Intervention Efficiency and Perturbation Validation Framework: Capacity-Aware and Robust Clinical Model Selection under the Rashomon Effect

Yuwen Zhang[1], **Viet Tran**[2], **Paul Weng**[1,3,4*]
[1]Duke Kunshan University (DKU), Kunshan, China
[2]University College Dublin, Dublin, Ireland
[3]Digital Innovation Research Center, DKU, Kunshan, China
[4]Jiangsu Provincial University Key (Construction) Laboratory for Smart Diagnosis and Treatment of Lung Cancer
yuwen.zhang@dukekunshan.edu, quoc.tran@ucdconnect.ie, paul.weng@dukekunshan.edu.cn

Abstract

In clinical machine learning, the coexistence of multiple models with comparable performance—a manifestation of the Rashomon Effect—poses fundamental challenges for trustworthy deployment and evaluation. Small, imbalanced, and noisy datasets, coupled with high-dimensional and weakly identified clinical features, amplify this multiplicity and make conventional validation schemes unreliable. As a result, selecting among equally-performing models becomes uncertain, particularly when resource constraints and operational priorities are not considered by conventional metrics like F1 score. To address these issues, we propose two complementary tools for robust model assessment and selection: Intervention Efficiency (IE) and the Perturbation Validation Framework (PVF). IE is a capacity-aware metric that quantifies how efficiently a model identifies actionable true positives when only limited interventions are feasible, thereby linking predictive performance with clinical utility. PVF introduces a structured approach to assess the stability of models under data perturbations, identifying models whose performance remains most invariant across noisy or shifted validation sets. Empirical results on synthetic and real-world healthcare datasets show that using these tools facilitates the selection of models that generalize more robustly and align with capacity constraints, offering a new direction of tackling the Rashomon Effect in clinical settings.

Code — https://github.com/YuwenZhang-Peter/PVF-IE

1 Introduction

Small-scale clinical prediction modeling often faces challenges of limited sample size, class imbalance, and high dimensionality, with few established predictors. Data scarcity arises from high collection costs, ethical constraints, and follow-up difficulties (Coors et al. 2017). Class imbalance is common since adverse events affect only a small fraction of patients, leading to misleading results if conventional metrics like accuracy are used (Davis and Maiden 2021; He and Garcia 2009). In such contexts, clinicians value precision and recall more than overall accuracy, especially when only a limited number of interventions can be performed (Martin et al. 2025; Li et al. 2025; Gao et al. 2021).

Because improving precision often reduces recall, an explicit trade-off is essential under constrained resources. The F1 score, as the harmonic mean of precision and recall, ignores true negatives and can favor suboptimal models when this trade-off is unclear (He and Garcia 2009). Threshold-independent metrics mitigate this issue but have drawbacks: AUC-ROC can overestimate performance on imbalanced data, while AUC-PR, though better for rare events, remains sensitive to prevalence and difficult to interpret clinically (Saito and Rehmsmeier 2015; Richardson et al. 2024). Metrics such as balanced accuracy, Cohen's kappa, and MCC have been proposed to address imbalance, with MCC noted for robustness (Chicco and Jurman 2020), yet all may yield ambiguous results requiring manual interpretation. Evaluating many models across several criteria is also computationally costly, motivating a measure that captures the precision–recall trade-off while reflecting capacity constraints.

High dimensionality further increases overfitting and obscures relevant predictors (Berisha et al. 2021; Bommert et al. 2022), motivating feature selection or dimensionality reduction (Jolliffe and Cadima 2016; Jia et al. 2022). Still, models like logistic regression, SVMs, k-nearest neighbors, and random forests can achieve similar accuracies while relying on distinct feature sets (Łukaszuk and Krawczuk 2024). This phenomenon, called the Rashomon Effect (Breiman 2001b; Rudin, Parr, and Seltzer 2024; Xin et al. 2022), reveals that many "equally good" but mutually-exclusive models may coexist. This multiplicity usually complicated interpretation and clinical trust.

Model evaluation is also hindered by instability across data splits: small datasets amplify variance, and similar scores can occur by chance (Cawley and Talbot 2010a). Traditional validation provides point estimates but neglects robustness, though model performance is known to fluctuate under realistic perturbations (Varoquaux 2018; Paschali et al. 2018; Arleto, Delfanti, and Nguyen 2023). Robustness to data shifts is now recognized as a core requirement for safe deployment in healthcare (Paschali et al. 2018; Pfohl et al. 2019; Heinrich et al. 2022).

To address these challenges, we propose two complementary tools: the *Intervention Efficiency (IE)* and the *Perturbation Validation Framework (PVF)*. IE quantifies how ef-

*Corresponding author.
Workshop on Navigating Model Uncertainty and the Rashomon Effect: From Theory and Tools to Applications and Impact (AAAI 2026)

fectively a model identifies minority cases when intervention capacity is limited, making the precision–recall tradeoff explicit and clinically meaningful. PVF stress-tests models across perturbed validation sets and selects the one with the most stable performance, emphasizing robustness and interpretability. While introduced alongside IE, PVF is fully compatible with conventional metrics such as F1 score or accuracy. Experiments on synthetic and real medical datasets show that PVF guides model selection toward greater generalization under both IE and traditional metrics. Together, IE and PVF provide a resource-aware, robustness-focused evaluation framework that facilitates model selection when the Rashomon Effect appears in small, imbalanced healthcare datasets.

2 Related Work

This section discusses issues related to the Rashomon Effect. We (i) examine the pitfalls of relying on averaged metrics for small, imbalanced clinical datasets and motivate selecting a single deployable model (2.1), and (ii) review perturbation-based validation for robust model selection while clarifying how our PVF differs from existing approaches (2.2).

2.1 Average Metrics vs. Single Deployable Model

Small, imbalanced datasets create high variance, overfitting, and instability in model and feature selection (Cawley and Talbot 2010b; Zhang et al. 2023; Xin et al. 2022), making it difficult to select one reliable model among many, particularly when the Rashomon Effect appears and causes multiplicity in models. In low-sample settings, small data perturbations can cause drastic changes in model structure and predictions, especially for unstable learners such as decision trees or stepwise regression (Breiman 1996a,b; Morozova et al. 2015; Meinshausen and Bühlmann 2010; Dietterich 2000). Under class imbalance, models may favor majority classes or overfit rare ones (Chawla et al. 2002; He and Garcia 2009; Buda, Maki, and Mazurowski 2018; van den Goorbergh et al. 2022).

Cross-validation (CV) is widely used to estimate model performance but performs unreliably on small, high-dimensional data. Specifically, CV-based selection often yields optimistic estimates and unstable rankings across splits (Varma and Simon 2006; Cawley and Talbot 2010b; Arlot and Celisse 2010; Varoquaux 2018; Zhang, Yang, and Wang 2021). Nested CV can reduce bias but remains computationally expensive and unstable when samples are few (Wainer and Cawley 2021). Moreover, conventional CV averages over models rather than assessing individual consistency. Likewise, although ensemble and bagging methods reduce variance and improve generalization (Breiman 1996a, 2001a), they typically produce complex, less interpretable systems prone to overfitting (Dietterich 2000; Bühlmann and Yu 2002; Opitz and Maclin 1999). In contrast, clinical use prioritizes transparent, single-model predictions for interpretability and deployment (Dong et al. 2020; Valente et al. 2021; Zadeh et al. 2023).

Our PVF can be applied on a certain train-validation split. Therefore, it can either be combined with cross-validation to improve the robustness of each validation, or as an alternative, help select a single deployable model. It can act as a promising approach to facilitate model selection when the Rashomon Effect appears.

2.2 Perturbation-Based Validation in Small, Imbalanced Clinical Datasets

Robustness evaluation often involves injecting random, non-adversarial noise. Classic studies added Gaussian noise to features to assess performance degradation (Egmont-Petersen, Talmon, and Hasman 1997; Shafiee et al. 2019; Parmar et al. 2015; Yang et al. 2020; Pfohl, Zhang, and Celi 2021; Kulkarni et al. 2021). Recent gene–expression experiments found that logistic regression retained the most stable feature sets, while Random Forests were least stable, despite comparable accuracy (Łukaszuk and Krawczuk 2024). This shows that conventional metrics alone can overlook robustness to noise.

For interpretable models, robustness work is emerging. For example, loss-function redesign improves single-tree resilience to label noise (Sztukiewicz, Good, and Dubrawski 2024), yet trees in small-sample settings can still memorize noisy outliers, and ensemble averaging may leave feature importance unstable (Soloff, Foygel Barber, and Willett 2024).

Perturbation-based validation explicitly tests stability against noise. Label-noise injection, as in Mutation Validation (MV) (Zhang et al. 2023), flips a subset of labels to measure performance drop. Feature noise assesses covariate sensitivity (Egmont-Petersen, Talmon, and Hasman 1997; Kulkarni et al. 2021). Resampling methods, such as bootstrap, repeated k-fold CV and bagging, perturb sample composition to produce performance distributions rather than single estimates (Breiman 1996a, 2001a; Soloff, Foygel Barber, and Willett 2024).

Our Perturbation Validation Framework (PVF) differs by applying feature-level noise only to the validation set, directly evaluating generalization under noisy conditions without retraining on corrupted data. Aggregating results from multiple perturbations identifies the model with the most consistent and stable predictions—a crucial aspect when addressing the Rashomon Effect.

We also considered label perturbation during validation. Prior work shows that robust models should withstand limited label noise in training (Zhang et al. 2023). However, introducing label noise in validation can distort comparisons. Flipped labels disproportionately penalize stronger models (which previously predicted correctly) while leaving weaker ones largely unaffected, compressing or inverting true performance gaps. In highly imbalanced datasets, even a few flipped labels can drastically alter metrics like precision and recall. Therefore, PVF excludes label perturbation in validation, despite its popularity for robustness testing during training.

3 Methodologies

This section has two parts: Section 3.1 proposes *Intervention Efficiency* (IE), comparing model-guided and random inter-

ventions under limited capacity; Section 3.2 introduces the *Perturbation Validation Framework* (PVF) for robust model assessment and model selection.

3.1 Intervention Efficiency (IE)

In many applied domains such as clinical follow-up, fraud investigation, and social services, only a fixed budget c of interventions can be performed. In practice, this means that only a subset of the at-risk population can be targeted due to limited capacity.

A benchmark policy, referred to as *uniform intervention*, randomly selects c individuals from the population without any predictive guidance. In contrast, a *model-guided intervention* uses a binary classifier to prioritize individuals for action. Two situations may occur:

1. The model predicts more positives than c, and only the top c flagged cases can be intervened upon.
2. The model predicts fewer positives than c, and after acting on all predicted positives, the remaining capacity is allocated through uniform intervention on the rest of the population.

Because uniform intervention represents a fixed baseline under identical problem setting and capacity constraint, it provides a natural reference for evaluating model-guided strategies.

We define the *Intervention Efficiency (IE)* as the ratio between the expected number of true positives captured by a model-guided intervention and that captured by uniform intervention, given an intervention capacity $\gamma = c/\beta$, where β denotes the population size. IE thus quantifies how effectively a predictive model utilizes limited resources to identify positive cases: higher IE indicates a greater number of true positives captured within the same budget.

In the following proposition, we show that the IE of a model f depends only on its precision p, recall r, the prevalence rate π (i.e., the percentage of positive cases in the population), and the intervention capacity γ.

Proposition 3.1 (Closed form of IE)**.** *Assume that there is a model f that can predict binary labels and it is used for model-guided intervention. The problem has an overall prevalence rate π and only a fraction γ of the whole population can be intervened due to limited capacity. The Intervention Efficiency of f on this problem under intervention capacity γ, denoted as $\mathrm{IE}_\gamma(f)$, satisfies*

$$\mathrm{IE}_\gamma(f) = \frac{sp + (\gamma - s)\,\dfrac{\pi - sp}{1 - s}}{\gamma\pi},$$

where p is the precision of f and $s = \min\!\left(\gamma, \frac{\pi r}{p}\right)$, with r being the recall of f.

We refer readers to Section A for the detailed proof of Proposition 3.1. In practice, we can approximate the true π by the prevalence rate $\hat{\pi}$ in the specific set of samples to which the model is applied. For instance, when estimating IE on a validation set, even with the whole dataset that can provide a "more accurate estimation" for π, we should use the $\hat{\pi}$ from the specific validation set rather than the prevalence rate from the whole dataset. The reason is that IE is tied to precision and recall, two values depending not only on the model but also largely on the set of samples that yields them, especially in small sample cases. In contrast, in large sample cases, $\hat{\pi}$ has been a good estimate of true π given enough samples. In sum, it is both reasonable and essential to estimate π, p and r on the same set to ensure they are aligned, since they are jointly dependent on each other. Practically, when computing IE on a specific set of samples $\mathcal{D}$, it not only depends on a model f and a γ of interest, but also on $\mathcal{D}$, where $\hat{\pi}$, $\hat{p}$ and $\hat{r}$ (estimators of π, p and r) are derived. Consequently, we can rewrite the formula for IE in real practice as:

$$\mathrm{IE}_\gamma(f, \mathcal{D}) = \frac{\hat{s}\hat{p} + (\gamma - \hat{s})\,\dfrac{\hat{\pi} - \hat{s}\hat{p}}{1 - \hat{s}}}{\gamma\hat{\pi}},$$

where $\mathcal{D}$ is a set of samples the model f is applied on for evaluation, $\hat{\pi}$ is the prevalence rate in $\mathcal{D}$, $\hat{p}$ is the precision of f when evaluated on $\mathcal{D}$, and $\hat{s} = \min\!\left(\gamma, \frac{\hat{\pi}\hat{r}}{\hat{p}}\right)$, with $\hat{r}$ being the recall of f when evaluated on $\mathcal{D}$.

As for γ, to deploy a model trained on distribution 1 in distribution 2, the IE used during model selection on distribution 1 should be equipped with the γ corresponding to distribution 2. This aligns selection with the intervention capacity of the target setting, improving generalization to the distribution 2 user case.

3.2 Perturbation Validation Framework (PVF)

During model selection, the goal is not only to identify a model that performs well on the validation set, but also one that remains robust under practical noise that may be present in the data. Perturbed Validation Framework (PVF) provides a protocol for robustness assessment and model selection. Starting from a fixed validation set, it generates multiple perturbed versions via adding noise on each sample, evaluates each candidate model on every perturbed set using a chosen metric, and aggregates the resulting scores into a single robustness-adjusted criterion.

Setup Let the original validation set be $\mathcal{D}_{\mathrm{val}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where each $(\boldsymbol{x}_i, y_i), i = 1, 2, ..., n$ is a multi-dimensional sample with features $\boldsymbol{x}_i$ and a binary label y_i, while n is the total number of samples in $\mathcal{D}_{\mathrm{val}}$. Let $\mathcal{F}$ be a finite set of trained candidate models. The goal is to select the most robust model $f \in \mathcal{F}$, denoted as $f^\star$, based on the model performance on $\mathcal{D}_{\mathrm{val}}$.

Perturbation mechanism Each feature of each sample is perturbed independently based on its type, with labels unchanged. Below, we give a couple of examples of perturbation.

1. For numeric feature x in $\boldsymbol{x}$, add small scalar noise drawn from a user-specified distribution. For example, in our experiments, we draw $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and set $x' = x + \varepsilon$.
2. For categorical (nominal) feature x in $\boldsymbol{x}$, with a user-controlled flip probability, change x to another category

sampled from a specified mass function over the remaining categories. For example, in our experiments, with probability $1-\xi$, keep $x' = x$. Otherwise, choose $x' \neq x$ uniformly from the remaining categories, i.e.,

$$\Pr(x' = a \mid x) = \xi \cdot \frac{1}{(|\mathcal{C}| - 1)}$$

for $a \neq x$, where $|\mathcal{C}|$ is the number of categories feature x can have.

3. For ordinal feature x in $\boldsymbol{x}$, move along a predefined order with probability mass concentrated on nearby categories, as controlled by a decay/temperature parameter. For example, in our experiments, with probability $1 - \xi$, keep $x' = x$. Otherwise, for a pre-determined distance $d(\cdot, \cdot)$ on the order, sample $x' \neq x$ with probability

$$\Pr(x' \mid x) = \xi \cdot \frac{\exp\{-\lambda\, d(x, x')\}}{\sum_{a \neq x} \exp\{-\lambda\, d(x, a)\}},$$

where a is any order x can have and λ is the decay parameter.

Constructing perturbed validation sets Suppose M perturbed validation sets are desired. Then, for each $m = 1, \ldots, M$, form the m-th perturbed validation set $\mathcal{D}_{\text{val}}^{(m)}$ by generating exactly k independent replicas of every $(\boldsymbol{x}_i, y_i)$ using the mechanism above. Thereby, $|\mathcal{D}_{\text{val}}^{(m)}| = kn$. The reason for having an integer k here is to make sure every sample is perturbed by equal times, making the overall distribution and class imbalance preserved. Also, k should be sufficiently large in order to let various perturbed versions of the original sample emerge in the perturbed validation sets, given different perturbation patterns for different features and a resulting complex composition of them.

Evaluation, aggregation, and selection Let Eval be any scalar metric that maps a model and a dataset to a performance score. For each candidate model $f \in \mathcal{F}$ and $m = 1, \ldots, M$, the performance of f on the m-th perturbed validation set $\mathcal{D}_{\text{val}}^{(m)}$ can be expressed as $P_f^{(m)} = \text{Eval}\big(f, \mathcal{D}_{\text{val}}^{(m)}\big)$. Aggregate $\boldsymbol{P}_f = (P_f^{(1)}, \ldots, P_f^{(M)})$ via $\mathcal{A}$ to obtain $A_f = \mathcal{A}(\boldsymbol{P}_f)$, where A_f is the aggregated performance score of f on the M perturbed validation sets and $\mathcal{A}$ is a user-specified aggregation function (e.g., the 25-th percentile of the M scores). The overall workflow for evaluating a single model is summarized in Figure 1.

Finally, select $f^\star \in \arg\max_{f \in \mathcal{F}} A_f$, as shown in Figure 2. The complete implementation pseudocode of combining PVF with any evaluation metric Eval, denoted as PVF-Eval, appears in Algorithm 1.

Hyperparameters and user-specified functions For clarity, we provide the following list of components in PVF that can be changed according to specific user cases:

(i) d: number of (or, which) features to be perturbed.

(ii) k: number of replicas per original sample (so each perturbed set has size kn).

(iii) M: number of perturbed validation sets.

(iv) A collection of perturbation controls (e.g., noise standard deviation, flipping probabilities).

(v) $\mathcal{A} : \mathbb{R}^M \to \mathbb{R}$: aggregation function (e.g., a lower quantile).

Since no retraining is required, the per-model computational complexity is $\mathcal{O}(Mknd)$, and across all Q models it is $\mathcal{O}(QMknd)$. We refer readers to Section B for theoretical guarantees of PVF.

Algorithm 1: PVF-Eval Model Selection

Require: Original validation set $\mathcal{D}_{\text{val}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$; candidate model set $\mathcal{F}$; evaluation metric EVAL; number of perturbed sets M; replicas per sample k; a collection of perturbation controls; aggregation function $\mathcal{A}$

Ensure: Selected most robust model $f^\star$ on $\mathcal{D}_{\text{val}}$

```
for all f ∈ F do
    for m = 1 to M do
        Build D_val^(m) by independently perturbing each feature of each (x_i, y_i) as specified in Section 3.2, creating k replicas per original sample.
        P_f^(m) ← EVAL(f; D_val^(m))
    end for
    A_f ← A(P_f^(1), ..., P_f^(M))
end for
f* ← arg max_{f∈F} A_f
return f*
```

4 Experiments

This section outlines the objectives, general setups, and overall findings of our experiments on both synthetic and real datasets. Detailed descriptions, full quantitative results, and sensitivity analyses are provided in the Appendix. Specifically, Section C present the detailed setups and results for the synthetic dataset experiments; Section D presents the detailed setups for real datasets; Section E reports a comprehensive hyperparameter sensitivity analysis.

4.1 Research Questions

To evaluate the effectiveness and robustness of our proposed method, we designed a set of experiments guided by the following four research questions:

Q1. Can the proposed PVF be more effective than the traditional method (model selection based on a single held-out validation set) under traditional metrics (here, F1 Score and accuracy) on synthetic data (Section 4.3 and Section C.2)?

Q2. Is PVF more effective than the traditional method under IE across different intervention fractions (γ) on synthetic data (Section 4.3 and Section C.2)?

Q3. On real clinical datasets, does PVF select models that perform better than those selected by the traditional approach under the same metrics used in the synthetic dataset experiments (Section 4.3)?

Q4. How sensitive is PVF to its hyperparameters, with a focus on the perturbation controls (Section E)?

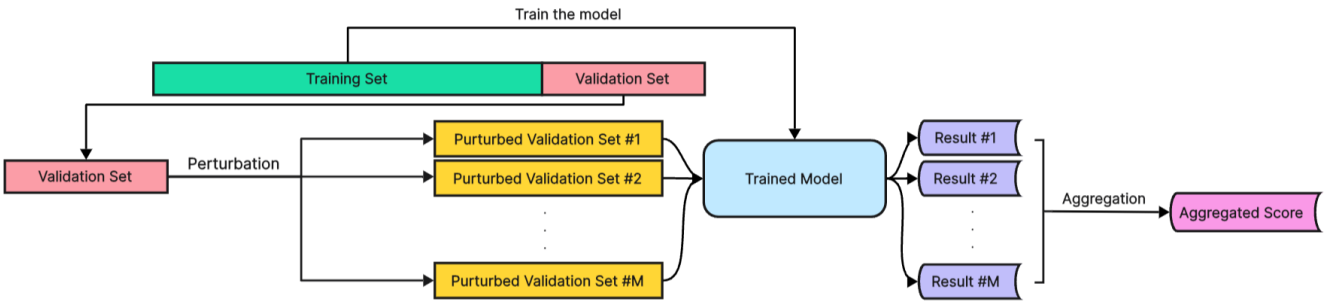

Figure 1: Overview of the Perturbation Validation Framework (PVF). After training a model on the training split, the original validation set is repeatedly perturbed (by adding noise to numeric features, changing categories for nominal features, or changing orders for ordinal features) to yield M validation sets. The trained model is evaluated on each perturbed validation set, producing M scores, which are then aggregated into a single robustness metric score.

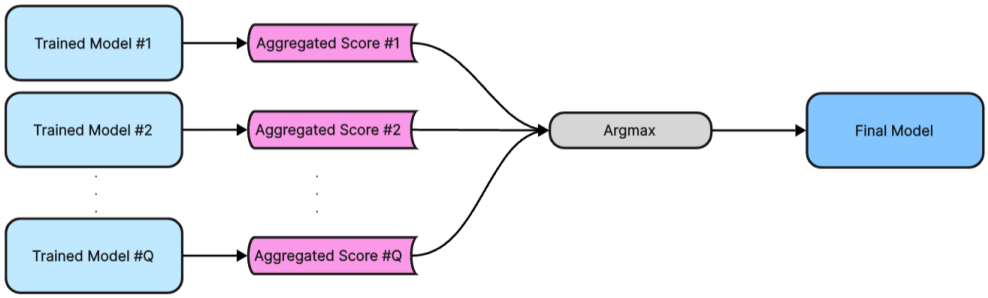

Figure 2: Model selection stage. Each candidate f yields a final aggregated score via PVF. The model $f^\star$ with the highest aggregated score is chosen.

These questions allow us to compare PVF against the traditional validation method, examine its performance under both synthetic and real clinical data, and assess its robustness to experimental settings.

4.2 General Experimental Setup

In this section, we introduce the general experimental setups for both the synthetic dataset experiments and the real dataset experiments, along with the rationales behind the designs.

We design our experiments around the task of model selection—choosing a single model from a pool of candidates based on its validation performance. This setting commonly arises in the presence of the Rashomon Effect. The aim is to evaluate whether PVF can yield more reliable choices than a conventional single-split validation. The reason for not including more sophisticated frameworks like cross-validation is that PVF can be applied on any single-split train-validation settings, thus can be combined with these frameworks, as stated in Section 2.1.

Evaluation is organized by fixing one metric at a time and comparing PVF against the traditional method under that metric. Intervention Efficiency (IE) is reported at different intervention capacity γ to represent different clinical resource levels. We also report F1 Score (or supplemented with accuracy) as a complementary analysis, given that these two traditional metrics are widely adopted and emphasized in many clinical applications. Note that the focus is on which selection strategy chooses better models given a chosen metric, instead of whether IE or any traditional metric (F1 Score, accuracy) is better for model selection, as the choice of performance metric is case-specific.

After selection, the chosen model is either assessed based on ground truth information or when it is not available, as an approximation of it, on a large external test set using the same metric as in validation. To improve reliability and reduce variance from any single split, each configuration is repeated extensively. Model classes and other dataset-specific design choices are kept simple and interpretable to mimic clinical use (Stiglic et al. 2020), and are described in subsequent subsections.

4.3 Overview of Experimental Results

This section presents an overview of our key findings, highlighting the probability of PVF outperforming the traditional validation method across all synthetic experimental configurations, as well as its benefits on real datasets after parameter tuning. Comprehensive descriptions of the experimental settings, configurations, detailed results, and sensitivity analyses are provided in the Appendix.

Synthetic Dataset Experiments As shown in Figure 3, PVF consistently outperforms the traditional validation method across most configurations (the configurations are elaborated in Section C.1). When $\gamma = 0.1$ and $\gamma = 0.3$,

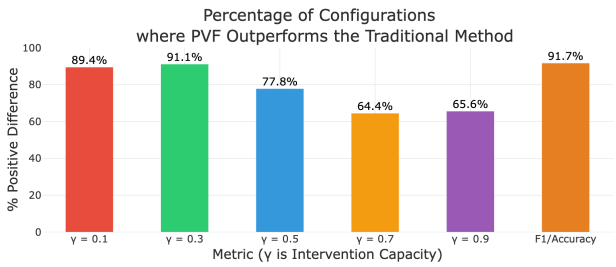

Figure 3: Percentage of configurations in synthetic experiments where PVF outperforms the traditional method across all metrics. "% Positive Difference" shows the percentage of configurations in which PVF more often selects the best model than the traditional method does across all repetitions.

PVF achieves superior selections in approximately 90% of all configurations, indicating PVF's advantage under limited intervention capacities. Although the advantage narrows as γ increases ($\gamma \geq 0.7$), PVF still maintains a clear lead.

Real-World Dataset Experiments In this subsection, similar to the synthetic dataset experiment, we also present the experimental results in the form of a comparison between the performance of PVF under σ (defined as in Section 3.2) tuning and that of the traditional method, respectively for the cervical cancer (Wolberg et al. 1993) and breast cancer (Fernandes, Cardoso, and Fernandes 2017) datasets. We refer the readers to Section D for the detailed setups of this part of experiments, including exactly how we calculate the performance gaps between PVF and the traditional method with repetitive experiments.

Results on cervical cancer data For the cervical cancer dataset, PVF achieves its largest performance gap with the traditional method with a small perturbation for IE ($\sigma = 0.01$) and a near-zero perturbation for F1 Score ($\sigma = 10^{-6}$), as shown in Figure 4. At the most stringent intervention capacity ($\gamma = 0.1$), PVF selects an externally better model in 60.0% of times, compared to 26.7% for the traditional method and 13.3% ties. As γ increases to 0.3–0.9, PVF continues to win more frequently, though the margin narrows: 43.3–46.7% times PVF wins versus 33.3–36.7% the traditional method does, with around 20% ties across these settings. For F1, PVF still leads with 50.0% wins, compared to 33.3% the traditional method does and 16.7% ties. These results indicate that once σ is tuned, PVF consistently yields more reliable model selection than the traditional method, with its strongest advantage at $\gamma = 0.1$, where false positives are extremely costly.

Results on breast cancer data For the breast cancer dataset, the perturbation scale that works best is much larger ($\sigma = 0.2$–0.3), as shown in Figure 5. At $\gamma = 0.1$, PVF wins in 52.0% of folds while the traditional method accounts for only 20.0%, with 28.0% ties. At $\gamma = 0.3$, the pattern is similar: 48.0% PVF, 20.0% traditional, and 32.0% ties. At higher intervention fractions ($\gamma = 0.5, 0.7, 0.9$) with $\sigma = 0.2$, results are more balanced, with PVF and the traditional method each winning 40.0% of folds and 20.0% ties. For F1 Score, the best σ is 0.2, and PVF wins in 52.0% of folds compared to 16.0% naive and 32.0% ties.

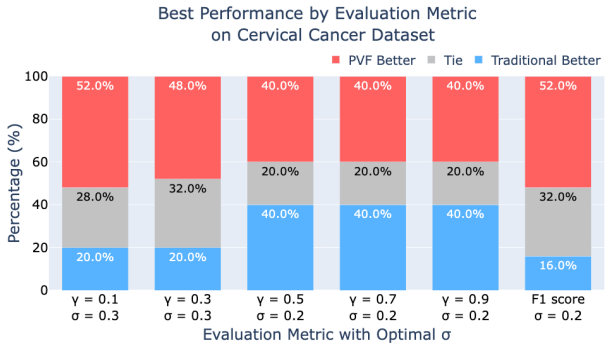

Figure 4: Largest performance gaps of PVF on the cervical cancer dataset under different evaluation metrics, when σ is tuned. PVF consistently selects better external models than the traditional method across all γ values, with its strongest advantage at $\gamma = 0.1$.

Real dataset result summary The perturbation level required for PVF to be most effective differs across datasets and evaluation metrics, but once tuned, PVF consistently outperforms or matches the traditional approach, with the largest benefits again observed at lower γ's.

Brief Conclusion of Experiments Across synthetic and real-data studies, PVF generally more often selects better models than the traditional single-split procedure. On synthetic datasets, PVF consistently improves selection under the traditional metrics (F1/accuracy) and IE, with the largest gains at smaller intervention capacities ($\gamma = 0.1, 0.3$) and F1 score/accuracy. On real datasets, PVF retains a clear advantage once the perturbation level is tuned, again with the strongest improvements at smaller capacities and F1 score. Section E further shows that σ is a key hyperparameter, and there is no universal best scale for different real datasets.

5 Conclusion

This work aim to address a persistent challenge in clinical machine learning: selecting a single, interpretable model when data are limited, class distributions are imbalanced, and measurements are noisy, potentially aggravating the Rashomon Effect. We propose *Intervention Efficiency* (IE), a capacity-aware evaluation metric that quantifies the expected benefit of a model when only a fixed fraction of individuals can receive an intervention, and *Perturbation Validation Framework* (PVF) that prioritizes candidate models whose performance remains strong under perturbations of the validation set. Together, IE and PVF facilitate model selection under operational constraints and robustness requirements that are typical of clinical deployment.

Empirical studies spanning an exhaustive synthetic enumeration and two clinical datasets indicate that the PVF

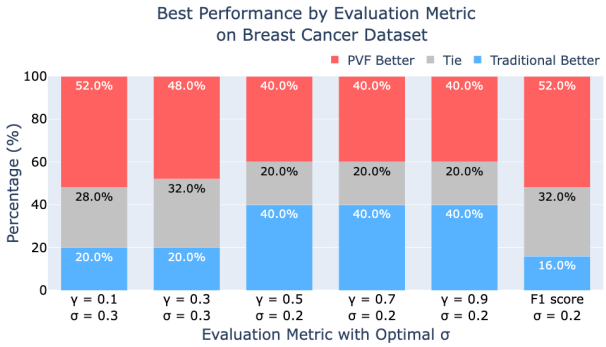

Figure 5: Largest performance gaps of PVF on the breast cancer dataset under different evaluation metrics, when σ is tuned. PVF tends to outperform the traditional method across lower γ values, specifically when $\gamma \in \{0.1, 0.3\}$, while tying with the traditional method at higher γ values.

more oftern identifies models that generalize better out of sample than a conventional single-split selection. The gains are most pronounced in the regimes with a low intervention capacity and when F1 score is the metric of interest.

As for limitations and future work, the selection and calibration of perturbation noise for PVF are consequential and may require domain-expert input to ensure that perturbations faithfully reflect clinical measurement variability and data quality. Otherwise, PVF may fail to outperform the traditional validation approach. Therefore, developing principled elicitation methods (e.g., protocol- or instrument-informed priors) and practical defaults for common data types would be crucial to improve applicability. On the theory side, there is substantial scope for formal analysis of PVF, including consistency guarantees and sensitivity to different performance metrics, especially the newly proposed IE. Such results would clarify applicable scenarios defined by data scale, noise, and resource limits. For IE, extensions to multiclass outcomes, integration of explicit cost, and consideration of fairness constraints remain important directions. Overall, coupling PVF with IE offers a practical, conceptual pathway and a future direction for navigating the Rashomon Effect in clinical machine learning, helping translate model multiplicity from a source of uncertainty into a framework for robust and resource-conscious model selection.

6 Acknowledgments

This work was supported by the Data+X (2024) project at DKU. We would also like to thank Prof. Dongmian Zou for valuable discussions on the theoretical aspects of the Intervention Efficiency and the Perturbation Validation Framework, and Prof. Chenkai Wu for his insightful suggestions on the Intervention Efficiency that helped us connect the proposed metric to practical interpretations.

References

Abdullah, T. A. A.; Zahid, M. S. M.; and Ali, W. 2021. A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions. *Symmetry*, 13(12): 2439.

Arleto, G.; Delfanti, L.; and Nguyen, H. 2023. A Scoping Review of Robustness Concepts for Machine Learning in Healthcare. *Journal of Biomedical Informatics*, 142: 104399.

Arlot, S.; and Celisse, A. 2010. A Survey of Cross-Validation Procedures for Model Selection. *Statistics Surveys*, 4: 40–79.

Berisha, V.; Krantsevich, C.; Hahn, P. R.; Turaga, P.; Liss, J.; and et al. 2021. Digital medicine and the curse of dimensionality. *NPJ Digital Medicine*, 4: 153.

Bommert, A.; Welchowski, T.; Schmid, M.; and Rahnenführer, J. 2022. Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. *Briefings in Bioinformatics*, 23(1): 1–13.

Breiman, L. 1996a. Bagging Predictors. *Machine Learning*, 24(2): 123–140.

Breiman, L. 1996b. Heuristics of Instability and Stabilization in Model Selection. *The Annals of Statistics*, 24(6): 2350–2383.

Breiman, L. 2001a. Random Forests. *Machine Learning*, 45(1): 5–32.

Breiman, L. 2001b. Statistical modeling: The two cultures. *Statistical Science*, 16: 199–231.

Buda, M.; Maki, A.; and Mazurowski, M. A. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106: 249–259.

Bühlmann, P.; and Yu, B. 2002. Analyzing Bagging. *The Annals of Statistics*, 30(4): 927–961.

Cawley, G. C.; and Talbot, N. L. 2010a. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70): 2079–2107.

Cawley, G. C.; and Talbot, N. L. C. 2010b. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, 11: 2079–2107.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16: 321–357.

Chen, J.; Aseltine, R. H.; Wang, F.; and Chen, K. 2022. Tree-Guided Rare Feature Selection and Logic Aggregation with Electronic Health Records Data. *Journal of the American Statistical Association*, 117(540): 1594–1608.

Chicco, D.; and Jurman, G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1): 6.

Christodoulou, E.; Ma, J.; Collins, G. S.; Steyerberg, E. W.; Verbakel, J. Y.; and Van Calster, B. 2019. A Systematic Review Shows No Performance Benefit of Machine Learning Over Logistic Regression for Clinical Prediction Models. *Journal of Clinical Epidemiology*, 110: 12–22.

Coors, M.; Bauer, L.; Edwards, K.; Erickson, K.; Goldenberg, A.; Goodale, J.; Goodman, K.; Grady, C.; Mannino, D.; Wanner, A.; Wilson, T.; Yarborough, M.; and Zirkle, M. 2017. Ethical Issues Related to Clinical Research and Rare Diseases: 15th Gordon L. Snider Critical Issues Workshop, April 1, 2016, Bethesda, Maryland. *Translational Science of Rare Diseases*, 2(3-4): 175–194.

Davis, K.; and Maiden, R. 2021. Addressing Accuracy Paradox Using Enhanced Weighted Performance Metric in Machine Learning. *Studies in Social Science Research*, 2(2): 1–10.

Dietterich, T. G. 2000. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*, 1–15. Springer.

Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; and Ma, Q. 2020. A Survey on Ensemble Learning. *Frontiers of Computer Science*, 14(2): 241–258.

Egmont-Petersen, M.; Talmon, J. L.; and Hasman, A. 1997. Robustness Metrics for Measuring the Influence of Additive Noise on Statistical Classifiers. *International Journal of Medical Informatics*, 46: 103–112.

Fernandes, K.; Cardoso, J.; and Fernandes, J. 2017. Cervical Cancer (Risk Factors). [Online; accessed October 2025].

Gao, R.; Tang, Y.; Khan, M. S.; Xu, K.; Paulson, A. B.; Sullivan, S.; Huo, Y.; Deppen, S.; Massion, P. P.; Sandler, K. L.; and Landman, B. A. 2021. Cancer Risk Estimation Combining Lung Screening CT with Clinical Data Elements. *Radiology: Artificial Intelligence*, 3(6): e210032.

He, H.; and Garcia, E. A. 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9): 1263–1284.

Heinrich, M.; Amos, C.; Kratzwald, B.; Kluth, O.; Scheipl, F.; and Percha, B. 2022. Assessing robustness of clinical prediction models under dataset shift. *Bioinformatics*, 38(10): 2841–2848.

Jia, W.; Sun, M.; Lian, J.; and Hou, S. 2022. Feature Dimensionality Reduction: A Review. *Complex & Intelligent Systems*, 8(3): 2663–2693.

Jolliffe, I. T.; and Cadima, J. 2016. Principal Component Analysis: A Review and Recent Developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065): 20150202.

Kulkarni, S. A.; Gurpur, V.; King, C.; and Koval, A. 2021. Impact of Gaussian Noise on Optimised Support Vector Machines Applied to Medicare Payment. *Informatica*, 45(4): 643–652.

Li, H.; Zou, R.; Xin, H.; He, P.; Xi, B.; Tian, Y.; Zhao, Q.; Yan, X.; Qiu, X.; Gao, Y.; Liu, Y.; Cao, M.; Chen, B.; Han, Q.; Chen, J.; Wang, G.; and Cai, H. 2025. Mortality Risk Prediction in Patients With Antimelanoma Differentiation-Associated, Gene 5 Antibody-Positive, Dermatomyositis-Associated Interstitial Lung Disease: Algorithm Development and Validation. *Journal of Medical Internet Research*, 27(1): e62836.

Martin, B.; Bennett, T. D.; DeWitt, P. E.; Russell, S.; and Sanchez-Pinto, L. N. 2025. Use of the Area Under the Precision–Recall Curve to Evaluate Prediction Models of Rare Critical Illness Events. *Pediatric Critical Care Medicine*. Epub ahead of print.

Meinshausen, N.; and Bühlmann, P. 2010. Stability Selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4): 417–473.

Morozova, O.; Levina, O.; Uusküla, A.; and Heimer, R. 2015. Comparison of Subset Selection Methods in Linear Regression in the Context of Health-Related Quality of Life and Substance Abuse in Russia. *BMC Medical Research Methodology*, 15: 71.

Nguyen, T.; Patel, A.; Cohen, S.; Chang, E. Y.; Bhattacharya, S.; Prakash, P.; Rao, R.; and Hu, X. 2023. Interpretable Machine Learning Models for Hospital Readmission Prediction. *BMC Medical Informatics and Decision Making*, 23(1): 175.

Ning, Y.; Li, S.; Ong, M. E. H.; Xie, F.; Chakraborty, B.; Ting, D. S. W.; and Liu, N. 2022. A Novel Interpretable Machine Learning System to Generate Clinical Risk Scores: An Application for Predicting Early Mortality or Unplanned Readmission. *JAMA Network Open*, 5(6): e2215360.

Opitz, D.; and Maclin, R. 1999. Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, 11: 169–198.

Parmar, C.; Grossmann, P.; Rietbergen, M. M.; Lambin, P.; and Aerts, H. J. W. L. 2015. Radiomic Machine-Learning Classifiers for Prognostic Biomarkers: Robustness to Magnetic Resonance Image Noise. *Radiology*, 279(2): 505–514.

Paschali, M.; Conjeti, S.; Navarro, F.; and Navab, N. 2018. Generalizability vs. robustness: Adversarial examples for medical imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 52–60.

Pfohl, S. R.; Shivade, C.; Malin, B.; Denny, J.; and Weber, G. M. 2019. Benchmarking clinical prediction and risk stratification models on publicly available electronic health records. *NPJ Digital Medicine*, 2(1): 2.

Pfohl, S. R.; Zhang, Z.; and Celi, L. A. 2021. Quantifying the Impact of Data Noise on Intensive Care Unit Mortality Prediction Models. *Critical Care Medicine*, 49(1): e9–e17.

Richardson, E.; Trevizani, R.; Greenbaum, J. A.; Carter, H.; Nielsen, M.; and Peters, B. 2024. The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns*, 5(6): 100994.

Rudin, C. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5): 206–215.

Rudin, C.; Parr, R.; and Seltzer, M. 2024. Position Paper: Amazing Things Come From Having Many Good Models. *Proceedings of Machine Learning Research*, 235: 1–6.

Saito, T.; and Rehmsmeier, M. 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3): e0118432.

Shafiee, S.; Kapralos, B.; Kay, J.; and Jaffray, D. A. 2019. Impact of Gaussian Noise on CT-Based Radiomic Features and Predictive Models. *Medical Physics*, 46(3): 1134–1146.

Soloff, J. A.; Foygel Barber, R.; and Willett, R. 2024. Bagging Provides Assumption-Free Stability. *Journal of Machine Learning Research*, 25(1): 6336–6370.

Stiglic, G.; Kocbek, P.; Fijacko, N.; Zitnik, M.; Verbert, K.; and Cilar, L. 2020. Interpretability of Machine Learning Based Prediction Models in Healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5): e1379.

Suttaket, T.; Harsha Vardhan, V.; and Kok, S. 2024. Interpretable Predictive Models for Healthcare via Rational Logistic Regression. *Journal of Biomedical Informatics*, 142: 104827.

Sztukiewicz, Ł.; Good, J. H.; and Dubrawski, A. W. 2024. Exploring Loss Design Techniques for Decision Tree Robustness to Label Noise. In *Proc. International Conference on Learning Representations (ICLR)*.

Tan, Y. S.; Singh, C.; Nasseri, K.; Agarwal, A.; Duncan, J.; Ronen, O.; Epland, M.; Kornblith, A.; and Yu, B. 2022. Fast Interpretable Greedy-Tree Sums (FIGS) for Clinical Decision Instruments. *Proceedings of the National Academy of Sciences*, 119(12): e2310151122.

Valente, F.; Henriques, J.; Paredes, S.; Rocha, T.; de Carvalho, P.; and Morais, J. 2021. Improving the compromise between accuracy, interpretability and personalization of rule-based machine learning in medical problems. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2132–2135.

van den Goorbergh, R.; van Smeden, M.; Timmerman, D.; and Van Calster, B. 2022. The Harm of Class Imbalance Corrections for Risk Prediction Models: Illustration and Simulation Using Logistic Regression. *Journal of the American Medical Informatics Association*, 29(10): 1787–1795.

Varma, S.; and Simon, R. 2006. Bias in Error Estimation When Using Cross-Validation for Model Selection. *BMC Bioinformatics*, 7(1): 91.

Varoquaux, G. 2018. Cross-validation failure: small sample sizes lead to large error bars. *NeuroImage*, 180: 68–77.

Wainer, J.; and Cawley, G. 2021. Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications*, 166: 114114.

Wolberg, W.; Mangasarian, O.; Street, N.; and Street, W. 1993. Breast Cancer Wisconsin (Diagnostic). [Online; accessed October 2025].

Xin, R.; Zhong, C.; Chen, Z.; Takagi, T.; Seltzer, M.; and Rudin, C. 2022. Exploring the Whole Rashomon Set of Sparse Decision Trees. In *Advances in Neural Information Processing Systems*, volume 35, 14071–14084.

Yang, R.; Yu, L.; Chen, H.; and Zhang, Y. 2020. Evaluating the Stability of ECG Signal Classification under Gaussian Noise. *Journal of Medical Systems*, 44(7): 122.

Zadeh, M. R.; Moghadamnia, M.; Rezaei, M.; and et al. 2023. A comparative study of explainable ensemble learning and logistic regression for in-hospital mortality prediction in the emergency department. *Scientific Reports*, 13: 17541.

Zhang, J. M.; Harman, M.; Guedj, B.; Barr, E. T.; and Shawe-Taylor, J. 2023. Model validation using mutated training labels: An exploratory study. *Neurocomputing*, 539: 126116.

Zhang, Y.; Yang, Y.; and Wang, L. 2021. Cross-Validation Failure: Small Sample Sizes Lead to Large Error Bars. *NeuroImage*, 210: 116471.

Łukaszuk, T.; and Krawczuk, J. 2024. Importance of Feature Selection Stability in Classifier Evaluation on High-Dimensional Genetic Data. *PeerJ*, 12: e18405.

A Proof of Proposition 3.1

Proof. We present the derivation in five steps including the analysis of the model's capability, the two operating regimes, the expression unification for the two regimes, and the transition from counting-based expression to the final ratio-based one. All notations and terms follow the same definitions in Proposition 3.1. For clarity, we additionally denote α as the total number of positives in the population, β as the population size, and c as the number of individuals that can be intervened. Specifically,

$$\pi = \frac{\alpha}{\beta}, \quad \gamma = \frac{c}{\beta}.$$

(i) Model capability. We first calculate how many individuals the model flags positive to attain the number of true positives. In the following analysis, TP, FP, TN, FN denote true positives, false positives, true negatives and false negatives in the confusion matrix. Since $r = TP/\alpha$, the model can capture at most $TP = \alpha\, r$ positives. By the definition of precision,

$$p = \frac{TP}{TP + FP} \quad \Longrightarrow \quad TP + FP = \frac{TP}{p} = \frac{\alpha\, r}{p}.$$

Therefore, the model labels $\frac{\alpha\, r}{p}$ individuals as positive and captures $\alpha\, r$ true positives in the whole population. Note that we may not be able to intervene on all these $\frac{\alpha\, r}{p}$ cases due to the fixed budget c, and this creates the two following regimes.

(ii) Regime A (scarce resources). If $\frac{\alpha\, r}{p} > c$, only c individuals can be intervened, which is insufficient for the model to realize its theoretical best in capturing all $\alpha\, r$ positives. Each model-guided intervention yields a true positive with probability p, so the expected number of positives captured by the model in this regime, denoted as T^{A}, satisfies

$$T^{\mathrm{A}} = c\, p.$$

By comparison, random selection yields $T_{\mathrm{random}} = c\,(\alpha/\beta)$ true positives in expectation. Hence, by definition, the IE value in this regime satisfies

$$IE^{\mathrm{A}} = \frac{T^{\mathrm{A}}}{T_{\mathrm{random}}} = \frac{c\, p}{c\,(\alpha/\beta)} = \frac{\beta\, p}{\alpha}.$$

(iii) Regime B (ample resources). If $\frac{\alpha\, r}{p} \leq c$, the budget suffices to treat all model-flagged positives. We intervene through a two-stage procedure. In stage 1, intervening on those $\frac{\alpha\, r}{p}$ cases predicted positive by the

model captures $T^{\mathrm{B},1} = \alpha\, r$ positives. The remaining budget is $c_{\mathrm{rem}} = c - \frac{\alpha\, r}{p}$. The remaining pool has

$$\beta_{\mathrm{rem}} = \beta - \tfrac{\alpha\, r}{p}, \quad \alpha_{\mathrm{rem}} = \alpha - \alpha\, r, \quad \pi_{\mathrm{rem}} = \frac{\alpha_{\mathrm{rem}}}{\beta_{\mathrm{rem}}},$$

where β_{rem} denotes the size of the residual population after removing the $\frac{\alpha r}{p}$ intervened, model-flagged individuals; α_{rem} is the number of true positives remaining in that residual population; and π_{rem} is the corresponding prevalence rate within this residual pool. We assume that the predictive model is uninformative on this residual pool. Therefore, in stage 2, the remaining c_{rem} interventions are allocated uniformly at random, yielding

$$T^{\mathrm{B},2} = c_{\mathrm{rem}}\, \pi_{\mathrm{rem}}$$

additional expected true positives. Consequently, the total number of positives captured by the model in regime B, denoted as T^{B}, satisfies

$$T^{\mathrm{B}} = T^{\mathrm{B},1} + T^{\mathrm{B},2} = \alpha\, r + \Big(c - \tfrac{\alpha\, r}{p}\Big)\, \frac{\alpha - \alpha\, r}{\beta - \frac{\alpha\, r}{p}},$$

and the IE value in this regime satisfies

$$IE^{\mathrm{B}} = \frac{\alpha\, r + \Big(c - \frac{\alpha\, r}{p}\Big)\, \frac{\alpha - \alpha\, r}{\beta - \frac{\alpha\, r}{p}}}{c\, \frac{\alpha}{\beta}}.$$

(iv) Unified expression. To avoid case-by-case handling, set $c' = \min\big(c, \frac{\alpha\, r}{p}\big)$. Then, the expected positives captured respectively by the model, denoted as T_{model}, and by random selection, are

$$T_{\mathrm{model}} = c'\, p + (c - c')\, \frac{\alpha - c'\, p}{\beta - c'}, \qquad T_{\mathrm{random}} = c\, \frac{\alpha}{\beta}.$$

Therefore, the counting-based expression of IE satisfies

$$\mathrm{IE}_\gamma(f) = \frac{T_{\mathrm{model}}}{T_{\mathrm{random}}} = \frac{c'\, p \;+\; \big(c - c'\big)\, \frac{\alpha - c'\, p}{\beta - c'}}{c\, \frac{\alpha}{\beta}}.$$

When $c < \frac{\alpha\, r}{p}$, this reduces to IE^{A}; otherwise it reduces to IE^{B}.

(v) Ratio-based expression. Recalling the expression of s and c', we can notice that

$$s = \min\Big(\gamma, \frac{\pi r}{p}\Big) = \frac{1}{\beta} \cdot \min\Big(c, \frac{\alpha r}{p}\Big) = \frac{c'}{\beta}.$$

To leverage this relationship and get rid of c', α and β in the formula of IE, we divide both the numerator and denominator of the current form with β to obtain the ratio-based expression:

$$\begin{aligned}
\mathrm{IE}_\gamma(f) &= \frac{\frac{T_{\mathrm{model}}}{\beta}}{\frac{T_{\mathrm{random}}}{\beta}} \\
&= \frac{\frac{c'\, p}{\beta} + \Big(\frac{c}{\beta} - \frac{c'}{\beta}\Big)\, \frac{\frac{\alpha}{\beta} - \frac{c'\, p}{\beta}}{1 - \frac{c'}{\beta}}}{\frac{c}{\beta} \cdot \frac{\alpha}{\beta}} \\
&= \frac{sp + (\gamma - s)\, \frac{\pi - sp}{1 - s}}{\gamma\pi},
\end{aligned}$$

whence the stated closed form for $\mathrm{IE}_\gamma(f)$.

□

B Theoretical Guarantees of PVF

In this section we provide theoretical guarantees for PVF. We formalize how the PVF score behaves as the number of perturbed validation sets, the number of perturbation replicas, and the validation set size grow, and we show that PVF asymptotically selects models that are optimal for a property determined by the perturbation mechanism and the aggregation rule.

B.1 Setup

We use the notation introduced in Section 3.2. The original validation set is

$$\mathcal{D}_{\mathrm{val}} = \{(x_i, y_i)\}_{i=1}^{n},$$

where n is the number of validation points, each x_i is a feature vector, and y_i is its label.

PVF uses a fixed perturbation mechanism that, given a feature vector x, produces a random perturbed feature vector X'. We denote this conditional distribution by

$$X' \mid X = x \sim K(\cdot \mid x).$$

In the implementation, K is specified by Gaussian noise for numeric features, label-flip noise for categorical features, and decay rules for ordinal features; for the theory, we only need that K is fixed and the perturbations are independent across samples and replicas.

From the original data distribution $\mathcal{P}$ on (X, Y) and the perturbation rule K, we define the induced perturbed distribution $\tilde{\mathcal{P}}$ on (X', Y) by

$$(X, Y) \sim \mathcal{P}, \qquad X' \mid X \sim K(\cdot \mid X), \qquad (X', Y) \sim \tilde{\mathcal{P}}.$$

For each perturbation index $m = 1, \ldots, M$, PVF constructs a perturbed validation set

$$\mathcal{D}_{\mathrm{val}}^{(m)} = \big\{(x_i^{(m,j)}, y_i)\big\}_{i=1,\ldots,n;\; j=1,\ldots,k},$$

where, for each fixed i, the replicas $\{x_i^{(m,j)}\}_{j=1}^k$ are i.i.d. drawn from $K(\cdot \mid x_i)$, and the labels y_i are unchanged.

For a model f in a finite candidate set F, the evaluation metric on each perturbed set is

$$P_f^{(m)} = \mathrm{Eval}(f, \mathcal{D}_{\mathrm{val}}^{(m)}),$$

and the vector of perturbed scores is

$$P_f = \big(P_f^{(1)}, \ldots, P_f^{(M)}\big) \in \mathbb{R}^M.$$

The PVF score of model f is obtained by an aggregation rule

$$A_f = A(P_f),$$

where $A\colon \mathbb{R}^M \to \mathbb{R}$ is chosen by the user (for example, A can be the mean or a lower quantile). PVF then selects

$$f^\star \in \arg\max_{f \in F} A_f.$$

We make the following assumptions, which will be referenced explicitly in the propositions and proofs.

(A1) $\mathcal{D}_{\mathrm{val}} = \{(x_i, y_i)\}_{i=1}^n$ is drawn i.i.d. from a distribution $\mathcal{P}$ on (X, Y).

(A2) For each feature vector x, the perturbation rule is given by a fixed conditional distribution $K(\cdot \mid x)$, and all perturbations across different samples, replicas, and perturbed sets are independent.

(A3) The set of candidate models F is a finite set.

(A4) For any fixed model f and any dataset D, the evaluation metric $\mathrm{Eval}(f, D)$ is a bounded function of empirical statistics of D. In particular, for the metrics used in the paper (IE, F1, accuracy), $\mathrm{Eval}(f, D)$ depends on D only through the empirical confusion matrix entries, which are empirical averages of bounded per-sample indicators. Generally, $\mathrm{Eval}(f, D)$ should be a metric that calculates a generalized mean of the performance on each data point.

(A5) There exists a functional $\mathcal{A}$ on probability distributions on $\mathbb{R}$ such that, if $Z_1, \ldots, Z_M$ are i.i.d. with common distribution L, then

$$A(Z_1, \ldots, Z_M) \xrightarrow[M\to\infty]{\text{a.s.}} \mathcal{A}(L).$$

This holds for standard aggregation rules including the sample mean, sample quantiles, the median, and more general continuous L-statistics.

We now state and prove the theoretical guarantees under these assumptions.

B.2 Convergence of PVF Scores

The first step is to understand the behavior of the PVF score A_f for a fixed model f as the sampling dimensions M, k, and n grow.

For each fixed n, k and realization of $\mathcal{D}_{\mathrm{val}}$, let $\mathcal{L}_{n,k}(f)$ denote the conditional distribution of $P_f^{(1)} = \mathrm{Eval}(f, \mathcal{D}_{\mathrm{val}}^{(1)})$ given $\mathcal{D}_{\mathrm{val}}$. Similarly, for each n, we will use $\mathcal{L}_{n,\infty}(f)$ for the limiting distribution of $P_f^{(1)}$ as $k \to \infty$.

Proposition B.1 (i.i.d. structure of perturbed scores)**.** *Suppose (A1) and (A2) hold, and fix $\mathcal{D}_{\mathrm{val}}$, n, k, and a model $f \in F$. Then, conditional on $\mathcal{D}_{\mathrm{val}}$, the scores $P_f^{(1)}, \ldots, P_f^{(M)}$ are i.i.d. random variables with common distribution $\mathcal{L}_{n,k}(f)$.*

Proof. Assumption (A2) states that, for each i, all perturbations $\{x_i^{(m,j)}\}$ across different m and j are independent given $\mathcal{D}_{\mathrm{val}}$ and follow the same distribution $K(\cdot \mid x_i)$. By construction, each $\mathcal{D}_{\mathrm{val}}^{(m)}$ is built from independent perturbations of $\mathcal{D}_{\mathrm{val}}$, and thus the sets $\mathcal{D}_{\mathrm{val}}^{(1)}, \ldots, \mathcal{D}_{\mathrm{val}}^{(M)}$ are conditionally independent and identically distributed given $\mathcal{D}_{\mathrm{val}}$. Assumption (A4) ensures that $\mathrm{Eval}(f, \cdot)$ is a deterministic and measurable function of its input dataset. Because determinism implies that $\mathrm{Eval}(f, \cdot)$ introduces no additional randomness beyond that already present in the dataset, applying $\mathrm{Eval}(f, \cdot)$ to each perturbed dataset $\mathcal{D}_{\mathrm{val}}^{(m)}$ will not create dependence between the resulting scores.

Formally, for any real intervals $I_1, \ldots, I_k$ and any indices $m_1, \ldots, m_k$, the events $\{P_f^{(m_i)} \in I_i\}$ satisfy

$$\{P_f^{(m_i)} \in I_i\} = \{\mathcal{D}_{\mathrm{val}}^{(m_i)} \in \mathrm{Eval}(f, \cdot)^{-1}(I_i)\},$$

where $\mathrm{Eval}(f, \cdot)^{-1}(I_i)$ is the preimage of I_i under the evaluation map. Since the perturbed datasets $\mathcal{D}_{\mathrm{val}}^{(1)}, \ldots, \mathcal{D}_{\mathrm{val}}^{(M)}$ are independent conditional on $\mathcal{D}_{\mathrm{val}}$ (by construction under Assumption (A2)), we have

$$\Pr\Big(P_f^{(m_1)} \in I_1, \ldots, P_f^{(m_k)} \in I_k \mid \mathcal{D}_{\mathrm{val}}\Big) = \prod_{i=1}^k \Pr\Big(P_f^{(m_i)} \in I_i \mid \mathcal{D}_{\mathrm{val}}\Big).$$

This shows that evaluating each $\mathcal{D}_{\mathrm{val}}^{(m)}$ through the deterministic map $\mathrm{Eval}(f, \cdot)$ preserves the i.i.d. structure established at the level of the perturbed datasets.

Therefore, the resulting scores $P_f^{(1)}, \ldots, P_f^{(M)}$ are conditionally i.i.d. given $\mathcal{D}_{\mathrm{val}}$ with some common distribution, which we denote by $\mathcal{L}_{n,k}(f)$. □

Interpretation: for a fixed validation set and perturbation parameters, PVF is aggregating many i.i.d. draws of a random performance score for each model under the chosen perturbation rule.

Proposition B.2 (Convergence in M)**.** *Suppose (A1), (A2), (A4), and (A5) hold, and fix $\mathcal{D}_{\mathrm{val}}$, n, k, and $f \in F$. Then*

$$A_f = A\big(P_f^{(1)}, \ldots, P_f^{(M)}\big) \xrightarrow[M\to\infty]{\text{a.s.}} \mathcal{A}\big(\mathcal{L}_{n,k}(f)\big).$$

Proof. By Proposition B.1, conditional on $\mathcal{D}_{\mathrm{val}}$, the scores $P_f^{(1)}, \ldots, P_f^{(M)}$ are i.i.d. with common distribution $\mathcal{L}_{n,k}(f)$. Assumption (A5) states that for any i.i.d. sequence from a distribution L, the aggregation rule satisfies

$$A(Z_1, \ldots, Z_M) \to \mathcal{A}(L) \quad \text{almost surely as } M \to \infty.$$

Applying this with $Z_m = P_f^{(m)}$ and $L = \mathcal{L}_{n,k}(f)$ yields the claimed convergence. □

Interpretation: when the validation set and perturbation settings are fixed, increasing the number of perturbed validation sets guarantees that PVF approximates the population-level value of the aggregation rule applied to the distribution of perturbed scores for each model.

Proposition B.3 (Convergence in k)**.** *Suppose (A1), (A2), and (A4) hold, and fix $\mathcal{D}_{\text{val}}$, n, and $f \in F$. Then, there exists a distribution $\mathcal{L}_{n,\infty}(f)$ such that*

$$\mathcal{L}_{n,k}(f) \xrightarrow[k\to\infty]{} \mathcal{L}_{n,\infty}(f),$$

and, if (A5) also holds,

$$\mathcal{A}\big(\mathcal{L}_{n,k}(f)\big) \xrightarrow[k\to\infty]{} \mathcal{A}\big(\mathcal{L}_{n,\infty}(f)\big).$$

Proof. Fix $\mathcal{D}_{\text{val}}$, n, and f. By Assumption (A2), for each original sample (x_i, y_i), the replicas $\{x_i^{(1,j)}\}_{j=1}^{k}$ are i.i.d. from $K(\cdot \mid x_i)$. Assumption (A4) states that $\text{Eval}(f, \mathcal{D}_{\text{val}}^{(1)})$ is a bounded function of empirical statistics of $\mathcal{D}_{\text{val}}^{(1)}$. By the law of large numbers, for the replicas, as $k \to \infty$, these empirical statistics of $\mathcal{D}_{\text{val}}^{(1)}$ converge almost surely to their conditional expectations given $\mathcal{D}_{\text{val}}$. Since $\text{Eval}(f, \cdot)$ is a deterministic function of these statistics, the distribution of $P_f^{(1)} = \text{Eval}(f, \mathcal{D}_{\text{val}}^{(1)})$ also converges to a limiting distribution, which we denote by $\mathcal{L}_{n,\infty}(f)$. Assumption (A5) then implies the convergence of $\mathcal{A}(\mathcal{L}_{n,k}(f))$ to $\mathcal{A}(\mathcal{L}_{n,\infty}(f))$. □

Interpretation: as the number of perturbation replicas per validation point grows, the randomness arising from finite perturbation sampling disappears, and the distribution of perturbed scores for each model stabilizes to a limit.

Proposition B.4 (Convergence in n)**.** *Suppose (A1), (A2), and (A4) hold. For each model $f \in F$, define*

$$J(f) := \text{Eval}(f, \tilde{\mathcal{P}}),$$

the performance value of f under the perturbed population distribution $\tilde{\mathcal{P}}$. Then, as $n \to \infty$,

$$\mathcal{L}_{n,\infty}(f) \xrightarrow[n\to\infty]{} \delta_{J(f)},$$

where $\delta_{J(f)}$ is the probability distribution that assigns probability 1 to the single point $J(f)$; equivalently, a random variable Z has distribution $\delta_{J(f)}$ if and only if

$$\Pr(Z = J(f)) = 1.$$

If (A5) also holds, then

$$\mathcal{A}\big(\mathcal{L}_{n,\infty}(f)\big) \xrightarrow[n\to\infty]{P} \mathcal{A}\big(\delta_{J(f)}\big) =: J_A(f).$$

Proof. Assumption (A1) implies that the empirical distribution of $\mathcal{D}_{\text{val}}$ converges to $\mathcal{P}$ as $n \to \infty$. Combined with the perturbation rule in (A2), this implies that the empirical distribution of perturbed samples in each $\mathcal{D}_{\text{val}}^{(1)}$ converges to $\tilde{\mathcal{P}}$. By Assumption (A4), $\text{Eval}(f, \mathcal{D}_{\text{val}}^{(1)})$ is a bounded function of empirical statistics of $\mathcal{D}_{\text{val}}^{(1)}$. These statistics converge to their expectations under $\tilde{\mathcal{P}}$, and hence $\text{Eval}(f, \mathcal{D}_{\text{val}}^{(1)})$ converges in probability to $\text{Eval}(f, \tilde{\mathcal{P}}) = J(f)$. Therefore, the distribution $\mathcal{L}_{n,\infty}(f)$ converges to $\delta_{J(f)}$. Assumption (A5) then yields $\mathcal{A}(\mathcal{L}_{n,\infty}(f)) \to \mathcal{A}(\delta_{J(f)})$. □

Interpretation: with a sufficiently large validation set, all randomness in the perturbed score for a fixed model collapses to a single deterministic value $J(f)$, and the limit of any reasonable aggregation of that score is just a deterministic function $J_A(f)$ of the same quantity.

Proposition B.5 (Unified convergence of PVF scores)**.** *Suppose (A1)–(A5) hold. Then, for any fixed model $f \in F$,*

$$A_f \xrightarrow[M,k,n\to\infty]{P} J_A(f),$$

where $J_A(f) = \mathcal{A}(\delta_{J(f)})$ and $J(f) = \text{Eval}(f, \tilde{\mathcal{P}})$.

Proof. For fixed $\mathcal{D}_{\text{val}}$, n, and k, Proposition B.2 implies

$$A_f \xrightarrow[M\to\infty]{\text{a.s.}} \mathcal{A}\big(\mathcal{L}_{n,k}(f)\big).$$

Proposition B.3 implies

$$\mathcal{A}\big(\mathcal{L}_{n,k}(f)\big) \xrightarrow[k\to\infty]{} \mathcal{A}\big(\mathcal{L}_{n,\infty}(f)\big),$$

and Proposition B.4 implies

$$\mathcal{A}\big(\mathcal{L}_{n,\infty}(f)\big) \xrightarrow[n\to\infty]{P} \mathcal{A}\big(\delta_{J(f)}\big) =: J_A(f).$$

Combining these convergences (almost sure in M, then in k, then in probability in n) yields the stated convergence of A_f to $J_A(f)$ in probability as all three variables grow. □

Interpretation: for each model, as we use more perturbed validation sets, more perturbation replicas, and more validation data, the PVF score converges to a deterministic limit that depends only on the underlying data distribution, the perturbation rule, the evaluation metric, and the aggregation rule.

B.3 Consistency of PVF Model Selection

We now show that, under the same assumptions, PVF consistently selects models that maximize the objective $J_A(f)$.

Define

$$J_A \colon F \to \mathbb{R}, \qquad f \mapsto J_A(f),$$

and the set of maximizers

$$F_A^\star := \{f \in F : J_A(f) = \max_{g \in F} J_A(g)\}.$$

Proposition B.6 (Consistency of PVF selection)**.** *Suppose (A1)–(A5) hold. Then the PVF-selected model $f^\star$ satisfies*

$$f^\star \xrightarrow[M,k,n\to\infty]{P} f_\infty^\star,$$

for some $f_\infty^\star \in F_A^\star$.

Proof. By Proposition B.5, for each fixed $f \in F$,

$$A_f \xrightarrow[M,k,n\to\infty]{P} J_A(f).$$

Assumption (A3) states that F is finite. Let $F = \{f_1, \ldots, f_Q\}$. For each $i \in \{1, \ldots, Q\}$, define

$$A_i := A_{f_i}, \qquad J_i := J_A(f_i).$$

Since convergence in probability holds for each coordinate individually, the whole score vector converges in probability:

$$(A_1, \ldots, A_Q) \xrightarrow[M,k,n\to\infty]{P} (J_1, \ldots, J_Q).$$

To analyze the argmax selection rule, let

$$F_A^\star := \{i : J_i = \max_{j=1,\ldots,Q} J_j\}$$

be the (nonempty) set of population maximizers. For any index $i \notin F_A^\star$, we have a strict gap

$$J_{i^\star} - J_i > 0 \qquad \text{for every } i^\star \in F_A^\star.$$

Define

$$\Delta := \min_{i \notin F_A^\star} (J_{i^\star} - J_i),$$

which is strictly positive.

Now, fix any $\varepsilon < \Delta/3$. We can see that, if all coordinates satisfy

$$|A_i - J_i| < \varepsilon \qquad \text{for all } i = 1, \ldots, Q,$$

then, for every $i^\star \in F_A^\star$ and $i \notin F_A^\star$,

$$A_{i^\star} \geq J_{i^\star} - \varepsilon > J_i + \Delta - \varepsilon > J_i + 2\varepsilon > A_i.$$

Hence, whenever the coordinatewise errors are smaller than ε, every truly optimal model has strictly larger estimated score than every non-optimal model, and thus the maximizer of the estimated scores must lie in the true maximizer set, ensuring that we get the true optimal model:

$$\arg\max_{i=1,\ldots,Q} A_i \subseteq F_A^\star.$$

Since

$$\Pr\Big(\max_i |A_i - J_i| > \varepsilon\Big) \longrightarrow 0,$$

it follows that

$$\Pr\Big(\arg\max_i A_i \subseteq F_A^\star\Big) \longrightarrow 1.$$

Any deterministic tie-breaking rule selects a single element of $\arg\max_i A_i$, so the selected model index converges in probability to a member of $F_A^\star$. Rewriting indices back in terms of models, we obtain:

$$f^\star \in \arg\max_{f \in F} A_f \quad \Longrightarrow \quad f^\star \xrightarrow{P} F_A^\star.$$

Thus any PVF selection rule converges in probability to a population maximizer of J_A. □

Interpretation: PVF is a consistent model selection procedure. In the limit of large validation sets and sufficient perturbation sampling, PVF will choose a model whose PVF score is maximal among all candidates.

B.4 PVF as Selection Toward a User-Specified Property

Finally, we interpret the limit objective J_A as an induced property determined by the design choices in PVF.

For a fixed perturbation rule K, evaluation metric Eval, and aggregation rule A, we define the induced property

$$\Phi_{A,K}(f) := J_A(f) = \mathcal{A}\big(\delta_{J(f)}\big), \qquad J(f) = \mathrm{Eval}(f, \tilde{\mathcal{P}}).$$

This property assigns to each model f a scalar that reflects its performance under the perturbed distribution $\tilde{\mathcal{P}}$, summarized through the aggregation rule A.

Proposition B.7 (PVF selects toward the induced property)**.** *Suppose (A1)–(A5) hold. Let $\Phi_{A,K}(f) := J_A(f)$ be the induced property described above. Then*

$$f^\star \xrightarrow[M,k,n\to\infty]{P} f_{\mathrm{opt}},$$

for some

$$f_{\mathrm{opt}} \in \arg\max_{f \in F} \Phi_{A,K}(f).$$

Proof. This is a direct restatement of Proposition B.6, since $\Phi_{A,K}(f)$ is defined to be $J_A(f)$ for each $f \in F$. □

Interpretation: once a user fixes a perturbation mechanism and an aggregation rule, these choices implicitly define a property $\Phi_{A,K}$ of interest (for example, robustness to specific perturbations, emphasis on lower-tail performance, or any other stability notion encoded by K and A). The propositions in this section show that, under the stated assumptions, PVF is guaranteed to select models that maximize this induced property.

C Detailed Setups and Results of Synthetic Dataset Experiments

In this section, we present the detailed experimental setup for the synthetic dataset experiment, along with analysis about its results under proper hyperparameter tuning.

C.1 Synthetic Dataset Experimental Setup

Data generation To simulate key attributes of clinical prediction tasks, we take into consideration class imbalance, limited sample size, measurement noise, and a preference for interpretable decision rules. Specifically, we create two classes in $\mathbb{R}^2$ (with independent Gaussian features (x_1, x_2)) that determine the true signal and three irrelevant independent Gaussian features (x_3, x_4, x_5) that disturb the feature space. Ground-truth 2D cluster 1 is centered at $(0, 0)$ and cluster 2 is at (μ, μ) with $\mu \in \{0.1, 0.3, \ldots, 2.9\}$ to span low-through-high separability. Class imbalance is fixed at 80% negatives and 20% positives. We study two dataset sizes, $n \in \{50, 100\}$, to reflect data-scarce settings.

Validation protocols Upon generating one synthetic dataset, it is split into 70% for training and 30% for validation using stratification to preserve class balance. We compare two model/feature selection strategies: PVF and the traditional method, based on whether they select the ground-truth feature pair (x_1, x_2).

To mitigate imbalance in training folds, we apply SMOTE oversampling (and minority duplication when the minority count is below two) before fitting models on the training set.

Candidate models Clinical settings prioritize interpretability, often favoring simpler models like logistic regression or decision trees (Abdullah, Zahid, and Ali 2021; Rudin 2019; Christodoulou et al. 2019). In practice, one can even decide the complexity of the model based on heuristic judgement on overfitting and underfitting. In this case, choosing the best model often reduces to choosing the right subset of features. This equivalence is clearest for bounded-depth decision trees or sparse logistic regression, whose predictive power hinges directly on the selected variables (Breiman 2001b; Chen et al. 2022; Ning et al. 2022; Tan et al. 2022; Nguyen et al. 2023; Suttaket, Harsha Vardhan, and Kok 2024). Following these ideas, we use logistic regression models trained on pairs of features only. This yields ten pairs and thus ten model candidates, one per distinct pair among the five features. By the Rashomon Effect, there might be more than one pair of features that achieve optimal or almost-optimal performances. For example, although the truly informative pair is (x_1, x_2), due to a small number of samples, (x_1, x_3) might accidentally performs equally well or even better than (x_1, x_2). Our design deliberately creates the possibility of encountering this situation under the Rashomon Effect, and aligns feature selection with model selection: the best model should be the one trained on the truly informative pair (x_1, x_2). It creates a clear, ground-truth notion of "correct selection" while keeping the interpretability of the selected model.

Model evaluation metrics The primary analysis uses Intervention Efficiency (IE) at different intervention capacities $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. For the complementary analysis under traditional metrics, we report F1 Scores when class separability is relatively low ($\mu \leq 2.5$) and accuracies when class separability is very high ($\mu > 2.5$). This intentionally favors the traditional method, because based on our empirical results, using F1 Score when $\mu \leq 2.5$ and accuracy when $\mu > 2.5$ makes the traditional method more likely to select the best model (truly informative feature pair). Therefore, we force PVF to adopt the same setting to stress-test it and avoid underestimating the traditional method by the "wrong" metric. Selection is always compared within a fixed metric: PVF versus traditional under IE, or PVF versus traditional under F1 Score/accuracy.

Hyperparameters, settings and repetitions We vary the dataset size ($n = 50$ or 100) and the separability parameter μ, which takes values uniformly spaced between 0.1 and 2.9 with a step size of 0.2, and perturbation Gaussian noise standard deviation $\sigma \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ as different perturbation controls. For each hyperparameter configuration ($2 \times 15 \times 6 = 180$ combinations) in Table 1, we generate 5000 independent synthetic datasets to obtain stable comparison results and uncertainty summaries. For other hyperparameters discussed in Section 3.2 while not in Table 1, we empirically set $d = 5$, $k = 7$ and $M = 100$ to balance the computational cost and perturbation effect. The aggregation function $\mathcal{A}$ of PVF is manually chosen to be the 25th percentile (Q1) of a model's validation scores across perturbed validation sets for demonstration, as it is well suited to high-stakes clinical contexts where worst-case generalization matters and they could correspond to the cases where all independent noises compound in a way that is most disturbing to the model's prediction, while not being too conservative. Note that the choice of aggregation function in other applications could be decided according to specific user requirements of the attributes of the selected model, as Proposition B.7 has stated that ideally PVF should be able to select the optimal model toward any user-defined property.

Selection method assessment For each synthetic dataset, we record the feature pair chosen by PVF and the traditional method. Specifically, the pair of features among the total 10 pairs that achieves the highest logistic regression performance on the perturbed validation sets/the original validation set is selected as the preferred pair for the corresponding method and synthetic dataset. The quality of a selection method is assessed by the frequency with which a method chooses the truly informative pair (x_1, x_2) out of 5000 repetitions. Specifically, for each hyperparameter configuration $h = (n, \mu, \sigma)$,the performance difference between the PVF and the traditional method is computed as

$$d_h = c_h^{\text{PVF}} - c_h^{\text{Trad}},$$

where c_h^{PVF} and c_h^{Trad} are the counts of selecting the truly informative pair (x_1, x_2) for configuration h.

Compute environment and runtime Synthetic sweeps were executed on a SLURM-managed computing cluster. Each job was allocated 11 CPU cores and 30 GB RAM within a fixed Python environment. Cluster nodes are equipped with 94-core x86-64 CPUs and about 740 GB RAM, running Linux kernel 5.14.0. The size-50 sweep (i.e., evaluating every hyperparameter configuration for data size 50) required 1642 minutes, and the size-100 sweep required 1598 minutes. Data generation time was negligible relative to model fitting. Scaling by cores, repetitions (5000), and hyperparameter settings (180), the inner-loop PVF selection for one synthetic dataset and one hyperparameter on a single CPU is about 1.2 seconds for all model evaluation metrics.

C.2 Synthetic Dataset Experimental Results

In this subsection, we present the experimental results in the form of a comparison between the performance of PVF with tuned σ and that of the traditional method. Across all hyperparameter configurations, statistical significance is confirmed using paired t–tests. During the experiments, we found that a key factor influencing the effectiveness of the proposed method is class separability, represented by the parameter μ. Hence, we categorized μ into three separability scenarios: low (0.1 – 0.9), moderate (1.1 – 1.9), and high (2.1 – 2.9). We aggregate the differences d_h's within each class separability range to get the mean difference for the corresponding range.

Table 1: Summary of hyperparameters and settings in the synthetic dataset experiment

Hyperparameters	**Values**
Original Data Size (n)	50, 100
Center of Cluster 2 (μ)	From 0.1 to 2.9 with step 0.2
Perturbation Noise Standard Deviation (σ)	1e-6, 1e-5, 1e-4, 1e-3, 0.01, 0.1

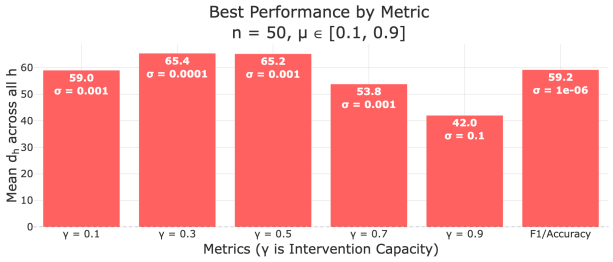

Figure 6: The best performance of PVF on synthetic datasets ($n = 50, \mu \in [0.1, 0.9]$) under different metrics, when σ is tuned. PVF identifies the truly informative feature pair more frequently across all evaluation metrics.

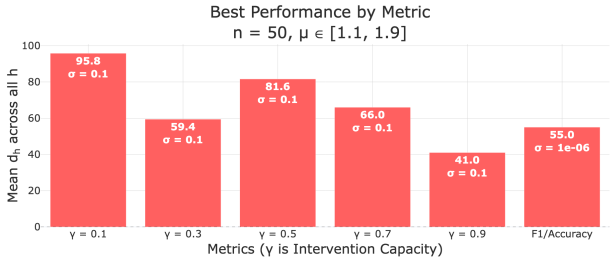

Figure 7: The best performance of PVF on synthetic datasets ($n = 50, \mu \in [1.1, 1.9]$) under different metrics, when σ is tuned. PVF continues to outperform the traditional method across all metrics, with notably higher improvements when $\gamma \in \{0.1, 0.5, 0.7\}$. Variability increases relative to low-separability conditions, reflecting more complex interactions between IE and classification difficulty.

Data size 50 Across all class separability, PVF always outperforms the traditional method no matter which model evaluation metric is used. The performance difference is overall more significant for lower γ's.

Specifically, when the class separability is low, PVF overall selects the truly informative feature pair 42 to 65 more times than the traditional method does, across all metrics (Figure 6). This consistently positive difference confirms the superiority of PVF over the traditional method in low-separability cases, regardless of the metric.

When the class separability is moderate, built on consistent positive differences, the value of difference fluctuates more, with larger values emerging at $\gamma = 0.1, 0.5, 0.7$ (Figure 7). This could be due to a higher variability in synthetic datasets and the resulting complex relationship between IE and the classification task. F1 score/accuracy exhibit substantial positive differences, confirming that PVF is promising when these two traditional metrics are of interest in moderate-separability cases.

When the class separability is high, we can witness a clearer pattern in the relationship between the performance difference and γ. That is, when γ increases, the positive difference shrinks almost monotonically (Figure 8), with smaller γ's (0.1, 0.3 and 0.5) enjoying a rising positive difference while larger γ's (0.7, 0.9) suffering from a loss in positive difference. PVF with F1 Score/accuracy still maintains a large edge in the positive performance. These observations confirms that under large separability circumstances, not only is PVF consistently better than the traditional method, but also it brings large improvements when the metric of interest is IE with lower γ's or F1 Score/accuracy.

Data size 100 Although generally PVF still outperforms the traditional method for most metrics when data size is 100, a small number of cases emerge where the traditional method becomes preferable. Specifically, with moderate separability and $\gamma = 0.7$ or 0.9, the traditional method selects the true feature pair a small number of times ($<$20) more than PVF (Figure 9).

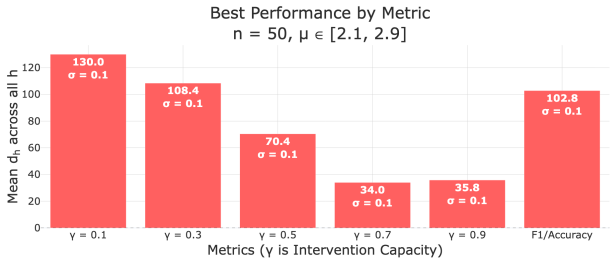

Figure 8: The best performance of PVF on synthetic datasets ($n = 50, \mu \in [2.1, 2.9]$) under different metrics, when σ is tuned. PVF still outperforms the traditional method, but the magnitude of improvement declines almost monotonically as γ increases.

Nevertheless, in all other cases, PVF consistently outperforms the traditional method (Figures 10 and 11). Besides the general trend that the positive edge is larger at smaller γ's as in data size 50, the strongest comparison occurs when the separability is high and the metrics of interest are IE with γ (0.1, 0.3 and 0.5) or F1 Score/accuracy (Figure 11). Overall, despite a small number of slightly underperforming cases, PVF under most metrics enjoys a large edge over the traditional method with data size 100.

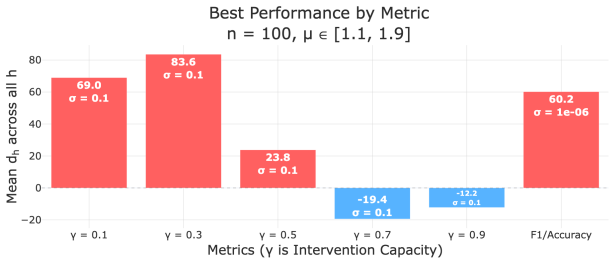

Figure 9: The best performance of PVF on synthetic datasets ($n = 100, \mu \in [1.1, 1.9]$) under different metrics, when σ is tuned. Although PVF generally outperforms the traditional method, a few cases where the traditional method selects the true feature pair slightly more often appear when $\gamma \in \{0.7, 0.9\}$

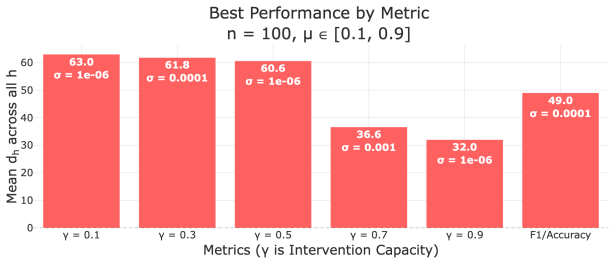

Figure 10: The best performance of PVF on synthetic datasets ($n = 100, \mu \in [0.1, 0.9]$) under different metrics, when σ is tuned. PVF consistently selects the informative feature pair more frequently across all metrics, with larger advantages at smaller γ values.

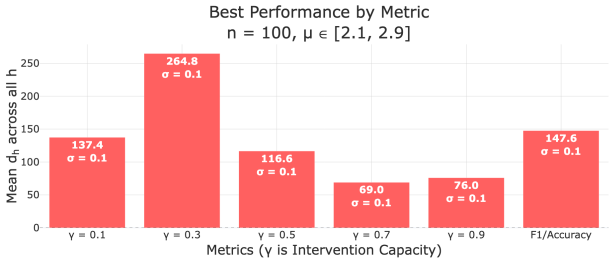

Figure 11: The best performance of PVF on synthetic datasets ($n = 100, \mu \in [2.1, 2.9]$) under different metrics, when σ is tuned. PVF exhibits its strongest improvements under high separability, especially when $\gamma \in \{0.1, 0.3, 0.5\}$ and for F1 Score/accuracy.

D Detailed Setups of Real Data Experiments

In this section, we present the detailed experimental setup for the real dataset experiments.

Datasets and preprocessing We evaluate on two public clinical datasets to ground the study in realistic settings. The Breast Cancer Wisconsin (Diagnostic) dataset (Wolberg et al. 1993) has 569 samples, 212 positives, and no missing values; no preprocessing beyond standard scaling is needed. The Cervical Cancer (Risk Factors) dataset (Fernandes, Cardoso, and Fernandes 2017) has 808 samples and substantial missing values. We discard features with more than 50% missing values and impute the remaining entries with k-nearest neighbors ($k = 5$). This yields 668 usable samples with 34 features. Among the four diagnostic targets, we use the Schiller test outcome because its higher prevalence provides adequate positive cases for stable validation.

Validation protocols, evaluation metrics, hyperparameters and settings To emulate small-cohort deployments and create multiple independent model selection tasks, we partition each dataset into several disjoint subsets, each containing 100 samples, and perform model selection within each subset. For the Breast Cancer dataset ($n = 569$), five non-overlapping 100-sample subsets are formed from 500 samples, while the remaining 69 samples are reserved as part of the external test data. For a given subset, the external test set consists of all remaining samples not included in that subset, i.e., $n - 100 = 469$ instances for the Breast Cancer dataset. Within each 100-sample subset, we further divide the data into five different train–validation splits, each time using a distinct split for validation while keeping the rest for training. Repeating this procedure five times reduces the influence of random partitioning within the subset and enables a more reliable assessment of whether one selection method consistently identifies better models across multiple validation scenarios and possibilities. This design focuses on comparing the stability and consistency of model selection performance, rather than evaluating the final predictive accuracy of any individual model.

The same configuration is applied to the Cervical Cancer dataset ($n = 668$), resulting in six disjoint 100-sample subsets. All train–validation divisions are stratified to preserve class ratios. We apply standard scaling to the training data and use the same transformation on validation data to (i) maintain comparability with the synthetic experiments and (ii) make the perturbation noise scale in PVF interpretable across datasets.

We compare PVF (using the 25th percentile aggregator) with the traditional method under the same evaluation metrics and hyperparameters as in the synthetic experiments, except for the perturbation controls. For the perturbation controls, we extend the previous range of $noise\_add\_std$ ($\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$) to include $\{2 \times 10^{-1}, 3 \times 10^{-1}, 4 \times 10^{-1}, 5 \times 10^{-1}\}$ to further examine robustness gains under stronger perturbations. Additionally, for ordinal and nominal features, we empirically set the feature-changing probability to 0.1 and the decay parameter λ for ordinal features to 0.1 (see Section 3.2 for details

on λ), leaving the exploration of alternative choices to future work.

Candidate models We use decision trees as the base model class to prioritize interpretability and to account for potential non-linear separability. Within each 100-sample subset, we generate a diverse pool of candidate models by repeatedly subsampling 70% of the training data and training one tree per subsample (maximum depth 4 for interpretability), yielding 100 candidates in total. This resampling procedure intentionally produces multiple models with similar overall accuracy but differing feature splits and decision paths, characterizing the presence of the the Rashomon Effect.

Selection method assessment For each validation split of each data subset and each model selection method, we pick one model from the candidate pool using the chosen metric based on validation performance. The selected model is then evaluated on the external test set associated with this piece using the same metric as in validation. This aligns validation with testing and approximates ground truth generalization. We repeat the procedure across subsets and splits, as explained in the last paragraph, and aggregate the testing performance w.r.t the model evaluation metrics, subsets and splits.

Compute environment and runtime Real-data experiments run locally on an Apple MacBook Pro (M4 Pro, 12-core CPU, 16-core GPU, 24 GB RAM) in a fixed Python environment. The two datasets are processed concurrently with a total time of 424.3 minutes.

E Sensitivity Analysis

In this section, we discuss how sensitive is PVF to the problem setting and the key hyperparameter in PVF: perturbation control, i.e., the standard deviation of the Gaussian perturbation noise, σ.

Synthetic Dataset Experiment Sensitivity Analysis In this subsection, we isolate the effect of the perturbation noise standard deviation σ by holding other settings fixed and tracking the mean selection difference within each class separability case.

Data size 50 We start with data size 50. First, when class separability is low, very small perturbations ($\sigma \leq 10^{-3}$) maintain a stable, positive advantage for PVF (Figure 12). In this range the perturbed validation replicates the original validation set closely while still eliminating some extent of occasionality of a single split. Increasing σ beyond this range can reduce the fidelity of the perturbed sets to the underlying problem, compressing score differences across model candidates and generally weakening the advantage, except for larger γ's (0.7 and 0.9). The drop is pronounced at $\sigma = 10^{-2}$ and makes the difference negative at $\sigma = 10^{-1}$ for F1 Score.

Second, at moderate class separability, performance difference often slightly decreases from a stable positive value when σ goes from a small value towards a moderate one (Figure 13), while the difference increases back to a higher value when σ reaches the largest value in the setting ($\sigma = 0.1$). This could be because: a small perturbation already has positive effects on selecting the best model because it stress-test the candidates; a slightly stronger perturbation enables reshuffling the rankings of the candidate models to a greater extent but not yet strong enough to consistently separate stable from unstable candidates, while the disturbance has exceeded the benefits brought by the stress test; at the upper end of the perturbation range, the stress test becomes the strongest, and somehow the advantage recovers.

Third, at high class separability, larger perturbations amplify PVF's advantage almost certainly for all metrics (with a small fluctuation in the curve for $\gamma = 0.9$), as shown in Figure 14. With a large perturbation, well-supported candidate models remain top of the rank across perturbed validation sets, while models whose strength is fraudulently exaggerated by some noisy patterns due to small sample size fall in rank easily.

Data size 100 With $n = 100$ and low separability, the small-σ region is not merely flat but consistently above zero (Figure 15), indicating a robust positive margin for PVF across all very small perturbations. The guidance here is to keep σ not too large, and every value in this wide range would be very positively effective. In contrast, when σ becomes too large under low separability, the same degradation appears as in $n = 50$.

In moderate-separability settings, for IE at more stringent capacities ($\gamma = 0.1, 0.3$), PVF's positive performance difference margin is relatively stable at small perturbations ($\sigma \leq 10^{-3}$), dips around $\sigma = 10^{-2}$, and then increases sharply at $\sigma = 10^{-1}$ (Figure 16). This agrees with the results from $n = 50$ that a weak stress test is already informative; while a mid-range σ can reshuffle the rankings of the candidate models to a larger extent without consistently separating stable candidates from unstable ones; and a larger σ provides a decisive stress that favors candidates whose rankings remain intact under huge perturbation.

For IE at $\gamma = 0.5$, the advantage is near zero or slightly negative at very small σ, then becomes positive only at the largest σ (10^{-1}), indicating that stronger stress is required before PVF separates candidates meaningfully in this setting.

For IE at $\gamma = 0.7$ and 0.9, PVF is negative everywhere to a small extent, especially at smaller σ's. IE at $\gamma = 0.9$ drops further at $\sigma = 10^{-2}$, with a large recovery at $\sigma = 10^{-1}$, while IE at $\gamma = 0.7$ simply recovers at $\sigma = 10^{-3}$. Here, when the intervention capacity is large, small perturbations add variation without enough discriminatory power over models. Under F1, PVF's margin slightly decrease with increasing σ, while remaining above zero. This indicates that PVF is still beneficial, and bad performances under IE at $\gamma = 0.7$ and 0.9 may simply due to IE's lack of capability of recognizing truly good models in large-capacity scenarios.

In high-separability settings, increasing σ raises the curves for both $n = 100$ and $n = 50$, with $n = 100$ exhibiting especially large gains at the top of the perturbation range Figure 17. This is consistent with the results from $n = 50$ that higher separability enables the stress test from PVF to

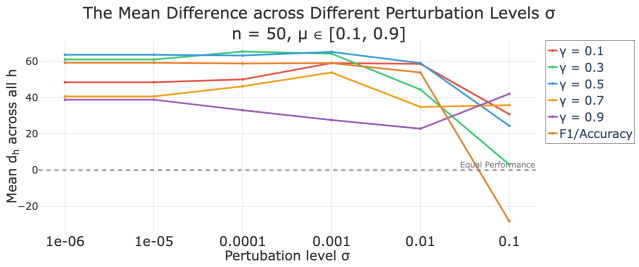

Figure 12: The performance difference between PVF and the traditional method on synthetic datasets ($n = 50, \mu \in [0.1, 0.9]$) across perturbation levels σ and evaluation metrics. PVF maintains a stable positive advantage at relatively small σ ($\sigma \leq 10^{-2}$), while larger perturbations slightly reduce its margin and can overturn it for F1 Score/accuracy.

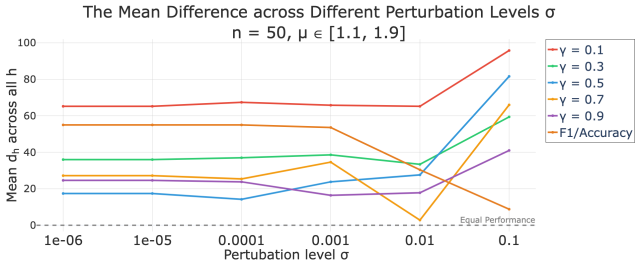

Figure 13: The performance difference between PVF and the traditional method on synthetic datasets ($n = 50, \mu \in [1.1, 1.9]$) across perturbation levels σ and evaluation metrics. The advantage of PVF dips as σ increases from small to moderate values, but recovers at the largest perturbation level ($\sigma = 0.1$).

be more informative and effective.

Summary of sensitivity analysis for synthetic dataset experiments Overall, for IE, small perturbations are preferred when class separability is low, and large perturbations become beneficial as separability increases. Also, IE with higher γ's generally enjoys less benefit from PVF than the IE with lower γ's does. For F1 Score, except for high-separability cases, keeping σ small always gives stably good results. Only when the separability is high does using large σ gives better results, with a steep improvement as for IE.

Real Dataset Experiment Sensitivity Analysis In this subsection, like in the synthetic sensitivity analysis subsection, we isolate the effect of the perturbation noise standard deviation σ by looking at the results of two datasets separately and examining PVF's performance under all possible choices of σ. The performance difference is defined as the proportion of repetitive experiments in which PVF selects a superior model based on the external test set minus the proportion in which the traditional method does, across all subsets and splits.

Cervical cancer Figure 18 shows that the perturbation level σ yielding the largest advantage for PVF depends on

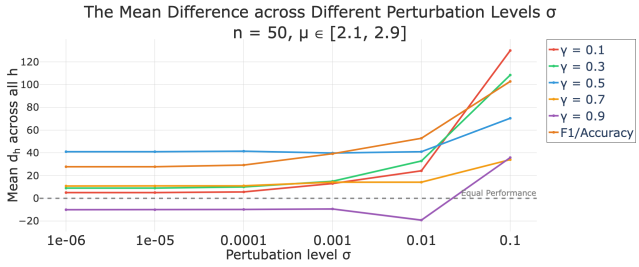

Figure 14: The performance difference between PVF and the traditional method on synthetic datasets ($n = 50, \mu \in [2.1, 2.9]$) across perturbation levels σ and metrics. Larger perturbations consistently amplify PVF's advantage, with only minor fluctuations when $\gamma = 0.9$.

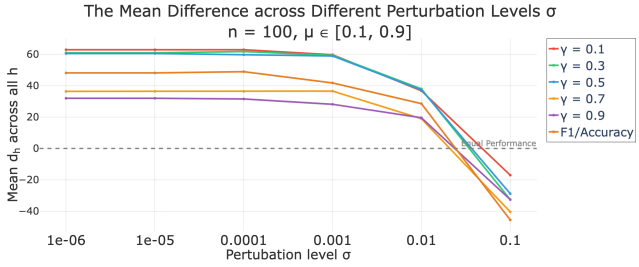

Figure 15: The performance difference between PVF and the traditional method on synthetic datasets ($n = 100, \mu \in [0.1, 0.9]$) across perturbation levels σ and metrics. PVF maintains a positive margin over small to moderate σ, while the largest large perturbation reverses the advantage, similarly to the $n = 50$ setting.

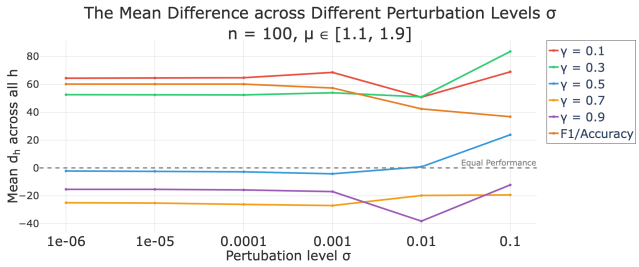

Figure 16: The performance difference between PVF and the traditional method on synthetic datasets ($n = 100, \mu \in [1.1, 1.9]$) across perturbation levels σ and metrics. PVF shows stable gains at small σ, dips at mid-range σ, and recovers at the largest perturbation level, with the pattern varying by evaluation metrics.

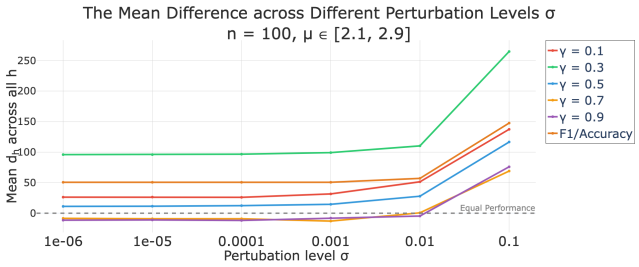

Figure 17: The performance difference between PVF and the traditional method on synthetic datasets ($n = 100, \mu \in [2.1, 2.9]$) across perturbation levels σ and metrics. Increasing σ steadily amplifies PVF's advantage, producing especially strong improvements at the largest perturbation level.

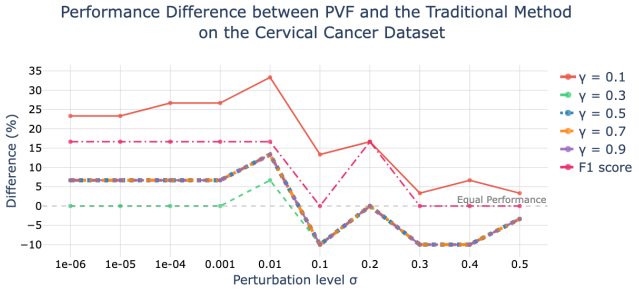

Figure 18: The performance difference between PVF and the traditional method on the cervical cancer dataset across perturbation levels σ and evaluation metrics. PVF achieves the most reliable improvements at small σ values (around 10^{-2}), while larger perturbations diminish or reverse its advantage, especially when $\gamma \geq 0.3$.

the evaluation setting. For the smallest capacity ($\gamma = 0.1$), PVF's margin is positive across the grid, increases with the rise of σ, and peaks at $\sigma = 0.01$, after which it declines as σ further increases. For broader capacities ($\gamma \geq 0.3$), the difference is small, none or even turns negative once $\sigma \geq 0.1$, indicating that large perturbations blur the ranking that IE induces and no longer help PVF distinguish candidates. For F1, the curve is flat and positive at very small σ and drops to zero near $\sigma = 0.1$, with a small recovery around $\sigma = 0.2$ and another dropp to zero if σ further increase. Overall, the most reliable gains for this dataset occur at small σ (near 10^{-2}), while larger perturbations reduce PVF's benefit.

Breast cancer Figure 19 exhibits a different profile. Very small perturbations ($\sigma \leq 10^{-4}$) yield ties across settings. As σ moves into the 10^{-3}–10^{-1} range, PVF's advantage increases, especially for $\gamma \in \{0.1, 0.3\}$ and F1 Score. For IE under small γ (0.1 and 0.3) and F1 Score, the largest differences all occur at $\gamma = 0.3$. Beyond 0.3, the curves either flatten or decrease slightly. For larger capacities ($\gamma \geq 0.5$), the advantage of PVF remains near zero at moderate σ and becomes negative only at the largest σ values. Thus, for this dataset, moderate to large perturbation depending on the metric of interest is required for PVF to reveal stable candi-

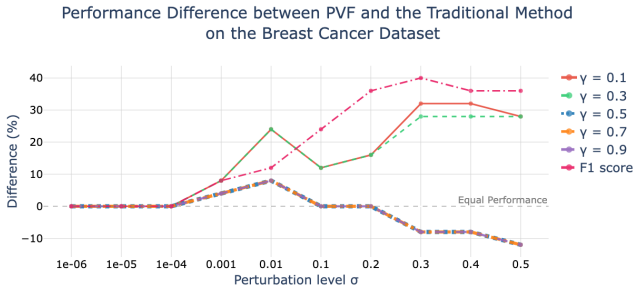

Figure 19: The performance difference between PVF and the traditional method on the breast cancer dataset across perturbation levels σ and evaluation metrics. PVF benefits most from moderate perturbations ($\sigma \in [10^{-3}, 10^{-1}]$), particularly for $\gamma \in \{0.1, 0.3\}$ and F1 Score, whereas very small or very large σ values can lead to ties or reduced gains depending on the metric.

dates. Too little perturbation produces ties, while excessive perturbation can erode the margin in the large-capacity settings.

Summary of sensitivity analysis for real dataset experiments The two datasets demand different perturbation scales: small σ is most effective for cervical cancer (with a clear optimum near 10^{-2} at $\gamma = 0.1$), whereas breast cancer benefits from large σ's (peaking near 0.3–0.4 for smaller capacities and F1 Score) or moderate σ's (peaking near 10^{-2} for larger capacities), depending on the metric of interest. These patterns underscore that σ should be tuned per dataset and evaluation setting rather than being kept as a fixed prior. With that being said, empirically, we can see that a moderate perturbation (e.g., $\gamma = 10^{-2}$) generally applies well to both datasets, which could be a starting point for other similar datasets.