DiPACE: Diverse, Plausible and Actionable Counterfactual Explanations

Anonymous submission

Abstract

As Artificial Intelligence (AI) is integrated further into critical aspects of our society, the need to understand and interpret the decision making of these models is critical. Counterfactual Explanations (CFX) are a popular method of achieving this understanding through "what if?" scenarios, showing how an instance should be modified in order to reach a more desirable outcome, providing cause-and-effect interpretation. The literature agrees on several desirable qualities of CFX; however, existing methods do not adequately strike a balance between these, but rather focus on optimizing just one or two aspects. While these methods can be useful for specific applications, we argue that a more well-rounded solution is necessary for general use, that can be fine-tuned by the user to emphasize the more important elements for a given application. Further, we introduce an enhanced version of our method, which applies additional penalty to each term depending on its adherence to user-defined threshold values, encouraging results that better meet their requirements. We demonstrate through experimentation with real-world datasets that our proposed framework provides the best balance of four key features of CFX: diversity, plausibility, proximity, and sparsity.

Code — available in the supplementary material, a github link will be provided on acceptance.

Introduction

In recent years, the importance of interpretability in artificial intelligence (AI) has grown, especially in critical areas such as healthcare (Yagin et al. 2023; Shin et al. 2023), finance (Babaei, Giudici, and Raffinetti 2023; El Qadi et al. 2023; Zhu et al. 2023), and disaster relief (Prasanth Kadiyala and Woo 2021; Sanderson et al. 2023a,b, 2024). Explainable AI (XAI) seeks to provide explanation of AI model behaviour, enhancing human understanding and trust in their implementation (Mirzaei et al. 2023). Counterfactual explanations (CFX) provide insight into how modifications to the input features can reverse the decision of an AI model, providing cause and effect understanding, as well as actionable interventions for undesirable outcome (Jiang et al. 2024). Effective CFX should be diverse, plausible, and feasible, to ensure the suggested interventions can be utilised (Guidotti 2022). Diverse CFX provides a range of potential changes, offering the user different options for achieving the desired outcome (Mothilal, Sharma, and Tan 2020). Plausibility ensures that the CFs are not just theoretically possible, but can also be achieved within real-world applications (Del Ser et al. 2024). Proximity ensures small changes are made to features (Wachter, Mittelstadt, and Russell 2018), while sparsity ensures that only a small number of changes are required (Tsiourvas, Sun, and Perakis 2024).

Wachter, Mittelstadt, and Russell (2018) introduced the notion of CFs as the closest instance in which the classification differs from the original instance. As such, this work focuses primarily on proximity, without addressing other potential real-world user needs. DiCE (Mothilal, Sharma, and Tan 2020) enhances the diversity of the generated CFs by returning a set of varied CF instances, as well as enabling user constraints to provide greater control of the results, ensuring their feasibility. DECE (Cheng, Ming, and Qu 2021) follows a similar approach to DiCE, but further ensures sparsity by restricting feature changes to the top k most influential features. Tsiourvas, Sun, and Perakis (2024) introduce a method encouraging plausibility through optimization of the local outlier factor, ensuring CFs align with the data manifold, addressing the need for changes that are realistic within the context of the data. These methods all make use of gradient-based optimization, which grants the explainer access to gradient information from the underlying model, enabling precise CFX results. In order for gradient descent optimization to take place, gradient information is required, meaning that the underlying model has to be differentiable. Gradient descent also often gets stuck within a local optima. To overcome these challenges, researchers have looked to alternative optimization stratgies, in particular genetic optimizers (Schleich et al. 2021), shortest path algorithms (Poyiadzi et al. 2020; Barzekar and McRoy 2023), and mixed integer programming (Tsiourvas, Sun, and Perakis 2024; Carrizosa, Ramírez-Ayerbe, and Romero Morales 2024; Kanamori et al. 2020; Russell 2019). Multi-Objective optimizers have also been implemented to improve the balancing of multiple competeing objectives (Dandl et al. 2020; Rasouli and Yu 2021; Rasouli and Chieh Yu 2024). While these methods have advanced the utility of CFX, their focus on optimizing just one or two characteristics is often at the expense of others. There remains the need for more holistic solutions that consider feasibility, diversity, and plausibility, to ensure high quality, actionable solutions. The use of

genetic optimization is beneficial for model-agnosticity and escaping local optima, however, this reduces the precision of the results by denying access to the inner working of the underlying model.

In this work, we aim to address these challenges by introducing DiPACE and DiPACE+, frameworks for generating diverse, plausible and actionable CFXs, enhacing the practical utility of CFX in real-world applications. Our study makes the following contributions:

(i) A novel loss function is proposed that explicitly optimizes CFs for diversity, plausibility, proximity, and sparsity, and enable users to fine tune the characteristics to meet diverse user requirements.

(ii) Perturbations are integrated into the optimization strategy to leverage gradient information, overcoming the challenge of gradient descent getting stuck in local optima.

(iii) Users are enabled to define the mutable and immutable features, acceptable ranges for continuous features, categories for categorical features, and directions for feature changes, ensuring feasibility.

(iv) We introduce DiPACE+, including additional penalties in the loss function for unacceptable characteristic values to encourage more aggressive optimization.

Methodology

DiPACE Framework

The goal of any CFX engine is to return a set of CF instances given a trained model and query instance, where the predicted output differs to that of the query instance. DiPACE aims to ensure these CF instances are diverse, plausible, and actionable. DiPACE uses gradient-descent optimization, requiring a differentiable model. For a model-agnostic solution, we propose training a surrogate differentiable model to approximate the behaviour of any underlying model. To overcome the challenge of local minima, we incorporate perturbation in the optimization strategy. We also present DiPACE+, including an additional penalty in the loss funciton for more aggressive optimization. Our framework is described in Algorithm 1.

Loss Function

The DiPACE loss function is designed to balance diversity, plausibility, proximity, and sparsity, ensuring feasible and realistic CFX. Recognising that users may prioritize certain characteristics, we enable weighting λ of each characteristic, enabling users to tune the results. Our loss function is expressed by Equation 1.

$$L = L_{pred}(f(c_i), y) + \lambda_1 \cdot L_{di}(C) + \lambda_2 \cdot L_{pl}(C, X) + \lambda_3 \cdot L_{pr}(C, x) + \lambda_4 \cdot L_{sp}(C, x) + L_{cat}(C) \quad (1)$$

Where C is a set of CF instances, c is a counterfactual instance, X is the observed dataset and x is the query instance.

Algorithm 1: DiPACE and DiPACE+

Input: Query Instance x, Model f, Hyperparameters θ
Parameter: Learning rate α, Weights λ, Thresholds τ, Scale Factors γ, Maximum Perturbation Attempts δ
Output: Counterfactual Set C

Let $t = 0$, $p = 0$, $C \leftarrow initialize\_CF(x)$.
while not converged **do**
  Compute loss components L_{pred}, L_{di}, L_{pl}, L_{pr}, L_{sp}, L_{cat}.
  if DiPACE+ **then**
    if $L_{di} < \tau_{di}$ **then**
      $L_{div} \leftarrow L_{div}(1 - \gamma_{pen})$
    end if
    if $L_{pl} > \tau_{pl}$ **then**
      $L_{pl} \leftarrow L_{pl}(1 + \gamma_{pen})$
    end if
    if $L_{pr} > \tau_{pr}$ **then**
      $L_{pr} \leftarrow L_{pr}(1 + \gamma_{pen})$
    end if
    if $L_{sp} > \tau_{sp}$ **then**
      $L_{sp} \leftarrow L_{sp}(1 + \gamma_{pen})$
    end if
  end if
  Compute total loss $L \leftarrow L_{pred} - \lambda_1 \cdot L_{di} + \lambda_2 \cdot L_{pl} + \lambda_3 \cdot L_{pr} + \lambda_4 \cdot L_{sp} + L_{cat}$.
  Compute gradients of L w.r.t C.
  Update C with gradient descent with learning rate α.
  Apply user defined constraints.
  if Convergence criteria met **then**
    break
  end if
  Increment t.
end while
if $L > \tau_{pert}$ **then**
  while $p < \delta$ **do**
    Perturb C by adding $N(0, \gamma_{pert})$
    Initialize $t = 0$
    Repeat steps 2 - 24.
    if $L \leq \tau_{pert}$ **then**
      return C
    end if
    Increment p.
  end while
end if
return C with lowest L.

In DiPACE+, an threshold τ is required for each characteristic. If the computed value exceeds τ, an additional penalty is applied, determined by γ, computed as in equation 2.

$$L_t = \begin{cases} L_t, & \text{if } L_t \geq \tau. \\ L_t\,(1 + \gamma)\,, & \text{otherwise.} \end{cases} \quad (2)$$

Diversity Diversity measures the average pairwise distance between each of the CF instances in the set, encouraging a broad range of possible outcomes, increasing the likelihood of an actionable result. We compute the diversity as the negative of the determinantal point process (DPP) of

a matrix of pairwise distances D_{ij} between CF instances in C.

$$L_{di} = -dpp\left(\frac{1}{1+\sum_{l=1}^{n}|c_{il}-c_{jl}|\cdot w_j}\right) \tag{3}$$

Where n is the number of features, and w is a vector assigning weight to each feature based on the inverse mean absolute deviation (MAD).

Plausibility Plausibility measures the average distance between the CF instance and the nearest k observed instances in X. This encourages CFs that are similar to the observed instances, ensuring that they are realistic.

$$L_{pl} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{k}\sum_{j=1}^{k}|c_i - x_{j_i}|\right) \tag{4}$$

Where m is the number of CFs in the set and x_{j_i} is an instance of X.

Proximity Proximity measures the distance between the query instance and a set of CF instances $(c_i, \ldots, c_m) \in C$. We calculate it as the mean element-wise absolute differences between the features of query instance x and CF instances C.

$$L_{pr} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}|c_{ij} - x_j|\cdot w_j \tag{5}$$

Sparsity Sparsity measures how many features change between the query instance and CF instances, encouraging CFs that differ in as few features as possible from the query instance.

$$L_{sp} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{I}(c_{ij} \neq x_j) \tag{6}$$

Handling Categorical Variables Categorical variables present a unique challenge in CFX due to their discreet nature. To maintain the integrity of the one hot encoding, we enforce a linear equality constraint in the loss function, ensuring that all levels of a categorical variable sum to 1. This constraint iterates over each categorical variable and computes the squared deviation of the sum of the returned probabilities for each CF instance's categories from 1.

$$L_{cat} = \sum_{v\in cat}\sum_{i=1}^{m}\left(\left(\sum_{j=v[0]}^{v[-1]} c_{ij}\right) - 1\right)^2 \tag{7}$$

Where $v \in cat$ is the range of indices for each categorical feature.

Optimization

We minimize our loss function with the Adam gradient descent-based optimizer. Gradient descent algorithms efficiently navigates complex loss landscapes and high-dimensional datasets, while allowing precise control of the optimization process, and user constraints. To escape from local optima, we introduce perturbation into the optimization process. If the overall loss exceeds a user-defined threshold, we perturb the CF by adding random noise sampled from a normal distribution to each feature that significantly deviates from the original instance, scaled by a user-defined scale factor. Optimization stops when the overall loss reaches the threshold, or a maximum number of perturbation attempts have been made without reduction of the loss.

User Constraints

Optimizing proximity, sparsity, plausibility, and diversity help to ensure a realistic and feasible CF, but other real-world consideration may need to be considered depending on the context, so need to be specified by the user. In the DiPACE framework, users can specify:

- Features to vary: list of features that can be varied, useful where only a select few features can feasibly be changed. Default: all features.
- Immutable features: a list of features that cannot be changed, useful where most features can be changed, but a select few cannot. Default: none.
- Feature ranges: a dictionary specifying acceptable ranges for certain continuous features. If unspecified, ranges are restricted by the minimum and maximum values in the dataset.
- Feature categories: a dictionary specifying acceptable categories for categorical features. If unspecified, features are restricted to the categories in the dataset.
- Feature directions: a dictionary specifying if a feature can only increase or decrease. If unspecified, features can change in either direction.

Evaluation

Datasets To evaluate our framework we use two datasets from the UCI Machine Learning Repository. The selected datasets are from diverse real-world problems, in which intervention could realistically be taken to have change the outcome, to properly demonstrate the potential impact of the work. Firstly, we use the heart disease dataset, which includes 13 features. 5 of the features are continuous: Age (F1), Trestbps (F2), Chol (F3), Thalach (F4), and Oldpeak (F5); and 8 are categorical: Sex (F6), Cp (F7), Restecg (F8), Exang (F9), Slope (F10), Ca (F11), and Thal (F12). We also use the credit approval dataset, which contains 14 features. 4 of the features are continuous: Age (F1), Debt (F2), Years Employed (F3), and Income (F4); and 10 are categorical: Gender (F5), Married (F6), Bank Customer (F7), Industry (F8), Ethnicity (F9), Prior Default (F10), Employed (F11), Credit Score (F12), Drivers License (F13), and Citizen (F14). Both datasets have a binary target variable (T).

Experimental Setup

To evaluate the performance of DiPACE+ we perform several experiments, using the same predictive model architecture, a multi-layer perceptron (MLP) with an input layer, a 64 neuron hidden layer, a 32 neuron hidden layer, and

an output layer with sigmoid activation. We set consistent values for λ for each term to ensure uniform comparison and tuneds these with a grid search of different values. For the heart disease dataset the optimal λ values are 0.5, and 0.3 for credit approval. For the heart disease dataset, we found γ_{pen} of 0.1, and τ_{pen} values of 0.5 for proximity, 0.4 for sparsity, 0.9 for diversity, and 1.5 for plausibility to encourage optimal results, with γ_{pert} of 0.7 and τ_{pert} of 1.0. For the credit approval dataset, we found τ_{pen} values of 0.2 for proximity, 0.2 for sparsity, 0.9 for diversity and 1.0 for plausibility, and 0.1 for γ_{pen} to be the most optimal values. A τ_{pert} value of 0.6 was used, and γ_{pert} value of 0.5.
In the final experiment we compare DiPACE+ with DiPACE and 3 state-of-the-art algorithms, Wachter (Wachter, Mittelstadt, and Russell 2018), DiCE (Mothilal, Sharma, and Tan 2020) and CARE (Rasouli and Chieh Yu 2024). Wachter is selected as a baseline, as it is the seminal work in CF generation, DiCE for its widespread use in CFX applications, and CARE as it represents a recent advancement in CFX. Each of these algorithms are implemented using their respective previously published python libraries.

Results and Analysis

Loss Function Ablation Study

To demonstrate that the proposed loss function best encourages a balanced CF result, we perform an ablation study of the four key characteristics, the results for which are shown in Table 1. We keep proximity in each combination, as its inclusion is fundamental to any CF generation algorithm. The combinations we consider are: (1) Proximity, (2) Proximity and Diversity, (3) Proximity and Plausibility, (4) Proximity and Sparsity, (5) Proximity, Diversity, and Plausibility, (6) Proximity, Diversity, and Sparsity, (7) Proximity, Plausibility, and Sparsity, and (8) Proximity, Diversity, Plausibility, and Sparsity.

	Div.	Plaus.	Prox.	Spars.	Conf.
Heart Disease					
1	0.74	4.33	0.14	0.30	0.71
2	0.92	5.67	0.19	0.31	0.50
3	0.88	2.21	0.41	0.38	0.61
4	0.15	2.76	0.09	0.24	0.74
5	0.68	7.01	0.56	0.57	0.59
6	0.92	5.20	0.48	0.53	0.67
7	0.89	6.68	0.16	0.27	0.56
8	0.96	1.38	0.49	0.36	0.73
Credit Approval					
1	0.01	2.31	0.04	0.09	0.60
2	0.94	6.79	0.18	0.21	0.66
3	0.92	1.51	0.21	0.23	0.72
4	0.79	4.29	0.10	0.15	0.58
5	0.94	0.79	0.28	0.30	0.71
6	0.93	3.97	0.11	0.14	0.72
7	0.92	1.75	0.19	0.21	0.76
8	0.92	0.92	0.18	0.20	0.78

Table 1: Ablation Study of Loss Function Terms for Each Dataset.

It can be observed here that the inclusion of all four characteristics yields the most balanced CF sets. For the heart disease dataset, a high diversity of 0.96, low plausibility of 1.38, reasonable proximity and sparsity scores of 0.49 and 0.36 respectively, and a high confidence score of 0.73 are achieved. Similarly, with the credit approval dataset a high diversity of 0.92, low plausibility of 0.92, good proximity and sparsity scores of 0.18 and 0.20, and the highest confidence score of 0.78 are achieved. Excluding characteristics leads to improved results for those remaining, highlighting the flexibility offered by DiPACE. A strong correlation between proximity and sparsity can be observed. When sparsity is included, proximity tends to improve. This becomes more complex where diversity is included, however, as the goal of diversity conflicts with that of both proximity and sparsity. The plausibility, proximity, and sparsity values are generally lower and less varied with the credit approval dataset in comparison to the heart disease dataset. This may be attributed to the greater proportion of continuous features in the heart disease dataset, which are more likely to change at a higher magnitude, highlighting the importance of control of the τ and δ values for different applications.

Optimization Strategy

Figure 1 shows the loss curves with the heart disease dataset, and Figure 2 with the credit approval dataset through optimization. These reveal that following each perturbation, the loss value for each component is further reduced, indicating the efficacy of this approach in escaping local optima.

The proximity, sparsity, plausibility, diversity, and confidence values for the CFs returned both with and without perturbations are presented in Table 2.

	Div.	Plaus.	Prox.	Spars.	Conf.
Heart Disease					
W/ Perturb.	0.96	1.38	0.49	0.36	0.73
W/o Perturb.	0.91	2.18	0.54	0.54	0.59
Credit Approval					
W/ Perturb.	0.92	0.92	0.18	0.20	0.78
W/o Perturb.	0.88	2.17	0.25	0.24	0.70

Table 2: Results with and without Perturbation in the Optimization Strategy for Each Dataset.

This further highlights the impact of including perturbation in the gradient descent process for reaching a more globally optimal solution, as it is able to achieve superior values for each metric. Existing methods of CF generation either rely on gradient-based optimization, without accounting for local optima, or use an alternative method, sacrificing the access to the inner working of the underlying model. The inclusion of perturbation also adds an additional element of stochasticity into the framework, further diversifying the results.

Qualitative Analysis

An example CF set generated by DiPACE+ is presented in Table 3 for heart disease prediction and Table 4 for credit

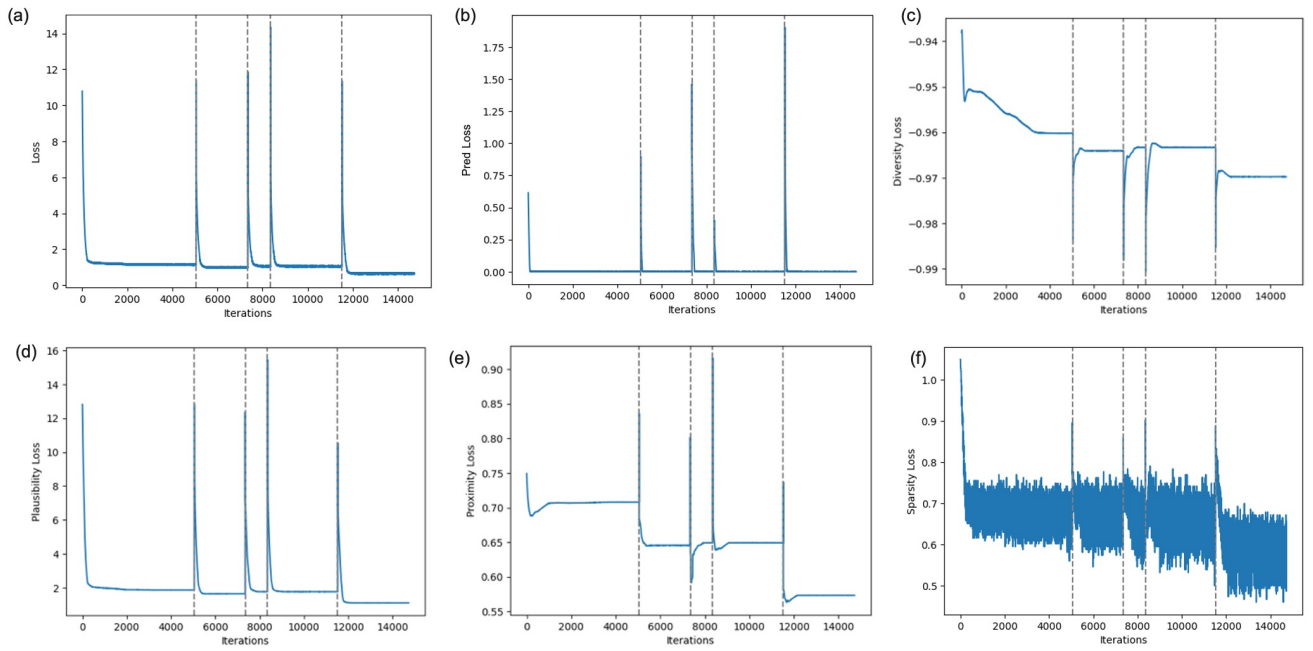

Figure 1: Loss Curves for (a) overall loss, (b) prediction loss, (c) diversity loss, (d) plausibility loss, (e) proximity loss, and (f) sparsity loss with Heart Disease Data. The vertical dashed lines represent the point of perturbation.

approval prediction.

	Query	CF Values				
F1	52	44	47	52	50	47
F2	172	129	138	130	129	130
F3	199	196	199	199	176	180
F4	162	150	156	152	144	150
F5	0.5	0.3	0.1	0.5	0.5	0
F6	1	1	1	1	1	1
F7	2	2	2	2	2	2
F8	1	0	0	1	0	0
F9	1	1	0	0	1	0
F10	0	0	0	0	0	0
F11	2	2	2	2	2	2
F12	0	0	0	0	0	0
F13	3	2	2	2	2	2
T	1	0	0	0	0	0

Table 3: Example CF Set for Heart Disease.

These results show that DiPACE captures intuitive feature changes necessary for prediction to be reversed. In Table 3, blood pressure and maximum heart rate are two key identifiers of heart health, and in both cases the value was reduced, indicating a healthier heart. Cholesterol and fasting blood sugar also have impact on cardiovascular health, and in most cases these values were reduced. Older age is commonly associated with higher risk of heart disease, and it is consistently lowered. This indicates that these features either have little impact on the outcome, or were already favourable.

In Table 4, Age, Years Employed and Credit Score generally increased and Employment Status changed from unemployed to employed. These changes align with a more reliable customer, making credit approval more likely. Interestingly, Debt is consistently higher in the CF instances than the original, and income does not in general increase. Prior Default also consistently changes from negative to positive. These results are not intuitive, as they would generally reflect a weaker client profile. This suggests that the model may be capturing some intricate feature dependencies where it may be acceptable for a client to have lower income, or more debt, where other factors are more favourable.

	Query	CF Values				
F1	34.1	35.5	34.8	44	34.6	39.7
F2	2.8	4	10.1	4	4	8.12
F3	2.5	7.4	3.3	3	6	2
F4	200	198	210.7	188.2	194.9	193.2
F5	1	1	1	0	1	1
F6	1	1	1	1	1	1
F7	1	1	1	1	1	1
F8	1	13	0	0	13	0
F9	0	1	0	0	0	0
F10	0	1	1	1	1	1
F11	0	1	1	0	1	1
F12	0	6	6	0	8	1
F13	1	1	0	1	0	1
F14	0	0	0	0	0	0
T	0	1	1	1	1	1

Table 4: Example CF Set for Credit Approval.

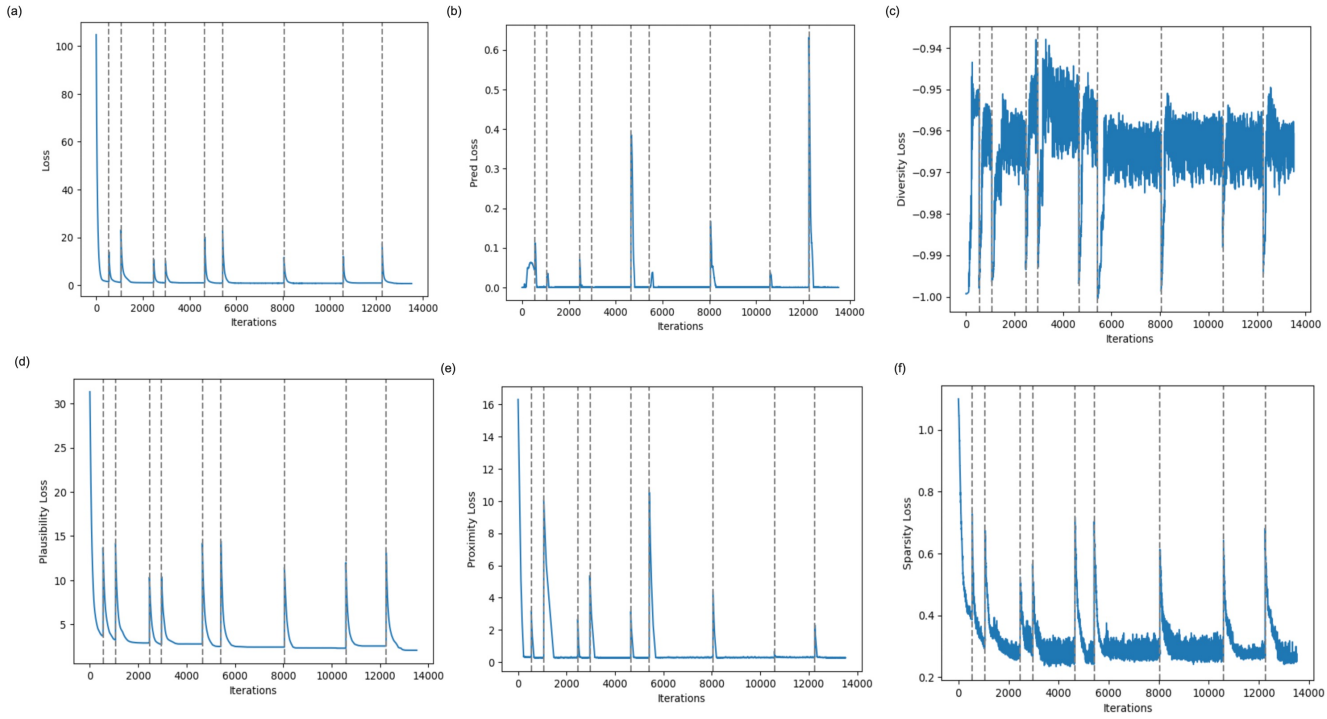

Figure 2: Loss Curves for (a) overall loss, (b) prediction loss, (c) diversity loss, (d) plausibility loss, (e) proximity loss, and (f) sparsity loss with Credit Approval Data. The vertical dashed lines represent the point of perturbation.

User Constraints

In the heart disease dataset, sex chest pain type and presence of exercise induced angina are immutable, as these are part of the patient's medical history or inherent to the patient. The age of the patient can increase, but cannot decrease. The other features are all mutable through lifestyle change. So, we set sex, cp and exang to immutable, and age to increase only. Maximum heart rate is calculated as $200 - age$, so we set the maximum value for thalach to 168 $(200-52)$, and the minimum to 94, which is the minimum value for this feature in the observed data.

In the credit approval dataset, gender, ethnicity and citizenship are inherent to the individual and cannot be change. Prior default is based on the history of the customer, so cannot be changed. Age and years employed can both only be increased. Industry can only be realistically changed to values in which the customer is qualified to work. We set gender, ethnicity, and citizen to immutable and age and years employed to increase only. The original value for industry is consumer discretionary (1), so we assume hypothetically that the client is qualified to work in consumer staples (2), industrials (7), real estate (10) and transport (12), as these have overlapping job opportunities.

Table 5 shows the quantitative performance where the constraints are applied against the unconstrained CF generation for both datasets. Tables 6 and 7 show example CF sets from each dataset, where the described constraints are applied.

It is evident that applying constraints improves the proximity and sparsity, but negatively impacts the diversity,

	Div.	Plaus.	Prox.	Spars.	Conf.
Heart Disease					
Constrained	0.60	1.85	0.19	0.27	0.61
Unconstrained	0.96	1.38	0.49	0.36	0.73
Credit Approval					
Constrained	0.79	1.98	0.12	0.17	0.68
Unconstrained	0.92	0.92	0.18	0.20	0.78

Table 5: Quantitative Results of Applying User Constraints with Heart Disease Data.

	Query	CF Values				
F1	52	56	67	64	60	54
F2	172	139	140	152	128	120
F3	199	199	199	198	199	199
F4	162	159	150	151	149	148
F5	0.5	0.4	0.5	0.5	0.4	0.5
F6	1	1	1	1	1	1
F7	2	2	2	2	2	2
F8	1	0	0	0	0	0
F9	1	0	0	0	0	0
F10	0	0	0	0	0	0
F11	2	1	1	1	1	1
F12	0	0	3	0	0	0
F13	3	3	3	3	3	3
T	1	0	0	0	0	0

Table 6: Example CF Set for Heart Disease with User Constraints.

	Query	CF Values				
F1	34.1	34.2	43.2	39.1	36.2	36.3
F2	2.8	4	1.8	6	5	4.3
F3	2.5	5	3.5	2.5	3.3	3.3
F4	200	196.8	197.2	196.7	179.3	215.4
F5	1	1	1	1	1	1
F6	1	1	1	1	1	1
F7	1	1	1	1	1	1
F9	1	12	1	2	1	2
F10	0	0	0	0	0	0
F11	0	0	0	0	0	0
F12	0	0	0	1	1	1
F13	0	0	0	12	0	0
F14	1	1	0	1	0	0
F15	0	0	0	0	0	0
T	0	1	1	1	1	1

Table 7: Example CF Set for Credit Approval with User Constraints.

	Div.	Plaus.	Prox.	Spars.	Conf.
Heart Disease					
DiPACE+	0.96	1.38	0.49	0.36	0.73
DiPACE	0.81	2.09	0.52	0.53	0.80
Wachter	0.31	16.05	0.02	0.17	0.75
DiCE	0.82	17.93	0.16	0.25	0.88
CARE	0.77	14.53	0.44	0.48	0.89
Credit Approval					
DiPACE+	0.92	0.92	0.18	0.20	0.78
DiPACE	0.87	1.55	0.30	0.31	0.82
Wachter	0.35	3.99	0.03	0.10	0.56
DiCE	0.84	3.81	0.12	0.18	0.64
CARE	0.68	2.91	0.18	0.56	0.66

Table 8: Comparison of DiPACE+ and DiPACE with Previous Work with Heart Disease Data.

plausibility and confidence of the generated CFs. Forcing values to stay fixed means that fewer features will change, therefore values also are closer to the query instance. Fixing certain values means that there is less variation between the instances in each set, and it is more challenging for the CFs to meet a realistic distribution, and to be confidently classified as the opposite class. This does, however, show that a set of CF instances can still be found while being constrained in various ways by the user. This demonstrates the practical utility of DiPACE+ for real-world application, where domain-specific practical or ethical considerations need to be taken into account.

Comparison of Algorithms

To further demonstrate the quality of DiPACE+ we benchmark its performance against DiPACE, and existing CFX algorithms. The comparative results are shown in Table 8.

Across both datasets, DiPACE+ performs best in diversity and plausibility, and achieves reasonable values for proximity and sparsity. In contrast, Wachter obtains the weakest diversity score, but the best proximity and sparsity scores. This is expected, as proximity is the only characteristic considered in the Wachter algorithm, so it is optimized more aggressively, while a diverse set of counterfactuals is not its focus. DiCE and CARE both achieve moderate diversity and proximity scores, both performing more poorly than DiPACE+ in diversity. DiCE performs better in proximity, while CARE is comparable to DiPACE+. Notably, all three existing algorithms achieve significantly weaker plausibility in comparison to both DiPACE and DiPACE+. This means that their results are less similar to the observed instances and, consequently, less realistic. This highlights the importance of explicitly optimizing for plausibility in CF generation if realistic CF instances are desired. Against the four key characteristics, DiPACE+ consistently outperform DiPACE, demonstrating the benefit of the additional penalty term. This does, however, come with a trade-off in the confidence of the predicted outcome. Overall, DiPACE+ achieves the most balanced performance across the metrics, demonstrating its ability to produce realistic and diverse sets of CFs while maintaining a good level of feasibility and prediction confidence.

Conclusions

In this study we have presented novel frameworks for generating diverse, plausible, and actionable CF sets, contributing towards the landscape of CFX and more broadly XAI development. Our approach integrates diversity, plausibility, proximity, and sparsity into its loss function, and utilises a novel optimization strategy, leveraging gradient-descent, with perturbations to escape local optima and further enhance diversity. Through experimentation, this work has empirically demonstrated the quality of our frameworks in their ability to balance these characteristics, while remaining flexible to diverse user requirements.

Our experimentation was conducted on two datasets, where it was observed that proximity and sparsity were easier to optimize on the credit approval dataset. Given the differing structures of the two datasets, we infer that a greater proportion of categorical features is responsible for this, however more rigorous experimentation on diverse data structures is needed to more comprehensively understand the impact. Additionally, evaluation is a well-known challenge in XAI, so consideration of a more comprehensive evaluation metric for feasibility that is not impacted in this way could be beneficial in future work.

While our perturbation-based optimization strategy effectively enables the escape of local optima, this method can significantly impact convergence time. Further refinement of the optimization strategy to improve the computational efficiency is an important direction for future work.

Finally, our work focuses mainly on the benefit of DiPACE for stakeholders looking for real-world intervention strategies. In the future, further consideration of its benefit to a wider range of stakeholders should be made. In particular, DiPACE can provide insights into model behaviour and potential biases, which can aid data scientists and machine learning engineers in refining their models.

References

Babaei, G.; Giudici, P.; and Raffinetti, E. 2023. Explainable fintech lending. *Journal of Economics and Business*, 125: 106126.

Barzekar, H.; and McRoy, S. 2023. Achievable Minimally-Contrastive Counterfactual Explanations. *Machine Learning and Knowledge Extraction*, 5(3): 922–936.

Carrizosa, E.; Ramírez-Ayerbe, J.; and Romero Morales, D. 2024. Generating collective counterfactual explanations in score-based classification via mathematical optimization. *Expert Systems with Applications*, 238: 121954.

Cheng, F.; Ming, Y.; and Qu, H. 2021. DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2): 1438–1447.

Dandl, S.; Molnar, C.; Binder, M.; and Bischl, B. 2020. Multi-objective counterfactual explanations. In *Proceedings of the International Conference on Parallel Problem Solving from Nature*, 448–469. Springer.

Del Ser, J.; Barredo-Arrieta, A.; Díaz-Rodríguez, N.; Herrera, F.; Saranti, A.; and Holzinger, A. 2024. On generating trustworthy counterfactual explanations. *Information Sciences*, 655: 119898.

El Qadi, A.; Trocan, M.; Diaz-Rodriguez, N.; and Frossard, T. 2023. Feature contribution alignment with expert knowledge for artificial intelligence credit scoring. *Signal, Image and Video Processing*, 17(2): 427–434.

Guidotti, R. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 36.

Jiang, J.; Leofante, F.; Rago, A.; and Toni, F. 2024. Robust Counterfactual Explanations in Machine Learning: A Survey. *arXiv preprint arXiv:2402.01928*.

Kanamori, K.; Takagi, T.; Kobayashi, K.; and Arimura, H. 2020. DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization. In *IJCAI*, 2855–2862.

Mirzaei, S.; Mao, H.; Al-Nima, R. R. O.; and Woo, W. L. 2023. Explainable AI Evaluation: A Top-Down Approach for Selecting Optimal Explanations for Black Box Models. *Information*, 15(1): 4.

Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–617. ACM.

Poyiadzi, R.; Sokol, K.; Santos-Rodriguez, R.; De Bie, T.; and Flach, P. 2020. FACE: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 344–350.

Prasanth Kadiyala, S.; and Woo, W. L. 2021. Flood Prediction and Analysis on the Relevance of Features using Explainable Artificial Intelligence. In *Proceedings of the 2nd Artificial Intelligence and Complex Systems Conference*, 1–6.

Rasouli, P.; and Chieh Yu, I. 2024. CARE: Coherent actionable recourse based on sound counterfactual explanations. *International Journal of Data Science and Analytics*, 17(1): 13–38.

Rasouli, P.; and Yu, I. C. 2021. Analyzing and Improving the Robustness of Tabular Classifiers using Counterfactual Explanations. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1286–1293. IEEE.

Russell, C. 2019. Efficient Search for Diverse Coherent Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 20–28. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Sanderson, J.; Mao, H.; Abdullah, M. A.; Al-Nima, R. R. O.; and Woo, W. L. 2023a. Optimal Fusion of Multispectral Optical and SAR Images for Flood Inundation Mapping through Explainable Deep Learning. *Information*, 14(12): 660.

Sanderson, J.; Mao, H.; Tengtrairat, N.; Al-Nima, R.; and Woo, W. 2024. Explainable Deep Semantic Segmentation for Flood Inundation Mapping with Class Activation Mapping Techniques. In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence*, volume 3 of *ICAART*, 1028–1035. Scitepress.

Sanderson, J.; Tengtrairat, N.; Woo, W. L.; Mao, H.; and Al-Nima, R. R. 2023b. XFIMNet: an Explainable deep learning architecture for versatile flood inundation mapping with synthetic aperture radar and multi-spectral optical images. *International Journal of Remote Sensing*, 44(24): 7755–7789.

Schleich, M.; Geng, Z.; Zhang, Y.; and Suciu, D. 2021. GeCo: Quality Counterfactual Explanations in Real Time. In *Proceedings of the VLDB Endowment*, 1681–1693.

Shin, H.; Park, J. E.; Jun, Y.; Eo, T.; Lee, J.; Kim, J. E.; Lee, D. H.; Moon, H. H.; Park, S. I.; Kim, S.; et al. 2023. Deep learning referral suggestion and tumour discrimination using explainable artificial intelligence applied to multiparametric MRI. *European Radiology*, 33: 1–12.

Tsiourvas, A.; Sun, W.; and Perakis, G. 2024. Manifold-Aligned Counterfactual Explanations for Neural Networks. In *International Conference on Artificial Intelligence and Statistics*, 3763–3771. PMLR.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law and Technology*, 31.

Yagin, F. H.; Cicek, İ. B.; Alkhateeb, A.; Yagin, B.; Colak, C.; Azzeh, M.; and Akbulut, S. 2023. Explainable artificial intelligence model for identifying COVID-19 gene biomarkers. *Computers in Biology and Medicine*, 154: 106619.

Zhu, X.; Chu, Q.; Song, X.; Hu, P.; and Peng, L. 2023. Explainable prediction of loan default based on machine learning models. *Data Science and Management*, 6.