RoCourseNet: Robust Training of a Prediction Aware Recourse Model

Hangzhi Guo, Feiran Jia, Jinghui Chen, Anna Squicciarini, and Amulya Yadav. 2023. RoCourseNet: Robust Training of a Prediction Aware Recourse Model. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM ’23), October 21–25, 2023, Birmingham, United Kingdom. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3583780.3615040

1 INTRODUCTION

Existing work in Explainable Artificial Intelligence (XAI) has been focused on developing techniques to interpret decisions made by black-box machine learning (ML) models [23, 24, 28, 39]. In particular, counterfactual (CF) explanation methods find a new counterfactual example cf(x), which is similar to input instance x but gets a different/opposite prediction from the ML model. Counterfactual explanation techniques [22, 33, 48, 51] are often preferred by human end-users because of their ability to provide actionable recourse to individuals who are negatively impacted by algorithm-mediated decisions. For example, CF explanation techniques can be used to provide algorithmic recourse for impoverished loan applicants who have been denied a loan by a bank’s ML algorithm, etc.

Most CF explanation techniques assume that the underlying ML model is stationary and does not change over time [3]. However, in practice, ML models are often updated regularly when new data is available to improve predictive accuracy on the new shifted data distribution. This shifted ML model might render previously recommended recourses ineffective [38], and in turn, diminish end users’ trust towards our system. For example, when providing a recourse to a loan applicant who was denied a loan by the bank’s ML algorithm, it is critical to ensure that the bank can honor that recourse and approve re-applications that fully follow recourse recommendations, even if the bank updates their ML model in the meantime. This necessitates the development of robust algorithms that can generate recourses that remain effective (or valid) for an end-user in the face of ML models being frequently updated. Figure 1 illustrates this challenge of generating robust recourses.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:
Hangzhi Guo, Feiran Jia, Jinghui Chen, Anna Squicciarini, and Amulya Yadav. 2023. RoCourseNet: Robust Training of a Prediction Aware Recourse Model. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM ’23), October 21–25, 2023, Birmingham, United Kingdom. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3583780.3615040

Note that counterfactual explanation [51] and algorithmic recourse [48] are closely related [46, 50]. Hence, we use these terms interchangeably.
Limitations of Prior Work. To our knowledge, only two studies [35, 47] propose methods to generate robust recourses. Unfortunately, both these studies suffer from two major limitations. First, both methods are based on strong modeling assumptions which degrade their effectiveness at finding robust recourses (as we show in Section 4). For example, Upadhyay et al. [47] assume that the ML model’s decision boundary can be locally approximated via a linear function, and adopt LIME [39] to find this linear approximation. However, recent works show that the local approximation generated from LIME is unfaithful [25, 40] and inconsistent [1, 45]. Similarly, Nguyen et al. [35] assumes that the underlying data distribution can be approximated via kernel density estimation [5]. However, kernel density estimation suffers from the curse of dimensionality [4], which performs exponentially worse with increasing dimensionality of data [10, 34]. This limits its usability for estimating data distributions in real-world high-dimensional datasets.

Second, these two techniques are post-hoc methods designed for use with proprietary black-box ML models whose training data and model weights are not available. However, with the advent of data regulations that enshrine the “Right to Explanation” (e.g., EU-GDPR [51]), service providers are required by law to communicate both the decision outcome (i.e., the ML model’s prediction) and its actionable implications (i.e., a recourse for this prediction) to an end-user. In these scenarios, the post-hoc assumption is overly limiting, as service providers can build recourse models that leverage the knowledge of their ML model to generate higher-quality recourses. In fact, prior work [19] has shown that post-hoc CF explanation approaches are unable to balance the cost-invalidity trade-off [38], which is an important consideration in generating recourses. To date, very little prior work departs from the post-hoc paradigm; [19] propose one such approach, unfortunately, it does not consider the robustness of generated recourses.

Contributions. We propose Robust ReCourse Neural Network (or RoCourseNet), a novel framework for generating recourses which: (i) departs from the paradigm of post-hoc explainability in generating recourses; while (ii) optimizing the robustness of recourse explanations. RoCourseNet presents four key contributions:

- (Formulation-wise) We formulate the robust recourse generation problem as a tri-level (min-max-min) optimization problem, which consists of two sub-problems: (i) a bi-level (max-min) problem which simulates a worst-case attacker to find an adversarially shifted ML model by explicitly simulating the worst-case data shift in the training dataset; and (ii) an outer minimization problem which simulates an ML model designer who wants to generate robust recourses against this worst-case bi-level attacker. Unlike prior approaches, our bi-level attacker formulation explicitly connects shifts in the underlying data distribution to corresponding shifts in the ML model parameters.

- (Methodology-wise) We propose RoCourseNet for solving our tri-level optimization problem for generating robust recourses. RoCourseNet relies on two key ideas: (i) we propose a novel Virtual Data Shift (VDS) algorithm to optimize for the inner bi-level (max-min) attacker problem, which results in an adversarially shifted ML model; and (ii) inspired by Guo et al. [19], RoCourseNet leverages a block-wise coordinate descent training procedure to optimize the robustness of generated recourses against these adversarially shifted models. Unlike prior methods [35, 47], our method requires no intermediate steps in approximating the underlying model or data distribution.

- (Experiment-wise) We conduct rigorous experiments on three real-world datasets to evaluate the robustness of several popular recourse generation methods under data shifts. Our results show that RoCourseNet generates highly robust CF explanations against data shifts, as it consistently achieves >96% robust validity, outperforming state-of-the-art baselines by ~10%.
(Framework-wise) Finally, we extend the RoCourseNet training as a generalized robust training framework to be used with any parametric post-hoc explanation method. By applying the RoCourseNet training, we witness ~25% robust validity improvements to existing CF parametric methods.

2 RELATED WORK

Counterfactual Explanation Techniques. A significant body of literature exists on CF explanation techniques, which focuses on generating recourses that lead to different (and often more preferable) predicted outcomes [21, 50, 51]. We categorize prior work on CF explanation techniques into non-parametric methods [21, 22, 33, 47–51], which aim to find recourses without involving parameterized models, and parametric methods [19, 31, 36, 54], which adopt parametric models (e.g., a neural network model) to generate recourses. In particular, our work is most closely related to CounterNet [19], which unlike post-hoc methods, jointly trains the predictive model and a CF explanation generator. This joint-training procedure leads to significantly better alignment between the generated predictions and corresponding CF explanations. However, all aforementioned CF explanation techniques (including CounterNet) do not optimize for robustness against adversarial model shifts. In contrast, our work devises a novel tri-level adversarial training approach to ensure the robustness of CF explanations generated by RoCourseNet.

Robustness in Recourse Explanations. Our method is closely related to the model shift problem in algorithmic recourse [38], i.e., how to ensure that the generated recourse is robust to shifts in the underlying predictive model. However, existing approaches [35, 47] rely on simplifying assumptions: (i) Upadhyay et al. [47] propose ROAR which relies on a locally linear approximation (via LIME [39]) to construct a shifted model, which is known to suffer from inconsistency [1, 45] and unfaithfulness issues [25, 40]. (ii) Similarly, Nguyen et al. [35] propose RBR which assumes that kernel density estimators can approximate the underlying data distribution. In particular, RBR uses Gaussian kernels for multivariate density estimation, which suffers from the curse of dimensionality [4, 10, 34]. In contrast, our work relaxes these assumptions by constructing adversarially shifted models via simulating the worst-case data shift, and conducting adversarial training for robust CF generation.

Orthogonal to our work, Pawelczyk et al. [37] analyze the model multiplicity problem, which studies the validity of recourses under different ML models trained on the same data, and Black et al. [6] propose methods to ensure consistency under the model multiplicity setting. In addition, some prior work focuses on ensuring robustness to small perturbations in the recourse features [12, 14, 32].

Adversarial training. We leverage adversarial robustness techniques to protect ML models from adversarial examples [8, 17, 30, 44, 53]. In addition, recent works [15, 16] also leverage adversarial training to defend against data poisoning [20] and backdoor attacks [41]. In general, adversarial training solves a bi-level (min-max) optimization problem. In our work, we formulate RoCourseNet’s objective as a tri-level (min-max-min) optimization problem, which we can decompose into a game played between a model designer and a worst-case (hypothetical) adversary. The inner worst-case data and model shifts are assumed to be generated by a bi-level worst-case attacker. The defender trains a robust CF generator against this bi-level attacker by following an adversarial training procedure.

3 ROCOURSENET: END-TO-END ROBUST RECURSE GENERATION

RoCourseNet is an end-to-end training framework for simultaneously generating accurate predictions and corresponding recourses (or CF explanations) that are robust to model shifts induced by shifts in the training dataset. We describe the RoCourseNet framework in two stages. First, we discuss the attacker’s problem: (i) we propose a novel bi-level attacker problem to find the worst-case data shift that leads to an adversarially shifted ML model; and (ii) we propose a novel Virtual Data Shift (VDS) algorithm for solving this bi-level attacker problem. Second, we discuss the defender’s problem: (i) we derive a novel tri-level learning problem based on the attacker’s bi-level problem; and (ii) we propose the RoCourseNet training framework for optimizing this tri-level optimization problem, which leads to the simultaneous generation of accurate predictions and robust recourses.

3.1 Virtual Data Shift: Constructing Worst-case Data Shifts

We define a predictive model \( f : X \to Y \). Let \( D = \{ (x_i, y_i) | i \in \{1, \ldots, N\} \} \) represent our training dataset containing \( N \) points. We denote \( f(x, \theta) \) as the prediction generated by predictive model \( f \) on point \( x \), parameterized by \( \theta \). Next, we denote \( \theta \) and \( \theta' \) as parameters of an (original) ML model \( f(\cdot; \theta) \) and its shifted counterpart \( f(\cdot; \theta') \), respectively. Let \( x^{cf} \) denote a CF explanation (or recourse) for input point \( x \). Finally, a recourse \( x^{cf} \) is valid iff it gets an opposite prediction from the original data point \( x \), i.e., \( f(x^{cf}, \theta) = 1 - f(x, \theta) \).

On the other hand, a recourse \( x^{cf} \) is robustly valid w.r.t. a shifted model \( f(\cdot; \theta') \) iff \( x^{cf} \) gets an opposite prediction from the shifted model (as compared to the prediction received by \( x \) on the original model), i.e., \( f(x^{cf}, \theta') = 1 - f(x, \theta) \). This definition aligns with the notion of robustness to model shifts in the literature [47]. Finally, \( \mathcal{L}(\cdot) \) represents a loss function formulation, e.g., binary cross-entropy, mean squared error, etc.

Model Shift as an optimization problem. To motivate the need of introducing our bi-level attacker problem, we first discuss an optimization problem for a worst-case attacker which directly perturbs model parameters to find an adversarially shifted model (denoted by \( f(\cdot; \theta'_{adv}) \)). The goal of the attacker is to find an adversarially shifted model which minimizes the robustness of the generated recourses. More formally, given our training dataset \( D \) and a CF explanation method (that can generate recourses \( x^{cf} \) for each \( (x, y) \in D \)), the attacker’s problem aims to find the worst-case shifted model \( f(\cdot; \theta'_{adv}) \) which minimizes the robust validity of the generated recourses, i.e., \( f(x^{cf}, \theta'_{adv}) \neq 1 - f(x, \theta) \). We can find an adversarial shifted model by solving Equation 1.

\[
\theta'_{adv} = \arg\max_{\theta' \in \mathcal{F}} \frac{1}{N} \sum_{(x_i, y_i) \in D} \left[ \mathcal{L}(f(x^{cf}, \theta'), 1 - f(x_i, \theta) ) \right]
\] (1)
where \( x^f \) correspond to CF explanations of input \( x \) produced by a CF generator, and \( \mathcal{F} = \{ \theta' \mid \theta + \delta \} \) denotes a plausible set of the parameters of all possible shifted models.

Unfortunately, it is non-trivial to construct a plausible model set \( \mathcal{F} \) by directly perturbing the ML model’s parameters \( \theta \), especially when \( f(\cdot; \theta) \) is represented using a neural network. Unlike a linear model, quantifying the importance of neurons is challenging [11, 26], which leads to difficulty in applying weight perturbations. To overcome these challenges, prior work [47] adopts a simplified linear model to approximate the target model, and perturbs this linear model accordingly. Unfortunately, this simplified local linear model introduces approximation errors into the system, which leads to poor performance (as shown in Section 4). Instead of directly perturbing the model’s weights, we explicitly consider a worst-case data shift, which then leads to an adversarial model shift.

### Data Shift as a bilevel optimization problem
We identify distributional data shifts as the fundamental cause of non-robust recourses. Unlike prior work, we propose a bi-level optimization problem for the attacker which explicitly connects shifts in the underlying training data to corresponding shifts in the ML model parameters. Specifically, in response to a shift in the training data (from \( D \) to \( D_{\text{shifted}} \)), defenders update (or shift) their predictive model by optimizing the prediction loss:

\[
0_{\text{opt}}^0 = \arg\max_{\theta'} \frac{1}{N} \sum_{(x_i, y_i) \in D_{\text{shifted}}} \left[ \mathcal{L}(f(x_i; \theta'), y_i) \right]. \tag{2}
\]

Crucially, this updated ML model \( f(\cdot; \theta_{\text{opt}}^0) \) (caused by the shifted data \( D_{\text{shifted}} \)) is the key to the non-robustness of recourses (i.e., \( f(x^f; \theta_{\text{opt}}^0) \neq 1 - f(x; \theta) \)). Therefore, to generate robust recourses, we optimize against an adversary who creates a worst-case data shift, which then leads to an adversarial model shift.

\[
\delta^* = \arg\max_{\delta} \frac{1}{N} \sum_{(x_i, y_i) \in D} \left[ \mathcal{L}(f(x_i; \theta^0_{\text{opt}}(\delta)), 1 - f(x_i; \theta)) \right] \tag{3}
\]

\[
s.t., \theta^0_{\text{opt}}(\delta) = \arg\max_{\theta'} \frac{1}{N} \sum_{(x_i, y_i) \in D} \left[ \mathcal{L}(f(x_i + \delta_i; \theta'), y_i) \right].
\]

where \( \delta_i \in \Lambda \) denotes the data shift for a single data point \( x_i \in D \), and \( \delta = \{ \delta_i \mid \forall (x_i, y_i) \in D \} \) denotes the data shift across all data points in the entire dataset. We define \( \Lambda \) as the \( \ell_\infty \)-norm ball \( \Lambda = \{ \delta \in \mathbb{R}^n \mid ||\delta||_\infty \leq \epsilon \} \). Intuitively, the outer problem minimizes the robust validity to construct the worst-case data shift \( \delta^* \), and the inner problem learns a shifted model \( f(\cdot; \theta^0_{\text{opt}}(\delta)) \) on the shifted dataset \( D^*_{\text{shifted}} \). Once we get the optimal \( \delta^* = (\delta_1^*, \delta_2^*, \ldots, \delta_N^*) \) by solving Equation 3, the worst-case shifted dataset \( D^*_{\text{shifted}} = \{ (x_i + \delta_i^*; y_i) \mid \forall (x_i, y_i) \in D \} \).

### Virtual Data Shift (VDS)
Unfortunately, solving Eq. 3 is computationally intractable due to its nested structure. To approximate this bi-level problem, we devise Virtual Data Shift (VDS) (Algorithm 1), a gradient-based algorithm with an unrolling optimization pipeline. At a high level, VDS iteratively approximates the inner problem by unrolling \( K \)-steps of gradient descent for each outer optimization step. Similar unrolling pipelines are adopted in many ML problems with a bi-level formulation [18, 43], e.g., meta-learning [13], hyperparameter search [29], and poisoning attacks [20].

Algorithm 1 layouts the VDS algorithm which outputs the worst-case data shift \( \delta^* \), and the corresponding shifted model \( f(\cdot; \theta^0_{\text{opt}}(\delta)) \). VDS makes two design choices. First, it uniformly randomizes the data shift \( \delta \sim \mathcal{U}(\epsilon, \epsilon) \), where \( \delta = \{ \delta_1, \ldots, \delta_N \} \) (Line 3), following practices of Wong et al. [53]. Uniform randomization is critical to adversarial model performance as it increases the smoothness of the objective function, leading to improved convergence of gradient-based algorithms [8]. Then, we iteratively solve this bi-level optimization problem via \( T \) outer attack steps. At each step, we first update the predictor \( f(\cdot; \theta^0) \) using the shifted data \( x + \delta \) via \( K \) unrolling steps of gradient descent (Line 6). Next, similar to the fast sign gradient method (FSGM) [17], we maximize the adversarial loss and project \( \delta \) into the feasible region \( \Lambda \) (Line 8). Crucially, when computing the gradient of adversarial loss (outer problem) w.r.t. data shift \( \delta \) (Line 8), we look ahead a few steps in the inner problem before back-propagating to the initial unrolling step. This approach stems from applying \( K \) unrolling steps of gradient descent, as opposed to full-blown gradient descent until convergence. Note that the gradient w.r.t. \( \delta \) depends on \( \theta(\delta) \), where \( \theta(\delta) \) is a function derived from LINE 5-7.

### 3.2 Block-wise Coordinate Descent with Adversarial Training

**Choice of CF Explanation Technique.** Note that our bi-level attacker formulation assumes that CF explanations \( x^f \) for all data points \( x_i \) are provided as input to the VDS algorithm. Thus, a key design choice inside the RoCourseNet framework is the selection of an appropriate CF explanation technique, which can be used to generate recourses for input data points in Algorithm 1.

As mentioned in Section 2, most existing CF explanation techniques follow the post-hoc paradigm, which makes them unsuitable for use inside the RoCourseNet framework for two reasons: (i) mis-aligned motivations: post-hoc CF explanation methods are mainly designed for use with proprietary black-box ML models whose training data and model weights are not available; instead, the VDS...
algorithm relies on having access to the training data. Thus, the motivations and use cases of VDS and post-hoc methods are misaligned. (ii) the post-hoc paradigm is overly limiting in many real-world scenarios. With the advent of data regulations that enshrine the “Right to Explanation” (e.g., EU-GDPR [51]), service providers are required by law to communicate both the decision outcome (i.e., the ML model’s prediction) and its actionable implications (i.e., a recourse for this prediction) to an end-user. In these scenarios, the post-hoc assumption is overly limiting. Service providers can build specialized CF explanation techniques that leverage the knowledge of their specific ML model to generate higher-quality recourses.

Motivated by these reasons, we choose CounterNet [19] as our CF explanation model of choice inside the RoCourseNet framework, as that is an end-to-end approach that departs from the post-hoc explanation paradigm by jointly training predictions and recourses. In fact, [19] show that CounterNet can better balance the cost-invalidation trade-off [38] than state-of-the-art post-hoc approaches.

RoCourseNet Objective Function. We describe how we combine the bi-level attacker problem with an outer minimization, which represents RoCourseNet’s tri-level objective function. Inspired by [19], RoCourseNet has three objectives: (i) high predictive accuracy - we expect RoCourseNet to output accurate predictions; (ii) high robust validity - we expect that generated recourses in RoCourseNet are robustly valid on shifted models; (iii) low proximity - we desire minimal changes required to modify input instance \( x \) to corresponding recourse \( x_{cf} \). Given these objectives, RoCourseNet solves the following min-max-min problem to optimize parameters for its predictor \( f(\cdot; \theta) \) and recourse generator \( g_{\cdot}(\cdot; \theta_g) \). Note that similar to [19], both \( f(\cdot; \theta) \) and \( g_{\cdot}(\cdot; \theta_g) \) are neural networks.

\[
\text{argmin}_{\theta, \theta_g} \frac{1}{N} \sum_{(x_i, y_i) \in D} \left[ \lambda_1 \cdot L\left(f(x_i; \theta), y_i\right) + \lambda_3 \cdot L\left(x_i, x_{cf}^i\right) \right] \quad \text{(Prediction Loss (L1))}
\]

\[
+ \max_{\delta \in \Delta} \frac{1}{N} \sum_{(x_i, y_i) \in D} \left[ \lambda_2 \cdot L\left(f\left(\frac{x_{cf}^i}{f(x_i; \delta); \theta'}\right), y_i\right) \right] \quad \text{(Proximity Loss (L2))}
\]

s.t., \( \theta' = \text{argmin}_{\delta} \frac{1}{N} \sum_{(x_i, y_i) \in D} \left[ L\left(f(x_i + \delta, \cdot; \theta') - y_i\right) \right] \),

\( x_{cf}^i = g(x_i; \theta_g) \).

\text{(4)}

RoCourseNet Training. A common practice in solving a min-max formulation is to first solve the inner maximization problem, and then solve the outer minimization problem [30]. Thus, we solve Eq. 4 as follows (see Algorithm 2): (i) To solve the inner bi-level problem, we find the worst-case model shift by solving Eq. 3 using VDS (Algorithm 1). (ii) To solve the outer minimization problem, we adopt block-wise coordinate descent optimization by distributing gradient descent backpropagation on the objective function into two stages (as suggested in [19]) - at stage one, we optimize for predictive accuracy, i.e., \( L_1 \) in Eq. 4 (Line 7), and at stage two, we desire minimal changes required to modify input instance \( x \), as the worst-case model shift case encompasses the unshifted model case. We also empirically validate this design choice in Section 4.

**Algorithm 2 Tri-level Robust CF Training**

1. **Hyperparameters**: learning rates \( \eta, \theta \) of epochs \( N \), maximum perturbation \( E \)
2. **Input**: dataset \((x, y) \in D\)
3. **Initialize**: \( \theta \)
4. for epoch = 1 \( \rightarrow \) N do
5. \( \epsilon = E \cdot \text{epoch}/N \) \( \triangleright \) Linearly schedule \( \epsilon \).
6. for each minibatch \( B \) in \( D \) do
7. \( \theta' \leftarrow \theta - \eta \cdot \nabla g_{\theta} L_1 \)
8. \( \theta_g' \leftarrow \text{VDS}(B, x_{cf}^i, \theta, \epsilon) \)
9. \( \theta_g \leftarrow \theta_g - \eta \cdot \nabla \theta_g (L_2 + L_3) \)
10. end for
11. end for

**Note**: The quality of CF explanations, i.e., \( \lambda_2 \cdot L_2 + \lambda_3 \cdot L_3 \) in Eq. 4 (Line 9). Note that in Line 9, the gradient of \( \lambda_2 \cdot L_2 + \lambda_3 \cdot L_3 \) is calculated using the adversarially shifted model parameters \( \theta' \) found by VDS in Line 8. [19] show that this block-wise coordinate descent algorithm efficiently optimizes the outer minimization problem by alleviating the problem of divergent gradients.

In addition, we linearly increase the perturbation constraints \( \epsilon \) for improved convergence of our robust CF generator. Intuitively, linearly increasing \( \epsilon \) values correspond to increasingly strong adversarial. Prior work in curriculum adversarial training [7, 52] suggests that adaptively adjusting the strength of the adversary improves the convergence of adversarial training (we verify this in Section 4).

**4 EXPERIMENTAL EVALUATION**

**Baselines.** To our knowledge, RoCourseNet is the first method that optimizes an end-to-end model for generating predictions and robust recourses. Hence, there exist no previous approaches which we can directly compare against. Nevertheless, we compare RoCourseNet against four state-of-the-art baselines.

- **VANILLA-CF** [51] is a popular post-hoc non-parametric method optimizing validity and proximity.
- **COUNTERNET** [19] is the first end-to-end model for simultaneously generating predictions and recourses, which inspires the development of RoCourseNet.
- **ROAR-LIME** [47] generates robust recourses by perturbing parameters of locally approximated linear models.
- **RBR** [35] generates robust recourses by a Gaussian kernel.

Other than CounterNet, all remaining baselines require a trained ML model as an input. Similar to [19], we train a neural network model as the base ML model for all baselines. For each dataset, we separately tune hyperparameters via grid search.

**Datasets.** To remain consistent with prior work on robust recourses [35, 47], we evaluate RoCourseNet on three benchmarked real-world datasets. Table 1 summarizes these three datasets.

- **Loan** [27] captures temporal shifts in loan application records. It predicts whether a business defaulted on a loan \((Y=1)\) or not \((Y=0)\). This large-sized dataset (i.e., ~450k data points) has loan approval records across the U.S. during 1994 to 2009 (i.e., \(D_{\text{loan}} = \{D_1, D_2, \ldots, D_k\} \), where each subset \(D_i\) corresponds to a particular year, and \(k = 16\) is the total number of years).
Table 1: Summary of Datasets used for Evaluation. Each dataset consists of \( k \) subsets. Size represents the number of data points for the entire dataset.

| Dataset      | Size | \( k \) | #Continuous | #Categorical |
|--------------|------|--------|-------------|--------------|
| Loan         | 449,152 | 16     | 7           | 5            |
| German Credit| 2,000  | 2      | 6           | 3            |
| Student      | 649   | 2      | 2           | 13           |

Finally, we use three metrics to evaluate a CF explanation: (i) \( \text{Validity} \) is the fraction of valid CF examples on the original model \( f(...) \). (ii) \( \text{Robust validity} \) is the fraction of robustly valid CF examples on a shifted predictive model \( f_{\theta'} \). We calculate the robust validity on all possible shifted models (as described above). (iii) \( \text{Proximity} \) is the \( l_1 \) distance between the input and the CF example. We report the averaged results across all subsets \( \mathcal{D}_{all} \) (see Table 2).

4.1 Experimental Results

Validity & Robust Validity. Table 2 compares the validity and robust validity achieved by RoCourseNet and other baselines. RoCourseNet achieves the highest \( \text{robust validity} \) on each dataset - it outperforms ROAR-LIME by 10% (the next best performing baseline), and consistently achieves at least 96.5% robust validity. This illustrates RoCourseNet’s effectiveness at generating highly robust recourses. Also, RoCourseNet achieves the highest \( \text{validity} \) on each dataset. This result shows that optimizing the worst-case shifted model (i.e., \( L_2 \) in Eq. 4) is sufficient to achieve high validity, without the need to explicitly optimize for an additional validity loss.

Proximity. Table 2 compares the proximity achieved by RoCourseNet and baselines. In particular, RoCourseNet performs exceedingly well on the Loan application dataset (our largest dataset with \approx 450k data points), as it is the second-best method in terms of proximity (just behind CounterNet), and outperforms all our post-hoc baseline methods (ROAR-LIME, and VanillaCF). Perhaps understandably, RoCourseNet achieves poorer proximity on the German Credit and Student dataset, given that the limited size of these datasets (less than 1000 data points) precludes efficient training.

Cost-Validity Trade-Off. We compare RoCourseNet, ROAR-LIME, and RBR (three recourse methods explicitly optimizing for distributional shift) in their trade-off between the cost (measured by proximity) and their original and robust validity [38]. For each method, we plot the Pareto frontier of the cost-validity trade-off as follows: (i) For RBR, we obtain the Pareto frontier by varying the ambiguity sizes \( \epsilon_1, \epsilon_2 \) (i.e., hyperparameters of RBR [35]); (ii) For ROAR-LIME and CounterNet, we obtain the Pareto frontier by varying the trade-off hyperparameter \( \lambda \) that controls the proximity loss term in their respective loss functions (e.g., \( \lambda_3 \) in Equation 4). Figure 3 shows that on the large-sized Loan dataset, RoCourseNet’s Pareto frontiers either dominate or are comparable to frontiers achieved by ROAR-LIME and RBR. On the other hand, we observe a clear trade-off on the German Credit and Student datasets, where RoCourseNet (and other methods) can increase their normal and robust validity, but only at the cost of a poorer proximity score. This is consistent with prior literature, which shows that proximity needs to be sacrificed to achieve higher validity [35].

4.2 Further Analysis

RoCourseNet Training Requires More Data. We further delve into the impact of data size on the cost-invalidity trade-off. Figure 5 illustrates the Pareto Frontier plot of RoCourseNet trained on various fractions of the Loan dataset. This figure shows that increasing the training data leads to a better balance between normal and robust cost-invalidity trade-offs. This supports our hypothesis.
that the underperformance of RoCourseNet on the German Credit and Student datasets can be attributed to their limited number of data points, making adversarial training more challenging. This result also echoes findings in prior literature [42], which shows that the sample complexity of robust learning can be significantly larger.

Evaluating VDS Attacker. We demonstrate the effectiveness of VDS on solving the inner maximization problem in Eq. 4, i.e., how often VDS succeeds in finding an adversarial shifted model $f(\cdot; \theta')$, such that the generated recourses $x^{cf}$ are not robustly valid. To evaluate the performance of VDS, we apply Algorithm 1 to find shifted model parameters $\theta'$, and compute the robust validity of all hold-out test-sets with respect to this shifted model.

Figure 4a and 4d compares the effectiveness of attacking RoCourseNet and CounterNet via the VDS algorithm, which highlights three important findings: (i) First, the VDS algorithm is effective in finding a shifted model which invalidates a given recourse - the average robust validity drops to 74.8% when the attacker steps $T = 20$, as compared to 99.8% validity on the original model. (ii) Figure 4a shows that when $T$ is increased, the robust validity of both CounterNet and RoCourseNet is degraded, which indicates that increasing attack steps improves the effectiveness of solving the bi-level problem in Eq. 3. Similarly, Figure 4d shows that increasing $E$ also improves effectiveness of solving Eq.3. (iii) Finally, RoCourseNet is more robust than CounterNet when attacked by the VDS algorithm, as RoCourseNet vastly outperforms CounterNet in robust validity (e.g., ~28%, ~10% improved robust validity when $T = 20$, $E = 0.5$ in Figure 4a and 4d, respectively).

Understanding the Tri-level Robust CF Training. We further analyze our tri-level robust training procedure. First, a stronger attacker (i.e., more effective in solving Eq. 3) leads to the training of a more robust CF generator. In Figure 4b, we observe that increasing $T$ (which results in a stronger attacker as shown in Figure 4a) leads to improved robust validity, which indicates a more robust CF generator. Similarly, Figure 4e illustrates that increasing $E$ values leads to improved robust validity. These results show that having an appropriately strong attacker (i.e., effectively optimizing Eq. 3) is crucial to train for a robust CF generator.

Epsilon Scheduler. Figure 6 illustrates the importance of linearly scheduling $\epsilon$ values inside VDS. By linearly increasing $\epsilon$, we observe ~0.88% improved robust validity (on average) compared to using a static $\epsilon$ during the entire adversarial training. This shows that this curriculum training strategy can boost the robustness of recourses.

Table 2: Evaluation of recourse robustness under model shift. It is desirable for recourse methods to have low proximity (prox.) with high validity (Val.) and high robust validity (Rob-Val.).

| Methods     | Loan  | German Credit | Student |
|-------------|-------|---------------|---------|
|             | Prox. | Val. | Rob-Val. | Prox. | Val. | Rob-Val. | Prox. | Val. | Rob-Val. |
| VANILLA CF  | 7.390 | 0.942 | 0.885   | 4.635 | 0.940 | 0.772   | 15.236 | 0.915 | 0.673   |
|             | ± 1.860 | ± 0.026 | ± 0.121 | ± 0.197 | ± 0.011 | ± 0.008 | ± 0.383 | ± 0.056 | ± 0.006 |
| COUNTERNET  | 6.746 | 0.964 | 0.639   | 5.719 | 0.960 | 0.706   | 18.619 | 0.967 | 0.859   |
|             | ± 0.723 | ± 0.085 | ± 0.222 | ± 0.130 | ± 0.028 | ± 0.074 | ± 0.131 | ± 0.033 | ± 0.071 |
| ROAR-LIME   | 7.648 | 0.934 | 0.918   | 4.862 | 0.910 | 0.792   | 11.931 | 0.967 | 0.820   |
|             | ± 1.951 | ± 0.024 | ± 0.066 | ± 0.117 | ± 0.0255 | ± 0.052 | ± 1.396 | ± 0.026 | ± 0.075 |
| RBR         | 11.71 | 0.824 | 0.821   | 6.005 | 0.524 | 0.586   | 26.255 | 0.660 | 0.611   |
|             | ± 1.633 | ± 0.130 | ± 0.132 | ± 2.099 | ± 0.148 | ± 0.046 | ± 3.089 | ± 0.014 | ± 0.073 |
| ROOCOURSENET | 6.611 | 0.996 | 0.969   | 6.903 | 0.978 | 0.964   | 16.508 | 1.000 | 1.000   |
|             | ± 0.418 | ± 0.002 | ± 0.106 | ± 0.250 | ± 0.002 | ± 0.006 | ± 0.281 | ± 0.000 | ± 0.000 |

Figure 3: Pareto frontiers of the cost-invalidity trade-off for ROAR-LIME, RBR, and ROOCOURSENET. Methods located in the upper-left region are preferable, as they exhibit a favorable balance between cost and invalidity. On the Loan (left) dataset, ROOCOURSENET exhibits a superior balance between cost and invalidity compared to ROAR-LIME and RBR. On the German Credit (middle) and Student (right) dataset, ROOCOURSENET achieves high normal and robust validity at the cost of proximity score.
We now discuss how the RoCourseNet framework is general enough when increasing training data.

Figure 4: The impact of the number of attacker steps $T$ (4a-4c) and max-perturbation $E$ (4d-4f) to robustness on the Loan dataset (↑ or ↓ means that higher or lower value is preferable, respectively).

Figure 5: The influence of dataset size on the cost-invalidity trade-off. We analyze RoCourseNet on various fractions of the Loan Dataset. The Pareto Frontier plot reveals improvement in both normal and robust cost-invalidity trade-offs when increasing training data.

Figure 6: Impact of epsilon scheduler on the robustness. The curriculum training boosts the robustness of recourses.

Note that the VDS algorithm and its corresponding adversarial training procedure (Algorithms 1 & 2) require access to (i) the training dataset $(x, y) \in \mathcal{D}$ and (ii) the weights of the predictive models $\mathbf{\theta}_f$. Therefore, any parametric CF explanation method that satisfies these assumptions can leverage these algorithms to improve their robustness to data shift.

Thus, while in this paper, we have made a conscious decision of choosing CounterNet as the CF explanation method of choice within the RoCourseNet framework (since CounterNet’s end-to-end architecture addresses the limitations of post-hoc approaches); in general, the RoCourseNet framework is model-agnostic, and it can work with any parametric model based CF explanation method.

5 GENERALIZING ROCOURSENET TO PARAMETRIC CF EXPLANATION METHODS

We now discuss how the RoCourseNet framework is general enough to be used with other parametric CF explanation methods; in fact, we illustrate how the RoCourseNet framework can be used to improve the robustness of any parametric CF explanation method.
Table 3: Evaluating robustness under model shift with generalized parametric methods on the *Adult* dataset. The robust training improves both validity and robust validity.

| Methods          | Metrics       |
|------------------|---------------|
|                  | Prox. | Val.  | Rob-Val. |
| CF Model         | 5.75 ± 0.647 | 0.826 ± 0.1825 | 0.607 ± 0.167 |
| Robust CF Model  | 0.739 ± 1.078 | 0.949 ± 0.022 | 0.906 ± 0.119 |
| VAE-CF           | 8.531 ± 1.241 | 0.745 ± 0.272 | 0.692 ± 0.251 |
| Robust VAE-CF    | 9.473 ± 0.962 | 0.818 ± 0.186 | 0.807 ± 0.189 |
| RoCourseNet      | 6.611 ± 0.418 | 0.996 ± 0.002 | 0.969 ± 0.106 |

We formulate a slightly altered version of Eq. 4 as the objective function to use the RoCourseNet framework:

\[
\min_{\theta, \theta_g} \frac{1}{N} \sum_{(x_i, y_i) \in D} \left[ \lambda_1 \cdot \mathcal{L}(x_i, g(x_i; \theta_g)) \right] \\
+ \max_{\delta, \hat{\delta}, \bar{\delta}} \frac{1}{N} \sum_{(x_i, y_i) \in D} \left[ \lambda_2 \cdot \mathcal{L}(f(x_i^T, \theta_{opt}(\delta)), 1 - f(x_i; \theta)) \right] \\
\text{s.t., } \theta_{opt}(\delta) = \arg\min_{\theta} \frac{1}{N} \sum_{(x_i, y_i) \in D} \left[ \mathcal{L}(f(x_i + \delta_i; \theta'), y_i) \right].
\]

where \(g(\cdot; \theta_g)\) represents a CF model \(g : X \rightarrow X^T\) parametrized by \(\theta_g\). Note that Eq. 5 is almost identical to Eq. 4, except that the prediction loss (\(\mathcal{L}_1\)) is ignored, as post-hoc parametric CF explanation methods do not jointly train predictions and CF explanations (instead, they assume access to a pre-trained ML model).

To solve Eq. 5, we can apply a similar strategy used while training RoCourseNet: for each mini-batch datapoints \(\{x^{(i)}, y^{(i)}\}_m\), (i) we solve the inner bi-level problem to obtain the worst-shift weight of the predictive model \(\theta^*_b\) by applying the VDS algorithm (outlined in Algorithm 1); (ii) next, we solve the outer minimization problem by optimizing against the shifted predictive model.

**Experiment Setting.** To demonstrate the generalizability of the RoCourseNet framework, we experiment with two post-hoc parametric CF explanation methods and their "robustified" models.

- **CF Model** is a multi-layer perceptron model which outputs recourses \(x^T\) given \(x\). This model is optimized for validity and proximity. We also train a ROBUST CF MODEL to generate robust recourses using the RoCourseNet framework with CF Model.
- **VAE-CF** is a well-known parametric method that uses variational auto-encoder (VAE) to generate recourses \(x^T\) given input \(x\). Similarly, ROBUST VAE-CF utilizes the RoCourseNet framework to generate robust recourses with CF Model.

For fair comparison, the architecture of CF Model is the same as the network that combines the encoder and CF generator. Our experiment is evaluated on the *Loan* dataset, which follows the same experiment settings in Section 4 (e.g., data partitioning, metrics, etc).

**Empirical Results.** Table 3 compares the (robust) CF MODEL and VAE-CF with RoCourseNet (the best performing method). The results demonstrate two key findings: (i) First, our proposed tri-level robust training (in Algorithm 2) is general purpose and can be plugged in as-is to improve the robustness of post-hoc parametric CF methods. In particular, applying robust training to both CF Model and VAE-CF improves the robust validity of these models by \(\sim 32.5\%\) and \(17.3\%\) (on average), respectively. Figure 7 further illustrates this finding, where robustly trained methods (via the generalized RoCourseNet framework) generate recourses with higher robust validity (i.e., the robust validity matrix has more darker colored elements). (ii) Additionally, the design of RoCourseNet, utilizing the design of CounterNet [19], proves to be effective in balancing the cost-validity trade-off, as we observe that RoCourseNet outperforms ROBUST CF MODEL and ROBUST VAE-CF in proximity, validity, and robust validity. The result shows the advantages of RoCourseNet’s joint training of prediction and robust recourses.

**6 CONCLUSION.** We present RoCourseNet, an end-to-end training framework to generate predictions and robust CF explanations. We formulate this robust end-to-end training as a tri-level optimization problem, and leverage novel adversarial training techniques to solve this problem. Empirical results show that RoCourseNet outperforms state-of-the-art baselines in robust validity, and achieves better balance on the cost-validity trade-off. We further demonstrate that the RoCourseNet training framework is generalizable to be applied with any parametric CF explanation method.

**ACKNOWLEDGMENTS.**

Guo and Yadav were supported by ARO Grant # W911NF-21-1-0047.
