instance of the dataset, has a label close to only half its neighbors. This implies a noisy model that predicts differently close instances. These two examples show the limitations of the *DL* models, but they also show that we can use the *cfe* to detect robustness failures in *tser*. Third, the threshold choice of $\varepsilon = 3$ is somewhat arbitrary and can have unexplored implications for the outcomes (85). In addition, our algorithms have primarily been tested on univariate time-series data, which could limit their applicability to more complex datasets. Finally, the effectiveness of *tsevor* heavily depends on the availability of a valid reference set, which means it may not be effective in scenarios where such data is lacking. Therefore, it is important to explore alternative strategies in those cases. By addressing these limitations, we can develop a more comprehensive and reliable framework for generating counterfactuals.

VI. OUTLOOK

As mentioned in the previous section, our approach can be applied to any *DL* technique to predict a continuous variable from time series data. In future work, we would like to improve our methods and recommendation pipeline further pipeline (cf. Fig. 1). We will focus on three main areas for improvement. Data Our current experiment uses physical activity data from 2003 to 2006. To improve our dataset, we plan to include data available until 2014. Additionally, we will enrich our analysis by incorporating diverse digital biomarkers such as heart rate and tension, shifting from univariate to multivariate time series. These enhancements are expected to improve the predictive power of our model, thereby enabling a deeper understanding of the four methods in various scenarios. *tsxair* Building on recent research by Letzgus et al. (109) and Shim et al. (13), we aim to enhance our *xai* framework. We will incorporate contextualized *xai* techniques and population clus- ters based on physical activity patterns, enhancing the explana- tion's interpretability. Furthermore, we will explore alternative mutation techniques suggested in the *tsevo* paper and leverage *llm* or text-based models for automated recommendations. These improvements could make our approach more robust and effective. Recommendations We plan to improve our rec- ommendation system by analyzing the extensive *tser* datasets compiled by Guijo et al. (28), which will provide valuable insights and enable us to integrate expert recommendations. Additionally, we will test our method with diverse datasets, especially those with well-established outcomes, to ensure that our counterfactual explanations align with common medical understanding. This validation step is crucial for enhancing the credibility and applicability of our approach in real-world scenarios. By considering these improvements, we aim to refine our methodology and advance the field of mobile health. We aim to foster trust, improve accuracy, and enhance the interpretability of our model's recommendations.

ACRONYMS

CFE Counterfactual Explanations. 7
TSC Time-Series Classification. 6
TSER Time Series Extrinsic Regression. 6
Wachte W Rachter Counterfactual for Time Series Extrinsic Regression. 10

REFERENCES

[1]. Fries James F., Koop C. Everett, Beadle Carson E., Cooper Paul P., England Mary Jane, Greaves Roger F., Sokolov Jacque J., Wright Daniel, and null null, "Reducing Health Care Costs by Reducing the Need and Demand for Medical Services," *New England Journal of Medicine*, vol. 329, no. 5, pp. 321–325, 1993. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJM199307293290506.

[2]. T. Bodenheimer, "High and Rising Health Care Costs. Part 1: Seeking an Explanation," Annals of Internal Medicine, vol. 142, pp. 847–854, May 2005. Publisher: American College of Physicians.

[3]. CDC, "Chronic Diseases," July 2022.

[4]. G. Huzooree, K. Kumar Khedo, and N. Joonas, "Perva- sive mobile healthcare systems for chronic disease mon- itoring," Health Informatics Journal, vol. 25, pp. 267–291, June 2019. Publisher: SAGE Publications Ltd.

[5]. U. Varshney, "Mobile health: Four emerging themes of research," Decision Support Systems, vol. 66, pp. 20–35, Oct. 2014.

[6]. M. Javaid, A. Haleem, S. Rab, R. Pratap Singh, and R. Suman, "Sensors for daily life: A review," Sensors International, vol. 2, p.

100121, Jan. 2021.

[7]. A. Coravos, S. Khozin, and K. D. Mandl, “Developing and adopting safe and effective digital biomarkers to improve patient outcomes,” npj Digital Medicine, vol. 2, pp. 1–5, Mar. 2019. Publisher: Nature Publishing Group.

[8]. F. Barata, J. Shim, F. Wu, P. Langer, and E. Fleisch, “The Bitemporal Lens Model—toward a holistic ap- proach to chronic disease prevention with digital biomarkers,” JAMIA Open, vol. 7, p. ooae027, Apr. 2024.

[9]. C. Vlachopoulos, P. Xaplanteris, V. Aboyans, M. Brod- mann, R. C´ıfkova´, F. Cosentino, M. De Carlo, A. Gallino, U. Landmesser, S. Laurent, J. Lekakis, D. P. Mikhailidis, K. K. Naka, A. D. Protogerou, D. Rizzoni, A. Schmidt-Trucksa¨ss, L. Van Bortel, T. Weber, A. Ya- mashina, R. Zimlichman, P. Boutouyrie, J. Cockcroft, M. O’Rourke, J. B. Park, G. Schillaci, H. Sillesen, and R. R. Townsend, “The role of vascular biomarkers for primary and secondary prevention. A position paper from the European Society of Cardiology Working Group on peripheral circulation: Endorsed by the As- sociation for Research into Arterial Structure and Physiology (ARTERY) Society,” Atherosclerosis, vol. 241, pp. 507–532, Aug. 2015.

[10]. T. V. Pyrkov, K. Slipensky, M. Barg, A. Kondrashin, B. Zhurov, A. Zenin, M. Pyatnitskiy, L. Menshikov, S. Markov, and P. O. Fedichev, “Extracting biological age from biomedical data via deep learning: too much of a good thing?,” Scientific Reports, vol. 8, p. 5210, Mar. 2018.

[11]. S. A. Rahman and D. A. Adjeroh, “Deep Learning using Convolutional LSTM estimates Biological Age from Physical Activity,” Scientific Reports, vol. 9, p. 11425, Aug. 2019. Publisher: Nature Publishing Group.

[12]. C. Tudor-Locke, C. L. Craig, W. J. Brown, S. A. Clemes, K. De Cocker, B. Giles-Corti, Y. Hatano, S. Inoue, S. M. Matsudo, N. Mutrie, J.-M. Oppert, D. A. Rowe, M. D. Schmidt, G. M. Schofield, J. C. Spence, P. J. Teixeira, M. A. Tully, and S. N. Blair, “How many steps/day are enough? for adults,” International Journal *of Behavioral Nutrition and Physical Activity*, vol. 8, p. 79, July 2011.

[13]. J. Shim, E. Fleisch, and F. Barata, “Wearable-based accelerometer activity profile as digital biomarker of inflammation, biological age, and mortality using hi- erarchical clustering analysis in NHANES 2011–2014,” Scientific Reports, vol. 13, p. 9326, June 2023. Publisher: Nature Publishing Group.

[14]. S. Saxena, M. Van Ommeren, K. C. Tang, and T. P. Armstrong, “Mental health benefits of phys- ical activity,” Journal of Mental Health, vol. 14, pp. 445–451, Jan. 2005. Publisher: Routledge eprint: https://doi.org/10.1080/09638230500270776.

[15]. H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, “Application of explainable artifi- cial intelligence for healthcare: A systematic review of the last decade (2011–2022),” Computer Methods and Programs in Biomedicine, vol. 226, p. 107161, Nov. 2022.

[16]. R. Byrne, Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. Aug. 2019. Pages: 6282.

[17]. S. Wachter, B. Mittelstadt, and C. Russell, “Counter- factual Explanations without Opening the Black Box: Automated Decisions and the GDPR,” Mar. 2018. arXiv:1711.00399 [cs].

[18]. T. Miller, “Explanation in artificial intelligence: In- sights from the social sciences,” Artificial Intelligence, vol. 267, pp. 1–38, Feb. 2019.

[19]. S.-I. Lee and E. J. Topol, “The clinical potential of counterfactual AI models,” The Lancet, vol. 403, p. 717, Feb. 2024. Publisher: Elsevier.

[20]. A. Theissler, F. Spinnato, U. Schlegel, and R. Guidotti, “Explainable AI for Time Series Classification: A Re- view, Taxonomy and Research Directions,” IEEE Ac- cess, vol. 10, pp. 100700–100724, 2022. Conference Name: IEEE Access.

[21]. A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laer- hoven, “Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks,” Sensors (Basel, Switzerland), vol. 19, p. 3079, July 2019.

[22]. Y. Hagiwara, H. Fujita, S. L. Oh, J. H. Tan, R. S. Tan, E. J. Ciaccio, and U. R. Acharya, “Computer-aided diagnosis of atrial fibrillation based on ECG Signals: A review,” Information Sciences, vol. 467, pp. 99–114, Oct. 2018.

[23]. O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial time series forecasting with deep learning : A systematic literature review: 2005–2019," Applied Soft Computing, vol. 90, p. 106181, May 2020.

[24]. C. W. Tan, C. Bergmeir, F. Petitjean, and G. I. Webb, "Time Series Extrinsic Regression," Feb. 2021. arXiv:2006.12672 [cs, stat].

[25]. H. Drucker, C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," Adv Neural Inform Process Syst, vol. 28, pp. 779–784, Jan. 1997.

[26]. Z. Wang, W. Yan, and T. Oates, "Time Series Classifica- tion from Scratch with Deep Neural Networks: A Strong Baseline," Dec. 2016. arXiv:1611.06455 [cs, stat].

[27]. C. W. Tan, C. Bergmeir, F. Petitjean, and G. I. Webb, "Monash University, UEA, UCR Time Series Extrinsic Regression Archive," Oct. 2020. arXiv:2006.10996 [cs, stat].

[28]. D. Guijo-Rubio, M. Middlehurst, G. Arcencio, D. F. Silva, and A. Bagnall, "Unsupervised Feature Based Algorithms for Time Series Extrinsic Regression," May 2023. arXiv:2305.01429 [cs, stat].

[29]. J. Rodr´ıguez, L. Kuncheva, and C. Alonso, "Rotation Forest: A New Classifier Ensemble Method," IEEE transactions on pattern analysis and machine intelli- gence, vol. 28, pp. 1619–30, Nov. 2006.

[30]. T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, and N. D´ıaz-Rodr´ıguez, "Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey," Apr. 2021. arXiv:2104.00950 [cs].

[31]. M. T. Ribeiro, S. Singh, and C. Guestrin, "Model- Agnostic Interpretability of Machine Learning," June 2016. arXiv:1606.05386 [cs, stat].

[32]. N. Strodthoff and C. Strodthoff, "Detecting and inter- preting myocardial infarction using fully convolutional neural networks," Physiological Measurement, vol. 40, p. 015001, Jan. 2019. arXiv:1806.07385 [cs, stat].

[33]. S. A. Siddiqui, D. Mercier, M. Munir, A. Dengel, and S. Ahmed, "TSViz: Demystification of Deep Learning Models for Time-Series Analysis," IEEE Access, vol. 7, pp. 67027–67040, 2019. arXiv:1802.02952 [cs].

[34]. H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Accurate and interpretable evalua- tion of surgical skills from kinematic data using fully convolutional neural networks," International Journal of Computer Assisted Radiology and Surgery, vol. 14, pp. 1611–1617, Sept. 2019.

[35]. F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. T. P. Hartono, R. Savitha, B. L. DeCost, S. I. P. Tian, G. Romano, A. G. Kusne, and T. Buonassisi, "Fast and interpretable classification of small X-ray diffrac- tion datasets using data augmentation and deep neural networks," Apr. 2019. arXiv:1811.08425 [cond-mat, physics:physics] version: 2.

[36]. R. Assaf, I. Giurgiu, F. Bagehorn, and A. Schumann, "MTEX-CNN: Multivariate Time Series EXplanations for Predictions with Convolutional Neural Networks," in 2019 IEEE International Conference on Data Mining (ICDM), pp. 952–957, Nov. 2019. ISSN: 2374-8486.

[37]. M. Munir, S. A. Siddiqui, F. Ku¨sters, D. Mercier, A. Dengel, and S. Ahmed, "TSXplain: Demystification of DNN Decisions for Time-Series using Natural Language and Statistical Features," vol. 11731, pp. 426– 439, 2019. arXiv:1905.06175 [cs].

[38]. S. Cho, G. Lee, W. Chang, and J. Choi, "Interpreta- tion of Deep Temporal Representations by Selective Visualization of Internally Activated Nodes," July 2020. arXiv:2004.12538 [cs].

[39]. A. Wolanin, G. Mateo-Garc´ıa, G. Camps-Valls, L. Go´mez-Chova, M. Meroni, G. Duveiller, Y. Liangzhi, and L. Guanter, "Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt," Environmental Research Letters, vol. 15, p. 024019, Feb. 2020. Publisher: IOP Publishing.

[40]. S. M. Lauritsen, M. E. Kalør, E. L. Kongsgaard, K. M. Lauritsen, M. J. Jørgensen, J. Lange, and B. Thiesson, "Early detection of sepsis utilizing deep learning on electronic health record event sequences," Artificial In- telligence in Medicine, vol. 104, p. 101820, Apr. 2020.

[41]. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Tor- ralba, "Learning Deep Features for Discriminative Lo- calization," Dec. 2015. arXiv:1512.04150 [cs].

[42]. K. Kashiparekh, J. Narwariya, P. Malhotra, L. Vig, and G. Shroff, "ConvTimeNet: A Pre-trained Deep Convo- lutional Neural Network for Time Series Classification," May 2019. arXiv:1904.12546 [cs, stat].

[43]. M. Guilleme´, V. Masson, L. Roze´, and A. Termier, "Agnostic Local Explanation for Time Series Classifi- cation," in 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), pp. 432–439, Nov. 2019. ISSN: 2375-0197.

[44]. Q. Pan, W. Hu, and J. Zhu, "Series Saliency: Temporal Interpretation for Multivariate Time Series Forecasting," Dec. 2020. arXiv:2012.09324 [cs].

[45]. A. A. Ismail, M. Gunady, H. C. Bravo, and S. Feizi, "Benchmarking Deep Learning Interpretability in Time Series Predictions," Oct. 2020. arXiv:2010.13924 [cs, stat].

[46]. S. Jain and B. C. Wallace, "Attention is not Explana- tion," May 2019. arXiv:1902.10186 [cs].

[47]. S. Wiegreffe and Y. Pinter, "Attention is not not Expla- nation," Sept. 2019. arXiv:1908.04626 [cs].

[48]. J. Crabbe´ and M. van der Schaar, "Explaining Time

[49]. A. Bibal, R. Cardon, D. Alfter, R. Wilkens, X. Wang, Series Predictions with Dynamic Masks," June 2021. arXiv:2106.05303 [cs].

[50]. Y. Zhao, J. Ren, B. Zhang, J. Wu, and Y. Lyu, "An explainable attention-based TCN heartbeats classifica- tion model for arrhythmia detection," Biomedical Signal Processing and Control, vol. 80, p. 104337, Feb. 2023.

[51]. F. Mujkanovic, V. Doskocˇ, M. Schirneck, P. Scha¨fer, and T. Friedrich, "timeXplain – A Framework for Explaining the Predictions of Time Series Classifiers," Nov. 2023. arXiv:2007.07606 [cs, stat].

[52]. P. Vinayavekhin, S. Chaudhury, A. Munawar, D. J. Agravante, G. De Magistris, D. Kimura, and R. Tachibana, "Focusing on What is Relevant: Time- Series Learning and Understanding using Attention," June 2018. arXiv:1806.08523 [cs].

[53]. W. Ge, J.-W. Huh, Y. R. Park, J.-H. Lee, Y.-H. Kim, and A. Turchin, "An Interpretable ICU Mortality Prediction Model Based on Logistic Regression and Recurrent Neural Networks with LSTM units," AMIA ... Annual Symposium proceedings. AMIA Symposium, vol. 2018, pp. 460–469, 2018.

[54]. F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate LSTM-FCNs for Time Series Classifica- tion," Neural Networks, vol. 116, pp. 237–245, Aug. 2019. arXiv:1801.04503 [cs, stat].

[55]. Y. Hao and H. Cao, "A New Attention Mechanism to Classify Multivariate Time Series," pp. 1971–1977, July 2020.

[56]. C. Schockaert, R. Leperlier, and A. Moawad, "Atten- tion Mechanism for Multivariate Time Series Recurrent Model Interpretability Applied to the Ironmaking Indus- try," July 2020. arXiv:2007.12617 [cs].

[57]. S. A. Siddiqui, D. Mercier, A. Dengel, and S. Ahmed, "TSInsight: A local-global attribution framework for interpretability in time-series data," Apr. 2020. arXiv:2004.02958 [cs, stat].

[58]. B. Lim, S. O. Arik, N. Loeff, and T. Pfis- ter, "Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting," Sept. 2020. arXiv:1912.09363 [cs, stat].

[59]. Y. S. Choi, S. Bae, J. H. Chang, S.-G. Kang, S. H. Kim, J. Kim, T. H. Rim, S. H. Choi, R. Jain, and S.-K. Lee, "Fully automated hybrid approach to predict the IDH mutation status of gliomas via deep learning and radiomics," Neuro-Oncology, vol. 23, pp. 304–313, Feb. 2021.

[60]. P. Gao, X. Yang, R. Zhang, and K. Huang, "Explain- able Tensorized Neural Ordinary Differential Equations forArbitrary-step Time Series Prediction," IEEE Trans- actions on Knowledge and Data Engineering, pp. 1–1, 2022. arXiv:2011.13174 [cs]. T. Franc¸ois, and P. Watrin, "Is Attention Explanation? An Introduction to the Debate," in Proceedings of the 60th Annual Meeting of the Association for Computa- tional Linguistics (Volume 1: Long Papers) (S. Muresan, P. Nakov, and A. Villavicencio, eds.), (Dublin, Ireland), pp. 3889–3900, Association for Computational Linguis- tics, May 2022.

[61]. P. Senin and S. Malinchik, "SAX-VSM: Interpretable Time Series Classification Using SAX and Vector Space Model," in 2013 IEEE 13th International Conference on Data Mining, pp. 1175–1180, Dec. 2013. ISSN: 2374- 8486.

[62]. T. L. Nguyen, S. Gsponer, I. Ilie, and G. Ifrim,

"Interpretable Time Series Classification using All- Subsequence Learning and Symbolic Representa- tions in Time and Frequency Domains," Aug. 2018. arXiv:1808.04022 [cs, stat].

[63]. S. El-Sappagh, J. Alonso, F. Ali, A. Ali, J.-H. Jang, and K. Kwak, "An Ontology-Based Interpretable Fuzzy Decision Support System for Diabetes Diagnosis," IEEE Access, vol. PP, pp. 1–1, July 2018.

[64]. J. Wang, Z. Peng, X. Wang, C. Li, and J. Wu, "Deep Fuzzy Cognitive Maps for Interpretable Multivariate Time Series Prediction," IEEE Transactions on Fuzzy Systems, vol. 29, pp. 2647–2660, Sept. 2021. Confer- ence Name: IEEE Transactions on Fuzzy Systems.

[65]. Y. Wang, R. Emonet, E. Fromont, S. Malinowski, E. Menager, L. Mosser, and R. Tavenard, "Learn- ing Interpretable Shapelets for Time Series Classifica- tion through Adversarial Regularization," June 2019. arXiv:1906.00917 [cs, stat].

[66]. P. Kidger, J. Morrill, and T. Lyons, "Generalised In- terpretable Shapelets for Irregular Time Series," May 2020. arXiv:2005.13948 [cs, stat].

[67]. G. Li, B. Choi, J. Xu, S. S. Bhowmick, K.-P. Chun, and G. L.-H. Wong, "Efficient Shapelet Discovery for Time Series Classification," IEEE Transactions on Knowledge and Data Engineering, vol. 34, pp. 1149–1163, Mar. 2022. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

[68]. L. Ye and E. Keogh, "Time series shapelets: a new primitive for data mining," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '09, (New York, NY, USA), pp. 947–956, Association for Computing Machinery, June 2009.

[69]. J. Lines, L. M. Davis, J. Hills, and A. Bagnall, "A shapelet transform for time series classification," in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '12, (New York, NY, USA), pp. 289–297, Asso- ciation for Computing Machinery, Aug. 2012.

[70]. A. H. Gee, D. Garcia-Olano, J. Ghosh, and D. Pay- darfar, "Explaining Deep Classification of Time- Series Data with Learned Prototypes," Sept. 2019. arXiv:1904.08935 [cs, stat].

[71]. B. Li, C. Jentsch, and E. Mu¨ller, "Prototypes as Expla- nation for Time Series Anomaly Detection," July 2023. arXiv:2307.01601 [cs].

[72]. C. Obermair, A. Fuchs, F. Pernkopf, L. Felsberger, A. Apollonio, and D. Wollmann, "Example or Proto- type? Learning Concept-Based Explanations in Time- Series," in Proceedings of The 14th Asian Conference on Machine Learning, pp. 816–831, PMLR, Apr. 2023. ISSN: 2640-3498.

[73]. S. Tonekaboni, S. Joshi, D. Duvenaud, and A. Gold- enberg, "Explaining Time Series by Counterfactuals," Sept. 2019.

[74]. E. Ates, B. Aksar, V. J. Leung, and A. K. Coskun, "Counterfactual Explanations for Multivariate Time Se- ries," in 2021 International Conference on Applied Arti- ficial Intelligence (ICAPAI), (Halden, Norway), pp. 1–8, IEEE, May 2021.

[75]. E. Delaney, D. Greene, and M. T. Keane, "Instance- based Counterfactual Explanations for Time Series Classification," June 2021. arXiv:2009.13211 [cs, stat].

[76]. J. Ho¨llig, C. Kulbach, and S. Thoma, "TSEvo: Evo- lutionary Counterfactual Explanations for Time Series Classification," in 2022 21st IEEE International Confer- ence on Machine Learning and Applications (ICMLA), pp. 29–36, Dec. 2022.

[77]. N. Mohammadi Foumani, L. Miller, C. W. Tan, G. I. Webb, G. Forestier, and M. Salehi, "Deep Learning for Time Series Classification and Extrinsic Regression: A Current Survey," ACM Computing Surveys, vol. 56, pp. 217:1–217:45, Apr. 2024.

[78]. M. T. Keane and B. Smyth, "Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI)," May 2020. arXiv:2005.13997 [cs].

[79]. E. M. Kenny and M. T. Keane, "Twin-systems to explain artificial neural networks using case-based rea- soning: comparative tests of feature-weighting methods in ANN-CBR twins for XAI," in Proceedings of the 28th International Joint Conference on Artificial In- telligence, IJCAI'19, (Macao, China), pp. 2708–2715, AAAI Press, Aug. 2019.

[80]. C. Nugent, D. Doyle, and P. Cunningham, "Gaining insight through case-based

explanation," Journal of Intelligent Information Systems, vol. 32, pp. 267–295, June 2009.

[81]. D. Leake and D. Mcsherry, "Introduction to the Special Issue on Explanation in Case-Based Reasoning," Arti- ficial Intelligence Review, vol. 24, pp. 103–108, Oct. 2005.

[82]. M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," Aug. 2016. arXiv:1602.04938 [cs, stat].

[83]. S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Nov. 2017. arXiv:1705.07874 [cs, stat].

[84]. T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "The Dangers of Post-hoc In- terpretability: Unjustified Counterfactual Explanations," July 2019. arXiv:1907.09294 [cs, stat].

[85]. T. Spooner, D. Dervovic, J. Long, J. Shepard, J. Chen, and D. Magazzeni, "Counterfactual Expla- nations for Arbitrary Regression Models," June 2021. arXiv:2106.15212 [cs].

[86]. R. K. Mothilal, A. Sharma, and C. Tan, "Explaining Machine Learning Classifiers through Diverse Counter- factual Explanations," in Proceedings of the 2020 Con- ference on Fairness, Accountability, and Transparency, pp. 607–617, Jan. 2020. arXiv:1905.07697 [cs, stat].

[87]. M. Mu¨ller, "Dynamic Time Warping," in Information Retrieval for Music and Motion, pp. 69–84, Berlin, Heidelberg: Springer, 2007.

[88]. R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. An- droz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, and E. Woods, "Tslearn, A Machine Learning Toolkit for Time Series Data," Journal of Machine Learning Research, vol. 21, no. 118, pp. 1–6, 2020.

[89]. A. Hagen, "DiCE: Counterfactual Explanations offer clarity in AI decision-making," Jan. 2020.

[90]. G. Forestier, F. Petitjean, H. A. Dau, G. I. Webb, and E. Keogh, "Generating Synthetic Time Series to Augment Sparse Datasets," in 2017 IEEE International Conference on Data Mining (ICDM), pp. 865–870, Nov. 2017. ISSN: 2374-8486.

[91]. M. Mitchell, An Introduction to Genetic Algorithms. The MIT Press, Mar. 1998.

[92]. CDC, "NHANES 2003-2004: Physical Activity Moni- tor Data Documentation, Codebook, and Frequencies," 2003.

[93]. CDC, "NHANES 2005-2006: Physical Activity Data Documentation, Codebook, and Frequencies," 2005.

[94]. G. E. P. Box and D. R. Cox, "An Analysis of Transfor- mations," Journal of the Royal Statistical Society. Series B (Methodological), vol. 26, no. 2, pp. 211–252, 1964. Publisher: [Royal Statistical Society, Wiley].

[95]. D. Kwon and D. W. Belsky, "A toolkit for quantification of biological age from blood chemistry and organ func- tion test data: BioAge," GeroScience, vol. 43, pp. 2795–2808, Dec. 2021.

[96]. CDC, "BIOPROd NHANES 2005-2006: Standard Bio- chemistry Profile Data Documentation, Codebook, and Frequencies," 2005.

[97]. CDC, "CRPd NHANES 2005-2006: C-Reactive Pro- tein (CRP) Data Documentation, Codebook, and Fre- quencies," 2005.

[98]. CDC, "BMXc, NHANES 2003-2004: Body Mea- sures Data Documentation, Codebook, and Frequen- cies," 2003.

[99]. CDC, "L10c NHANES 2003-2004: Glycohemoglobin Data Documentation, Codebook, and Frequencies," 2003.

[100]. CDC, "BPQc NHANES 2003-2004: Blood Pressure & Cholesterol Data Documentation, Codebook, and Frequencies," 2003.

[101]. CDC, "L25c NHANES 2003-2004: Complete Blood Count with 5-part Differential - Whole Blood Data Documentation, Codebook, and Frequencies," 2003.

[102]. P. Klemera and S. Doubal, "A new approach to the con- cept and computation of biological age," Mechanisms of Ageing and Development, vol. 127, pp. 240–248, Mar. 2006.

[103]. C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal Convolutional Networks: A Unified Approach to Action Segmentation," Aug. 2016. arXiv:1608.08242 [cs].

[104]. S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evalu- ation of Generic Convolutional and Recurrent Networks for Sequence Modeling," Apr. 2018. arXiv:1803.01271 [cs].

[105]. D. P. Kingma and J. Ba, "Adam: A Method

for Stochas- tic Optimization," Jan. 2017. arXiv:1412.6980 [cs].

[106]. Z. Wang, I. Miliou, I. Samsten, and P. Papapetrou, "Counterfactual Explanations for Time Series Forecast- ing," Oct. 2023. arXiv:2310.08137 [cs].

[107]. M. D´ıaz-Lozano, D. Guijo-Rubio, P. A. Gutie´rrez, A. M. Go´mez-Orellana, I. Tu´n˜ez, L. Ortigosa-Moreno, A. Romanos-Rodr´ıguez, J. Padillo-Ruiz, and C. Herva´s-Mart´ınez, "COVID-19 contagion forecasting framework based on curve decomposition and evolutionary artifi- cial neural networks: A case study in Andalusia, Spain," Expert Systems with Applications, vol. 207, p. 117977, Nov. 2022.

[108]. F. Hamman, E. Noorani, S. Mishra, D. Magazzeni, and S. Dutta, "Robust Counterfactual Explanations for Neural Networks With Probabilistic Guarantees," Mar. 2024. arXiv:2305.11997 [cs, math, stat].

[109]. S. Letzgus, P. Wagner, J. Lederer, W. Samek, K.-R. Mu¨ller, and G. Montavon, "Toward Explainable AI for Regression Models," IEEE Signal Processing Magazine, vol. 39, pp. 40–58, July 2022. arXiv:2112.11407 [cs, stat].