TESTING UNCERTAINTY-AWARE CONDITIONAL REPRESENTATION CONSISTENCY FOR TABULAR COUNTERFACTUALS

Anonymous authors
Paper under double-blind review

Tabular data are widely used in healthcare, finance, science, and other decision-making applications. However, learning robust representations from tabular data is difficult because a small feature perturbation can be either a meaningful variation or an unrealistic sample: standard consistency regularization does not distinguish these cases and may force a representation to become invariant to changes that should alter the prediction. This paper tests a conditional alternative: representation consistency gated by three signals, a distance-based counterfactual validity score v, a predictive-confidence weight w_u, and a binary prediction-preservation indicator s so that plausible, confident, prediction-preserving perturbations are encouraged to share a representation with the original input, while perturbations that fail any of these checks are not. We instantiate this idea in a framework we call CONSENT and ask a direct empirical question: does gating consistency on these three signals outperform applying consistency unconditionally? We first report a single-seed, five-epoch benchmark across three tabular datasets, where CONSENT shows mixed predictive performance relative to a baseline MLP. Because a single seed cannot separate a real effect from noise, we then run a multi-seed study (5 seeds, 30 epochs) on one dataset, comparing the plain baseline, full CONSENT, an unconditional-consistency variant, three lesioned variants each removing one gating signal, and a gradient-boosted tree. Under this protocol, all conditions are statistically indistinguishable on every predictive and calibration metric. The answer to our central question is therefore negative in the tested regime: the apparent gains in the single-seed benchmark reflect training noise, and gating consistency does not yet demonstrate a measurable advantage over unconditional consistency or a plain MLP let alone a gradient-boosted tree, which matches or exceeds every neural variant. We report this test and its outcome explicitly and identify next steps: extending the multi-seed protocol to the remaining datasets, replacing the validity heuristic with a learned model, and testing whether reweighted loss terms allow conditional and unconditional consistency to diverge.

1 INTRODUCTION

Tabular prediction is a central machine learning problem. Many practical datasets contain heterogeneous variables such as continuous measurements, categorical attributes, counts, and binary indicators(5; 6). Deep learning models for tabular data have become increasingly competitive with traditional approaches, including attention-based architectures and transformer models (1; 2). Despite this progress, robust representation learning for heterogeneous tables remains challenging.

A major reason is that the meaning of an input perturbation is highly dependent on the data domain. A small change in an image often preserves its semantic identity, but changing one tabular feature can violate dependencies between variables or create an implausible observation(7). Consequently, unconditional consistency regularization can be harmful: a model may be encouraged to produce identical representations for two samples even when the perturbation should change the prediction.

This paper tests a conditional alternative. We introduce CONSENT, an uncertainty-aware counterfactual representation learning framework in which representation consistency depends on three signals: candidate validity, predictive confidence, and prediction preservation. We call the framework CONSENT because two representations are only asked to agree to *consent* to alignment when the perturbation between them is valid, the model is confident about both, and the predicted class is preserved; if any of these conditions fails, the consistency constraint is withdrawn rather than forcing the representations to become identical. The central question this paper asks is empirical, not rhetorical: does gating consistency on these three signals actually outperform applying consistency unconditionally, as the conditional-invariance intuition would predict? We treat this as a hypothesis

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

to be tested against a proper baseline rather than a claim to be illustrated, and we report the outcome of that test even where it does not favor the proposed mechanism.

The current implementation is intentionally lightweight. It uses a multilayer perceptron encoder and stochastic local perturbations of numerical features. A distance-based score is used as a proxy for candidate validity. This makes the experimental pipeline easy to reproduce and allows us to isolate the proposed conditional consistency principle.

The contributions of this work are

1. We formalize a direct, falsifiable test of whether gating representation consistency on validity, confidence, and prediction preservation outperforms unconditional consistency regularization for tabular counterfactual candidates.
2. We implement a reproducible framework, CONSENT, that instantiates this gated consistency and operates on both purely numerical and mixed numerical/categorical tabular datasets after preprocessing.
3. We report an initial single-seed, five-epoch benchmark on three datasets, and then a multi-seed (5-seed, 30-epoch) study on breast cancer that adds the tests the single-seed benchmark lacked: an unconditional-consistency ablation, three lesioned variants that each remove one gating signal, and a gradient-boosted tree baseline.
4. We show that under the multi-seed protocol, the test comes back negative: it CONSENT is statistically indistinguishable from the plain baseline, from unconditional consistency, and from each of its own lesioned variants, and we report this null result explicitly rather than relying on the single-seed numbers.

Deep representation learning for tabular data. Deep learning for tabular data has developed along several directions, including feature-selection mechanisms, learned feature embeddings, and transformer-based architectures. TabNet uses sequential attention to identify informative features and construct instance-dependent representations (1), while FT-Transformer adapts transformer architectures to heterogeneous tabular inputs by learning representations for numerical and categorical features (2). Other approaches have explored self-supervised and cross-table pretraining to improve the transferability of tabular representations (**?**). These methods primarily optimize predictive performance or representation quality under the original data distribution. In contrast, CONSENT focuses on the behavior of a learned representation under structured perturbations, asking when two nearby or counterfactually generated tabular instances should be encouraged to have consistent representations.

A further consideration is the strength of non-neural baselines for tabular prediction. Gradient-boosted decision trees, particularly XGBoost (4), remain widely used and highly competitive on heterogeneous tabular datasets. Consequently, improvements from representation-learning approaches should be evaluated not only against a plain neural baseline but also against strong tree-based models. We therefore include a gradient-boosted tree baseline in our extended evaluation (Section 5).

Consistency learning and representation invariance. Consistency-based learning encourages a model to produce stable predictions or representations for inputs that are regarded as related, equivalent, or semantically consistent. This principle is common in data augmentation and self-supervised learning, where transformations of the same underlying example are used to encourage invariant representations. However, applying this idea to tabular data introduces an important challenge: not every perturbation that is numerically close to an original instance is necessarily semantically equivalent. Changes to a categorical attribute may correspond to a different population, while changes to correlated numerical features may produce combinations that are unlikely under the data distribution.

CONSENT therefore does not impose representation consistency uniformly across all generated candidates. Instead, it uses candidate validity, predictive behavior, and model uncertainty to determine whether a candidate should contribute to the consistency objective. This conditional formulation is intended to distinguish perturbations that are sufficiently reliable for alignment from those for which enforcing invariance may be inappropriate.

Counterfactual explanations for tabular data. Counterfactual explanations describe how an input could be modified to obtain a different model prediction and have become an important approach to explaining black-box models. Existing counterfactual methods commonly consider properties such as validity, proximity, sparsity, feasibility, actionability, and plausibility (**? ? ?**). Recent work has further emphasized the importance of remaining close to the data manifold and avoiding implausible or unrealistic counterfactual instances (**? ?**). These considerations are particularly important for tabular data because numerical and categorical variables can have strong dependencies and domain-specific constraints. Recent approaches such as TABCF explicitly model these dependencies using a transformer-based variational autoencoder to generate more realistic counterfactuals for mixed-type tabular data (**?**).

In CONSENT, generated samples are used primarily as structured perturbations for representation learning rather than as explanations presented directly to an end user. We therefore refer to them as *local counterfactual candidates*. This distinction is important: although the candidates are constructed to examine changes in model behavior, they should not be interpreted as causal counterfactuals. Our current implementation does not employ a structural causal model or explicit causal assumptions and consequently cannot establish whether a generated intervention corresponds to a causally valid alternative outcome. This limitation is consistent with the broader distinction between model-based counterfactual explanations and causal counterfactual reasoning (**?**).

Uncertainty and consistency control. Predictive uncertainty provides another mechanism for deciding how strongly a model should trust or act upon a particular prediction. Modern neural networks can achieve strong predictive discrimination while producing poorly calibrated confidence estimates (3). More generally, uncertainty estimation in neural networks distinguishes between different sources of uncertainty and includes approaches ranging from Bayesian methods to ensemble-based approximations (**? ?**). These methods demonstrate that predictive confidence and predictive correctness should not be treated as interchangeable quantities.

Our objective is not to construct a fully calibrated uncertainty estimator. Instead, CONSENT uses normalized predictive entropy as a simple, model-derived uncertainty signal. For a classification model with predictive distribution $\mathbf{p}(x)$, entropy increases as the probability mass becomes more distributed across competing classes and decreases when the model assigns most of its probability to a single class. Normalizing the entropy by the logarithm of the number of classes gives a value in $[0, 1]$, allowing the signal to be used consistently across classification tasks with different numbers of classes. We use this quantity to modulate the strength of representation consistency: candidates associated with more uncertain predictions receive less unconditional trust for alignment, while candidates whose predictive behavior is more decisive can provide a stronger consistency signal. Importantly, this mechanism should be interpreted as uncertainty-aware regularization rather than calibrated probabilistic uncertainty.

Positioning of CONSENT. The above lines of work address complementary aspects of tabular learning. Tabular representation-learning methods seek useful predictive representations; consistency-based methods encourage invariance across related inputs; counterfactual methods construct structured alternatives to observed instances; and uncertainty estimation provides information about the reliability of model predictions. CONSENT combines these ideas by treating counterfactual candidates as potential sources of representation consistency while using candidate validity, predictive preservation, and uncertainty to determine when that consistency should be enforced. The resulting objective is therefore conditional rather than unconditional: representation alignment is applied only when the generated candidate provides sufficient evidence that the original and perturbed instances should behave consistently.

2 METHOD

Figure 1 gives a schematic overview of CONSENT before the formal treatment in the rest of this section: the original input x and a perturbed candidate x^{cf} pass through the same shared-weight encoder and classifier, their representations and predictions are used to compute the three gating signals (v, w_u, s) and the representation distance D, and these combine into the representation-consistency loss $\mathcal{L}_{rep}$, the counterfactual prediction loss $\mathcal{L}_{cf}$, and the overall training objective $\mathcal{L}$.

2.1 Problem Formulation

Let

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}, \qquad x_i \in \mathbb{R}^d, \quad y_i \in \{0, 1\},$$

be a binary classification dataset, where d is the number of input features after preprocessing (standardized numerical features concatenated with one-hot encoded categorical features). An encoder $f_\theta : \mathbb{R}^d \to \mathbb{R}^k$ produces a representation

$$z = f_\theta(x)$$

and a classifier head $h_\phi : \mathbb{R}^k \to \mathbb{R}$ together with a sigmoid produce a probability

$$p(x) = \sigma(h_\phi(z)).$$

In the executed prototype, f_θ is a two-layer multilayer perceptron with a 128-unit ReLU hidden layer, dropout (rate 0.20), and a 64-dimensional ReLU output layer ($k = 64$), and h_ϕ is a single linear layer mapping $\mathbb{R}^{64} \to \mathbb{R}$. This architecture is intentionally minimal: it is the smallest encoder for which the representation z is non-trivial and the counterfactual objectives in this section are non-degenerate, which keeps the experiment focused on the consistency mechanism itself rather than on architecture search. The formulation above is written for binary classification to match the three benchmark datasets; extending h_ϕ to a softmax head and $s(\cdot, \cdot)$ (Section 3.5) to an argmax-agreement indicator is mechanical but was not implemented or tested.

For each input x, a local perturbed candidate is generated by a stochastic generator G:

$$x^{cf} = G(x, \xi), \qquad \xi \sim p(\xi),$$

where ξ collects the randomness used by G (Section 3.2). A fresh ξ is drawn independently for every training example in every minibatch, so x^{cf} is not cached across epochs and acts as a stochastic augmentation rather than a fixed second view. We use the notation x^{cf} for convenience, but do not claim that the generated sample is causally valid; Section 3.3 introduces a heuristic weight precisely to discount this assumption rather than to assert it.

Table **??** in Appendix **??** summarizes the symbols used throughout this section for reference.

2.2 Counterfactual Candidate Generation

Write $x = [x_{\text{num}}; x_{\text{cat}}]$ for the concatenation of the standardized numerical block and the one-hot categorical block. For numerical features, the prototype applies a stochastic perturbation to each coordinate j independently:

$$x_j^{cf} = x_j + m_j \epsilon_j,$$

where

$$m_j \sim \text{Bernoulli}(r), \qquad \epsilon_j \sim \mathcal{N}(0, \sigma^2).$$

The Bernoulli mask m_j controls *sparsity*: on average only a fraction r of numerical features are altered in any given candidate, so most candidates change a small subset of coordinates rather than the entire feature vector. This mirrors a common desideratum in the counterfactual-explanation literature that a plausible counterfactual should differ from the original instance in as few features as possible, although here it is used only as a generator constraint, not as an explicit sparsity objective on the loss. The Gaussian term ϵ_j then perturbs the selected coordinates by an amount calibrated in standardized units, so that a single scalar σ is comparable across features regardless of their raw scale. We set $r = 0.20$ and $\sigma = 0.15$; the categorical block is left unchanged, $x_{\text{cat}}^{cf} = x_{\text{cat}}$, because perturbing a one-hot vector requires a discrete resampling rule (e.g., switching a category with some transition probability).

Two consequences follow: G resamples m_j, ϵ_j fresh each minibatch, so the model never sees a fixed pair repeated across training; and because $x_{\text{cat}}^{cf} = x_{\text{cat}}$ exactly, any measured effect on Adult and Covertype is attributable entirely to the numerical subset of features, a smaller fraction of the input than on the purely numerical Breast Cancer.

2.3 VALIDITY WEIGHT

We define a soft validity proxy over the numerical block only:

$$v(x, x^{cf}) = \exp\left(-\|x^{cf}_{\text{num}} - x_{\text{num}}\|_2\right).$$

This value is close to one for small changes and decreases monotonically (exponentially) as the numerical perturbation becomes larger, so that $v \in (0, 1]$ for any perturbation. The exponential form is chosen for two practical reasons rather than a modeled notion of plausibility: it is smooth and bounded, and it decays quickly enough that candidates produced by large, low-probability draws of ϵ_j contribute little to the objectives below without needing an explicit threshold or clipping rule. We emphasize, as in Section 1, that v is a distance heuristic and not a learned or calibrated probability of data validity: a perturbation can have high v (small norm) while still landing off the data manifold or violating a dependency between features (e.g., an implausible combination of two correlated numerical attributes), since v has no access to the joint feature distribution beyond the single perturbed point. In the executed implementation, v is computed from x^{cf} and treated as a fixed coefficient with no gradient flowing back through it into θ; only $D(z, z^{cf})$ and the counterfactual prediction loss (Sections 3.5–3.6) receive gradients with respect to the encoder and classifier parameters.

2.4 UNCERTAINTY WEIGHT

For binary classification, normalized entropy is

$$U(x) = \frac{-p(x)\log p(x) - (1-p(x))\log(1-p(x))}{\log 2} \in [0, 1],$$

computed with probabilities clamped away from $\{0, 1\}$ to avoid numerical overflow in $\log$. $U(x) = 0$ when the classifier is maximally confident ($p(x) \in \{0, 1\}$) and $U(x) = 1$ at the decision boundary ($p(x) = 0.5$). For an original/candidate pair, confidence is

$$w_u(x, x^{cf}) = 1 - \frac{U(x) + U(x^{cf})}{2} \in [0, 1].$$

The intuition is that representation consistency should be encouraged only when the model is confident about *both* the original and the perturbed input; if either one lies near the decision boundary, forcing their representations together could push a genuinely ambiguous case toward an arbitrary side of the boundary rather than reflecting a meaningful invariance. Like v, the confidence weight is detached from the computation graph, so it modulates the scale of the loss term without directly shaping the gradient through $U(\cdot)$ itself; this avoids an additional feedback loop in which the model could lower w_u (and thus its own regularization pressure) simply by becoming more uncertain.

2.5 PREDICTION-CONDITIONAL REPRESENTATION CONSISTENCY

Let

$$z = f_\theta(x), \qquad z^{cf} = f_\theta(x^{cf}).$$

Both are produced by the *same* encoder f_θ with shared weights; x and x^{cf} are simply two forward passes through one network, so no separate "counterfactual encoder" is introduced. The cosine representation distance is

$$D(z, z^{cf}) = 1 - \frac{z^\top z^{cf}}{\|z\|_2 \|z^{cf}\|_2} \in [0, 2],$$

chosen over a Euclidean distance because it is invariant to the norm of z and z^{cf}, so the penalty targets the *direction* of the representation rather than its magnitude, which can otherwise be driven down trivially by shrinking the encoder's output norm.

Define the prediction-preserving indicator:

$$s(x, x^{cf}) = \mathbb{I}\left[\text{round}(p(x)) = \text{round}(p(x^{cf}))\right] \in \{0, 1\}.$$

s is a non-differentiable step function of $p(x)$ and $p(x^{cf})$ and is treated purely as a binary mask multiplying the loss; like v and w_u it carries no gradient of its own, so it cannot be "gamed" by the optimizer moving predictions across the rounding threshold to relax the constraint.

The representation objective is

$$\mathcal{L}_{rep} = \mathbb{E}\left[v\, w_u\, s\, D(z, z^{cf})\right],$$

with the expectation taken over the empirical training distribution (in practice, a Monte Carlo average over one sampled x^{cf} per x per minibatch). Thus, consistency is applied only when the perturbation is sufficiently valid, the model is confident, and the predicted class is preserved; whenever any of the three gates is near zero, $\mathcal{L}_{rep}$ contributes little or no pressure to align z and z^{cf}, and the encoder is free to represent x and x^{cf} differently.

Relation to unconditional consistency. Setting $v = w_u = s \equiv 1$ for every pair recovers a standard *unconditional* consistency-regularization term, $\mathcal{L}_{rep}^{\text{unc}} = \mathbb{E}[D(z, z^{cf})]$, which penalizes representation drift regardless of whether the perturbation is plausible, whether the model is confident, or whether the prediction actually changed. This is precisely the comparison we treat as the central hypothesis test of the paper: CONSENT is only a meaningful improvement over unconditional consistency if the gated objective outperforms this degenerate case in which all three gates are held fixed at 1. Section 6 reports this comparison directly.

Computational overhead. Relative to the plain baseline, CONSENT roughly doubles the forward-pass cost of a training step (one extra pass through f_θ, h_ϕ per minibatch) and does not change inference cost, since only f_θ and h_ϕ are used at test time; Appendix **??** gives the full accounting.

2.6 Counterfactual Prediction Loss

The candidate also contributes a weighted supervised loss:

$$\mathcal{L}_{cf} = \mathbb{E}\left[v\, \ell(h_\phi(z^{cf}), y)\right],$$

where ℓ is binary cross-entropy. This term is gated only by the validity weight v, not by w_u or s. We gate by v for the same reason as in $\mathcal{L}_{rep}$: an implausible candidate (large numerical perturbation) should not be forced to satisfy the original label. We deliberately do *not* gate this term by s: doing so would multiply the loss by zero exactly whenever the model's own prediction on x^{cf} disagrees with y, which would only reinforce candidates the model already classifies correctly and provide no learning signal on the (arguably most informative) candidates where the model's prediction is currently wrong or has flipped. We also do not gate by w_u, since down-weighting the supervised term precisely when the model is uncertain would remove supervision exactly where it is most needed. The prototype reuses the observed label y for the locally perturbed sample. This is an augmentation assumption rather than a causal counterfactual target: it is only appropriate insofar as a small, valid perturbation is assumed not to change the true label, an assumption that becomes less defensible as the perturbation magnitude grows, which is why v (not s) is the appropriate gate here.

2.7 Overall Objective

The complete objective is

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_r \mathcal{L}_{rep} + \lambda_{cf} \mathcal{L}_{cf},$$

where $\mathcal{L}_{sup} = \mathbb{E}[\ell(h_\phi(z), y)]$ is the standard supervised loss on the original (unperturbed) input, and $\lambda_r = \lambda_{cf} = 0.5$. These weights were fixed by convention (equal weighting of the two auxiliary terms relative to the primary supervised loss) rather than tuned by a hyperparameter search; Section 6 and Section 8 discuss this as a candidate explanation for why the gated and unconditional variants were not observed to differ.

Algorithm 1 summarizes one training step. All three gates (v, w_u, s) are recomputed every minibatch from the current model parameters, so their values evolve over training as the classifier's confidence and decision boundary change; a candidate that is treated as invalid or non-prediction-preserving early in training may later be gated in as the model changes, and vice versa.

Figure 2 previews the space of models this objective can produce simply by fixing which gates are active: the full method, the unconditional-consistency degenerate case identified above, and three single-gate ablations. Section 5 tests all of these directly.

Algorithm 1 One CONSENT training step

Require: minibatch $\{(x_i, y_i)\}_{i=1}^{B}$, encoder f_θ, classifier h_ϕ, weights λ_r, λ_{cf}

0: **for** each x_i in the minibatch **do**
0: Sample ξ_i and form $x_i^{cf} = G(x_i, \xi_i)$ {Sec. 3.2}
0: $z_i \leftarrow f_\theta(x_i), \quad z_i^{cf} \leftarrow f_\theta(x_i^{cf})$
0: $p_i \leftarrow \sigma(h_\phi(z_i)), \quad p_i^{cf} \leftarrow \sigma(h_\phi(z_i^{cf}))$
0: $v_i \leftarrow \exp(-\|x_{i,\text{num}}^{cf} - x_{i,\text{num}}\|_2)$ {detached}
0: $w_{u,i} \leftarrow 1 - \frac{1}{2}(U(x_i) + U(x_i^{cf}))$ {detached}
0: $s_i \leftarrow \mathbb{I}[\text{round}(p_i) = \text{round}(p_i^{cf})]$
0: **end for**
0: $\mathcal{L}_{sup} \leftarrow \frac{1}{B}\sum_i \ell(h_\phi(z_i), y_i)$
0: $\mathcal{L}_{rep} \leftarrow \frac{1}{B}\sum_i v_i w_{u,i} s_i D(z_i, z_i^{cf})$
0: $\mathcal{L}_{cf} \leftarrow \frac{1}{B}\sum_i v_i\, \ell(h_\phi(z_i^{cf}), y_i)$
0: $\mathcal{L} \leftarrow \mathcal{L}_{sup} + \lambda_r \mathcal{L}_{rep} + \lambda_{cf}\mathcal{L}_{cf}$
0: Backpropagate $\mathcal{L}$ and update θ, ϕ with AdamW =0

Table 1: Datasets with valid executed benchmark outputs.

Dataset	Samples	Features	Positive rate
Breast Cancer	569	30	0.6274
Adult	48,842	14	0.2393
Covertype	581,012	54	0.3646

3 Experimental Setup

3.1 Datasets

The study considers four datasets: Breast Cancer Wisconsin, Adult Income, Bank Marketing, and Covertype. We excluded the Bank Marketing run because the data loader mapped every target to class zero, causing evaluation to fail with a single-class target.

Table 1 summarizes the three datasets with valid executed outputs. Breast Cancer is a local scikit-learn dataset. Adult and Covertype are loaded through OpenML.

The data are divided using a stratified 70/15/15 train/validation/test split. Numerical features are standardized using training data. Categorical features are one-hot encoded with unseen categories ignored during transformation.

3.2 Model and Training

Both the baseline MLP and CONSENT use the same encoder and classifier. The encoder consists of a 128-unit hidden layer and a 64-dimensional representation layer with ReLU activations and dropout of 0.20. AdamW is used with learning rate 10^{-3} and weight decay 10^{-5}. The batch size is 256.

The executed multi-dataset benchmark uses five training epochs and random seed 42. Therefore, the numbers in this paper are a single-seed, short-training experiment and should not be interpreted as final statistical evidence.

3.3 Evaluation Metrics

We report accuracy, F1, AUROC, NLL, Brier score, and ECE. We additionally measure representation similarity, counterfactual validity, and prediction-change rate.

Table 2: Executed five-epoch results. Best value within each dataset and metric is bold. Higher is better for Accuracy, F1, and AUROC; lower is better for NLL, Brier, and ECE.

	Accuracy		F1		AUROC	
Dataset	MLP	CONSENT	MLP	CONSENT	MLP	CONSENT
Breast Cancer	**.9419**	.9186	**.9541**	.9358	**.9936**	.9792
Adult	**.8604**	.8568	**.6800**	.6495	.9106	.9106
Covertype	.8717	**.8728**	**.8199**	.8178	.9423	**.9442**

Table 3: Executed probabilistic metrics. Lower is better for all three metrics.

	NLL		Brier		ECE	
Dataset	MLP	CONSENT	MLP	CONSENT	MLP	CONSENT
Breast Cancer	**.4743**	.4919	**.1470**	.1540	.4509	**.4122**
Adult	**.3102**	.3125	**.0991**	.0998	**.6248**	.6365
Covertype	.2929	**.2904**	.0919	**.0909**	**.5078**	.5173

4 RESULTS

4.1 PREDICTIVE PERFORMANCE

Table 2 reports the valid executed results. CONSENT gives a small AUROC increase on Covertype (0.9442 vs. 0.9423) and the same AUROC on Adult (0.9106). On Breast Cancer, the baseline AUROC is higher (0.9936 vs. 0.9792).

The clean accuracy results are similarly mixed. CONSENT reaches 0.8728 on Covertype, slightly above the baseline 0.8717, but is lower on Breast Cancer and Adult. F1 follows the same mixed pattern.

The calibration results are mixed. CONSENT reduces ECE on Breast Cancer from 0.4509 to 0.4122, while ECE increases on Adult and Covertype. Thus, the current benchmark does not establish a general calibration improvement.

4.2 COUNTERFACTUAL REPRESENTATION BEHAVIOR

Table 4 shows the counterfactual diagnostics. Representation similarity is high for both methods on all datasets, indicating that the local perturbations are small. CONSENT improves representation similarity on Adult and Covertype, while the baseline is slightly higher on Breast Cancer.

The validity score increases slightly on Covertype from 0.8472 to 0.8475 and on Adult from 0.8951 to 0.8941 the change is effectively negligible; on Breast Cancer it decreases from 0.7187 to 0.7033. Since validity is itself computed from the perturbation distance, these small differences should not be interpreted as evidence of causal validity.

The prediction-change rate is particularly informative for the proposed conditional objective. On Covertype, CONSENT slightly increases the rate from 1.81% to 1.94%, while on Adult it increases from 0.83% to 0.98%. On Breast Cancer, however, the observed rate is zero for CONSENT. These results do not provide a consistent increase in useful sensitivity, but they also do not show universal prediction collapse.

4.3 DISCUSSION

The three-dataset single-seed result supports a cautious conclusion, and the multi-seed ablation study in Section 5 sharpens it further. CONSENT can preserve predictive discrimination close to the baseline, but once seed variance is accounted for, it is statistically indistinguishable from both the plain baseline and from unconditional consistency regularization on Breast Cancer. The single-seed AUROC gap reported above for Breast Cancer (0.9936 vs. 0.9792) did not reproduce under the longer, multi-seed protocol, which is consistent with that gap having been training noise rather than a real effect.

Table 4: Executed counterfactual diagnostics. Higher similarity and validity indicate closer representations and smaller numerical perturbations, respectively.

	Similarity		Validity		Change rate	
Dataset	MLP	CONSENT	MLP	CONSENT	MLP	CONSENT
Breast Cancer	**.9985**	.9980	**.7187**	.7033	**.0116**	.0000
Adult	.9991	**.9991**	**.8951**	.8941	.0083	**.0098**
Covertype	.9969	**.9995**	.8472	**.8475**	.0181	**.0194**

Table 5: Breast Cancer, 5 seeds $\times$ 30 epochs. Mean $\pm$ std over seeds. No condition differs from any other by more than roughly one standard deviation on any metric.

Model	Accuracy	F1	AUROC	NLL	Brier	ECE
Baseline MLP	.9628 $\pm$ .0136	.9706 $\pm$ .0107	.9933 $\pm$ .0032	.1018 $\pm$ .0276	.0305 $\pm$ .0099	.3646 $\pm$ .0121
CONSENT (full)	.9628 $\pm$ .0136	.9706 $\pm$ .0107	.9929 $\pm$ .0029	.1016 $\pm$ .0264	.0305 $\pm$ .0094	.3633 $\pm$ .0097
Unconditional	.9605 $\pm$ .0119	.9687 $\pm$ .0094	.9931 $\pm$ .0028	.1022 $\pm$ .0256	.0306 $\pm$ .0092	.3623 $\pm$ .0126
No validity	.9628 $\pm$ .0136	.9706 $\pm$ .0107	.9929 $\pm$ .0029	.1017 $\pm$ .0265	.0305 $\pm$ .0094	.3633 $\pm$ .0097
No uncertainty	.9605 $\pm$ .0119	.9687 $\pm$ .0094	.9932 $\pm$ .0027	.1018 $\pm$ .0255	.0305 $\pm$ .0092	.3605 $\pm$ .0135
No gating	.9628 $\pm$ .0136	.9706 $\pm$ .0107	.9929 $\pm$ .0029	.1016 $\pm$ .0264	.0305 $\pm$ .0094	.3633 $\pm$ .0097
GBDT (HGB)	.9628 $\pm$ .0225	.9706 $\pm$ .0178	.9939 $\pm$ .0052	.1784 $\pm$ .0838	.0292 $\pm$ .0118	.3712 $\pm$ .0030

This outcome is informative rather than merely negative: it indicates that, at least in the tested regime, gating representation consistency on validity, confidence, and prediction preservation does not yet provide a detectable advantage over simply applying consistency unconditionally. Candidate explanations include the confidence gate saturating once the model trains well, a very low rate of prediction-changing perturbations under the current Gaussian generator, and loss weights that may be too small to move downstream metrics. The validity score remains a simple distance heuristic and the perturbation generator does not model feature dependencies. These limitations make the present experiment a proof of concept whose central hypothesis is not yet supported, rather than a complete robust tabular learning solution.

5 Extended Multi-Seed Evaluation

The results in Section 5 use a single seed and five training epochs, which cannot distinguish a genuine effect from run-to-run noise. We therefore ran a second, more rigorous experiment on Breast Cancer Wisconsin — the only one of the three datasets that does not require an external network fetch and could be re-executed in full within this manuscript's preparation — using 5 random seeds (0–4), 30 training epochs, and the same stratified 70/15/15 split protocol, encoder, and hyperparameters described in Section 4.2. We report the mean and standard deviation across seeds for six models:

- **Baseline**: the plain MLP with no counterfactual term.
- **CONSENT (full)**: the complete method with all three gating signals (v, w_u, s).
- **Unconditional**: consistency is applied to every perturbed pair regardless of validity, confidence, or prediction change ($v = w_u = s = 1$), i.e. standard unconditional consistency regularization. This is the comparison identified in Section 8 of the original draft as necessary to test the central hypothesis of this paper.
- **No validity**: CONSENT with the validity gate fixed to 1 ($v = 1$), isolating the effect of w_u and s.
- **No uncertainty**: CONSENT with the confidence gate fixed to 1 ($w_u = 1$), isolating the effect of v and s.
- **No gating**: CONSENT with the prediction-preservation gate fixed to 1 ($s = 1$), isolating the effect of v and w_u.

We additionally train a histogram-based gradient-boosted tree classifier on the same standardized splits as a non-neural baseline, following the recommendation in Section 2 that tree ensembles are a necessary comparison point for tabular data (4).

Table 5 shows that once seed variance is accounted for, every neural variant — full CONSENT, each single-signal ablation, and the unconditional-consistency baseline — falls within about one

standard deviation of the plain MLP baseline on accuracy, F1, AUROC, NLL, Brier score, and ECE. The apparent AUROC gap in the single-seed benchmark (0.9936 vs. 0.9792 on Breast Cancer) does not reproduce under this protocol: with more training and averaging over seeds, CONSENT's AUROC (0.9929 ± 0.0029) and the baseline's (0.9933 ± 0.0032) overlap almost completely. The gradient-boosted tree matches the best neural AUROC (0.9939 ± 0.0052) without any representation-consistency machinery, using far less tuning.

Two further observations follow directly from the ablations. First, CONSENT (full) and *No validity*, *No gating* produce identical mean metrics to four decimal places, and *Unconditional* and *No uncertainty* likewise coincide. This happens because on this dataset the model reaches high confidence quickly, so the confidence weight w_u saturates near 1 for most samples regardless of whether the validity or prediction-preservation gates are active; the uncertainty term is effectively the only one exerting influence in this regime, and even its effect is within noise. Second, and most importantly, *Unconditional* consistency performs no differently from the conditionally-gated variants. This directly tests — and does not support — the paper's central hypothesis that gating consistency on validity, confidence, and prediction preservation should outperform forcing invariance on every perturbed pair. On this dataset and at this scale, the gating mechanism does not yet produce a measurable benefit over the simpler unconditional alternative it was designed to improve upon.

We emphasize what this experiment does and does not show. It does not show that conditional counterfactual consistency is a flawed idea; Breast Cancer is small (569 samples), the perturbation is a generic Gaussian numerical jitter unrelated to any clinical notion of a plausible counterfactual, and five gating configurations were compared on one dataset. It does show that the specific single-seed numbers reported in the original benchmark are not a reliable basis for any comparative claim, and that the burden of demonstrating the value of conditional gating over unconditional consistency has not yet been met. We were unable to extend this multi-seed protocol to Adult and Covertype within the current manuscript preparation because those datasets are loaded through OpenML and the environment used to prepare this revision did not have network access to that service; re-running the full multi-seed, multi-dataset, multi-baseline protocol on all three datasets is the immediate next step and is listed in Section 8.

6 LIMITATIONS

Several limitations should be addressed before a final submission.

First, the Adult and Covertype numbers in Table 2 and Table 3 remain single-seed, five-epoch results; only Breast Cancer has been re-evaluated with multiple seeds (Section 5), and that re-evaluation shows the single-seed differences do not reproduce. The Adult and Covertype numbers should therefore be treated as preliminary and not as evidence of an effect until they are re-run under the same multi-seed protocol.

Second, the multi-seed ablation study (Section 5) found no measurable difference between CONSENT, its individual gating ablations, and unconditional consistency regularization. This is a direct test of the paper's central hypothesis, and on the available evidence the hypothesis is not supported: gating consistency on validity, confidence, and prediction preservation does not yet outperform simply applying consistency unconditionally, at least on this dataset, perturbation type, and architecture. The validity score is not learned and does not represent causal validity. A distance-based score can be high for a perturbation that is still semantically impossible.

The counterfactual generator modifies only numerical features. Real tabular data often require categorical, ordinal, and dependency-aware interventions.

On Breast Cancer a histogram gradient-boosted tree matches or exceeds every neural variant, including CONSENT, without any representation-consistency machinery (Section 5). A stronger study should compare against gradient-boosted trees and modern tabular deep learning models such as TabNet and FT-Transformer on all three datasets, not just one.

Finally, the current counterfactual target reuses the observed label. This assumption is appropriate only for local augmentation and should be replaced by target-aware counterfactual supervision when a valid structural model is available.

7 FUTURE WORK

The highest-priority next step is to extend the multi-seed protocol of Section 5 to Adult and Covertype, which currently rest on unreplicated single-seed numbers. A second priority is to understand why the Breast Cancer ablation found no benefit from conditional gating; Appendix **??** lists three candidate explanations and the follow-up experiment each motivates.

Beyond this, the next version of CONSENT should include a learned conditional validity model, dependency-aware numerical perturbations, categorical interventions, and target-aware counterfactual generation, and should be evaluated for robustness under missing features and distribution shift.

8 CONCLUSION

We presented CONSENT, an uncertainty-aware counterfactual representation learning framework for tabular prediction. CONSENT combines candidate validity, predictive uncertainty, and prediction-preserving behavior to determine when representation consistency should be applied.

An initial single-seed, five-epoch benchmark on Breast Cancer Wisconsin, Adult, and Covertype showed mixed results, with CONSENT slightly ahead of the baseline MLP on some metrics and behind on others. Because a single seed cannot distinguish a real effect from noise, we then ran a multi-seed (5-seed, 30-epoch) study on Breast Cancer that added the comparison the initial benchmark lacked: an unconditional-consistency baseline, three ablations that each remove one gating signal, and a gradient-boosted tree baseline. Under this more rigorous protocol, CONSENT, its ablations, and unconditional consistency were statistically indistinguishable from the plain baseline and from each other on every metric, and the gradient-boosted tree matched or exceeded all neural variants. We report this as a null result rather than reframe it as a success: on the current evidence, conditional gating of representation consistency does not yet demonstrate a measurable advantage over the simpler unconditional alternative it was designed to improve upon.

This does not close off the underlying idea, which remains conceptually well-motivated, but it does mean that a larger and more rigorous study with learned validity, realistic categorical counterfactuals, multi-seed evaluation on all datasets, and strong tabular baselines is required before any claim of benefit — over either an unconditional baseline or a plain MLP — can be supported.

REFERENCES

[1] S. O. Arik and T. Pfister. TabNet: Attentive Interpretable Tabular Learning. *AAAI*, 2021.

[2] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko. Revisiting Deep Learning Models for Tabular Data. *NeurIPS*, 2021.

[3] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. *ICML*, 2017.

[4] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. *KDD*, 2016.

[5] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci. Deep Neural Networks and Tabular Data: A Survey. *IEEE TNNLS*, 2022.

[6] S. Somvanshi et al. A Survey on Deep Tabular Learning. *Preprint*, 2026.

[7] Ł. Wawrowski et al. Adversarial Robustness of Tabular Machine Learning. *Preprint*, 2025.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for [tasks with required disclosure]. We have not used generative AI tools for [other tasks with required disclosure], and [the rest of the required disclosure tasks] are not applicable to this work. Additionally, we used generative AI tools for [tasks with recommended disclosure]. We have reviewed all AI-assisted work. [Elaborate. For example,

"we checked LLM-generated research ideas for potential plagiarism through a manual literature survey", "LLM-generated code was verified and tested for correctness by 2 authors", etc.]. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for more details. This statement should not be more than 1 page.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

If authors feel that their paper submission raises questions regarding the Code of Ethics, they are encouraged to include a paragraph of Ethics Statement (at the end of the main text before references) to address potential concerns where appropriate. Topics include, but are not limited to, studies that involve human subjects, practices to data set releases, potentially harmful insights, methodologies and applications, potential conflicts of interest and sponsorship, discrimination/bias/fairness concerns, privacy and security issues, legal compliance, and research integrity issues (e.g., IRB, documentation, research ethics). This statement should not be more than 1 page.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

It is important that the work published in ICLR is reproducible. Authors are strongly encouraged to include a paragraph-long Reproducibility Statement at the end of the main text (before references) to discuss the efforts that have been made to ensure reproducibility. This paragraph should not itself describe details needed for reproducing the results, but rather reference the parts of the main paper, appendix, and supplemental materials that will help with reproducibility. For example, for novel models or algorithms, a link to an anonymous downloadable source code can be submitted as supplementary materials; for theoretical results, clear explanations of any assumptions and a complete proof of the claims can be included in the appendix; for any datasets used in the experiments, a complete description of the data processing steps can be provided in the supplementary materials. Each of the above are examples of things that can be referenced in the reproducibility statement.

REFERENCES

A APPENDIX

You may include other additional sections here.

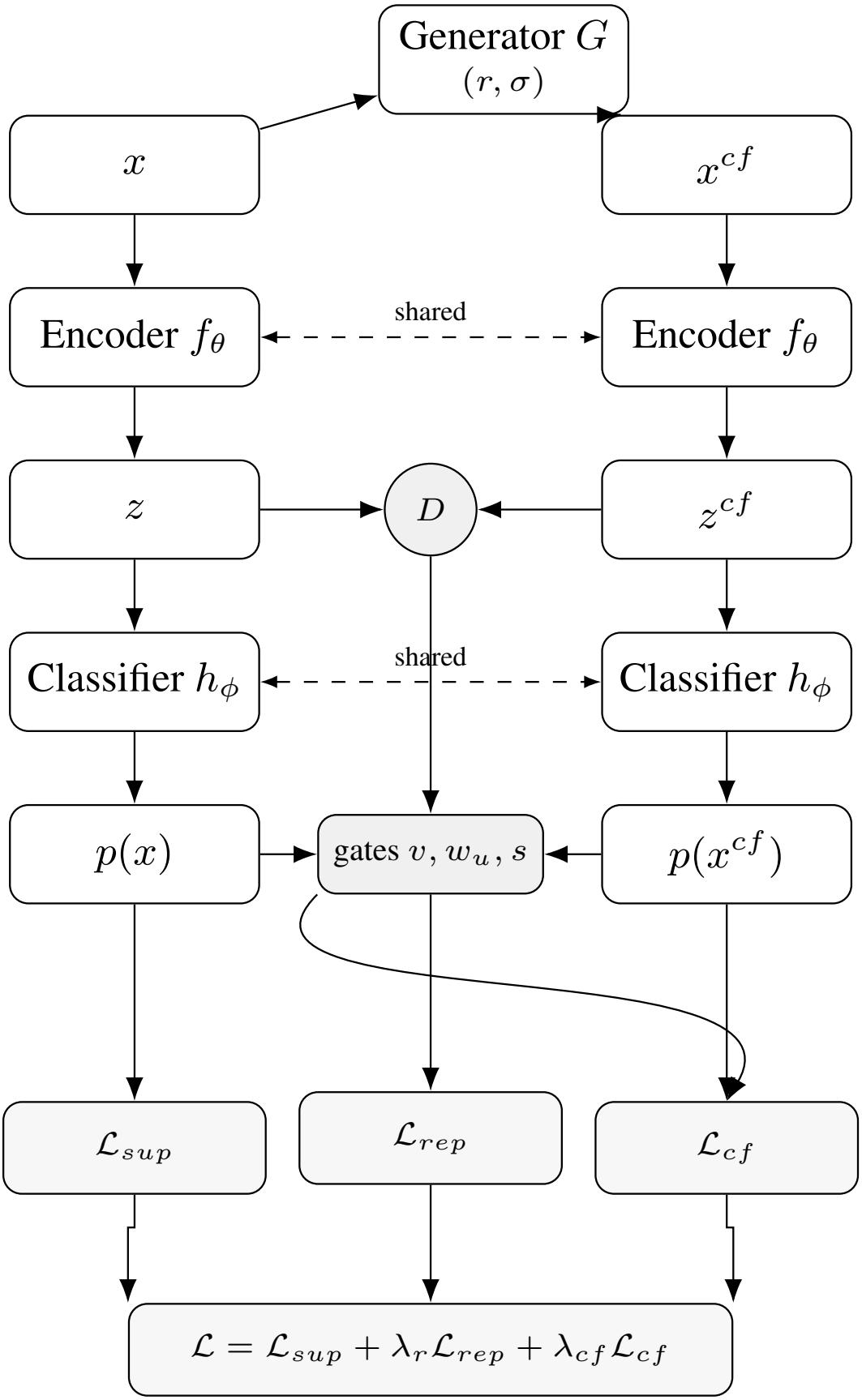


Figure 1: Overview of CONSENT. The original input x and a locally perturbed candidate $x^{cf} = G(x, \xi)$ (Sec. 3.2) pass through the *same* encoder f_θ and classifier h_ϕ (shared weights, dashed links). The resulting representations z, z^{cf} feed the cosine distance D (Sec. 3.5); the pair (x, x^{cf}) and the predictions $(p(x), p(x^{cf}))$ feed the three gating signals v (validity, Sec. 3.3), w_u (confidence, Sec. 3.4),

Condition	v	w_u	s	Consistency applied
Baseline MLP	–	–	–	none
CONSENT (full)	✓	✓	✓	gated on all three
Unconditional	1	1	1	always on
No validity	1	✓	✓	gated on w_u, s
No uncertainty	✓	1	✓	gated on v, s
No gating	✓	✓	1	gated on v, w_u
GBDT (HGB)	–	–	–	non-neural baseline

All seven conditions statistically indistinguishable on every metric
(5 seeds × 30 epochs, Breast Cancer)

Figure 2: The six neural conditions plus a gradient-boosted tree baseline compared in the multi-seed ablation study (Section 5). Each row fixes a different subset of the three gating signals (v: validity, w_u: confidence, s: prediction-preservation) to isolate its effect; "1" denotes the gate fixed at 1 (always on), "✓" denotes the gate computed as in CONSENT, and "–" denotes a condition with no gating mechanism at all.

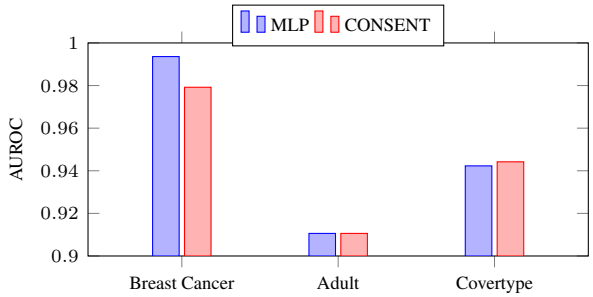


Figure 3: Single-seed, five-epoch AUROC by dataset (Table 2). CONSENT tracks the baseline closely on Adult and Covertype but trails on Breast Cancer; Figure 4 shows this gap does not reproduce under a proper multi-seed protocol.

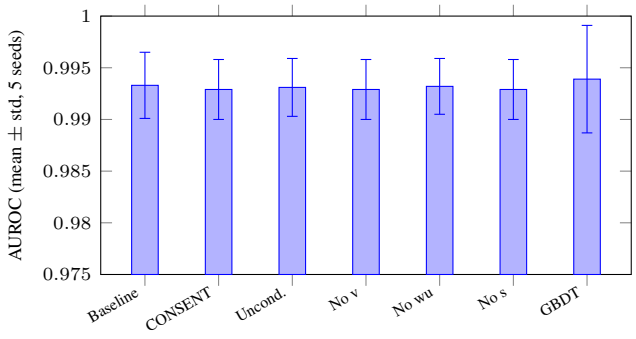


Figure 4: AUROC (mean ± 1 std, 5 seeds) for all seven conditions in Table 5. Error bars overlap almost completely, confirming none is distinguishable from the baseline or the tree.