Classifier Reconstruction Through Counterfactual-Aware Wasserstein Prototypes

Anonymous Authors[1]

Abstract

Counterfactual explanations provide actionable insights by identifying minimal input changes required to achieve a desired model prediction. Beyond their interpretability benefits, counterfactuals can also be leveraged for model reconstruction, where a surrogate model is trained to replicate the behavior of a target model. In this work, we demonstrate that model reconstruction can be significantly improved by recognizing that counterfactuals, which typically lie close to the decision boundary, can serve as informative—though less representative—samples for both classes. This is particularly beneficial in settings with limited access to labeled data. We propose a method that integrates original data samples with counterfactuals to approximate class prototypes using the Wasserstein barycenter, thereby preserving the underlying distributional structure of each class. This approach enhances the quality of the surrogate model and mitigates the issue of decision boundary shift, which commonly arises when counterfactuals are naively treated as ordinary training instances. Empirical results across multiple datasets show that our method improves fidelity between the surrogate and target models, validating its effectiveness. The code is available at `https://anonymous.4open.science/r/ce_reconstruction-8C2D`.

1. Introduction

Counterfactual explanations have gained prominence as a growing research field (Wachter et al., 2017; Guidotti, 2024; Barocas et al., 2020; Verma et al., 2024) for offering actionable insights into altering model outcomes—for instance, suggesting a $10K income increase to secure a loan. Notably, counterfactuals can inadvertently expose model internals, creating a delicate balance between transparency and privacy (Aïvodji et al., 2020; Wang et al., 2022; Dissanayake & Dutta, 2024). This poses risks in Machine Learning as a Service platforms, where adversaries might exploit counterfactual queries to replicate models via surrogate training, a tactic known as model extraction (Gong et al., 2020). Companies must therefore assess information leakage when deploying counterfactuals. Conversely, model reconstruction can empower applicants to gauge approval odds without formally submitting sensitive data—crucial in scenarios with limited attempts (e.g., credit applications affecting scores). This work formalizes the fidelity of model reconstruction from counterfactual queries.

Prior methods treat counterfactuals as labeled data to train surrogate models (Aïvodji et al., 2020). While effective for balanced queries near the decision boundary, imbalanced datasets induce decision boundary shifts, where surrogate boundaries diverge from the target model's (Dissanayake & Dutta, 2024). This stems from margin-based generalization (Shokri et al., 2021) and worsens with one-sided counterfactuals (e.g., only reject-to-accept modifications). Two-sided queries could mitigate this (Wang et al., 2022), but such strategies fail when only one-sided counterfactuals are available—a common yet challenging scenario (e.g., loan rejections lacking accept-to-reject examples). Counterfactual Clamping loss function (Dissanayake & Dutta, 2024) offers an approach to reconstruct the classifier using neural networks by adjusting the classic entropy loss. However, when applied to scenarios with limited query samples, there's a heightened risk of overfitting.

We propose a surrogate model based on prototype-driven classification to address the challenge of limited access to queries from the original classifier and its counterfactuals. Counterfactuals are assigned a soft label of 0.5, which helps address class imbalance while simultaneously mitigating issues arising from limited sample sizes. We adopt an approach by leveraging Wasserstein barycenters to represent soft prototypes for each class. This method captures the underlying structure of each class distribution in a data-efficient manner, making it particularly suitable for low-data

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

regimes.

2. Related Work

2.1. Counterfactual Explanations in Classification

Counterfactual explanations, first formalized by Wachter et al. (2017), have become a cornerstone of explainable AI. Counterfactual generation is typically formalized as a constrained optimization problem. Given a model f, an input x, and a desired prediction y', the goal is to find a counterfactual x' such that:

$$\min_{x'} \ \mathcal{L}(f(x'), y') + \lambda \cdot d(x, x')$$

where $\mathcal{L}$ is a loss function encouraging the prediction $f(x')$ to match y', $d(x, x')$ is a distance metric, and λ is a regularization parameter. This framework is widely used across methods for producing counterfactuals that are both faithful and interpretable.

Notably, Van Looveren & Klaise (2021) develop prototype-based counterfactuals using standard distance metrics. Our methodology, however, integrates counterfactual constraints directly into the barycenter optimization process.

2.2. Prototype-Based Classification

Prototype-based classification is a foundational approach in machine learning, where each class is represented by one or more prototypes—representative examples that encapsulate the characteristics of the class. Classification decisions are made by comparing new instances to these prototypes using a suitable distance metric. Biehl et al. (2016) provide a comprehensive overview of prototype-based models, discussing extensions like relevance learning, which adapts the distance metric to emphasize relevant features for classification. In the context of few-shot learning, where the goal is to classify instances with limited labeled examples, Prototypical Networks have gained prominence. Snell et al. (2017) propose learning a metric space where classification is performed by computing distances to class prototypes, defined as the mean of embedded support examples. Prototype-based methods are particularly advantageous in scenarios with limited data. By summarizing each class with representative prototypes, these methods can generalize effectively even when only a few labeled examples are available.

2.3. Wasserstein Barycenters in Machine Learning

The Wasserstein distance, rooted in optimal transport theory, provides a meaningful metric for comparing probability distributions by considering the geometry of the underlying space.

The *2-Wasserstein distance* W_2 between two probability measures μ and ν on a metric space (M, d) is defined as:

$$W_2^2(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} d(x, y)^2 \, d\gamma(x, y)$$

Here: $\Gamma(\mu, \nu)$ denotes the set of all couplings (i.e., joint distributions) with marginals μ and ν; $d(x, y)$ is the distance between points x and y in the space M.

Wasserstein barycenters extend this concept by offering a method to compute the "average" of multiple probability distributions, preserving the intrinsic geometric relationships among them. Such barycenters have been effectively utilized as prototypes in various machine learning applications.

Given a set of probability measures $\mu_1, \mu_2, \ldots, \mu_N$ defined on a metric space (M, d), and associated non-negative weights $\lambda_1, \lambda_2, \ldots, \lambda_N$ summing to 1, the *2-Wasserstein barycenter* ν^* is defined as the probability measure that minimizes the weighted sum of squared 2-Wasserstein distances to the given measures:

$$\nu^* = \arg \min_{\nu \in \mathcal{P}_2(M)} \sum_{i=1}^{N} \lambda_i W_2^2(\nu, \mu_i)$$

Here: $\mathcal{P}_2(M)$ denotes the set of probability measures on M with finite second moments. $W_2(\nu, \mu_i)$ represents the 2-Wasserstein distance between ν and μ_i.

In the context of prototype-based classification, utilizing Wasserstein barycenters allows for the aggregation of class-specific data distributions into a single representative prototype. This approach is particularly beneficial in scenarios with limited data, as it provides a robust summary that accounts for the variability within each class. Furthermore, the Wasserstein barycenter's sensitivity to the underlying geometry of the data ensures that the prototypes maintain the structural relationships present in the original distributions. Zen & Ricci (2011) use Earth Mover's Distance based prototypes which are actually Wasserstein barycenters.

3. Method

3.1. Problem Setting

We examine a binary classifier $m : \mathbb{R}^d \to [0, 1]$ that processes input vectors $\boldsymbol{x} \in \mathbb{R}^d$ and generates probability scores. The final classification decision $\hat{y}$ is made by applying a 0.5 threshold to the model's output probability: $\hat{y}_m = \mathbb{I}_{[0,1]}[m(\boldsymbol{x}) \geq 0.5]$, where $\mathbb{I}_{[0,1]}[\cdot]$ is the binary indicator function. The model is accompanied by a counterfactual generator g_m that creates perturbed examples lying close to the classification boundary. In our framework, we assume access to a pre-trained target model m that responds to user queries with prediction outputs. The counterfactual mechanism g_m is specifically activated for cases where m

predicts class 0 (i.e., when $m(\boldsymbol{x}) < 0.5$). The primary objective is to learn an approximation model $\hat{m}$ that replicates the target model's behavior while minimizing the number of necessary queries. To reconstruct the model m, we reformulate the problem over a feature space $\mathcal{X} \subseteq \mathbb{R}^d$ and a structured label space $\mathcal{Y} = \{0, 0.5, 1\}$. Here, label 0 corresponds to samples from class 0, label 1 to samples from class 1, and label 0.5 represents counterfactual or ambiguous examples that blend characteristics of both classes. These counterfactuals are assumed to lie near the boundary and are treated as soft samples of both classes. The dataset for class 0 is denoted by $\mathcal{D}_0 = \{(x_i^{(0)}, 0)\}_{i=1}^{N_0}$, where $x_i^{(0)} \sim \mathbb{P}_0$. Similarly, the dataset for class 1 is $\mathcal{D}_1 = \{(x_i^{(1)}, 1)\}_{i=1}^{N_1}$, with $x_i^{(1)} \sim \mathbb{P}_1$. The counterfactual dataset is given by $\mathcal{D}_{cf} = \{(x_i^{(cf)}, 0.5)\}_{i=1}^{N_{cf}}$, where $x_i^{(cf)} \sim \mathbb{P}_{cf}$ represents samples that mix features from both classes.

3.2. Model Reconstruction using Wasserstein Barycenters as Prototypes

To represent each class in a way that incorporates both original and counterfactual samples, we compute a soft prototype for each class using the Wasserstein barycenter. For every class $c \in \{0, 1\}$, we define a barycenter distribution $\mathbb{Q}_c$ as the solution to the following optimization problem:

$$\mathbb{Q}_c = \arg\min_{\mathbb{Q} \in \mathbb{P}(\mathcal{X})} \left(W_2^2(\mathbb{Q}, \mathbb{P}_c) + \lambda_c W_2^2(\mathbb{Q}, \mathbb{P}_{cf}) \right) \quad (1)$$

Here, $W_2^2(\cdot, \cdot)$ denotes the squared 2-Wasserstein distance between probability distributions, which measures the optimal transport cost of transforming one distribution into another. Intuitively, it captures both the amount of mass that needs to be moved and the distance over which it must be moved, making it a meaningful geometric measure of similarity between distributions. The coefficient $\lambda_c = 0.5$ balances the influence of the original class distribution $\mathbb{P}_c$ and the counterfactual distribution $\mathbb{P}_{cf}$ in determining the barycenter $\mathbb{Q}_c$.

To ensure that the counterfactual samples are symmetrically situated between the class prototypes, we introduce a regularization term:

$$\mathcal{R}(\mathbb{Q}_0, \mathbb{Q}_1) = \left(W_2(\mathbb{Q}_0, \mathbb{P}_{cf}) - W_2(\mathbb{Q}_1, \mathbb{P}_{cf}) \right)^2 \quad (2)$$

This term penalizes asymmetry in the distances from the counterfactual distribution to the class barycenters, encouraging the counterfactuals to be equidistant from both sides of the decision boundary.

The overall objective function combines the class-wise barycenter losses with the symmetry regularization:

$$\min_{\mathbb{Q}_0, \mathbb{Q}_1} \sum_{c \in \{0,1\}} \left(W_2^2(\mathbb{Q}_c, \mathbb{P}_c) + \lambda_c W_2^2(\mathbb{Q}_c, \mathbb{P}_{cf}) \right) + \gamma \mathcal{R}(\mathbb{Q}_0, \mathbb{Q}_1) \quad (3)$$

Here, $\gamma > 0$ is a regularization coefficient that controls the strength of the symmetry constraint. The result is a pair of barycentric prototypes that not only reflect their respective class distributions but also respect the structure of the decision boundary as implied by the counterfactuals.

Given the learned barycenters $\mathbb{Q}_0$ and $\mathbb{Q}_1$, we define a classification rule for assigning labels to new inputs. For a test sample $x \in \mathcal{X}$, we compare its distance to the two barycenters using the Wasserstein-2 metric. Let δ_x denote the Dirac delta distribution centered at x. Then, the predicted label $\hat{y}(x)$ is given by:

$$\hat{y}_{\hat{m}}(x) = \begin{cases} 0 & \text{if } W_2(\delta_x, \mathbb{Q}_0) < W_2(\delta_x, \mathbb{Q}_1) - \tau \\ 1 & \text{if } W_2(\delta_x, \mathbb{Q}_1) < W_2(\delta_x, \mathbb{Q}_0) - \tau \end{cases} \quad (4)$$

The threshold parameter $\tau \geq 0$ introduces a margin to avoid overly confident predictions near the decision boundary.

We employ fidelity (Aïvodji et al., 2020) as a metric to evaluate the performance of model reconstruction. Given a target model m and a reference dataset $\mathcal{D}_{\text{ref}}$, the fidelity of a surrogate model $\hat{m}$ is defined as

$$\text{Fid}_{m, \mathcal{D}_{\text{ref}}}(\hat{m}) = \frac{1}{|\mathcal{D}_{\text{ref}}|} \sum_{\boldsymbol{x} \in \mathcal{D}_{\text{ref}}} \mathbb{I}_{[0,1]} \left[\hat{y}_m(\boldsymbol{x}) = \hat{y}_{\hat{m}}(\boldsymbol{x}) \right].$$

4. Experiment

We conduct a series of experiments to evaluate the effectiveness of our proposed method for reconstructing binary classifiers using Wasserstein barycenters as class prototypes. We update our Wasserstein barycenters with the method *Barycenters with free support* in python package POT (Flamary et al., 2021). We compare our approach against two SOTA methods: the first is an attack technique for classifier reconstruction without primer knowledge to training data, as described in Aïvodji et al. (2020), referred to as *Baseline 1* where counterfactuals are treated as normal samples; the second is a method that modifies the standard entropy loss to incorporate counterfactual explanations, proposed by Dissanayake & Dutta (2024), referred to as *Baseline 2*. The target classifiers are trained using logistic regression on a training dataset $\mathcal{D}_{\text{train}}$, which remains unknown during the reconstruction phase. To generate initial one-sided counterfactuals, we adopt the Minimum Cost Counterfactuals (MCCF) method (Wachter et al., 2017). We evaluate reconstruction performance using the *fidelity* metric described in Section 3.2, which measures the agreement between the predictions of the target model and the reconstructed (surrogate) model. Fidelity is assessed over a set of test samples drawn from the original data manifold, and used as the reference dataset $\mathcal{D}_{\text{ref}}$. Experiments are conducted on four datasets: Adult Income, COMPAS, DCCC, and HELOC (see Appendix for details). Table 1 summarizes the *fidelity* results

Table 1. Average fidelity with 500, 400, 300 queries on the datasets.

	Query Size (500)			Query Size (400)			Query Size (300)		
	Baseline 1	Baseline 2	Ours	Baseline 1	Baseline 2	Ours	Baseline 1	Baseline 2	Ours
Adult In.	91 ± 3.2	94 ± 3.2	96 ± 2.5	89 ± 3.5	92 ± 3.5	94 ± 2.8	87 ± 3.8	90 ± 3.8	93 ± 3.2
COMPAS	92 ± 3.2	94 ± 2.0	96 ± 2.3	90 ± 3.5	92 ± 2.3	94 ± 2.6	88 ± 3.8	90 ± 2.6	94 ± 3.0
DCCC	89 ± 8.9	91 ± 0.9	97 ± 1.5	87 ± 9.2	89 ± 1.2	95 ± 1.8	85 ± 9.5	87 ± 1.5	93 ± 2.1
HELOC	91 ± 4.7	93 ± 2.2	95 ± 2.0	89 ± 5.0	91 ± 2.5	93 ± 2.3	87 ± 5.3	89 ± 2.8	93 ± 2.6

across these datasets. In all cases, our method achieves either superior or comparable fidelity relative to Baseline 1 and Baseline 2. Notably, our approach demonstrates a clear advantage when the query size is small, highlighting its efficiency in low-query regimes.

Other Architectures for Reconstruction: We also compare our method to Baseline 2 using neural network surrogate models of varying complexity, as shown in Table 3 in the Appendix. As the neural network architecture in Baseline 2 becomes more complex, our method consistently maintains strong performance, whereas Baseline 2 exhibits noticeable degradation. This is particularly striking given that more complex neural networks theoretically have greater capacity to approximate arbitrary functions. The decline in Baseline 2's performance is especially evident under limited query budgets, where its reliance on training a neural network surrogate leads to overfitting due to insufficient data.

Other Counterfactual Generation Techniques: We investigate the impact of various counterfactual generation methods, focusing on attributes such as sparsity, actionability, realism, and robustness. To generate sparse counterfactuals, we employ an L_1-norm as the cost function, promoting minimal feature changes. For actionable counterfactuals, we utilize the DiCE framework (Mothilal et al., 2020), which allows the specification of immutable features to ensure feasibility. Realistic counterfactuals, those that lie close to the data manifold, are produced by identifying the nearest neighbor from the desired class and by applying the C-CHVAE method (Pawelczyk et al., 2020), which leverages variational autoencoders. To enhance robustness, we generate counterfactuals using the ROAR approach (Upadhyay et al., 2021), designed to maintain validity under model shifts. We assess the effectiveness of these methods through attack performance evaluations on the Adult dataset (Table 2). The quality of counterfactual examples significantly influences the fidelity of model reconstruction. High-quality counterfactuals—those that are plausible and lie close to the data manifold enable surrogate models to more accurately approximate the decision boundaries of the original model. C-CHVAE does not perform well, which we attribute to the generally weaker performance of generative models on tabular datasets.

Table 2. Fidelity with different counterfactual methods on Adult dataset

	Query Size (500)	**Query Size (400)**	**Query Size (300)**
MCCF	96	94	93
DiCE	96	94	92
1-Nearest-Neightbor	97	91	89
ROAR	95	93	91
C-CHVAE	83	79	77

5. Discussion and Future Work

This work demonstrates that Wasserstein barycenters provide a robust framework for classifier reconstruction, particularly in scenarios with limited data and the presence of counterfactual examples. By integrating optimal transport theory, our approach captures nuanced relationships between data distributions, enabling the formation of soft prototypes that effectively represent both labeled data and counterfactuals. This representation-centric perspective is particularly beneficial when dealing with small or noisy samples, where traditional methods may falter. Analyses reveal that our method maintains high fidelity in model reconstruction. Notably, in small data regimes, where overfitting and poor generalization are prevalent concerns, our approach exhibits superior stability and flexibility compared to baseline methods. This performance underscores the significance of incorporating geometric and distributional considerations into the reconstruction process.

Future work should aim to quantitatively investigate the impact of sample size on the fidelity of model reconstruction. Additionally, it would be valuable to explore alternative representations of the dataset—for instance, using prototype representations other than the barycenter—to potentially improve performance. Our current approach assumes no prior knowledge of the original model; however, incorporating such prior information could substantially enhance reconstruction quality. Advancing research in these directions will contribute to developing more data-efficient and interpretable model extraction techniques for real-world applications.

References

Aïvodji, U., Bolot, A., and Gambs, S. Model extraction from counterfactual explanations. *CoRR*, abs/2009.01884, 2020.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Propublica compas recidivism risk score data and analysis, 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barocas, S., Selbst, A. D., and Raghavan, M. The hidden assumptions behind counterfactual explanations and principal reasons. In *FAT**, pp. 80–89. ACM, 2020.

Becker, B. Adult data set. https://archive.ics.uci.edu/dataset/2/adult, 1996.

Biehl, M., Hammer, B., and Villmann, T. Prototype-based models for the supervised learning of classification schemes. *Proceedings of the International Astronomical Union*, 12(S325):129–138, 2016.

Dissanayake, P. and Dutta, S. Model reconstruction using counterfactual explanations: A perspective from polytope theory. In *NeurIPS*, 2024.

FICO. Explainable machine learning challenge dataset, 2018. URL https://community.fico.com/s/explainable-machine-learning-challenge.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Gong, X., Wang, Q., Chen, Y., Yang, W., and Jiang, X. Model extraction attacks and defenses on cloud-based machine learning models. *IEEE Commun. Mag.*, 58(12): 83–89, 2020.

Guidotti, R. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Min. Knowl. Discov.*, 38(5):2770–2824, 2024.

Mothilal, R. K., Sharma, A., and Tan, C. Explaining machine learning classifiers through diverse counterfactual explanations. In *FAT**, pp. 607–617. ACM, 2020.

Pawelczyk, M., Broelemann, K., and Kasneci, G. Learning model-agnostic counterfactual explanations for tabular data. In *WWW*, pp. 3126–3132. ACM / IW3C2, 2020.

Shokri, R., Strobel, M., and Zick, Y. On the privacy risks of model explanations. In *AIES*, pp. 231–241. ACM, 2021.

Snell, J., Swersky, K., and Zemel, R. S. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Upadhyay, S., Joshi, S., and Lakkaraju, H. Towards robust and reliable algorithmic recourse. In *NeurIPS*, pp. 16926–16937, 2021.

Van Looveren, A. and Klaise, J. Interpretable counterfactual explanations guided by prototypes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):9566–9574, 2021.

Verma, S., Boonsanong, V., Hoang, M., Hines, K., Dickerson, J., and Shah, C. Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Comput. Surv.*, 56(12):312:1–312:42, 2024.

Wachter, S., Mittelstadt, B. D., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017.

Wang, Y., Qian, H., and Miao, C. Dualcf: Efficient model extraction attack from counterfactual explanations. In *FAccT*, pp. 1318–1329. ACM, 2022.

Yeh, I.-C. Default of credit card clients dataset, 2016. URL https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

Zen, G. and Ricci, E. Earth mover's prototypes: A convex learning approach for discovering activity patterns in dynamic scenes. In *CVPR*, pp. 3225–3232. IEEE Computer Society, 2011.

A. Experimental Setup and Supplementary Findings

A.1. Description of Real-World Benchmark Datasets

To assess the fedelity, we employed four publicly available tabular datasets: Adult Income, COMPAS, DCCC, and HELOC. Below are their key characteristics:

Adult Income (Becker, 1996): Derived from the 1994 U.S. Census, this dataset captures demographic and financial attributes such as education level, marital status, age, and annual earnings. The classification task involves predicting whether an individual's income exceeds \$50,000 (denoted as $y = 1$). The original dataset consists of 32,561 entries, with 24,720 labeled as $y = 0$ and 7,841 as $y = 1$. To balance the classes, we randomly selected 7,841 samples from $y = 0$, resulting in a final dataset of 15,682 entries. The dataset includes 6 numerical and 8 categorical features, with the latter converted to integer encodings. All features were normalized to $[0, 1]$.

Home Equity Line of Credit (HELOC): This dataset records credit risk assessments for customers seeking home equity loans (FICO, 2018). It comprises 10,459 entries, each with 23 numerical features. The prediction target, "is_at_risk," identifies customers likely to default. The dataset is moderately imbalanced, with 5,000 samples for $y = 0$ and 5,459 for $y = 1$. For our experiments, we used a subset of 10 key features, including "estimate_of_risk," "net_fraction_of_revolving_burden," and "percentage_of_legal_trades," all scaled to $[0, 1]$.

COMPAS: Developed to study racial bias in recidivism prediction algorithms, this dataset contains 6,172 entries with 20 numerical features (Angwin et al., 2016). The target variable, "is_recid," divides the data into 3,182 ($y = 0$) and 2,990 ($y = 1$) samples. Feature values were normalized to $[0, 1]$.

Default of Credit Card Clients (DCCC): This dataset tracks credit card payment behaviors in Taiwan (Yeh, 2016), with the goal of predicting defaults ("default.payment.next.month"). The original dataset has 30,000 entries (23,364 for $y = 0$, 6,636 for $y = 1$). To address class imbalance, we randomly subsampled 6,636 instances from $y = 0$. Categorical features were integer-encoded, and all attributes were normalized to [0,1].

A.2. Experiment Details

All experiments were conducted on a machine equipped with an NVIDIA RTX 3090 GPU. The regularization coefficient γ is set to be 0.3.

Table 3. Average fidelity with 500, 400, 300 queries on the datasets. Baseline 2.1 has hidden layers with neurons (20, 10, 5) and Baseline 2.2 has hidden layers with neurons (20, 10)

	Query Size (500)			Query Size (400)			Query Size (300)		
	Baseline 2.1	Baseline 2.2	Ours	Baseline 2.1	Baseline 2.2	Ours	Baseline 2.1	Baseline 2.2	Ours
Adult In.	92 ± 4.1	94 ± 3.2	96 ± 2.5	87 ± 3.9	92 ± 3.5	94 ± 2.8	80 ± 3.8	90 ± 3.8	93 ± 3.2
COMPAS	90 ± 3.0	94 ± 2.0	96 ± 2.3	86 ± 3.9	92 ± 2.3	94 ± 2.6	82 ± 5.8	90 ± 2.6	94 ± 3.0
DCCC	88 ± 8.3	91 ± 0.9	97 ± 1.5	84 ± 7.2	89 ± 1.2	95 ± 1.8	81 ± 5.5	87 ± 1.5	93 ± 2.1
HELOC	89 ± 4.5	93 ± 2.2	95 ± 2.0	87 ± 3.1	91 ± 2.5	93 ± 2.3	82 ± 3.3	89 ± 2.8	93 ± 2.6