Data as a Lever: A Neighbouring Datasets Perspective on Predictive Multiplicity

Prakhar Ganesh[1], **Hsiang Hsu**[2], **Golnoosh Farnadi**[1]
[1]McGill University & Mila
[2]JPMorgan Chase GT Applied Research
prakhar.ganesh@mila.quebec, hsiang.hsu@jpmchase.com, farnadig@mila.quebec

Abstract

Multiplicity—the existence of distinct models with comparable performance—has received growing attention in recent years. While prior work has largely emphasized modelling choices, the critical role of data in shaping multiplicity has been comparatively overlooked. In this work, we introduce a neighbouring datasets framework to examine the most granular case: the impact of a single-data-point difference on multiplicity. Our analysis yields a seemingly counterintuitive finding: neighbouring datasets with greater inter-class distribution overlap exhibit lower multiplicity. This reversal of conventional expectations arises from a shared Rashomon parameter, and we substantiate it with rigorous proofs.

Building on this foundation, we extend our framework to two practical domains: active learning and data imputation. For each, we establish natural extensions of the neighbouring datasets perspective, conduct the first systematic study of multiplicity in existing algorithms, and finally, propose novel multiplicity-aware methods, namely, *multiplicity-aware data acquisition* strategies for active learning and *multiplicity-aware data imputation* techniques.

1 Introduction

Predictive multiplicity refers to the phenomenon of a set of "good models" (the Rashomon set), typically defined as models whose performance exceeds a given threshold (the Rashomon parameter), learning distinct decision boundaries and therefore producing conflicting predictions for the same individual (Marx, Calmon, and Ustun 2020; Black, Raghavan, and Barocas 2022; Breiman 2001).

Multiplicity has been a point of concern for many, as decisions that affect individuals lack adequate justification when a model is chosen arbitrarily from the Rashomon set (Black, Raghavan, and Barocas 2022; Gomez et al. 2024; Watson-Daniels et al. 2024; Sokol et al. 2024). At the same time, multiplicity is also championed as a counterbalance to monoculture, where reliance on a single dominant system can systematically deny individuals access to critical resources, and multiplicity can introduce much needed diversity (Creel and Hellman 2022; Jain et al. 2024; Jain, Creel, and Wilson 2024; Kleinberg and Raghavan 2021). Recent work by Gur-Arieh and Lee (2025) brings together these two strands of research by identifying distinct settings in which one might prefer consistency versus arbitrariness.

Irrespective of the direction, controlling multiplicity requires understanding how developer choices shape downstream outcomes (Ganesh, Taik, and Farnadi 2025). While existing work has primarily examined how choices during model training influence predictive multiplicity (Black, Raghavan, and Barocas 2022), the role of data processing remains largely overlooked. This gap may stem from the difficulty of mapping how data processing decisions affect downstream models without actually training them (Koh et al. 2019), or from the prevailing norm in the literature of relying on already processed datasets rather than questioning the processing choices themselves (Paullada et al. 2021).

Consider, for example, a task with missing values for predicting an individual's income (Ding et al. 2021). Using our multiplicity-aware imputation methods (more details in §5), we find that the choice of imputation can shift downstream multiplicity from 14% to 24%, i.e., up to 10% of the dataset is affected by this one data processing choice. Income predictors are used in applications such as loan approval or hiring, where controlled arbitrariness can be helpful to prevent monoculture (Gur-Arieh and Lee 2025). Thus, a poor imputation choice can potentially result in a blanket rejection for up to 10% of the data, not recoverable during model training. Clearly, choices made during data processing play a significant role in downstream multiplicity.

While recent frameworks like dataset multiplicity (Meyer, Albarghouthi, and D'Antoni 2023) study noise in the data while keeping the training pipeline fixed, we argue that isolating either model or dataset multiplicity will give us an incomplete picture. In our work, we instead focus on how different data processing choices—creating *neighbouring datasets*—affect model multiplicity (see Figure 1).

The perspective of neighbouring datasets, inspired by the literature in differential privacy (Dwork 2006), pops up repeatedly and naturally in many data processing scenarios, such as data acquisition for active learning (Ren et al. 2021; Aggarwal et al. 2014), data imputation (Miao et al. 2022), and handling outliers (Aguinis, Gottfredson, and Joo 2013), among others. Data processing rarely transforms a dataset entirely; instead, it introduces incremental changes that can still have significant downstream effects. For instance, con-

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

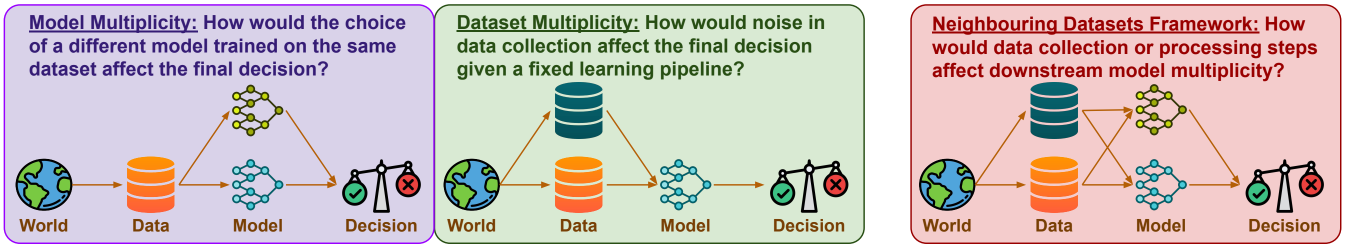

Model Multiplicity: How would the choice of a different model trained on the same dataset affect the final decision?
World
Data
Model
Decision
Dataset Multiplicity: How would noise in data collection affect the final decision given a fixed learning pipeline?
World
Data
Model
Decision
Neighbouring Datasets Framework: How would data collection or processing steps affect downstream model multiplicity?
World
Data
Model
Decision

Figure 1: Our neighbouring datasets framework alongside model and dataset multiplicity frameworks.

sider data imputation, where different techniques may fill the missing values in distinct ways. However, the majority of the data is not missing and thus remains unchanged. Hence, data imputation can be seen as a choice between various *neighbouring datasets*.

Contributions. By framing our study through the lens of neighbouring datasets, we provide a unified framework that accommodates many frequently studied problems in data processing, allowing us to systematically examine developers' choices and their influence on multiplicity. We then apply the insights derived from this perspective to two well-established subdomains of data processing, active learning and data imputation, highlighting the trends of downstream multiplicity as well as designing new algorithms offering control over multiplicity. More specifically, our main contributions are,

1. **Neighbouring Datasets Framework:** A novel unified framework to study the impact of various data processing choices on multiplicity (§3). We formalize neighbouring datasets for deeper theoretical insights in controlled settings and practical extensions in real-world applications.
2. **Reversed Multiplicity Trends under a Shared Rashomon Parameter:** Theoretical insights into neighbouring datasets and multiplicity reveal a surprising result: under a shared Rashomon parameter, less separability leads to lower multiplicity (§4). This reverses expected trends based on prior work (Watson-Daniels, Parkes, and Ustun 2023; Semenova et al. 2024). Without contradicting existing literature, this reversal occurs due to the use of a shared Rashomon parameter across neighbouring datasets, highlighting how these frameworks fail to capture multiplicity trends in data processing.
3. **Multiplicity and Data Imputation:** We investigate data imputation from the lens of neighbouring datasets, performing the first empirical study of multiplicity in data imputation, as well as using our theoretical insights to propose new multiplicity-aware data imputation algorithms (§5). Our experiments reveal consistent trends of less separability leading to lower multiplicity, even beyond the assumptions of our theoretical analysis, further strengthening the value of our framework.
4. **Multiplicity and Active Learning:** We repeat our study for another important data processing task, data acquisition for active learning, and observe a similar set of contributions and trends as in data imputation (§A).

2 Related Work

Multiplicity and Rashomon Sets. The literature on multiplicity has grown rapidly (Ganesh, Taik, and Farnadi 2025), with a particular focus on predictive multiplicity (Marx, Calmon, and Ustun 2020; Cooper, Frankle, and De Sa 2022; Watson-Daniels et al. 2024). Through extension to new forms of multiplicity (Watson-Daniels et al. 2023, 2024; Hsu et al. 2024b), development of better tools for auditing and quantifying multiplicity (Hsu et al. 2024b; Kissel and Mentch 2024; Zhong et al. 2024; Xin et al. 2022; Hsu et al. 2024a; Ganesh 2024), and deeper investigations into the benefits and harms of multiplicity (Black, Raghavan, and Barocas 2022; Rudin et al. 2024; Gur-Arieh and Lee 2025), it is evident that multiplicity has become a valuable lens for understanding the ambiguity inherent in learning pipelines.

Yet, despite growing interest, most research continues to concentrate only on modeling decisions during learning (Ganesh, Taik, and Farnadi 2025). In contrast, our work joins a smaller but emerging thread of research that aims to uncover the inherent multiplicity in the datasets themselves (Meyer, Albarghouthi, and D'Antoni 2023; Cavus and Biecek 2024; Semenova et al. 2024; Watson-Daniels, Parkes, and Ustun 2023).

Data and Multiplicity. Meyer, Albarghouthi, and D'Antoni (2023) proposed a framework for dataset multiplicity, showing how noisy data can introduce multiplicity. However, while their focus lies in aggregating variance across datasets using a fixed learning pipeline, we instead investigate and minutely compare variations across datasets and their relationship with downstream multiplicity under changing learning pipelines.

The works closely related to our theoretical analysis are those of Semenova et al. (2024); Watson-Daniels, Parkes, and Ustun (2023). Semenova et al. (2024) demonstrate that noisier tasks, i.e., tasks with higher inter-class distribution overlap, exhibit higher multiplicity. Watson-Daniels, Parkes, and Ustun (2023) provide similar insights on the low separability of a task as a potential cause of multiplicity. Interestingly, our examination of neighbouring datasets under a shared Rashomon parameter reverses these trends (Semenova et al. 2024; Watson-Daniels, Parkes, and Ustun 2023). This is because existing frameworks are designed to compare distinct tasks, and not neighbouring datasets within a single task. Our framework addresses this gap, enabling the study of how data processing affects multiplicity.

On the empirical side, the study by Cavus and Biecek (2024) is most related to our work. They conduct a large-

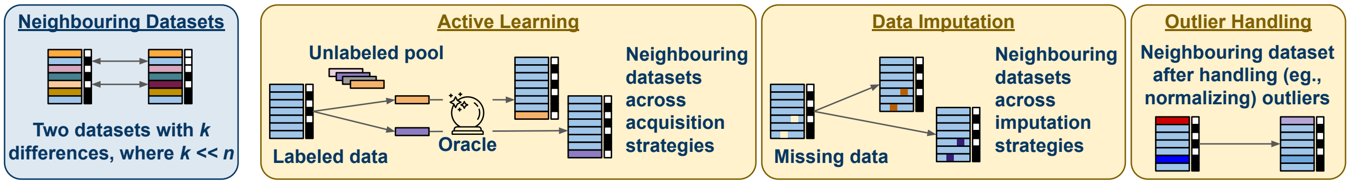

Neighbouring Datasets
Two datasets with k differences, where $k << n$
Active Learning
Unlabeled pool
Labeled data
Oracle
Neighbouring datasets across acquisition strategies
Data Imputation
Missing data
Neighbouring datasets across imputation strategies
Outlier Handling
Neighbouring dataset after handling (eg., normalizing) outliers

Figure 2: Examples of neighbouring datasets in several data preparation and processing pipelines.

scale empirical analysis of data balancing methods and their effect on multiplicity. We provide a similar analysis for data acquisition and imputation techniques. Furthermore, drawing on our theoretical insights, we also introduce *multiplicity-aware* data processing, which can achieve the lowest (or highest) multiplicity while preserving accuracy.

Active Learning and Data Imputation. In this work, we study two components of data processing from the lens of neighbouring datasets, namely active learning and data imputation. Active learning focuses on selecting the data points to label (Ren et al. 2021; Aggarwal et al. 2014), recognizing that labeling is often expensive. On the other hand, data imputation deals with the issue of missing data (Miao et al. 2022). Together, they represent decisions that developers must navigate during data collection and preparation. Although both fields have rich histories of research, to the best of our knowledge, we are the first to study their impact on multiplicity.

3 Neighbouring Datasets

When preparing data, developers routinely make decisions that involve choosing between neighbouring datasets. Examples include: active learning (Ren et al. 2021), where the new data points to label are chosen while the rest of the dataset remains unchanged; data imputation (Miao et al. 2022), where a few missing values are filled leading to datasets varied in only those data points; and handling outliers (Neale 2016), where normalizing only affects outliers (see Figure 2).

Making these choices with an awareness of multiplicity can allow developers to understand and control the downstream trends. Thus, studying multiplicity for neighbouring datasets can enable multiplicity-aware data collection and preparation practices from the outset and lead to informed decision-making.

3.1 Preliminaries: Rashomon Set and Multiplicity

Consider a supervised learning setup, with data distribution $\mathbb{D} \equiv \mathbb{X} \times \mathbb{Y}$, where $\mathbb{X}$ represents the feature distribution and $\mathbb{Y}$ represents the label distribution. We sample two datasets independently from the distribution $\mathbb{D}$, the train dataset $D_{train} \equiv (X_{train}, Y_{train}) \sim \mathbb{D}$ and the test dataset $D_{test} \equiv (X_{test}, Y_{test}) \sim \mathbb{D}$. Given a loss function $L(\theta, D)$ for the parameter vector θ on the dataset D, and the Rashomon parameter ϵ, the Rashomon set is defined as (Hsu and Calmon 2022):

Definition 3.1 (Rashomon Set)**.** The set of all parameter vectors $\Theta \equiv \{\theta_1, \theta_2, ...\}$, such that the loss defined by $L(\theta_i, D_{train})$ for each parameter vector in the set is less than a given threshold ϵ, i.e.,

$$\Theta_{(D_{train},\epsilon)} \equiv \{\theta_i \mid L(\theta_i, D_{train}) \leq \epsilon\} \tag{1}$$

The Rashomon set is the set of models that achieve similar loss on the training dataset. We will omit the subscript and refer to the Rashomon set as simply Θ for brevity. We can then quantify multiplicity as $M(\Theta, D_{test})$, where $M()$ is a multiplicity metric that maps the Rashomon set and the test dataset to a score between 0 and 1, representing the severity of prediction conflicts. For instance, we can quantify predictive multiplicity for classification by defining ambiguity (Marx, Calmon, and Ustun 2020) $M^A()$ as:

Definition 3.2 (Ambiguity)**.** The ambiguity of a prediction problem over the Rashomon set Θ is the proportion of points in the test dataset D_{test} that can be assigned a conflicting prediction between two classifiers in the Rashomon set, i.e., $\theta_i, \theta_j \in \Theta$:

$$M^A(\Theta, D_{test}) = \frac{1}{|D_{test}|} \sum\nolimits_{x \in D_{test}} \max\nolimits_{\theta_i,\theta_j \in \Theta} \mathbb{1}[\theta_i(x) \neq \theta_j(x)] \tag{2}$$

We will denote multiplicity as M_Θ (for example, ambiguity as M^A_Θ) for brevity. We make a distinction between the Rashomon set created on the train dataset D_{train} and the multiplicity measured on the test dataset D_{test}. This is different from the tradition of measuring multiplicity on the train dataset itself (Marx, Calmon, and Ustun 2020). We argue that this distinction is important in practice, as the phenomenon of several models achieving similar loss and thus forcing an arbitrary choice by the developer occurs during training, while its impact and hence the multiplicity is felt when the model is deployed.

3.2 k-Neighbouring Datasets

Definition 3.3 (k-Neighbouring Datasets)**.** Two datasets D^1, D^2 of same size, i.e., $|D^1| = |D^2| = n$ are k-neighbouring if they differ in exactly k data points, i.e.,

$$|D^1| = |D^2| = n \quad \text{and} \quad \left|\{i : D^1_i \neq D^2_i\}\right| = k \ll n \tag{3}$$

Here, the size of a dataset $|D|$ represents the number of data points present in the dataset.

Objective: As previously discussed, the formulation of k-neighbouring datasets extends naturally to various data preparation decisions, where the developer has to choose between several neighbouring datasets. The objective, thus, is to facilitate a multiplicity-aware choice in such scenarios. More formally, given two k-neighbouring datasets D^1_{train}, D^2_{train}, and the Rashomon sets on these datasets denoted by $\Theta^1 \equiv \Theta_{(D^1_{train},\epsilon)}, \Theta^2 \equiv \Theta_{(D^2_{train},\epsilon)}$, we aim to compare the multiplicity due to these datasets on a common test set D_{test}, i.e., compare the values M_{Θ^1} and M_{Θ^2}.

4 Higher Overlap leads to a Smaller Rashomon Set

Data-driven learning methods typically rely on implicitly approximating the underlying distribution. As a result, learning a classifier is tightly coupled with learning the empirical distribution. Intuitively, when the distributions of various classes in a dataset exhibit greater overlap than those of its neighbouring datasets, the decision boundary becomes more ambiguous and can lead to higher error rates. With a fixed Rashomon parameter ϵ, under appropriate assumptions, such a shift can exclude some models from the Rashomon set, thereby reducing its size and, in turn, reducing multiplicity under any metric that is monotonic within the Rashomon set (Ganesh, Taik, and Farnadi 2025).

Note, it is vital to emphasize that the insights presented in our work are based on comparisons between neighbouring datasets. This framing is important because it allows us to apply a shared fixed threshold ϵ across datasets. At first glance, our claim may seem counterintuitive, as higher overlap and higher error rates are typically associated with higher multiplicity (Semenova et al. 2024; Watson-Daniels, Parkes, and Ustun 2023). However, this is because when comparing different tasks, the Rashomon sets are defined using a task-dependent threshold ϵ, hence leading to the trends seen in the literature. In contrast, our analysis focuses on neighbouring datasets for the same task, where we argue that the threshold for what constitutes a "good model" should not vary due to data processing choices. In other words, the threshold for a good model remains anchored to the task itself [1]. As we will demonstrate, under this constraint, *higher overlap leads to a smaller Rashomon set.*

4.1 Theoretical Insights for Binary Classification

Consider two 1-neighbouring training datasets D^1_{train}, D^2_{train}. The learning task is binary classification, i.e., $Y^1_{train}, Y^2_{train} \in \{0,1\}^n$. Thus, each dataset contains two classes, i.e., $D^i_{train} \equiv 0^i_{train} \cup 1^i_{train}$, where $[0/1]^i_{train} \equiv \{(x_j, y_j) \mid (x_j, y_j) \in D^i_{train}$ and $y_j = [0/1]\}$. The overlap between the two classes is measured using the overlapping coefficient defined as:

Definition 4.1 (Overlapping Coefficient (Inman and Bradley Jr 1989))**.** The overlapping coefficient (OVL) between two probability distributions P, Q is defined as:

$$OVL(P,Q) = \sum_x min(P(x), Q(x)) \tag{4}$$

$$\text{or } \ OVL(P,Q) = \int_x min(P(x), Q(x))\, dx \tag{5}$$

depending on whether the distributions are discrete or continuous. The overlapping coefficient is the complement to total variation distance (TVD) (Dudley 2018), i.e., $OVL + TVD = 1$. We will write the overlapping coefficient between the two classes as $OVL^i_{train} = OVL(0^i_{train}, 1^i_{train})$.

Under the assumptions of a 0-1 loss function,

Theorem 4.1. Given two 1-neighbouring binary classification datasets D^1_{train}, D^2_{train} which, without loss of generality, differ only at the index 0, i.e., $(x^1_0, y^1_0) \neq (x^2_0, y^2_0)$ and $(x^1_j, y^1_j) = (x^2_j, y^2_j)\ \forall j \neq 0$, and adhere to the following assumptions:

1. Loss of all models in the Rashomon set is higher on one differing data point over another, i.e.,

$$\begin{aligned} &L(\theta, (x^1_0, y^1_0)) \geq L(\theta, (x^2_0, y^2_0)) \\ &\forall \theta \in \Theta_{(D^1_{train}, \epsilon)} \cup \Theta_{(D^2_{train}, \epsilon)} \end{aligned} \tag{6}$$

2. Loss of the Bayes optimal models θ^*_1, θ^*_2 follow the same trend as the Rashomon set, i.e.,

$$L(\theta^*_1, (x^1_0, y^1_0)) \geq L(\theta^*_2, (x^2_0, y^2_0)) \tag{7}$$

then we can say that the overlapping coefficient between the two classes will be higher for this dataset, i.e., $OVL^1_{train} \geq OVL^2_{train}$, and the resulting Rashomon set for this dataset under a common threshold ϵ will be a subset of the Rashomon set for the other dataset, i.e., $\Theta_{(D^1_{train}, \epsilon)} \subseteq \Theta_{(D^2_{train}, \epsilon)}$.

Proof Sketch. We first show that for neighbouring datasets, the Bayes optimal loss is proportional to the overlapping coefficient, under the assumption of identical class priors. Thus, we say that the overlapping coefficient is higher for the dataset with the higher Bayes optimal loss. We then use the loss relationship in the first assumption to show that any model in the Rashomon set of the higher-loss dataset also belongs to the Rashomon set of the lower-loss dataset, but not vice-versa, creating a subset relationship. Complete proof can be found in the Appendix (§B).

Interpreting the Assumptions. The assumptions together state that one of the datapoints differing between neighbouring datasets is harder to classify than the other, and that all good models and both Bayes optimal models agree on this. The assumption fails when both differing datapoints lie in the ambiguous region near the decision boundary. A tighter Rashomon parameter ϵ (i.e., a smaller ϵ) makes the ambiguous region smaller, increasing the likelihood that the assumption holds.

Note that if the Bayesian optimal models are in the Rashomon set, the second assumption becomes redundant. In other words, for any hypothesis class expressive enough to include the Bayesian optimal, the second assumption can be dropped.

4.2 Extending to k-Neighbouring Datasets

Our theoretical discussion has focused on 1-neighbouring datasets, which enabled us to provide a rigorous proof for the downstream multiplicity based on the precise relationship between neighbouring datasets. However, in practice, we are

[1] A recent work by Ganesh, Taik, and Farnadi (2025) argues for a broader definition of the Rashomon set, incorporating all decisions made during model development, including even data processing. Under this perspective, the different Rashomon sets across neighbouring datasets in our work can be seen as subsets of one larger Rashomon set. Although we do not adopt this perspective, since we compare data processing choices and their effects, it still offers a useful intuition to the reader for using a fixed threshold across neighbouring datasets.

unlikely to encounter datasets that differ by only a single data point. Instead, we typically face the more general and realistic case of k-neighbouring datasets. While our previous sets of proofs do not work directly in this setting, we propose the following conjecture:

Conjecture 4.1. Given two k-neighbouring binary classification datasets D^1_{train}, D^2_{train} of size n, with $k \ll n$, if the overlapping coefficient between the two classes in higher for one dataset, i.e., without loss of generality $OVL^1_{train} \geq OVL^2_{train}$, then the resulting multiplicity for this dataset under a common threshold ϵ will be a lower than the other dataset, i.e., $M_{\Theta^1} \leq M_{\Theta^2}$.

In addition to generalizing from 1-neighbouring datasets to k-neighbouring datasets, we also shift our focus from the Rashomon set to the resulting multiplicity. Interestingly, the conjecture remains provable under strong assumptions—specifically, if the assumptions of Theorem 4.1 hold across all k differing data points (see §B for details). However, as k increases, such an assumption becomes increasingly unrealistic. Instead, we draw on our previous observations that a greater overlap between datasets is likely to increase the error across most models within the Rashomon set. As a result, given a fixed Rashomon parameter ϵ, we expect lower multiplicity in datasets with higher overlap compared to their neighbours. We will support these claims through empirical evidence on two data processing tasks as case studies: data acquisition in active learning (§A) and data imputation (§5).

5 Multiplicity and Data Imputation

With an understanding of how neighbouring datasets influence multiplicity, we extend our discussion to data imputation. We empirically evaluate several data imputation algorithms, alongside our own multiplicity-aware techniques. Our results reveal a negative correlation between the overlapping coefficient and the resulting multiplicity, as well as the success of our techniques in achieving the lowest (or highest) multiplicity without sacrificing accuracy.

5.1 Neighbouring Datasets in Data Imputation

Data imputation fills the missing values in a dataset to best reflect what the real values might have been. It is a necessary step before learning, as most models cannot handle data with missing values (Miao et al. 2022). Given a dataset D^{mis} with missing values $S^{mis} = \{ij | D^{mis}_{ij} = \phi\}$, a data imputation algorithm fills them with a set of non-empty values $S^{imp} = \{s_{ij} | ij \in S^{mis} \text{ and } s_{ij} \neq \phi\}$. The final imputed dataset can be defined as $D^{imp} = D^{mis} \oplus S^{imp}$, where the $\oplus$ operator represent filling the values missing in D^{mis} with values from S^{imp}. We define $|D^{mis}| = n$ and $|S^{mis}| = s$.

Two different imputation techniques may fill the missing values in different ways while the rest of the dataset remains unchanged, and the resulting imputed dataset will be k-neighbouring datasets, where $k \leq s$. Thus, we argue that the choice between data imputation techniques can also be seen as a choice between neighbouring datasets, fitting within our broader discussion.

5.2 Experiment Setup and Algorithms

Before jumping into the empirical results, we provide an overview of the experiment setup, as well as define our multiplicity-aware data imputation algorithms.

Dataset. We use three different datasets, ACSIncome (Ding et al. 2021), ACSEmployment (Ding et al. 2021), and Bank Customer Churn dataset (Topre 2025), to ensure the robustness of our findings. Due to limited space, we focus on the ACSIncome dataset in the main paper, while additional results and details of the experiment setup are delegated to the Appendix (§E).

We first divide the dataset into train and test sets, with a ratio of $[0.8, 0.2]$. Next, we randomly remove r fraction of values from the train set, giving us D^{miss}. The pipeline is repeated 10 times, while sticking with the same test set.

Models. We use LogisticRegression (LR), RandomForest (RF), and Multi-Layer Perceptron (MLP) with a single hidden layer of size 10, three model classes of varying complexity. We use RF as our default setup, while additional results for LR and MLP are in the Appendix (§E). To evaluate multiplicity, for each dataset, we train a total of 100 models and then select the Rashomon set. As discussed during formalization (Definitions 3.1, 3.2), the creation of the Rashomon set is done using model loss on the train set, while all evaluations are performed on the test set.

Evaluation Metrics. We use accuracy (0-1 scale) as a performance measure and ambiguity (Definition 3.2) as a measure of multiplicity. More details are in the Appendix (§E).

Baseline Algorithms and Multiplicity-Aware Data Imputation. We study five commonly used baselines in imputation: (a) Mean (Miao et al. 2022), filling with the mean of the feature, (b) Median (Miao et al. 2022), filling with the median of the feature, (c) Mode (Miao et al. 2022), filling with the mode of the feature, (d) kNN (Altman 1992), using k-nearest neighbours algorithm to find 5 neighbours and fill with the mean of their value, and (e) MICE (Van Buuren and Groothuis-Oudshoorn 2011), learning predictors of a feature using other features, one at a time, and improving iteratively.

In addition, our multiplicity-aware imputation techniques include: (a) MultLow, which checks the confidence of the data point for all five baseline imputations and chooses the one with the least confidence, repeating for all missing values, and (b) MultHigh, which instead chooses the one with the highest confidence. To get the confidence scores, we train a single model on the mean-imputed dataset. Our multiplicity-aware imputation algorithms use existing imputation techniques and choose between them for every missing value. We provide pseudocode in the Appendix (§C).

5.3 Stronger Control with More Missing Values

We start with the relationship between overlapping coefficients and the resulting multiplicity for data imputation. Average correlation scores across all random seeds are reported in Figure 3 (standard deviations are present in the Appendix). We see a clear negative correlation on average, supporting our hypothesis that higher overlap leads to lower multiplicity (Conjecture 4.1). Moreover, the correlations are stronger for LR and MLP, which may be attributed to a

Model	0.01	0.02	0.03	0.04	0.05	0.10	0.15	0.20	0.25
	Missing Data Ratio								
RandomForest	-0.34	-0.30	-0.19	-0.22	-0.48	-0.28	-0.09	-0.17	-0.42
LogisticRegression	-0.41	-0.53	-0.72	-0.75	-0.85	-0.84	-0.77	-0.73	-0.70
MultiLayerPerceptron	-0.22	-0.42	-0.13	-0.38	-0.33	-0.60	-0.28	-0.36	-0.41

Figure 3: Correlation between the overlapping coefficient and resulting multiplicity.

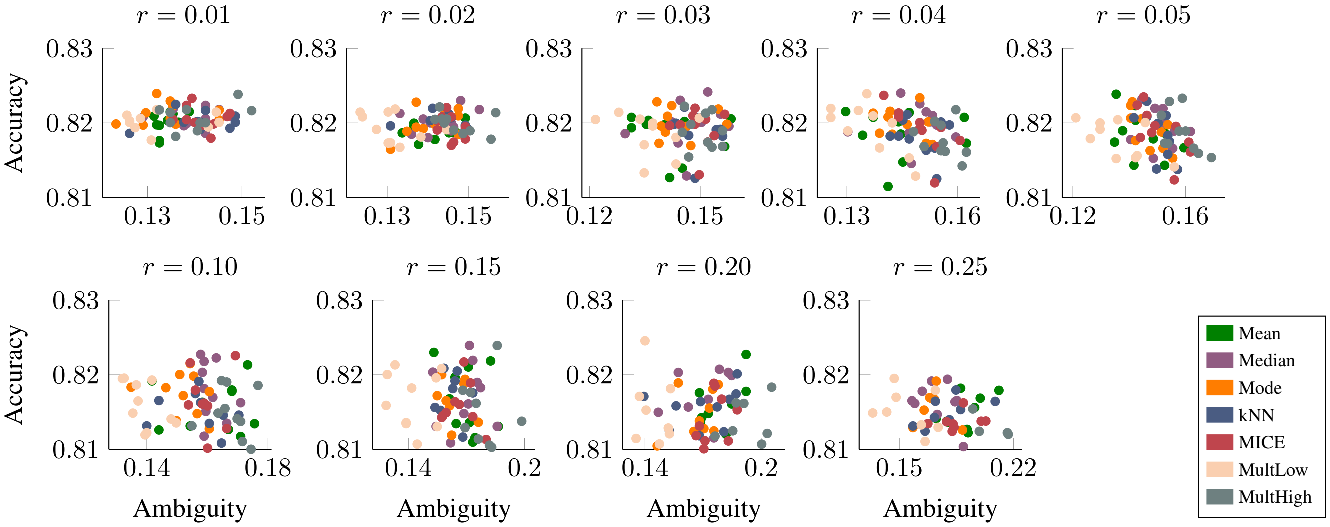


Figure 4: Accuracy and ambiguity for various data imputation strategies across varying values of missing data ratio r. MultLow (and MultHigh) algorithms stand out more for higher values of missing data ratio r, highlighting that a large amount of missing data can make the imputation more steerable.

poorer approximation of the true Rashomon set using only 100 models for RF.

The most intriguing results, however, come from our multiplicity-aware algorithms. In Figure 4, we present the average accuracy and resulting multiplicity across all random seeds and imputation techniques, evaluated over varying levels of missing data. Our techniques consistently achieve the lowest (or highest) multiplicity, but what stands out is that these trends become more pronounced at higher missing value ratios. With more missing data, the number of plausible imputations—and thus neighbouring datasets—increases. This leads to many neighbouring datasets varying substantially in downstream multiplicity, making our multiplicity-aware methods more valuable.

6 Multiplicity and Active Learning

Due to limited space, we delegate the entire analysis for active learning to the appendix (§A). However, we note that we find similar trends for active learning as data imputation, i.e., we find a high negative correlation between the overlapping coefficient and multiplicity, as well as effective control over multiplicity using our multiplicity-aware techniques.

7 Conclusion and Future Work

In this work, we introduced a neighbouring datasets framework to study the impact of data processing on multiplicity, offering a practical lens on the interplay between dataset and multiplicity. Our framework captures a wide range of data processing scenarios, provides theoretical insights into the relationship between neighbouring datasets and multiplicity, and reveals a surprising trend supported by rigorous proofs. We also demonstrated its utility through data imputation.

Looking ahead, an important avenue for future research is establishing a formal connection between our neighbouring dataset framework and differential privacy, which could yield valuable theoretical and practical insights in the future. Another promising direction involves revisiting the definition of neighbouring datasets, as alternatives based on L_1/L_2 distances may offer a closer alignment with robustness literature. This perspective opens up opportunities to study the influence of distribution shifts and adversarial data on multiplicity through the same lens of neighbouring datasets.

Acknowledgments

Funding support for project activities has been partially provided by Canada CIFAR AI Chair, Google award, FRQNT, and NSERC Discovery Grants program. We also express our gratitude to Compute Canada for their support in providing facilities for our evaluations.

Disclaimer

This paper was prepared by Hsiang Hsu prior to his employment at JPMorgan Chase & Co.. Therefore, this paper is not a product of the Research Department of JPMorgan Chase & Co. or its affiliates. Neither JPMorgan Chase & Co. nor any of its affiliates makes any explicit or implied representation or warranty and none of them accept any liability in connection with this paper, including, without limitation, with respect to the completeness, accuracy, or reliability of the information contained herein and the potential legal, compliance, tax, or accounting effects thereof. This document is not intended as investment research or investment advice, or as a recommendation, offer, or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction.

References

Aggarwal, C. C.; Kong, X.; Gu, Q.; Han, J.; and Yu, P. S. 2014. Active learning: A survey. In *Data classification*, 599–634. Chapman and Hall/CRC.

Aguinis, H.; Gottfredson, R. K.; and Joo, H. 2013. Best-practice recommendations for defining, identifying, and handling outliers. *Organizational research methods*, 16(2): 270–301.

Altman, N. S. 1992. An Introduction to Kernel and Nearest-Neighbor. *The American Statistician*, 46(3): 175–185.

Black, E.; Raghavan, M.; and Barocas, S. 2022. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 850–863.

Breiman, L. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3): 199–231.

Cavus, M.; and Biecek, P. 2024. Investigating the Impact of Balancing, Filtering, and Complexity on Predictive Multiplicity: A Data-Centric Perspective. *arXiv preprint arXiv:2412.09712*.

Cooper, A. F.; Frankle, J.; and De Sa, C. 2022. Non-Determinism and the Lawlessness of Machine Learning Code. In *Proceedings of the 2022 Symposium on Computer Science and Law*, 1–8.

Creel, K.; and Hellman, D. 2022. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy*, 52(1): 26–43.

Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34: 6478–6490.

Dudley, R. M. 2018. *Real analysis and probability*. Chapman and Hall/CRC.

Dwork, C. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, 1–12. Springer.

Ganesh, P. 2024. An Empirical Investigation into Benchmarking Model Multiplicity for Trustworthy Machine Learning: A Case Study on Image Classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4488–4497.

Ganesh, P.; Taik, A.; and Farnadi, G. 2025. Systemizing Multiplicity: The Curious Case of Arbitrariness in Machine Learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Gomez, J. F.; Machado, C.; Paes, L. M.; and Calmon, F. 2024. Algorithmic Arbitrariness in Content Moderation. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2234–2253.

Gur-Arieh, S.; and Lee, C. 2025. Consistently Arbitrary or Arbitrarily Consistent: Navigating the Tensions Between Homogenization and Multiplicity in Algorithmic Decision-Making. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 3336–3349.

Hsu, H.; Brugere, I.; Sharma, S.; Lecue, F.; and Chen, R. 2024a. Rashomongb: Analyzing the rashomon effect and mitigating predictive multiplicity in gradient boosting. *Advances in Neural Information Processing Systems*, 37: 121265–121303.

Hsu, H.; and Calmon, F. 2022. Rashomon capacity: A metric for predictive multiplicity in classification. *Advances in Neural Information Processing Systems*, 35: 28988–29000.

Hsu, H.; Li, G.; Hu, S.; and Chen, C.-F. 2024b. Dropout-Based Rashomon Set Exploration for Efficient Predictive Multiplicity Estimation. In *The Twelfth International Conference on Learning Representations*.

Inman, H. F.; and Bradley Jr, E. L. 1989. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics-theory and Methods*, 18(10): 3851–3874.

Jain, S.; Creel, K.; and Wilson, A. 2024. Scarce Resource Allocations That Rely On Machine Learning Should Be Randomized. *arXiv preprint arXiv:2404.08592*.

Jain, S.; Suriyakumar, V.; Creel, K.; and Wilson, A. 2024. Algorithmic Pluralism: A Structural Approach To Equal Opportunity. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 197–206.

Kissel, N.; and Mentch, L. 2024. Forward stability and model path selection. *Statistics and Computing*, 34(2): 82.

Kleinberg, J.; and Raghavan, M. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22): e2018340118.

Koh, P. W. W.; Ang, K.-S.; Teo, H.; and Liang, P. S. 2019. On the accuracy of influence functions for measuring group effects. *Advances in neural information processing systems*, 32.

Marx, C.; Calmon, F.; and Ustun, B. 2020. Predictive multiplicity in classification. In *International Conference on Machine Learning*, 6765–6774. PMLR.

Meyer, A. P.; Albarghouthi, A.; and D'Antoni, L. 2023. The dataset multiplicity problem: How unreliable data impacts predictions. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 193–204.

Miao, X.; Wu, Y.; Chen, L.; Gao, Y.; and Yin, J. 2022. An experimental survey of missing data imputation algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 35(7): 6630–6650.

Neale, J. 2016. Iterative categorization (IC): a systematic technique for analysing qualitative data. *Addiction*, 111(6): 1096–1106.

Paullada, A.; Raji, I. D.; Bender, E. M.; Denton, E.; and Hanna, A. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11).

Ren, P.; Xiao, Y.; Chang, X.; Huang, P.-Y.; Li, Z.; Gupta, B. B.; Chen, X.; and Wang, X. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9): 1–40.

Rudin, C.; Zhong, C.; Semenova, L.; Seltzer, M.; Parr, R.; Liu, J.; Katta, S.; Donnelly, J.; Chen, H.; and Boner, Z. 2024. Amazing things come from having many good models. *arXiv preprint arXiv:2407.04846*.

Semenova, L.; Chen, H.; Parr, R.; and Rudin, C. 2024. A path to simpler models starts with noise. *Advances in neural information processing systems*, 36.

Seung, H. S.; Opper, M.; and Sompolinsky, H. 1992. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, 287–294.

Sokol, K.; Kull, M.; Chan, J.; and Salim, F. 2024. Cross-model Fairness: Empirical Study of Fairness and Ethics Under Model Multiplicity. *ACM Journal on Responsible Computing*, 1(3): 1–27.

Topre, G. 2025. Bank Customer Churn Dataset. Accessed: 2025-09-20.

Van Buuren, S.; and Groothuis-Oudshoorn, K. 2011. mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45: 1–67.

Watson-Daniels, J.; Barocas, S.; Hofman, J. M.; and Chouldechova, A. 2023. Multi-target multiplicity: Flexibility and fairness in target specification under resource constraints. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 297–311.

Watson-Daniels, J.; Calmon, F. d. P.; D'Amour, A.; Long, C.; Parkes, D. C.; and Ustun, B. 2024. Predictive churn with the set of good models. *arXiv preprint arXiv:2402.07745*.

Watson-Daniels, J.; Parkes, D. C.; and Ustun, B. 2023. Predictive multiplicity in probabilistic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10306–10314.

Xin, R.; Zhong, C.; Chen, Z.; Takagi, T.; Seltzer, M.; and Rudin, C. 2022. Exploring the whole rashomon set of sparse decision trees. *Advances in neural information processing systems*, 35: 14071–14084.

Zhong, C.; Chen, Z.; Liu, J.; Seltzer, M.; and Rudin, C. 2024. Exploring and interacting with the set of good sparse generalized additive models. *Advances in neural information processing systems*, 36.

A Multiplicity and Active Learning

We now turn to our second application: data acquisition for active learning, repeating the k-neighbouring dataset formulation and the empirical study on existing algorithms alongside our own multiplicity-aware techniques. Combined with the results from data imputation, these studies underscore the practical utility of our framework in analyzing and guiding developer decisions during data processing.

A.1 Neighbouring Datasets in Active Learning

In active learning, we have access to a large pool of unlabeled data, and the objective is to selectively acquire a small subset of the potentially most informative data points to be labeled, known as data acquisition. Typically, active learning begins with a small labeled dataset D^0_{lab} and a large pool of unlabeled points X^0_{unlab}. At each timestep t, the algorithm uses the current labeled dataset D^t_{lab} and the remaining unlabeled pool X^t_{unlab} to select a batch of points $X^t_{or} \subset X^t_{unlab}$, to be labeled by the oracle. Once labeled, these are added to the labeled dataset, i.e., $D^{t+1}_{lab} = D^t_{lab} + (X^t_{or}, Y^t_{or})$. We define the initial labeled set size $|D^0_{lab}| = n$, and $|X^t_{or}| = q$ points are labeled at each step.

Over a total of T steps, two different active learning algorithms may choose distinct sequences of points to label. It is easy to see that the resulting labeled datasets can be considered k-neighbouring datasets with $k \leq Tq$. Thus, we argue that the choice between active learning strategies can also be seen as a choice between neighbouring datasets.

A.2 Experiment Setup and Algorithms

We use the same experiment setup as in data imputation, but with the following differences (more details in §D),

Dataset. We focus on the ACSIncome dataset in this section, while additional results are delegated to §D.

After dividing the dataset into train and test sets, we sample n points randomly from the train set that will serve as our D^0_{lab}, and test three different values of $n \in \{500, 1000, 2000\}$. The rest of the train set is our unlabeled pool of data. We run various active learning algorithms with a query size $q = 100$ for a total of $T = 5$ steps. The complete pipeline starting from sampling D^0_{lab} is repeated 10 times, while sticking with the same test set.

Baseline Algorithms and Multiplicity-Aware Data Acquisition. We study three common baselines for active learning: (a) Random (Aggarwal et al. 2014), data points to be labeled are chosen at random, (b) Confidence (Aggarwal et al. 2014), data points with the lowest prediction confidence are chosen, and (c) Committee (Seung, Opper, and Sompolinsky 1992), data points with the most conflicting predictions from a committee of 100 models trained on the current labeled data are chosen.

In addition, we propose two new data acquisition algorithms: (a) MultLow, which trains a committee of models on the labeled data and chooses the data points with low confidence in all models of the committee, and (b) MultHigh, which is similar but instead chooses the data points with high confidence in all models of the committee. Pseudocode for both algorithms is in the Appendix (§C).

RF	Number of Steps t				
Size of D^0_{lab} (n)	1	2	3	4	5
500	-0.21	-0.34	-0.12	-0.01	-0.02
1000	-0.35	-0.28	-0.20	-0.20	-0.25
2000	-0.41	-0.31	-0.27	-0.22	-0.19

LR	Number of Steps t				
Size of D^0_{lab} (n)	1	2	3	4	5
500	-0.73	-0.72	-0.61	-0.51	-0.57
1000	-0.84	-0.78	-0.61	-0.52	-0.44
2000	-0.80	-0.77	-0.64	-0.60	-0.69

MLP	Number of Steps t				
Size of D^0_{lab} (n)	1	2	3	4	5
500	-0.54	-0.35	-0.31	-0.38	-0.25
1000	-0.63	-0.62	-0.44	-0.35	-0.29
2000	-0.85	-0.65	-0.48	-0.43	-0.28

Figure 5: Spearman's rank correlation coefficients between the overlap and resulting multiplicity.

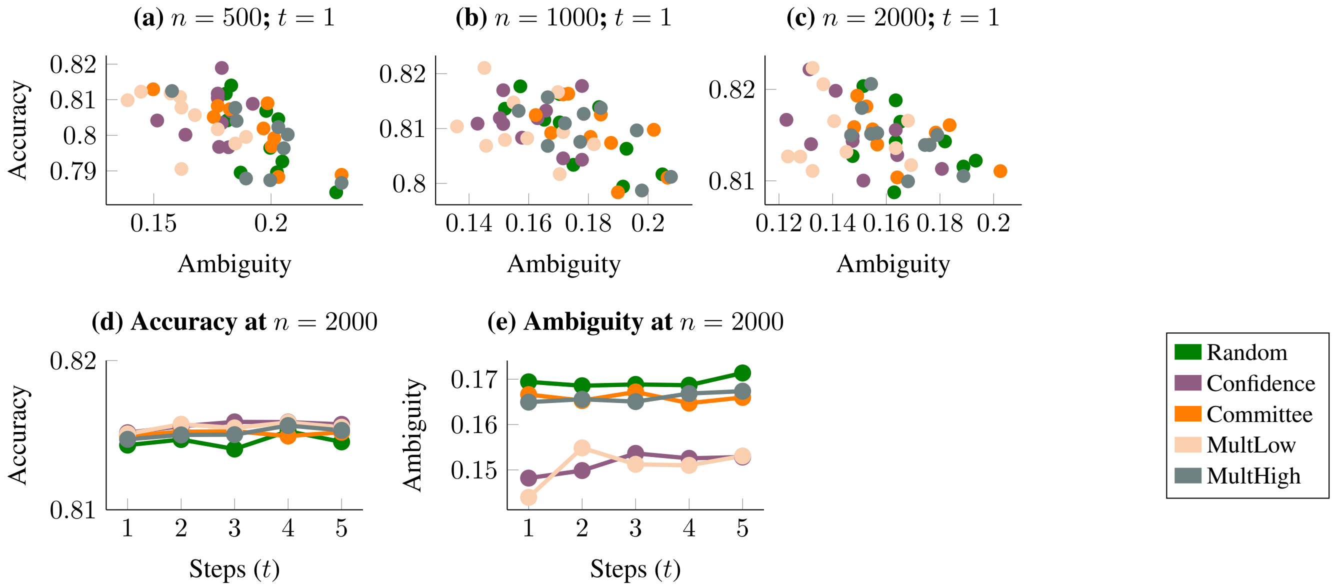


Figure 6: **(a, b, c)** Accuracy and ambiguity across various strategies for one step of data acquisition. We see clear trends of our MultLow (and MultHigh) approach(es) getting the lowest (and highest) multiplicity, while maintaining similar accuracies. **(d, e)** Accuracy and ambiguity across multiple steps of data acquisition. Similar trends persist across multiple steps.

A.3 Controlling Multiplicity During Active Learning

We start by examining the relationship between the overlapping coefficient and multiplicity for varying initial labeled sizes (n) and active learning steps (t), across all algorithms. Average correlation scores across all random seeds are reported in Figure 5 (standard deviations are present in the Appendix). We see a clear negative correlation on average, supporting our hypothesis that higher overlap leads to lower multiplicity (Conjecture 4.1). Unsurprisingly, the correlation is stronger when n is large or t is small, i.e., settings where $k \ll n$. Moreover, the correlations are stronger for LR and MLP, which may be attributed to a poorer approximation of the true Rashomon set using only 100 models for RF.

Moving beyond the overall correlation, we next analyze the trends exhibited by each algorithm separately in Figure 6. Our algorithms, MultLow and MultHigh, consistently achieve the lowest (and highest) multiplicity. Even in scenarios where our theoretical assumptions do not hold—such as when n is small or t is large—the efficacy of our algorithms, MultLow and MultHigh, indicates that our insights extend well beyond strict theoretical settings. This robustness highlights the practical utility of our approach across a broader range of real-world scenarios involving neighbouring datasets.

B Proof for Theorem 4.1 and Insights for Conjecture 4.1

Step 1: Bayes optimal 0-1 loss in terms of Overlapping Coefficient The Bayes optimal classifier minimizes the 0-1 loss, predicting the class with the higher posterior probability at each x. So the Bayes classifier $\theta^*(x)$ predicts:

$$\theta^*(x) = \arg\max_{y \in \{0,1\}} P_y(x).$$

Thus, the Bayes 0-1 loss L^* can be expressed as the expected probability of misclassification:

$$L^* = \mathbb{E}_x\left[\min(P_0(x), P_1(x))\right] = \sum_x \min(\pi_0 P_0(x), \pi_1 P_1(x)).$$

where π_0, π_1 are class priors for both classes. Assuming identical class priors, we can simplify it as,

$$L^* = \frac{1}{2}\sum_x \min(P_0(x), P_1(x)) = \frac{1}{2} OVL(P_0, P_1).$$

Hence, for the two neighbouring datasets:

$$L_1^* = \frac{1}{2}OVL^1_{\text{train}}, \quad L_2^* = \frac{1}{2}OVL^2_{\text{train}}.$$

Therefore, if $L_1^* \geq L_2^*$, then:

$$OVL^1_{\text{train}} \geq OVL^2_{\text{train}}.$$

Step 2: Overlapping coefficient is higher for $D^1_{\textbf{train}}$ We know from our second assumption that:

$$L(\theta_1^*, (x_0^1, y_0^1)) \geq L(\theta_2^*, (x_0^2, y_0^2)).$$

Since all other training examples are shared between the two datasets, the only difference in their total empirical losses lies in this one datapoint. Thus, the empirical loss satisfies:

$$L_1^* \geq L_2^*,$$

From the derivation in Step 1, we thus get:

$$OVL^1_{\text{train}} \geq OVL^2_{\text{train}}.$$

Step 3: Subset relationship from loss dominance Now let $\theta \in \Theta_{(D^1_{\text{train}},\epsilon)}$, i.e., $L_{D^1_{\text{train}}}(\theta) \leq \epsilon$. Since D^1_{train} and D^2_{train} differ in only one datapoint, we can write:

$$L_{D^1_{\text{train}}}(\theta) = \frac{1}{n}\left(L(\theta, (x_0^1, y_0^1)) + \sum_{j=1}^{n-1} L(\theta, (x_j, y_j))\right),$$

$$L_{D^2_{\text{train}}}(\theta) = \frac{1}{n}\left(L(\theta, (x_0^2, y_0^2)) + \sum_{j=1}^{n-1} L(\theta, (x_j, y_j))\right).$$

Subtracting:

$$\begin{aligned} &L_{D^1_{\text{train}}}(\theta) - L_{D^2_{\text{train}}}(\theta) \\ &= \frac{1}{n}\left(L(\theta, (x_0^1, y_0^1)) - L(\theta, (x_0^2, y_0^2))\right) \geq 0, \end{aligned}$$

by the assumed loss inequality in our first assumption. Therefore:

$$L_{D^2_{\text{train}}}(\theta) \leq L_{D^1_{\text{train}}}(\theta) \leq \epsilon \quad \Rightarrow \quad \theta \in \Theta_{(D^2_{\text{train}},\epsilon)}.$$

Since this holds for all $\theta \in \Theta_{(D^1_{\text{train}},\epsilon)}$, but not necessarily the other way around, we conclude:

$$\Theta_{(D^1_{\text{train}},\epsilon)} \subseteq \Theta_{(D^2_{\text{train}},\epsilon)}.$$

B.1 Extension to k-neighbouring datasets

Let D^1_{train} and D^2_{train} be k-neighbouring binary classification datasets, i.e., they differ at k indices $\mathcal{I} = \{i_1, \ldots, i_k\}$, such that for all $j \in \mathcal{I}$, we have:

$$(x_j^1, y_j^1) \neq (x_j^2, y_j^2),$$

and for all other $j \notin \mathcal{I}$, the examples are shared:

$$(x_j^1, y_j^1) = (x_j^2, y_j^2).$$

Suppose further that the per-point loss dominance condition holds at all differing indices:

$$\begin{gathered} L(\theta, (x_j^1, y_j^1)) \geq L(\theta, (x_j^2, y_j^2)) \\ \forall \theta \in \Theta_{(D^1_{\text{train}},\epsilon)} \cup \Theta_{(D^2_{\text{train}},\epsilon)}, \quad \forall j \in \mathcal{I}. \end{gathered}$$

and

$$L(\theta_1^*, (x_j^1, y_j^1)) \geq L(\theta_2^*, (x_j^2, y_j^2)) \quad \forall j \in \mathcal{I}.$$

Then the overlapping coefficient between the classes is higher for D^1_{train}, i.e.,

$$OVL^1_{\text{train}} \geq OVL^2_{\text{train}},$$

and the Rashomon set satisfies:

$$\Theta_{(D^1_{\text{train}},\epsilon)} \subseteq \Theta_{(D^2_{\text{train}},\epsilon)}.$$

To prove this, we can simply decompose the k-neighbouring datasets into a sequence of k consecutive 1-neighbouring transitions:

$$D^1_{\text{train}} = D^{(0)} \to D^{(1)} \to \cdots \to D^{(k)} = D^2_{\text{train}},$$

where each $D^{(t)}$ and $D^{(t+1)}$ differ at exactly one datapoint (x_t, y_t), and the loss dominance condition holds at each step. From Theorem 4.1, each such one-step transition satisfies:

$$\Theta_{(D^{(t)},\epsilon)} \subseteq \Theta_{(D^{(t+1)},\epsilon)}.$$

Applying this sequentially:

$$\Theta_{(D^1_{\text{train}},\epsilon)} = \Theta_{(D^{(0)},\epsilon)} \subseteq \cdots \subseteq \Theta_{(D^{(k)},\epsilon)} = \Theta_{(D^2_{\text{train}},\epsilon)}.$$

Thus,

$$\Theta_{(D^1_{\text{train}},\epsilon)} \subseteq \Theta_{(D^2_{\text{train}},\epsilon)}.$$

C Pseudocode for Algorithms

C.1 MultLow and MultHigh for Data Acquisition

Algorithm 1: MultLow for Data Acquisition

```
Require: Labeled dataset $L$, unlabeled dataset $U$, query size $Q$, committee size $K$
Train a committee of $K$ models $\{M_1, M_2, \ldots, M_K\}$ on $L$
Initialize $S \leftarrow \emptyset$
for each $x \in U$ do
    Compute confidence $c_i(x)$ from each model $M_i$
    Compute maximum confidence across all models: $c_{\max}(x) = \max_i c_i(x)$
end for
Select bottom-$Q$ points with lowest $c_{\max}(x)$ values
$S \leftarrow$ selected points
return $S$
```

Algorithm 2: MultHigh for Data Acquisition

```
Require: Labeled dataset $L$, unlabeled dataset $U$, query size $Q$, committee size $K$
Train a committee of $K$ models $\{M_1, M_2, \ldots, M_K\}$ on $L$
Initialize $S \leftarrow \emptyset$
for each $x \in U$ do
    Compute confidence $c_i(x)$ from each model $M_i$
    Compute minimum confidence across all models: $c_{\min}(x) = \min_i c_i(x)$
end for
Select top-$Q$ points with highest $c_{\min}(x)$ values
$S \leftarrow$ selected points
return $S$
```

C.2 MultLow and MultHigh for Data Imputation

Algorithm 3: MultLow for Data Imputation

```
Require: Dataset with missing values $D$, set of baseline imputations $\{I_{mean}, I_{median}, \dots, I_{mice}\}$
Train a model $C$ on mean-imputed version of $D$
Initialize $D' \leftarrow D$
for each record $r$ with missing values in $D$ do
    for each imputation method $I_j$ do
        Compute imputed record $r_j$ using $I_j$
        Compute confidence score $c_j = C(r_j)$
    end for
    Select $r^* = r_j$ with lowest confidence $c_j$
    Fill $r$ in $D'$ with $r^*$
end for
return $D'$
```

Algorithm 4: MultHigh for Data Imputation

```
Require: Dataset with missing values $D$, set of baseline imputations $\{I_{mean}, I_{median}, \dots, I_{mice}\}$
Train a model $C$ on mean-imputed version of $D$
Initialize $D' \leftarrow D$
for each record $r$ with missing values in $D$ do
    for each imputation method $I_j$ do
        Compute imputed record $r_j$ using $I_j$
        Compute confidence score $c_j = C(r_j)$
    end for
    Select $r^* = r_j$ with highest confidence $c_j$
    Fill $r$ in $D'$ with $r^*$
end for
return $D'$
```

D Additional Results for Active Learning

In this appendix section, we provide detailed results across all datasets for active learning. We find similar trends across various datasets and models as seen in the main paper.

D.1 Experiment Setup Details

Folktables Subset. We use the "New Mexico" state subset for both ACSIncome and ACSEmployment throughout the paper.

Choosing Rashomon parameter ϵ. To make sure the Rashomon set contains enough models for each algorithm in our setup while keeping the threshold tight, the value of ϵ is chosen to be the smallest value possible such that there are at least 50 models in the Rashomon set for each setup. The ϵ value is chosen separately for each setting, i.e., each random seed, initial dataset size, and number of steps; but is shared between all different algorithms, i.e., a common Rashomon parameter ϵ is used across algorithms for any particular setting.

D.2 All Results for ACSIncome Dataset

Here, we provide detailed results for active learning on the ACSIncome dataset. First, we restate the results in Figure 5, along with the standard deviations recorded separately, present in Figure 7. We then repeat the experiments in Figure 6(d, e) for LR and MLP models, and the results are presented in Figure 8.

D.3 All Results for ACSEmployment Dataset

Here, we provide detailed results for active learning on the ACSEmployment dataset. First, we repeat the experiments in Figure 5, along with the standard deviations recorded separately, present in Figure 9. We then repeat the experiments in Figure 6(d, e) for all three model types, and the results are presented in Figure 10.

D.4 All Results for Bank Dataset

Here, we provide detailed results for active learning on the Bank dataset. First, we repeat the experiments in Figure 5, along with the standard deviations recorded separately, present in Figure 11. We then repeat the experiments in Figure 6(d, e) for all three model types, and the results are presented in Figure 12.

E Additional Results for Data Imputation

In this appendix section, we provide detailed results across all datasets for data imputation. We find similar trends across various datasets and models as seen in the main paper.

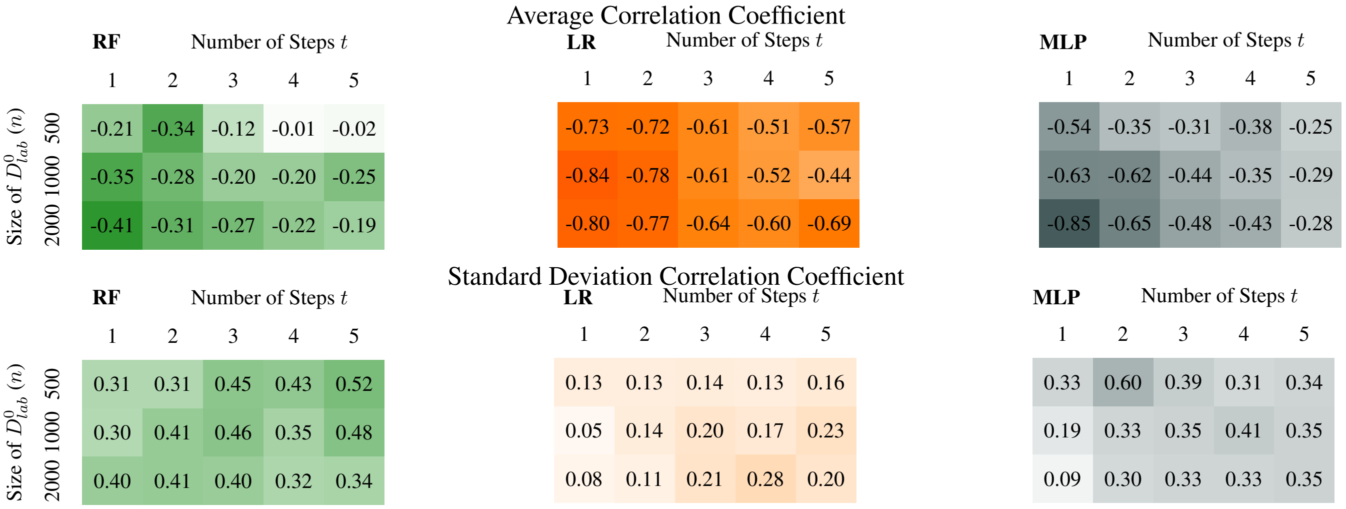
E.1 Experiment Setup Details

Folktables Subset. Same details as above in §D

Choosing Rashomon parameter ϵ. Same details as above in §D.

E.2 All Results for Bank Dataset

Here, we first repeat the experiments in Figure 3 for the Bank dataset and provide the results in Figure 13. Next, we repeat the experiments in Figure 4 for the Bank dataset using RandomForests, and the results are presented in Figure 14.



Average Correlation Coefficient

RF

Size of D_{lab}^0 (n)	Number of Steps t				
	1	2	3	4	5
500	-0.21	-0.34	-0.12	-0.01	-0.02
1000	-0.35	-0.28	-0.20	-0.20	-0.25
2000	-0.41	-0.31	-0.27	-0.22	-0.19

LR

Size of D_{lab}^0 (n)	Number of Steps t				
	1	2	3	4	5
500	-0.73	-0.72	-0.61	-0.51	-0.57
1000	-0.84	-0.78	-0.61	-0.52	-0.44
2000	-0.80	-0.77	-0.64	-0.60	-0.69

MLP

Size of D_{lab}^0 (n)	Number of Steps t				
	1	2	3	4	5
500	-0.54	-0.35	-0.31	-0.38	-0.25
1000	-0.63	-0.62	-0.44	-0.35	-0.29
2000	-0.85	-0.65	-0.48	-0.43	-0.28

Standard Deviation Correlation Coefficient

RF

Size of D_{lab}^0 (n)	Number of Steps t				
	1	2	3	4	5
500	0.31	0.31	0.45	0.43	0.52
1000	0.30	0.41	0.46	0.35	0.48
2000	0.40	0.41	0.40	0.32	0.34

LR

Size of D_{lab}^0 (n)	Number of Steps t				
	1	2	3	4	5
500	0.13	0.13	0.14	0.13	0.16
1000	0.05	0.14	0.20	0.17	0.23
2000	0.08	0.11	0.21	0.28	0.20

MLP

Size of D_{lab}^0 (n)	Number of Steps t				
	1	2	3	4	5
500	0.33	0.60	0.39	0.31	0.34
1000	0.19	0.33	0.35	0.41	0.35
2000	0.09	0.30	0.33	0.33	0.35

Figure 7: Average and standard deviation of spearman's rank correlation coefficients between the overlap and resulting multiplicity for the ACSIncome dataset.

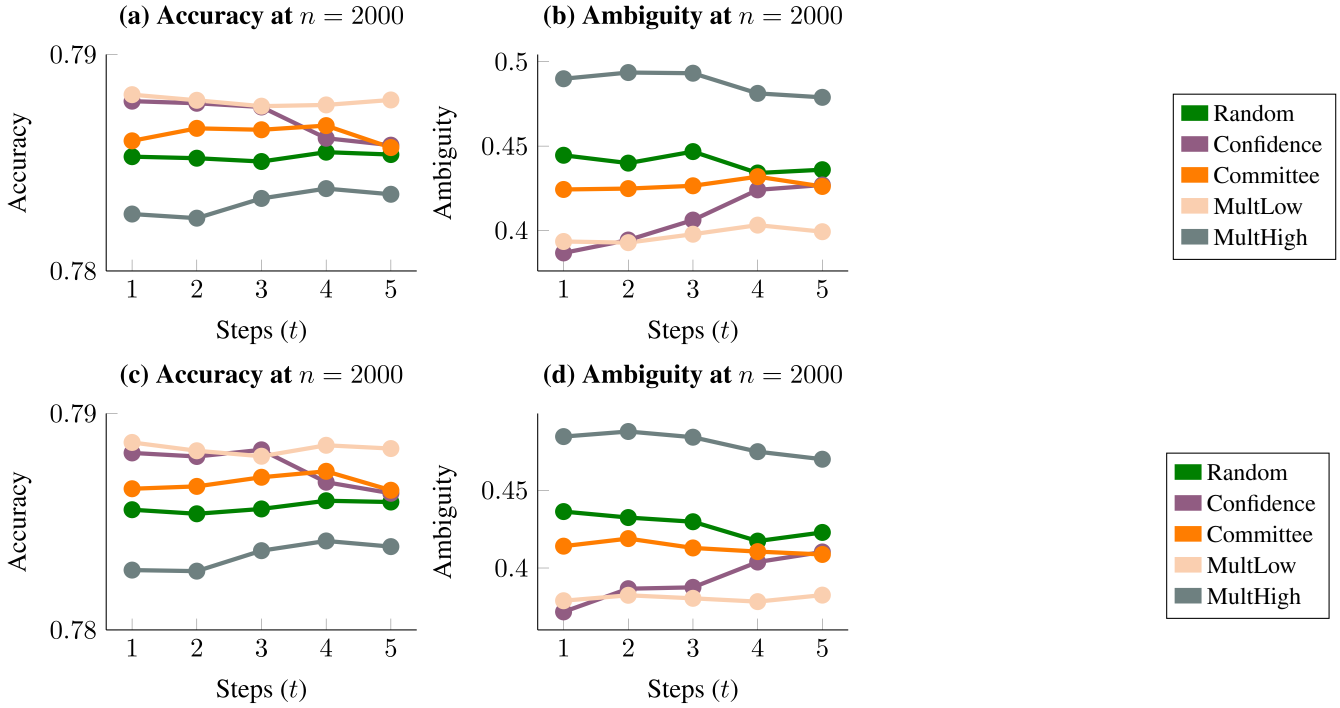
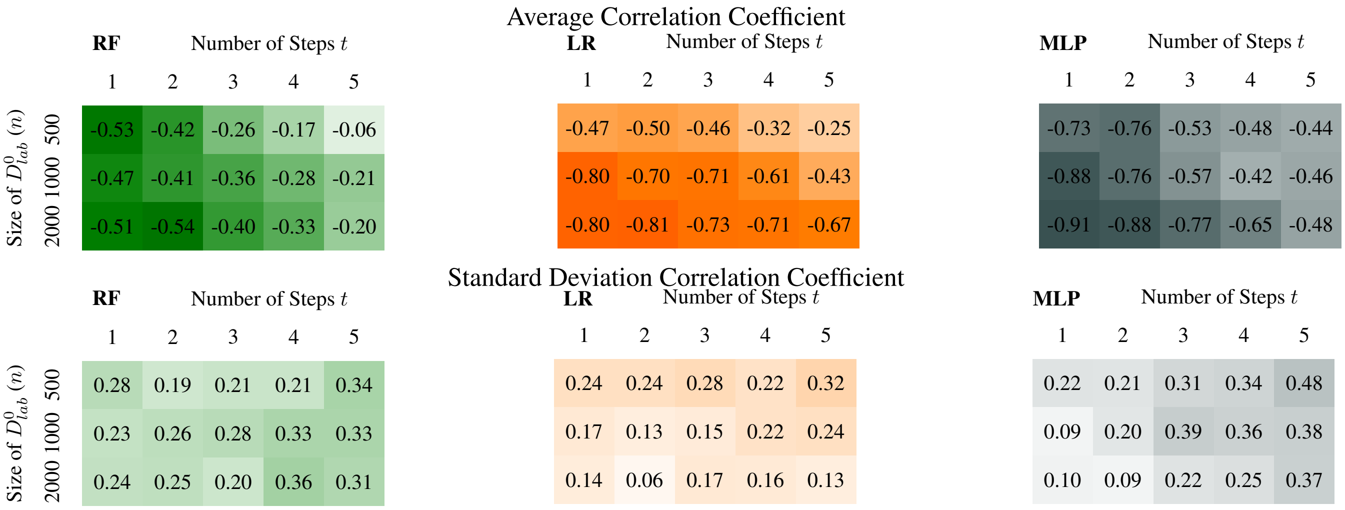


Figure 8: Accuracy and ambiguity across multiple steps of data acquisition for LR (top, (a), (b)) and MLP (bottom, (c), (d)) models for ACSIncome dataset. Similar trends persist across multiple steps of active learning.



Average Correlation Coefficient

RF

Size of D^0_{lab} (n) \ Number of Steps t	1	2	3	4	5
500	-0.53	-0.42	-0.26	-0.17	-0.06
1000	-0.47	-0.41	-0.36	-0.28	-0.21
2000	-0.51	-0.54	-0.40	-0.33	-0.20

LR

Size of D^0_{lab} (n) \ Number of Steps t	1	2	3	4	5
500	-0.47	-0.50	-0.46	-0.32	-0.25
1000	-0.80	-0.70	-0.71	-0.61	-0.43
2000	-0.80	-0.81	-0.73	-0.71	-0.67

MLP

Size of D^0_{lab} (n) \ Number of Steps t	1	2	3	4	5
500	-0.73	-0.76	-0.53	-0.48	-0.44
1000	-0.88	-0.76	-0.57	-0.42	-0.46
2000	-0.91	-0.88	-0.77	-0.65	-0.48

Standard Deviation Correlation Coefficient

RF

Size of D^0_{lab} (n) \ Number of Steps t	1	2	3	4	5
500	0.28	0.19	0.21	0.21	0.34
1000	0.23	0.26	0.28	0.33	0.33
2000	0.24	0.25	0.20	0.36	0.31

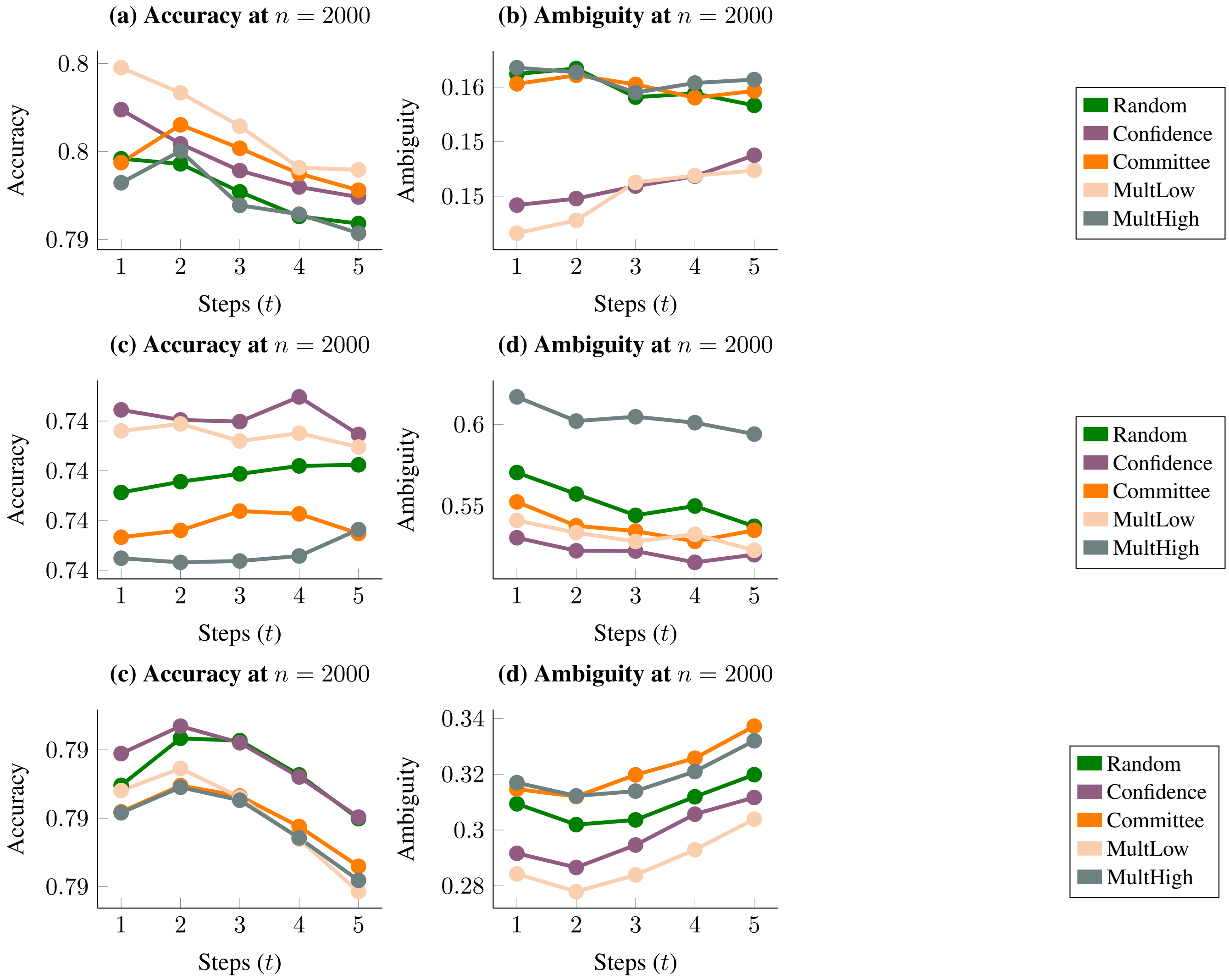
LR

Size of D^0_{lab} (n) \ Number of Steps t	1	2	3	4	5
500	0.24	0.24	0.28	0.22	0.32
1000	0.17	0.13	0.15	0.22	0.24
2000	0.14	0.06	0.17	0.16	0.13

MLP

Size of D^0_{lab} (n) \ Number of Steps t	1	2	3	4	5
500	0.22	0.21	0.31	0.34	0.48
1000	0.09	0.20	0.39	0.36	0.38
2000	0.10	0.09	0.22	0.25	0.37

Figure 9: Average and standard deviation of spearman's rank correlation coefficients between the overlap and resulting multiplicity for the ACSEmployment dataset.



(a) Accuracy at $n = 2000$

(b) Ambiguity at $n = 2000$

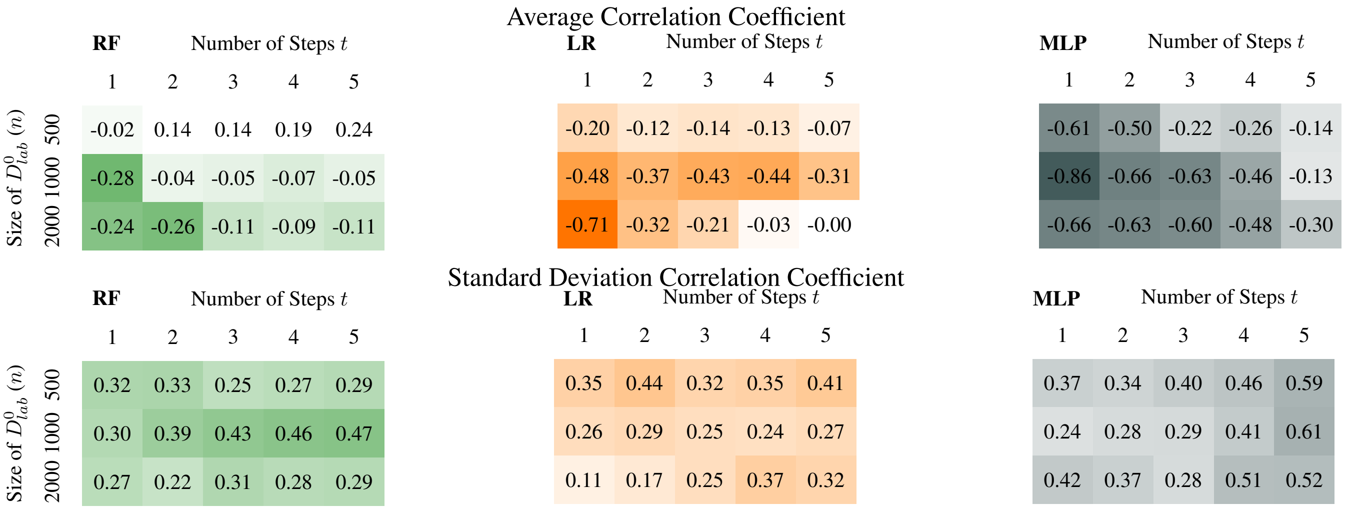
(c) Accuracy at $n = 2000$

(d) Ambiguity at $n = 2000$

(c) Accuracy at $n = 2000$

(d) Ambiguity at $n = 2000$

Figure 10: Accuracy and ambiguity across multiple steps of data acquisition for RF (top, (a), (b)), LR (middle, (c), (d)) and MLP (bottom, (e), (f)) models for ACSEmployment dataset. Similar trends persist across multiple steps of active learning.



Average Correlation Coefficient

RF

Size of D^0_{lab} (n)	Number of Steps t 1	2	3	4	5
500	-0.02	0.14	0.14	0.19	0.24
1000	-0.28	-0.04	-0.05	-0.07	-0.05
2000	-0.24	-0.26	-0.11	-0.09	-0.11

LR

Size of D^0_{lab} (n)	Number of Steps t 1	2	3	4	5
500	-0.20	-0.12	-0.14	-0.13	-0.07
1000	-0.48	-0.37	-0.43	-0.44	-0.31
2000	-0.71	-0.32	-0.21	-0.03	-0.00

MLP

Size of D^0_{lab} (n)	Number of Steps t 1	2	3	4	5
500	-0.61	-0.50	-0.22	-0.26	-0.14
1000	-0.86	-0.66	-0.63	-0.46	-0.13
2000	-0.66	-0.63	-0.60	-0.48	-0.30

Standard Deviation Correlation Coefficient

RF

Size of D^0_{lab} (n)	Number of Steps t 1	2	3	4	5
500	0.32	0.33	0.25	0.27	0.29
1000	0.30	0.39	0.43	0.46	0.47
2000	0.27	0.22	0.31	0.28	0.29

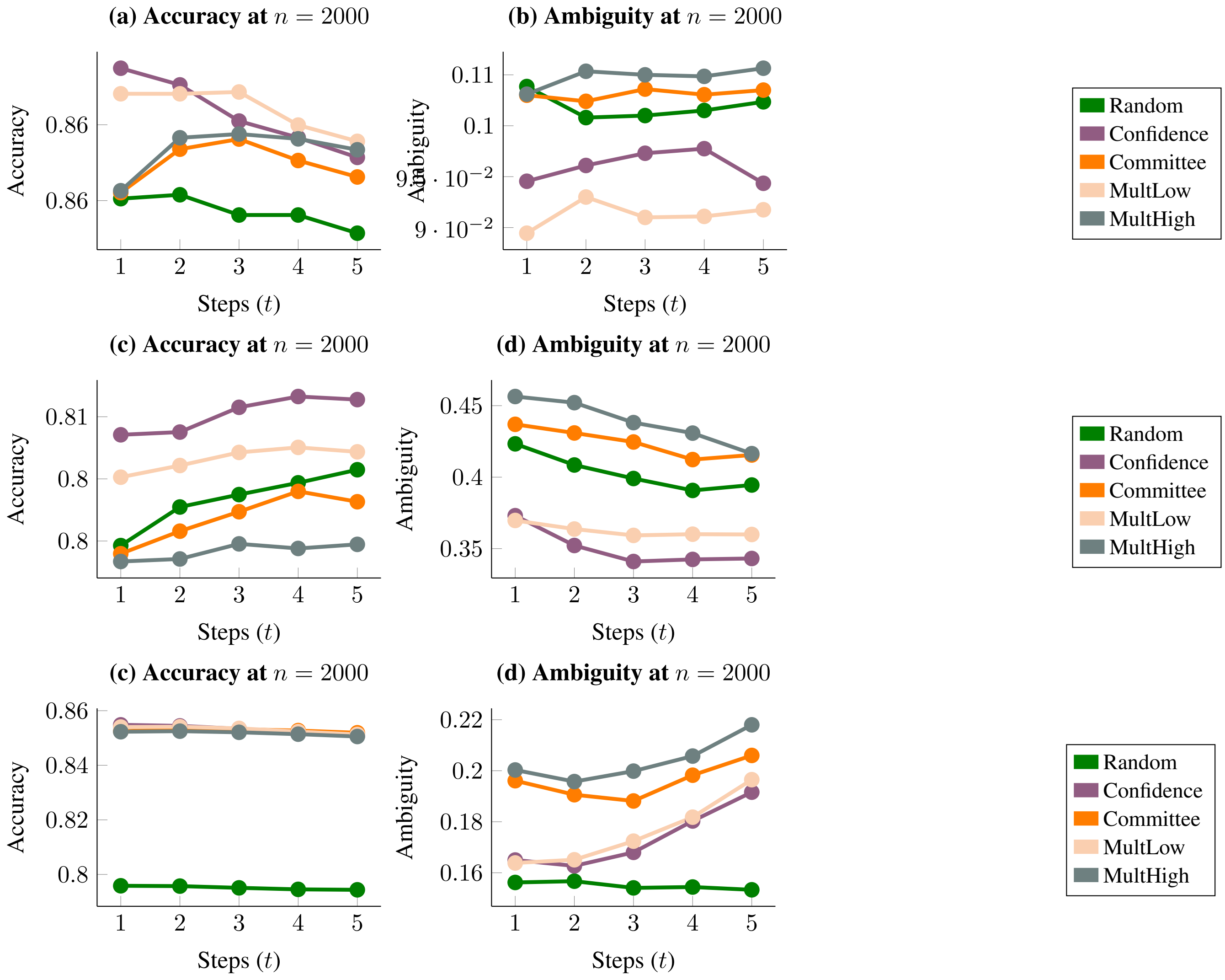
LR

Size of D^0_{lab} (n)	Number of Steps t 1	2	3	4	5
500	0.35	0.44	0.32	0.35	0.41
1000	0.26	0.29	0.25	0.24	0.27
2000	0.11	0.17	0.25	0.37	0.32

MLP

Size of D^0_{lab} (n)	Number of Steps t 1	2	3	4	5
500	0.37	0.34	0.40	0.46	0.59
1000	0.24	0.28	0.29	0.41	0.61
2000	0.42	0.37	0.28	0.51	0.52

Figure 11: Average and standard deviation of spearman's rank correlation coefficients between the overlap and resulting multiplicity for the Bank dataset.



(a) Accuracy at $n = 2000$

(b) Ambiguity at $n = 2000$

(c) Accuracy at $n = 2000$

(d) Ambiguity at $n = 2000$

(c) Accuracy at $n = 2000$

(d) Ambiguity at $n = 2000$

Figure 12: Accuracy and ambiguity across multiple steps of data acquisition for RF (top, (a), (b)), LR (middle, (c), (d)) and MLP (bottom, (e), (f)) models for Bank dataset. Similar trends persist across multiple steps of active learning.

	Missing Data Ratio								
Model	0.01	0.02	0.03	0.04	0.05	0.10	0.15	0.20	0.25
RandomForest	-0.50	-0.49	-0.37	-0.40	-0.29	-0.01	0.17	0.32	0.41
LogisticRegression	0.17	-0.21	-0.27	-0.44	-0.69	-0.56	-0.57	-0.49	-0.16
MultiLayerPerceptron	-0.26	-0.21	-0.08	-0.19	-0.18	-0.12	-0.05	-0.17	-0.44

Figure 13: Correlation between the overlapping coefficient and resulting multiplicity for the Bank dataset.

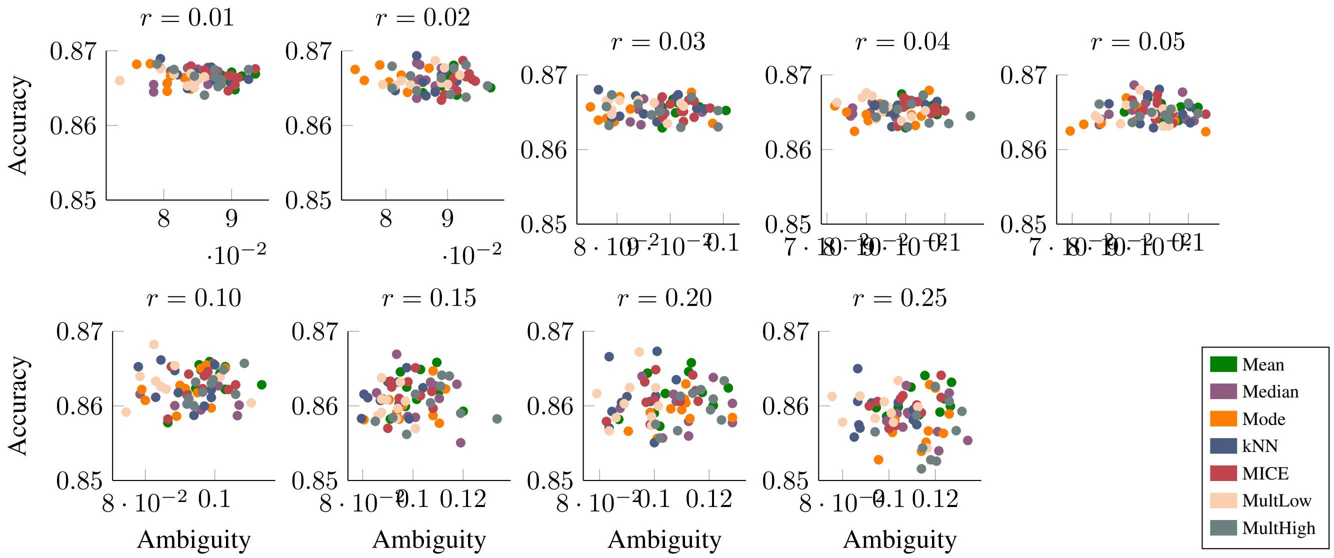


Figure 14: Accuracy and ambiguity for various data imputation strategies across varying values of missing data ratio r for Bank dataset.