Counterfactual Explanations of Neural Network-Generated Response Curves

Giorgio Morales and John Sheppard
Gianforte School of Computing
Montana State University
Bozeman, MT 59717

2023 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-8867-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IJCNN54540.2023.10191746

Abstract—**Response curves exhibit the magnitude of the response of a sensitive system to a varying stimulus. However, response of such systems may be sensitive to multiple stimuli (i.e., input features) that are not necessarily independent. As a consequence, the shape of response curves generated for a selected input feature (referred to as "active feature") might depend on the values of the other input features (referred to as "passive features"). In this work we consider the case of systems whose response is approximated using regression neural networks. We propose to use counterfactual explanations (CFEs) for the identification of the features with the highest relevance on the shape of response curves generated by neural network black boxes. CFEs are generated by a genetic algorithm-based approach that solves a multi-objective optimization problem. In particular, given a response curve generated for an active feature, a CFE finds the minimum combination of passive features that need to be modified to alter the shape of the response curve. We tested our method on a synthetic dataset with 1-D inputs and two crop yield prediction datasets with 2-D inputs. The relevance ranking of features and feature combinations obtained on the synthetic dataset coincided with the analysis of the equation that was used to generate the problem. Results obtained on the yield prediction datasets revealed that the impact on fertilizer responsivity of passive features depends on the terrain characteristics of each field.**

Index Terms—**Counterfactual explanations, response curves, deep regression, explainable machine learning.**

I. Introduction

Response curves are tools that allow for the analysis of the responsivity of a sensitive system to a particular stimulus; we refer to that stimulus as the "active feature". Specifically, a response curve is defined as a curve that exhibits the various values taken by the response of the system to all admissible values of the active feature.

Typically, analysis of response curves is focused on their shape rather than on their absolute values. For instance, in pharmacology, the shape of a drug's dose–response curve reflects the strength of the drug [1], [2]. Likewise, in agriculture, nitrogen fertilizer-yield response (N-response) curves estimate the crop yield based on different fertilizer inputs. The shape of such curves can help determining the economic optimum nitrogen rate (EONR), defined as the nitrogen rate beyond which there is no actual profit for the farmers [3].

Motivating the current work, response values displayed by response curves may depend, not only on the relationship between the response variable and the active feature, but also on other stimuli, which we will refer to as "passive features". Traditionally, such response curves are fitted using univariate (non-)linear regression models between the active feature and the response variable. Such models could be based on parametric response functions that assume plateau-type, quadratic, and exponential behavior, in the case of N-response curves [3], [4], or sigmoidal, U-shape, and hill functions, in the case of dose-response curves [2], [5].

The traditional approaches assume that, given an active feature, the shape of its response curves is independent of other variables that could possibly affect the absolute response value. In the case of agricultural applications, fitting a single N-response curve for an entire field implies that the field is homogeneous and behaves similarly at every location. Nevertheless, recent works suggest that the N-response functional form varies across the fields given the variability of factors such as terrain slope and soil composition [6], [7].

In this work, we build upon the conclusion that the shape of response curves might be dependent not only on the active feature but also on the interaction of the active feature with the passive features. We present a method to derive non-parametric response curves from observed data. That is, we learn the functional form of the response curves instead of using parametric curve-fitting approaches. To do this, we train multivariate regression neural networks (NNs) that act as mappings from the feature space and the response value space. Thus, they are used to generate approximated response curves given an active feature.

The high complexity of the non-linear functions learned by NNs often prevents humans from explaining their behavior, which is why they are usually referred to as "black-box models." Given that the importance of NNs in inference tasks has grown rapidly, the area of explainable machine learning (XML) has gained more interest in recent years. XML aims to allow humans to identify cause-effect relationships between inputs and outputs of black-box models [8]. In concordance with this, we propose a *post-hoc* explainability method that allows for understanding the impact that interacting passive features have on the shape of NN-generated response curves.

Thus, we propose a method that generates counterfactual explanations (CFEs) for each sample to find the minimum number of passive features we need to modify for the response curve of the counterfactual sample to show a change in responsivity with respect to the original one. The CFE generation problem is posed as a multi-objective optimization

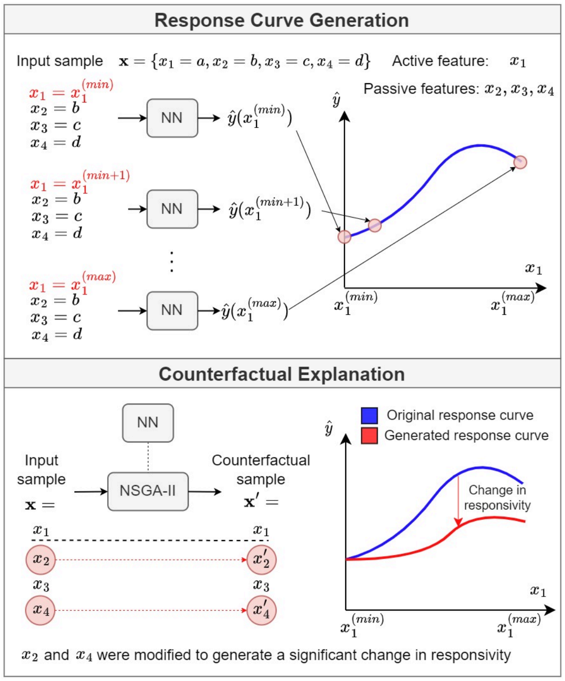


Fig. 1. Overview of counterfactual response curve generation methodology.

(MOO) problem and is solved using a genetic algorithm-based approach. The upper and lower regions of Fig. 1 illustrate the response curve generation and counterfactual response curve generation processes, respectively, of a given sample. Here, the Non-dominated Sorting Genetic Algorithm II (NSGA-II) represents the selected genetic algorithm. Finally, we repeat this process for each sample in a dataset to derive global relevance scores, not only for each passive feature but also for specific combinations of them. Our specific contributions are summarized as follows:

1) Our main contribution is a CFE generation method posed as a MOO problem that, given a sample, finds the passive features with the greatest impact on the responsivity of the active feature.
2) Considering a multivariate regression problem, this is the first work that analyzes the influence of a set of passive features over the shape of the response curves generated for the response variable and a selected active feature.
3) We present a method to generate aligned approximate response curves using regression NNs.
4) We provide global scores that assess the relevance of individual features and combinations of features.

II. Related Work

Counterfactual explanations are used to identify "actionable knowledge," i.e., knowledge about causal dependencies between inputs and outputs of a system [9]. Such explanations contribute to understanding what could be changed in the input to achieve a desired outcome. Most CFE methods focus on classification problems where the objective is to find CFEs that produce class labels different from those that were originally predicted [10]. These methods can be categorized as model-agnostic or model-specific. The former are based on general principles that use outputs of already fitted black-box models [9], [11], [12], where the latter are specific to a particular class of methods, such as differentiable models [13] or tree-based ensembles [14].

To date, little attention has been given to counterfactual methods for regression. Spooner *et al.* [15] proved that verifying the existence of counterfactuals is NP-complete. Regression problems may deal with multiple-dimensional real outputs and, thus, it is not always possible to have efficient algorithms for establishing the existence of counterfactuals with a desired target value. However, it is not always necessary to find CFEs that lead to specific target values. For example, Kommiya Mothilal *et al.* [16] proposed a method that assesses individual feature relevance from CFEs. They consider that a feature is more relevant than others for a prediction task if it is changed more often when generating CFEs.

On the other hand, analyses of systems' response curves consider that the shape of the response curve can be described solely by the relationship between the response variable and the active feature [3], [4]. This requires the strong assumption that the explanatory features are mutually independent. If this held, the set of passive features would only shift the response curves vertically but would not alter their shape. However, there is no guarantee for this assumption to always hold, as seen in some precision agriculture applications [6], [7]. Thus, we argue that it is essential to acknowledge all possible sources of variability in analyzing the shape of response curves. To the best of our knowledge, no work has been published on this type of analysis.

In this paper, we explore the use of CFEs to analyze the impact of passive features on the shape of response curves. It is worth noting that Schwab *et al.* [17] proposed a method to learn counterfactual representations for estimating individual dose-response curves using NNs. They aimed to estimate what would have happened if a different treatment (dose) had been given to a patient, hence their objective was to produce accurate response estimates across the entire range of all available treatment options. In this context, the term "counterfactual outcome" simply refers to the estimated response value obtained for dosages different from the one that the patient actually received (i.e., estimated response curves). This differs from our perspective significantly. Unlike previous counterfactual methods that aim to obtain a different response value, our objective is to generate counterfactual samples for a subset of the input features (i.e., passive features) in order to alter the response curve's shape. In other words, our CFEs aim to change the system's responsivity; that is, the way it reacts to all admissible values of the active feature.

III. Methodology

Let $f(\cdot)$ denote the underlying function of a system whose response y is sensitive to n stimuli (i.e., input features) $\mathbf{x} = \{x_1, \ldots, x_n\}$, such that $y = f(\mathbf{x}) + \varepsilon$. Here, ε is a random variable that represents the error term. For convenience, we first consider the case where each input and output is one-dimensional ($x_i \in \mathbb{R}$, $\forall i \in [1, n]$ and $y \in \mathbb{R}$). Later in Sec. IV-B, we extend this approach to consider two-dimensional inputs.

Without loss of generality, suppose we select the s-th feature ($s \in [1, n]$) as the active feature. The set of remaining features (i.e., passive features) is denoted as $\mathbf{p} = \mathbf{x} \setminus \{x_i\}$. Thus, a response curve generated for the s-th feature, $R_s(\mathbf{x})$, consists of the set of values taken by the response y for all admissible values of x_s (bounded by $x_s^{\min} \leq x_s \leq x_s^{\max}$) as follows:

$$R_s(\mathbf{x}) = \{f(\mathbf{x}|x_s = x_s^{\min}), \ldots, f(\mathbf{x}|x_s = x_s^{\max})\}.$$

A. Response Curve Generation

In many real-life settings, the underlying function $f(\cdot)$ cannot be retrieved directly; therefore, we need to approximate it based on observed data. Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be a data set with N training samples, where each sample is denoted as $\mathbf{x}_j = \{x_{j1}, \ldots, x_{jn}\}$, and $\mathbf{y} = \{y_1, \ldots, y_N\}$ is the set of corresponding target observations. We construct a NN regression model that captures the association between $\mathbf{X}$ and $\mathbf{y}$. Its computed function is denoted as $\hat{f}(\cdot)$, and $\boldsymbol{\theta}$ denotes its weights. Thus, given an input $\mathbf{x}_j$, the target estimate is computed as $\hat{y}_j = \hat{f}(\mathbf{x}_j, \boldsymbol{\theta})$.

The network $\hat{f}$ is trained to reduce the mean square error of the estimations such that the parameters $\boldsymbol{\theta}$ are obtained by the following optimization: $\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{N} \sum_{j=1}^{N} (\hat{y}_j - y_j)^2$. Hence, once the network is trained, and assuming that it captures the underlying causal structure of the problem sufficiently well, it can be used to generate an approximate response curve, $\hat{R}_s(\mathbf{x}_j)$, for a given active feature and input sample $\mathbf{x}_j$:

$$\hat{R}_s(\mathbf{x}_j) = \{\hat{f}(\mathbf{x}_j|x_{js} = x_s^{\min}), \ldots, \hat{f}(\mathbf{x}_j|x_{js} = x_s^{\max})\}. \quad (1)$$

Note that we are not interested in the absolute estimated response values when comparing the shape of two or more curves. Thus, we get rid of any vertical shifts and obtain the aligned approximate response curve $\tilde{R}_s(\mathbf{x}_j)$ by subtracting from $\hat{R}_s(\mathbf{x}_j)$ its minimum value:

$$\tilde{R}_s(\mathbf{x}_j) = \hat{R}_s(\mathbf{x}_j) - \min(\hat{R}_s(\mathbf{x}_j)). \quad (2)$$

Finally, we create the set $\mathbf{R}_s$ that consists of the approximate response curves of all samples in $\mathbf{X}$; that is, $\mathbf{R}_s = \{\tilde{R}_s(\mathbf{x}_0), \ldots, \tilde{R}_s(\mathbf{x}_N)\}$.

It is worth mentioning that, previously, we also experimented with other types of classifiers (i.e., support vector machines and random forests) to use in the response curve generation process. However, due to the fast convergence rates and the substantial improvements in performance, we decided to focus on feedforward NNs (FNNs) and convolutional neural networks (CNNs) for 1-D and 2-D regression, respectively, over the other types of models.

B. Functional Principal Component Analysis

The set $\mathbf{R}_s$ can be interpreted as a set of functional data where each of its samples corresponds to an approximated response curve. Recall that one of our objectives is to generate counterfactual response curves whose shapes differ with respect to the original ones. As such, we need to determine a distance metric that conveys the difference in shape between two functional data samples.

In that sense, we use functional principal component analysis (fPCA), which is a tool to extract the dominant modes of variation of functional data [18]. This approach allows us to obtain a reduced set of K orthonormal functions $\{\xi_1, \ldots \xi_K\}$ so that each curve in $\mathbf{R}_s$ is approximated as an expansion of these basis functions as $\tilde{R}_s(\mathbf{x}_j) \approx \sum_{k=1}^{K} v_k^{(s)}(\mathbf{x}_j)\, \xi_k$, where $v_k^{(s)}(\mathbf{x}_j)$ is the value of the k-th principal component corresponding to the response curve generated for $\mathbf{x}_j$ considering the s-th feature as the active feature.

A *functional* principal component (fPC) represents a distinct curve pattern. Hence, two curves with different shapes will be encoded using different fPC values. We consider using $K = 3$ fPCs, as they were sufficient to explain at least 99.5% of the variance of the datasets used in our experiments. We define our distance metric $d_s(\mathbf{x}_j, \mathbf{x}_q)$, calculated between the transformed response curves generated for $\mathbf{x}_j$ and $\mathbf{x}_q$, as:

$$d_s(\mathbf{x}_j, \mathbf{x}_q) = \sqrt{\sum_{k=1}^{K=3} \left(v_k^{(s)}(\mathbf{x}_j) - v_k^{(s)}(\mathbf{x}_q)\right)^2}. \quad (3)$$

C. Counterfactual Explanations for Response Curves

Given a sample $\mathbf{x}_j$, let $\mathbf{x}'_j = \{x'_{j1}, \ldots, x'_{jn}\}$ denote a counterfactual explanation and let $\tilde{R}_s(\mathbf{x}'_j)$ be its corresponding aligned approximated response curve. More specifically, we define $\mathbf{x}'_j$ as a data point whose response curve $\tilde{R}_s(\mathbf{x}'_j)$ has a different shape from that of $\tilde{R}_s(\mathbf{x}_j)$ (i.e., it shows different responsivity), such that $d_s(\mathbf{x}_j, \mathbf{x}'_j) \geq \epsilon$, where ϵ is a hyperparameter threshold. The CFE $\mathbf{x}'_j$ is obtained by introducing perturbations to the set of passive features:

$$x'_{ji} = \begin{cases} x_{ji} + \Delta_{ji}, & \text{if } i \neq s \\ x_{ji}, & \text{if } i = s. \end{cases}$$

Moreover, we aim to find the minimum subset of passive features that should be affected by small perturbations so that the CFE's responsivity is sufficiently different from that of the original one. As such, the counterfactual search problem can be cast as a MOO problem that is solved for each sample $\mathbf{x}_j$ as follows:

$$\min_{\mathbf{x}'_j} \mathbf{g}(\mathbf{x}_j) = \min_{\mathbf{x}'_j} \left(g_1(\mathbf{x}_j, \mathbf{x}'_j), g_2(\mathbf{x}_j, \mathbf{x}'_j), g_3(\mathbf{x}_j, \mathbf{x}'_j)\right), \quad (4)$$

where $g_1(\mathbf{x}_j, \mathbf{x}'_j)$, $g_2(\mathbf{x}_j, \mathbf{x}'_j)$, and $g_3(\mathbf{x}_j, \mathbf{x}'_j)$ are independent objective functions whose goals may contradict.

The first objective maximizes the distance between the transformed response curves using Eq. 3:

$$g_1(\mathbf{x}_j, \mathbf{x}'_j) = \begin{cases} -d_s(\mathbf{x}_j, \mathbf{x}'_j), & \text{if } d_s(\mathbf{x}_j, \mathbf{x}'_j) < \epsilon \\ -\epsilon, & \text{if } d_s(\mathbf{x}_j, \mathbf{x}'_j) \geq \epsilon. \end{cases} \quad (5)$$

The second objective function is used to minimize the number of features that are modified, which are calculated using the

L_0 norm of $(\mathbf{x}_j - \mathbf{x}'_j)$ so that:

$$g_2(\mathbf{x}_j, \mathbf{x}'_j) = ||\mathbf{x}_j - \mathbf{x}'_j||_0 = \sum_{i=1}^{n} \mathbb{I}_{x_{ji} \neq x'_{ji}}. \tag{6}$$

Finally, the third objective function minimizes the distance between $\mathbf{x}_j$ and $\mathbf{x}'_j$:

$$g_3(\mathbf{x}_j, \mathbf{x}'_j) = \frac{1}{n} \sum_{i=1}^{n} \delta_G(\mathbf{x}_{ji}, \mathbf{x}'_{ji}), \tag{7}$$

where δ_G represents the Gower distance, which takes into account that passive features can be numerical or categorical:

$$\delta_G(\mathbf{x}_{ji}, \mathbf{x}'_{ji}) = \begin{cases} \frac{1}{r_j}(|\mathbf{x}_j - \mathbf{x}'_j|), & \text{if numerical} \\ -\epsilon, & \text{if categorical,} \end{cases} \tag{8}$$

and r_i indicates the range of values of the i-th feature. All of the datasets used in this work consist of numerical features only (i.e., integer and real-valued); however, the use of the Gower distance in Eq. 7 allows us to consider more general cases.

Note that calculating $g_1(\mathbf{x}_j)$, $g_2(\mathbf{x}_j)$, and $g_3(\mathbf{x}_j)$ involves generating the aligned approximated response curves $\tilde{R}_s(\mathbf{x}_j)$ and $\tilde{R}_s(\mathbf{x}'_j)$. Some approaches optimize the CFE objective functions as part of a differentiable loss function [13]; however, those approaches cannot be applied in our case given that generating an aligned approximated response curve is not a differentiable process (see Eq. 1 and 2).

Similar to the work proposed by Dandl *et al.* [11], we use the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [19] to solve the mixed-variable optimization problem given in Eq. 4. NSGA-II is an elitist genetic algorithm that finds Pareto non-dominated solutions to MOO problems and uses a crowding distance measure to maintain diversity in subsequent generations. NSGA-II was preferred over the follow-on NSGA-III since it has shown superior performance on optimization problems with three objectives [20]. Note that even though we use a similar optimization framework to that of Dandl *et al.* [11], our objectives are designed to alter the shape of the response curve of a given sample (Fig. 1), while theirs are designed to alter a single response value in the context of classification or regression problems.

Let us consider a population size of T_0 CFE candidates, from which NSGA-II may select T non-dominated solutions ($T \leq T_0$) denoted as $\{\mathbf{x'}_j^{(1)}, \ldots, \mathbf{x'}_j^{(T)}\}$. For each solution, we calculate its performance $z_t = \{g_1(\mathbf{x'}_j^{(t)}), g_2(\mathbf{x'}_j^{(t)}), g_3(\mathbf{x'}_j^{(t)})\}$. The objective space is defined as the set of T three-dimensional points $\mathbf{Z} = \{z_1, \ldots, z_T\}$. A solution z_t is said to dominate another solution z_q ($z_t \preceq z_q$) if it is no worse than z_q (i.e., $g_1(\mathbf{x'}_j^{(t)}) \leq g_1(\mathbf{x'}_j^{(q)})$, $g_2(\mathbf{x'}_j^{(t)}) \leq g_2(\mathbf{x'}_j^{(q)})$, and $g_3(\mathbf{x'}_j^{(t)}) \leq g_3(\mathbf{x'}_j^{(q)})$) and it is strictly better than z_q in at least one objective. By definition, the set of non-dominated solutions $\mathbf{Z}$ constitutes a Pareto set on the Pareto front.

The remaining question is how to select the best solution from $\mathbf{Z}$. Recall that we are mainly interested in minimizing the number of modified features that are sufficient to alter the responsivity of $\mathbf{x}_j$ such that $d_s(\mathbf{x}_j, \mathbf{x}'_j) \geq \epsilon$. Thus, from the subset of solutions in the Pareto set that produces the lowest g_1 value, we select the solution that yields the lowest g_2 value. Note that, according to this criterion, we could reconfigure our MOO problem to optimize g_1 and g_2 only, or g_1 and g_3 only (selecting the solution with the fewest changes). Nevertheless, in practice, we noticed substantially faster convergence rates by optimizing the three functions simultaneously, as they were shown to guide the search more effectively.

D. Local and Global Explanations

Two types of explanation are sought: local and global. Local explanations convey which passive features have the greatest impact on the responsivity of a given sample, while global explanations allow for the identification of the passive features with the highest impact on the shape of the response curves generated for a sensitive system in general.

Given an input $\mathbf{x}_j$, its local explanation α_j is the set of passive features that were modified during the generation of the counterfactual sample $\mathbf{x}'_j$:

$$\alpha_j = \{i \mid (x_{ji} \in \mathbf{x}_j) \wedge (x'_{ji} \in \mathbf{x}'_j) \wedge (x_{ji} \neq x'_{ji})\}. \tag{9}$$

On the other hand, the global explanation of a sensitive system is twofold. First, we assess individual feature relevance r_i by providing the ratio of times that a feature was modified during the CFE generation process of all samples in $\mathbf{X}$:

$$r_i = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}_{i \,\in\, \alpha_j}. \tag{10}$$

We acknowledge that passive features are not necessarily independent and thus individual relevance scores are not enough to understand how they interact. For that reason, we also report the five most repeated feature combinations. By doing so, we identify which features react together and which feature combinations are the most effective. We restrict the number of reported feature combinations to five for conciseness, as explanations becomes more manageable for humans when smaller result sets are provided [8]. However, note that more feature combinations might need to be provided when dealing with problems with several passive features (the datasets used in this work require less than eight).

IV. Experimental Results

For our experiments, we evaluated our approach on a synthetic dataset with 1-D inputs and two real-world crop yield prediction datasets with 2-D outputs. As stated previously, no other works have been published on the analysis of the impact that a set of passive features have on the shape of response curves generated for an active feature. Therefore, we were unable to include other methods for comparison.

We also note that the described approach should not be confused with a sensitivity analysis, which is used to study how the different values of a set of independent variables affect the response variable. In contrast, we study how the different values of a subset of the input features (i.e., the passive features) affect how the response variable reacts to the entire range of admissible values of a selected feature (i.e., the active feature). Furthermore, we do not assume independence among features; thus, we report the feature combinations with the greatest responsivity impact in addition to the estimated individual feature relevance values.

A 10-fold cross-validation (CV) design was used with all datasets. Having selected an active feature, the trained network was used to generate the response curves for all samples in the dataset along with their corresponding CFEs. We argue that including samples from the training set in this process does not lead to biased results. The reason is that, when generating a response curve, we synthesize samples that were not observed either in the training set or in the validation set. We considered that the resulting global explanations would be more consistent if they were calculated using as many approximated response curves as possible. Thus, these curves were used to produce local and global explanations, and we produced an independent set of explanations for each of the ten folds, given each fold could yield different curves since we have no ground truth for the curves themselves. Even so, if the data available is sufficiently large and diverse, it is reasonable to expect similar curves and explanation results from the ten models. The implementation code is available at https://github.com/GiorgioMorales/ResponsivityAnalysis.

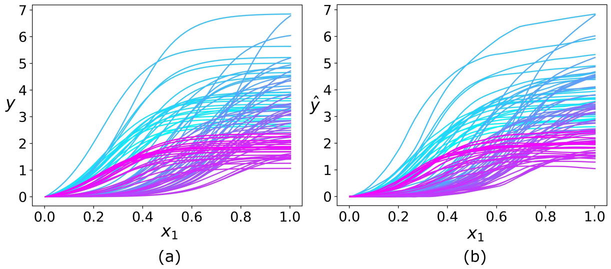
A. Synthetic Dataset

Validation of the analysis proposed in this work is challenging if the target response curves are unknown, which is the case when working with our real-world agricultural applications. For this reason, we created a synthetic dataset consisting of 10,000 samples derived from the following multiple nonlinear regression problem with five input features:

$$y = \text{sigmoid}((10x_1 - 5) + x_2)x_3^2 x_4 + 10x_5, \qquad (11)$$

where $x_1 \sim U(0,1)$, $x_2 \sim U(-3,3)$, $x_3 \sim U(1,2)$, $x_4 \sim U(1,2)$, and $x_5 \sim U(0,2)$. We used x_1 as the active feature ($s=1$).

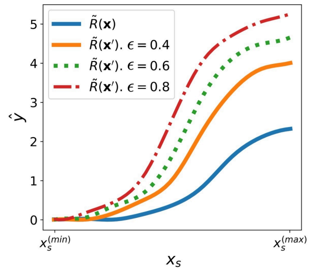
For these experiments, we trained a feed-forward neural network with two hidden layers, each with 100 nodes. We calculated the mean square error (MSE) on the validation sets after CV to analyze regression performance. The resulting average MSE and standard deviation were $6.78 \times 10^{-3} \pm 1.32 \times 10^{-4}$. Fig. 2 shows the aligned ground-truth response curves (generated from Eq. 11) and the aligned approximated response curves using the NN trained during the first CV iteration, for 100 random samples.



(a)

(b)

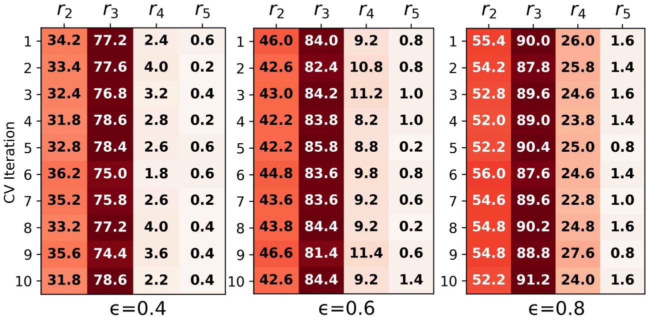
Fig. 2. Response curve generation from the synthetic dataset. **(a)** Ground-truth response curves. **(b)** NN-generated response curves.



Passive features

	Original	$\epsilon = 0.4$	$\epsilon = 0.6$	$\epsilon = 0.8$
x_2	−1.87	−1.11	−0.66	−0.06
x_3	1.32	1.77	1.90	1.99
x_4	1.30	1.30	1.30	1.30
x_5	0.62	0.62	0.62	0.62

Fig. 3. Example of the counterfactual response curves generated using $\epsilon = 0.4, 0.6,$ and 0.8 for a sample of the synthetic dataset.



	ϵ=0.4				ϵ=0.6				ϵ=0.8			
	r_2	r_3	r_4	r_5	r_2	r_3	r_4	r_5	r_2	r_3	r_4	r_5
1	34.2	77.2	2.4	0.6	46.0	84.0	9.2	0.8	55.4	90.0	26.0	1.6
2	33.4	77.6	4.0	0.2	42.6	82.4	10.8	0.8	54.2	87.8	25.8	1.4
3	32.4	76.8	3.2	0.4	43.0	84.2	11.2	1.0	52.8	89.6	24.6	1.6
4	31.8	78.6	2.8	0.2	42.2	83.8	8.2	1.0	52.0	89.0	23.8	1.4
5	32.8	78.4	2.6	0.6	42.2	85.8	8.8	0.2	52.2	90.4	25.0	0.8
6	36.2	75.0	1.8	0.6	44.8	83.6	9.8	0.8	56.0	87.6	24.6	1.4
7	35.2	75.8	2.6	0.2	43.6	83.6	9.2	0.6	54.6	89.6	22.8	1.0
8	33.2	77.2	4.0	0.4	43.8	84.4	9.2	0.2	54.8	90.2	24.8	1.6
9	35.6	74.4	3.6	0.4	46.6	81.4	11.4	0.6	54.8	88.8	27.6	0.8
10	31.8	78.6	2.2	0.4	42.6	84.4	9.2	1.4	52.2	91.2	24.0	1.6

Fig. 4. Individual relevance of passive features (in %) of the synthetic dataset.

For the CFE generation process, we used a population size of $T = 50$ samples for NSGA-II and 100 iterations. Eq. 5 specifies that two response curves show different responsivity if their distance in the transformed space (after using fPCA) is greater than a threshold ϵ. Since selecting ϵ is subjective, we considered using multiple threshold values and evaluating the consistency of the results. Intuitively, higher ϵ leads to bigger differences between the shapes of two response curves; thus, the number of modified passive features might increase. In the future, we plan to replace this threshold with tests that determine if the responsivity of two or more response curves is statistically significant. For this dataset, we considered three thresholds: $\epsilon \in \{0.4, 0.6, 0.8\}$. These values were chosen because they allowed us to see how the feature relevance values change as the ϵ values increase.

For example, Fig. 3 shows the counterfactual response curves generated for the first sample of the dataset ($j = 1$) using the network trained during the first CV iteration. Given this input, the local explanation is given by $\alpha_1 = \{2, 3\}$ for the three selected thresholds. In other words, it was sufficient to modify the second and third features, x_2 and x_3 to alter the responsivity of the selected sample, regardless of the threshold.

We repeated this process for all the samples to obtain global explanations. Fig. 4 shows the individual feature relevance values achieved for all CV iterations and ϵ values. The remaining global explanations are given by the most frequent feature combinations, which are reported in Table I. Specifically, we counted the combination of features that were most repeated across all the CV iterations and selected the top five. For each selected combination, we calculated the ratio of times it appeared in each CV iteration and reported the mean ratio

TABLE I

TOP-FIVE FEATURE COMBINATIONS – SYNTHETIC DATASET

ϵ	0.4		0.6		0.8	
#	Comb.	% Rep.	Comb.	% Rep.	Comb.	% Rep.
1	$[x_2, x_3]$	10.9 ± 0.6	$[x_2, x_3]$	27.2 ± 1.2	$[x_2, x_3]$	36.6 ± 1.2
2	$[x_3, x_4]$	2.7 ± 0.7	$[x_3, x_4]$	8.3 ± 0.9	$[x_3, x_4]$	15.5 ± 1.0
3	$[x_3, x_5]$	0.2 ± 0.1	$[x_2, x_3, x_4]$	0.6 ± 0.4	$[x_2, x_3, x_4]$	6.4 ± 0.8
4	$[x_2, x_5]$	0.1 ± 0.1	$[x_2, x_4]$	0.4 ± 0.2	$[x_2, x_4]$	2.2 ± 0.5
5	—	—	$[x_3, x_5]$	0.3 ± 0.2	$[x_3, x_4, x_5]$	0.4 ± 0.2

along with its corresponding standard deviation.

B. Yield Prediction Dataset

To consider the usefulness of our approach in a real-world setting, we analyzed data collected on a crop yield prediction problem, which is one of the main tasks of precision agriculture. Accurate and reliable crop yield prediction provides farmers with tools to make informed decisions, such as determining the nitrogen fertilizer rates needed in specific regions of their fields to maximize their profits [6].

We used an early-yield prediction dataset of winter wheat we curated and presented in previous work [21]. The early-yield prediction is posed as a regression problem where the explanatory variables are represented by a set of features obtained during the growing season (March):

1) N: Nitrogen rate applied previously (lb/ac).
2) A: Topographic aspect (radians).
3) S: Topographic slope (degrees).
4) TPI: Topographic position index.
5) P: Prior year precipitation (mm).
6) VH and VV: Backscattering coefficients obtained from synthetic aperture radar (SAR) images from Sentinel-I.

The response variable corresponds to the yield value in bushels per acre (bu/ac), measured during the harvest season (August). Hence, the data acquired in March is used to predict crop yield values in August of the same year.

Each sample $\mathbf{x}_j = \{x_{j1}, \dots x_{jn}\}$ is represented as a spatial data cube of 5×5 pixels with seven features or channels ($n = 7$); that is, $x_{ji} \in \mathbb{R}^{5 \times 5}$, $\forall i \in [1, n]$. Each pixel represents a region of $10 \times 10\,\mathrm{m}$ of the field. The output represents the yield value corresponding to the central pixel of the input sample ($y_j \in \mathbb{R}$). To tackle this regression problem, we trained a convolutional neural network. In particular, we use the Hyper3DNetReg network architecture we proposed in [21]. It is a 3D–2D CNN specifically designed to predict the yield values of small spatial neighborhoods of a field. For our experiments, we used data collected from two winter wheat fields, which we refer to as "Field A" and "Field B". Data from three growing seasons were collected for each field (2016, 2018, and 2020).

The selected active feature of this problem was N (i.e., fertilizer input, $s = 1$), as we are interested in the analysis of the N-response curves. The remaining six features constituted the set of passive features. Important factors such as EONR depend directly on the shape of these curves. For instance, EONR is traditionally found as the fertilizer rate at which the first derivative of the N-response curve is equal to a common yield-nitrogen price ratio [3].

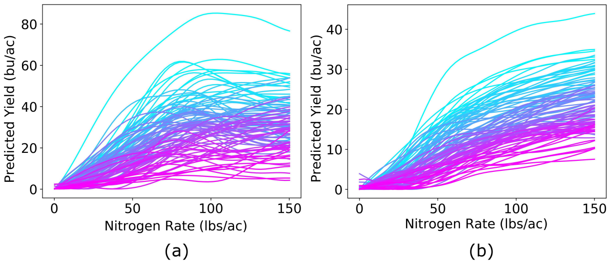


Fig. 5. N-response curve generation using 100 random samples from the two yield prediction datasets. **(a)** Field A. **(b)** Field B.

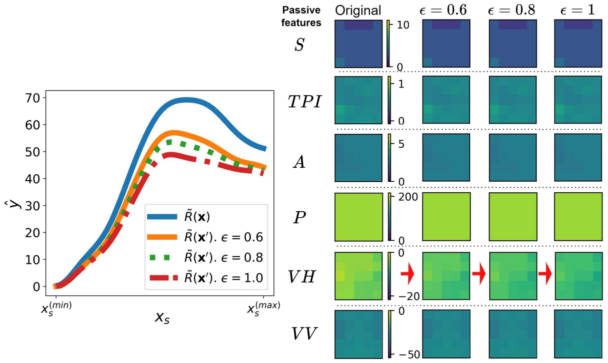
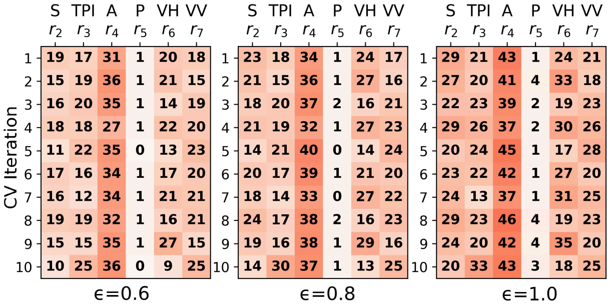


Fig. 6. Example of the counterfactual N-response curves generated using $\epsilon = 0.6, 0.8, \text{and}\, 1.0$ for a sample of field A.



	ϵ=0.6						ϵ=0.8						ϵ=1.0					
	S	TPI	A	P	VH	VV	S	TPI	A	P	VH	VV	S	TPI	A	P	VH	VV
CV Iteration	r_2	r_3	r_4	r_5	r_6	r_7	r_2	r_3	r_4	r_5	r_6	r_7	r_2	r_3	r_4	r_5	r_6	r_7
1	19	17	31	1	20	18	23	18	34	1	24	17	29	21	43	1	24	21
2	15	19	36	1	21	15	21	15	36	1	27	16	27	20	41	4	33	18
3	16	20	35	1	14	19	18	20	37	2	16	21	22	23	39	2	19	23
4	18	18	27	1	22	20	21	19	32	1	27	23	29	26	37	2	30	26
5	11	22	35	0	13	23	14	21	40	0	14	24	20	24	45	1	17	28
6	17	16	34	1	17	20	20	17	39	1	21	20	23	22	42	1	27	20
7	16	12	34	1	21	21	18	14	33	0	27	22	24	13	37	1	31	25
8	19	19	32	1	16	21	24	17	38	2	16	23	29	23	46	4	19	23
9	15	15	35	1	27	15	19	16	38	1	29	16	24	20	42	4	35	20
10	10	25	36	0	9	25	14	30	37	1	13	25	20	33	43	3	18	25

Fig. 7. Rounded individual relevance of passive features (in %) of field A.

After CV, the average validation MSE and standard deviation were 147.25±8.17 and 50.02±4.29 for fields A and B, respectively. Fig. 5.a and Fig. 5.b show the aligned approximated response curves generated for 100 random samples of field A and B, respectively, using the Hyper3DNetReg network. For these datasets, we experimented with $\epsilon \in \{0.4, 0.6, 0.8\}$. Fig.6 shows an example of the response curves generated for the first sample of field A ($j = 1$) using the network from the first CV iteration. Here, the local explanation is given by $\alpha_1 = \{6\}$, for the three selected thresholds. That is, we only needed to decrease the value of the sixth feature, VH, to alter the responsivity of the selected sample. Counterfactual response curves and local explanations generated for samples of field B were similar but were omitted due to space limitations.

Figs. 7 and 8 show the individual feature relevance values achieved for all CV iterations and ϵ values for fields A and B, respectively. Tables II and III show the mean percent of time

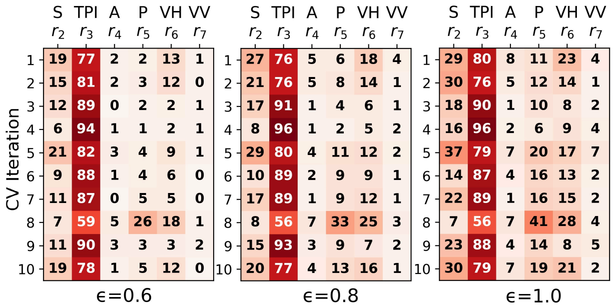


Fig. 8. Rounded individual relevance of passive features (in %) of field B.

TABLE II
TOP-FIVE FEATURE COMBINATIONS – FIELD A DATASET

ϵ	0.6		0.8		1.0	
#	Comb.	% Rep.	Comb.	% Rep.	Comb.	% Rep.
1	$[A, VV]$	0.6 ± 0.5	$[A, VV]$	1.8 ± 0.8	$[A, VV]$	4.3 ± 1.2
2	$[A, VH]$	0.6 ± 0.3	$[A, VH]$	1.8 ± 0.7	$[A, S]$	3.9 ± 1.1
3	$[TPI, VV]$	0.5 ± 0.3	$[S, VH]$	1.6 ± 0.5	$[A, VH]$	3.3 ± 0.4
4	$[TPI, VH]$	0.5 ± 0.3	$[A, TPI]$	1.6 ± 0.9	$[S, VH]$	3.2 ± 0.7
5	$[S, TPI]$	0.5 ± 0.3	$[TPI, VH]$	1.5 ± 0.7	$[A, TPI]$	3.0 ± 1.4

TABLE III
TOP-FIVE FEATURE COMBINATIONS – FIELD B DATASET

ϵ	0.6		0.8		1.0	
#	Comb.	% Rep.	Comb.	% Rep.	Comb.	% Rep.
1	$[TPI, S]$	4.4 ± 1.8	$[TPI, S]$	7.2 ± 2.7	$[TPI, S]$	9.7 ± 3.7
2	$[TPI, VV]$	2.1 ± 1.2	$[TPI, VV]$	5.0 ± 1.3	$[TPI, VV]$	6.5 ± 2.2
3	$[TPI, P]$	1.9 ± 1.5	$[TPI, P]$	3.9 ± 1.7	$[TPI, P]$	5.5 ± 2.0
4	$[TPI, A]$	1.0 ± 0.8	$[TPI, A]$	1.4 ± 0.5	$[S, P]$	2.3 ± 1.2
5	$[S, VV]$	0.6 ± 0.5	$[S, P]$	1.2 ± 0.8	$[TPI, A]$	1.7 ± 0.7

that a top-five feature combinations appeared across CV for fields A and B, respectively.

V. DISCUSSION

Although often unobservable, response curves of a system can be approximated using NNs. Fig. 2 suggests that the NN trained on the synthetic dataset learned to generate approximate response curves that were similar in shape to the ground-truth response curves. Even though ground-truth N-response curves are unobservable, we argue that the curves shown in Fig. 5 are sound. For example, previous works have described N-response curves as sigmoid-like curves [4], similar to most of the curves of Fig. 5.b. Other works have also considered quadratic functions [3] that account for the apparent decrease in yield response after reaching a certain saturation point, which can be seen in most of the curves of Fig. 5.a. Note that our analysis, as any other explainability method, allows for the explanation of the function approximated by the model. Specifically, we aim to identify the features (and feature combinations) that the model considers having the most responsivity impact.

The regression problem introduced in Eq. 11 consists of four passive features. x_2 alters the shape of the response curves by stretching them horizontally. x_3 multiplies the sigmoid function, as well as x_4; however, since it is squared, it is expected to have more impact on the shape of the response curves. In addition, a small change in x_3 produces a vertical stretching so that the resulting distance between the modified and original response curve is greater than the one produced by x_2 when modified by the same amount. From Fig. 4, we verified that x_3 has greater responsivity impact than x_4 and x_2, as it was more often modified by our CFE generation process. Finally, x_5 is independent of the rest of the features, which implies that it only shifts the response curves vertically but does not alter their shape. Again from Fig. 4, we verified that x_5 was assigned relevance values near 0%. Hence, our method found that the feature with the greatest responsivity impact is x_3, followed in relevance by x_2 and x_4. We conclude that the results obtained by our method coincide with those obtained by the analysis the equation used to generate the dataset. Also, Table I shows that the most effective feature combination is x_2 and x_3, followed by x_3 and x_4.

It is important to point out that we observed that the relevance ranking of features remained constant for different values of ϵ. For instance, Fig. 4 shows the ranking of features (from the most relevant to the least relevant) is: $[x_3, x_2, x_4, x_5]$ for the three tested ϵ values. Similarly, Fig. 7 shows that, for the three tested ϵ values in field A, A is the most relevant feature while the relevance scores of S, TPI, VH, and VV are comparable, and P is the least relevant feature. A similar behavior is observed in Fig. 8. This is meaningful because it suggests that the selection of ϵ is not crucial when finding the passive features with greater global relevance.

Note in Fig. 4 that the resulting relevance scores are similar for all CV iterations. This indicates that the NN model learns similar functions across different iterations. This is also the case for Fig. 7 and 8, although they show greater variation due to overfitting issues caused by limited data set sizes and lack of data variability in some of the training folds.

Furthermore, we found that the relevance values increase as ϵ increases. This is because, given a sample, if there is an increment in the desired distance threshold, it is likely that the number of modified passive features will grow. For example, Table I shows that the most repeated feature combination, (x_2, x_3), occurs 10.9% of the time when $\epsilon = 0.4$ but 36.6% when $\epsilon = 0.8$. This also happens for fields A and B, as seen in Tables II and III. However, the most repeated feature combination for field A, (A, VV), occurs only 0.6% of the time when $\epsilon = 0.6$ and 4.3% when $\epsilon = 1$. This means that changes in individual features have greater impact on the models trained for field A. For instance, in Fig. 6, it was enough to reduce the VH value to alter the shape of the response curve. This is because VH is related to the moisture content, meaning by lower VH values indicate less moisture in the soil. Thus it is reasonable to expect the soil to be less responsive to higher amounts of nitrogen.

From Figs. 7 and 8, we see that feature relevance values are different for both fields. The main reason is that field A is located on steep abrupt terrain while field B is not. As a consequence, the model trained for field A learned that the aspect A (i.e., the slope orientation) is the most relevant feature. Interestingly, in terrain with varying elevations located in the Northern Hemisphere, regions that are facing north and east have limited sunlight during the day and are more prone to snow retention. These are factors that may affect the responsiveness of the fertilizer. On the other hand, field B has an almost constant elevation, so A is not an important factor. The model trained for this field learned that the TPI, which is related to the ruggedness of the terrain, is the most relevant feature. The slope S, which influences fertilizer runoff, which in turn affects the responsiveness of the fertilizer, is the second most important factor for this field. Finally, P was assigned a low responsivity impact on both fields. This seems counterintuitive considering that precipitation is a critical factor for crop production. This would suggest that P is independent of the other features and, as in x_5 from the synthetic problem, only shifts the response curves vertically but does not affect their shape.

VI. Conclusion

The analysis of feature response curves often ignores other explanatory variables as a source of variability in the shape of the curves. Acknowledging all relevant variables, quantifying their responsivity impact, and understanding how they interact, may improve the accuracy of important applications such as drug dose optimization and N-fertilizer optimization.

We presented a method that generates approximate response curves for a selected active feature using neural networks. Our approach then estimates the impact that a set of passive features have on the shape of the response curves by generating counterfactual explanations. Experimental results on a synthetic dataset coincide with expectations following the analysis of the equation that was used to generate the training data. Experiments on two crop yield prediction datasets found that the factor with the greatest responsivity impact on N-response curves was the terrain aspect, for one of the studied winter wheat fields, and the topographic position index, for the other field. While perhaps already understood by farmers and agronomists, this analysis confirms that the models are reflecting what science would expect in crop production. Future work will focus on designing tests that will determine if the responsivity of two or more response curves is statistically significant. We also plan to use equation discovery approaches to approximate parametric equations for crop production.

Acknowledgements

The authors wish to thank the team members of the On-Field Precision Experiment (OFPE) project for their comments throughout the development of this work. We would also like to thank Jordan Schupbach for providing advice on the experimental design. This research was supported by a USDA-NIFA-AFRI Food Security Program Coordinated Agricultural Project (Accession Number 2016-68004-24769), and also by the USDA-NRCS Conservation Innovation Grant from the On-farm Trials Program (Award Number NR213A7500013G021).

References

[1] M. Fallahi-Sichani, S. Honarnejad, L. Heiser, J. Gray, and P. Sorger, "Metrics other than potency reveal systematic variation in responses to cancer drugs," *Nature Chemical Biology*, vol. 9, no. 11, pp. 708–714, 2013.

[2] G. Di Veroli, C. Fornari, I. Goldlust, G. Mills, S. Koh, J. Bramhall, F. Richards, and D. Jodrell, "An automated fitting procedure and software for dose-response curves with multiphasic features," *Scientific Reports*, vol. 5, p. 14701, 2015.

[3] D. G. Bullock and D. S. Bullock, "Quadratic and quadratic-plus-plateau models for predicting optimal nitrogen rate of corn: A comparison," *Agronomy Journal*, vol. 86, no. 1, pp. 191–195, 1994.

[4] K. B. Watkins, J. A. Hignight, R. J. Norman, T. L. Roberts, N. A. Slaton, C. E. Wilson, and D. L. Frizzell, "Comparison of economic optimum nitrogen rates for rice in Arkansas," *Agronomy Journal*, vol. 102, no. 4, pp. 1099–1108, 2010.

[5] D. Wang, J. Hensman, G. Kutkaite, T. S. Toh, A. Galhoz, J. R. Dry, J. Saez-Rodriguez, M. J. Garnett, M. P. Menden, F. Dondelinger *et al.*, "A statistical framework for assessing pharmacological responses and biomarkers using uncertainty estimates," *Elife*, vol. 9, p. e60352, 2020.

[6] B. Maxwell, P. Hegedus, P. Davis, A. Bekkerman, R. Payn, J. Sheppard, N. Silverman, and C. Izurieta, "Can optimization associated with on-farm experimentation using site-specific technologies improve producer management decisions?" in *14th International Conference on Precision Agriculture*, 2018.

[7] G. Morales, J. Sheppard, A. Peerlinck, P. Hegedus, and B. Maxwell, "Generation of site-specific nitrogen response curves for winter wheat using deep learning," in *15th International Conference on Precision Agriculture*, 2022.

[8] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, 2021.

[9] N. Wiratunga, A. Wijekoon, I. Nkisi-Orji, K. Martin, C. Palihawadana, and D. Corsar, "DisCERN: Discovering counterfactual explanations using relevance features from neighbourhoods," in *IEEE 33rd International Conference on Tools with Artificial Intelligence*, 2021, pp. 1466–1473.

[10] S. Barocas, A. D. Selbst, and M. Raghavan, "The hidden assumptions behind counterfactual explanations and principal reasons," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2020, p. 80–89.

[11] S. Dandl, C. Molnar, M. Binder, and B. Bischl, "Multi-objective counterfactual explanations," in *Parallel Problem Solving from Nature – PPSN XVI*, T. Bäck, M. Preuss, A. Deutz, H. Wang, C. Doerr, M. Emmerich, and H. Trautmann, Eds., 2020, pp. 448–469.

[12] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "Face: Feasible and actionable counterfactual explanations," in *Proceedings of the Conference on AI, Ethics, and Society*, 2020, p. 344–350.

[13] A. Van Looveren and J. Klaise, "Interpretable counterfactual explanations guided by prototypes," in *Machine Learning and Knowledge Discovery in Databases. Research Track*, N. Oliver, F. Pérez-Cruz, S. Kramer, J. Read, and J. A. Lozano, Eds. Springer International Publishing, 2021, pp. 650–665.

[14] R. R. Fernández, I. Martín de Diego, V. Aceña, A. Fernández-Isabel, and J. M. Moguerza, "Random forest explainability using counterfactual sets," *Information Fusion*, vol. 63, pp. 196–207, 2020.

[15] T. Spooner, D. Dervovic, J. Long, J. Shepard, J. Chen, and D. Magazzeni, "Counterfactual explanations for arbitrary regression models," 2021. [Online]. Available: https://arxiv.org/abs/2106.15212

[16] R. Kommiya Mothilal, D. Mahajan, C. Tan, and A. Sharma, "Towards unifying feature attribution and counterfactual explanations: Different means to the same end," in *Proceedings of the Conference on AI, Ethics, and Society*, 2021, p. 652–663.

[17] P. Schwab, L. Linhardt, S. Bauer, J. M. Buhmann, and W. Karlen, "Learning counterfactual representations for estimating individual dose-response curves," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 5612–5619, Apr. 2020.

[18] J. Ramsay and B. Silverman, *Functional Data Analysis*. Springer, 2005.

[19] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[20] H. Ishibuchi, R. Imada, Y. Setoguchi, and Y. Nojima, "Performance comparison of NSGA-II and NSGA-III on various many-objective test problems," in *IEEE Congress on Evolutionary Computation (CEC)*, 2016, p. 3045–3052.

[21] G. Morales, J. W. Sheppard, P. B. Hegedus, and B. D. Maxwell, "Improved yield prediction of winter wheat using a novel two-dimensional deep regression neural network trained via remote sensing," *Sensors*, vol. 23, no. 1, p. 489, jan 2023.