Learning under Localized Minority Imbalance

Amin Hosseininasab
Warrington College of Business
University of Florida
Gainesville, FL 32611
a.hosseininasab@ufl.edu

Steven M. Shugan
Warrington College of Business
University of Florida
Gainesville, FL 32611
sms@ufl.edu

Abstract

Class-imbalance methods implicitly assume that the minority class is uniformly undersampled relative to the majority class. However, in many real-world settings, minority instances may be disproportionately under-observed in certain regions of the feature space. For example, small businesses that go bankrupt may disappear from records, while those that survive remain visible, making bankruptcy appear less common among small firms than it actually is. This gives rise to localized minority imbalance (LMI), a challenge that is often overlooked and extends beyond general class-count imbalance. We show that under LMI, existing imbalance mitigation techniques can fit observation-induced biases in the training data and generalize poorly to under-observed regions of the true minority distribution. To address this, we propose a tree-based stratified approach that recursively partitions the feature space with the goal of reducing within-stratum LMI distortion. For each resulting stratum, we pair its majority instances with the full observed minority set and train a base classifier to create an ensemble. Extensive experiments over benchmark tabular datasets simulated with LMI show that our stratified ensembling approach outperforms popular and state-of-the-art imbalance mitigation techniques. We also introduce a gold-standard evaluation protocol that uses unbiased test sets, and demonstrate that conventional hold-out evaluation from the same LMI-biased data can substantially mislead performance. Overall, our results highlight that the cause of imbalance is as important as the correction method.

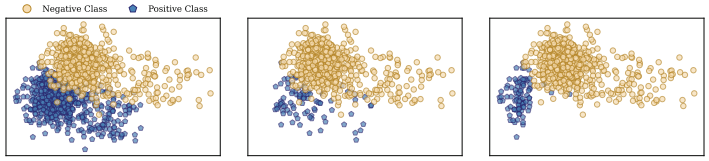
1 Introduction

Class imbalance is prevalent in real-world applications and can substantially affect supervised learning performance. When the distribution of classes is highly uneven, algorithms learn to prioritize the majority class and become poor detectors of the minority class. This is problematic, especially in social sciences and applications where the minority class is the primary focus of inference. Examples are customer churn, purchase conversion, fraud detection, and medical diagnosis.

Most existing imbalance methods inherently assume that the relative observation propensity of the two classes is uniform across the feature space. Under this condition, imbalance induces a constant shift in the observed class log-odds, which can be effectively mitigated using existing techniques. However, in practice, the relative observation propensity may vary across feature values, a condition that we term *Localized Minority Imbalance* (LMI).

Under LMI, certain regions of the minority class are disproportionately unobserved—beyond what their natural frequency in the data-generating distribution would suggest. This imbalance occurs precisely due to their class and feature values, resulting in selection bias, covariate shift, or similar paradigms. For example, the bankrupt firms we observe (minority) may not represent all bankruptcies, *since* small failed firms often vanish from records. This is unlike small non-bankrupt firms which remain visible. Similarly, a demographic with severe depression (minority) may be less likely than

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Balanced Dataset. (b) Class-count Imbalance. (c) LMI.

Figure 1: A balanced dataset with corresponding class-count and LMI scenarios. Both imbalanced datasets have the same number of instances, but positive instances with higher x-axis values are less likely observed under LMI, making them less representative of their true distribution.

other demographics to respond to questions about emotional well-being, particularly due to their level of mental health.

LMI can arise whenever the process that reveals, records, labels, or retains positive outcomes varies across feature-defined subpopulations. Examples include differential reporting, diagnosis or detection, survey response, record retention, and outcome ascertainment. Importantly, because the true posterior and the feature-dependent observation mechanism are not separately identifiable from the observed data alone (§4.1), the empirical prevalence of LMI cannot generally be estimated without external information. We therefore do not claim that LMI is universal. Rather, the framework is intended for applications in which the data-collection process provides a substantive reason to suspect that minority observability differs across feature regions.

The scarcity of minority observations can make the consequences of LMI particularly severe. Smaller samples of the population have a higher risk of underrepresenting or even excluding less common subpopulations. As demonstrated in Figure 1, this can lead to a biased representation of the minority population, and adds an additional challenge on the already difficult class-count imbalance problem. Our work exposes this overlooked bias and reframes imbalance learning through the lens of LMI.

We first formalize this seemingly overlooked type of imbalance, and show that under LMI, existing imbalance mitigation techniques can reinforce the very biases they aim to correct. In particular, they tend to overfit the majority class in regions where the minority population is less-observed, and therefore, generalize poorly to the true population.

We then show that LMI distortion can be mitigated by a careful selection and pairing of majority subsets from under-observed minority regions. Training over the entire observed dataset leads to abundant evidence for the majority class in parts of the feature space where minority instances truly occur but are less-observed. As a result, majority points distort classifiers towards predicting their class. However, by restricting the majority class to a carefully chosen subset according to LMI regions, we can reduce precisely the source of distortion that exists over the entire dataset.

Using this as motivation, we develop a novel tree-based procedure that seeks to generate strata with reduced within-stratum LMI distortion. Our Stratified Ensembling algorithm (SE) is guided by a custom objective function that uses observable parent-child prevalence changes as a surrogate for LMI distortion and recursively partitions the feature space while this objective improves. The procedure therefore seeks strata with reduced within-stratum LMI distortion, without requiring identification of the underlying observation mechanism. This then allows us to train a local classifier on each stratum, and apply existing imbalance mitigation techniques within the resulting ensemble.

While stratifying the feature space leads to lower LMI severity in each stratum, it also introduces higher variance due to a smaller and region-specific sample of data. This adversely affects the generalization of each local classifier to the full data distribution. For an effective bias-variance tradeoff, we propose to (1) combine the majority instances of each stratum with the entire set of minority instances, and (2) ensemble all classifiers for the final prediction task.

First, pairing all minority instances increases the effective sample size of the scarce class in every base classifier, leading to decision boundaries that are less sensitive to region-specific minorities. This is especially important under LMI, where local minority subsets can be rare and highly idiosyncratic. Second, restricting the majority class to each stratum preserves the intended localization—each local classifier sees a simpler, more homogeneous slice of the majority distribution—mitigating

the imbalance effect of the majority class in strata with under-observed minority instances. Lastly, ensembling across strata further lowers variance by averaging out stratum-specific estimation noise.

We compare our approach to popular and state-of-the-art class imbalance techniques over several base classifiers and benchmark datasets, and show that under LMI our approach achieves higher MCC, F1 on unbiased test sets. On the other hand, when imbalance is class-count, existing methods perform the best. SE is thus most appropriate when the data-collection process provides a substantive reason to expect feature-dependent minority observability. When a less-biased validation sample is available, SE and standard methods can be compared directly on that sample, while in its absence, SE should be interpreted as a robustness analysis rather than as a data-only diagnosis of LMI.

Lastly, we address an issue in evaluating imbalance methods under LMI: standard hold-out splits drawn from LMI datasets are biased and mislead performance assessments. Specifically, standard hold-out sampling evaluates performance primarily on the observed minority regions rather than the true target population. Therefore, evaluating imbalance mitigation techniques without considering the source of imbalance can lead to misleading conclusions, and even favor methods that overfit the selection pattern instead of learning the true data distribution.

To address this, we propose a gold-standard validation protocol that requires the test set to be LMI-free. While such balanced reference test sets may not always be readily available, they can be constructed when LMI is suspected in many realistic settings. For example, researchers can collect targeted labels, use external cohorts, follow up on delayed outcomes, or audit data-collection procedures to balance observation in LMI-suspected regions. Nonetheless, our goal in this paper is not to propose a universally adoptable evaluation standard. Rather, we seek to highlight the potential pitfalls of conventional evaluation practices when class imbalance arises from LMI, and showcase how our evaluation protocol can reveal improvements that translate to better coverage of the minority subgroups suppressed by LMI. Our contributions can be summarized into four parts:

- We formalize LMI and provide a theoretical analysis showing how partitioning the feature space into strata that approximate class-count imbalance can help mitigate its effects.
- We develop a novel SE procedure as a first step toward addressing LMI. We emphasize that our approach provides an approximate strategy for mitigating LMI, and also do not claim to improve over existing methods for general class-count imbalance.
- We empirically show that SE outperforms popular and state-of-the-art imbalance mitigation techniques, which while suited to class-count imbalance, fail to generalize under LMI.
- We show that standard hold-out sampling from LMI data misleads performance, and introduce an evaluation protocol that uses unbiased test sets for LMI correction.

2 Related Work

A broad category of imbalance techniques falls into data-level, cost-sensitive, and hybrid strategies (see Altalhan et al. (2025) for a recent review). Data-level methods include oversampling the minority class, such as SMOTE (Chawla et al., 2002) and its variants (Pradipta et al., 2021), which can mitigate imbalance but risk reinforcing minority class bias under LMI. On the other hand, undersampling methods like Cluster Centroids (Yen & Lee, 2009) may lose important structure that can aid in mitigating LMI. Cost-sensitive learning modifies the weight of instances in classification loss functions to emphasize the minority class during training, such as focal loss (Lin et al., 2017), class-balanced loss (Cui et al., 2019), and meta reweighting (Ren et al., 2018; Shu et al., 2019), but face similar limitations to oversampling under LMI.

Ensemble and hybrid methods combine multiple techniques to leverage their strengths, and have gained popularity in recent years (see Khan et al. (2024) for a review). However, most ensembles rely on random undersampling, are designed mainly to reduce variance, and do not explicitly address LMI. In contrast, our method specifically promotes diversity across the ensemble to learn informative contrasts between majority and minority instances in under-observed minority regions.

Other imbalance methods focus on establishing rigorous guarantees for metrics relevant under class imbalance (e.g., MCC, F1) (Diochnos & Trafalis, 2021; Narasimhan, 2018) or AUC-PR directives (Eban et al., 2017; Qi et al., 2021). Similarly, other work addresses performance disparities between hard and easy classes through margin regularization (Zhu et al., 2026). While these approaches

target metric optimization or class-wise generalization disparities, they do not directly addresses the feature-dependent observation bias underlying LMI. Our approach is complementary and can be applied in conjunction.

Another related stream is when the distribution of the test set is considered to *shift* from the training set. Such shifts are studied in the context of class imbalance with both covariate and label shifts (Tan et al., 2020; López et al., 2014; Yu et al., 2025), domain adaptation (Lee et al., 2024; Fu et al., 2025), and concept drift (Palli et al., 2024). These works focus on adaptation, reweighting, or tracking changes over time, often using external information and access to unlabeled target-domain data. In contrast, we assume a single underlying population but a biased observation mechanism, no prior knowledge on the underlying distribution, and no access to auxiliary unlabeled data. As the training minority class-conditional is itself distorted in our framework, shift/adaptation methods do not address the core challenge of LMI.

Similarly, distributionally robust optimization (Namkoong & Duchi, 2016; Sagawa et al., 2019) has gained attention in training robust classifiers that extend to unseen environments. While these methods also seek robustness to distributional differences, they typically optimize against worst-case shifts over an uncertainty set. In contrast, our work targets a specific LMI bias that is feature-dependent and aims to correct that structured under-observation rather than generic distribution shift.

A related literature examines long-tailed classification, where class frequencies are highly unequal. The literature can be categorized into re-balancing (Ren et al., 2020; Kang et al., 2019; Bhat et al., 2025), logit and prior correction (Menon et al., 2020), and decoupled or multi-expert training (Kang et al., 2019; Wang et al., 2020). The core goal of this stream is better tail generalization without sacrificing head performance. While these methods address predictive degradation under imbalance, they typically assume imbalance is global and tail classes are *uniformly* underrepresented across the feature space (that is, class-count imbalance). Therefore, these methods mainly correct frequency-driven optimization and calibration, not the LMI distortion that we study. Our work is thus complementary, and particularly relevant when long tails are LMI.

3 Problem Formalization

We consider a supervised classification task over an *observed* dataset $\mathcal{D}^{\text{obs}} = \{(x_i, y_i)\}_{i=1}^{N}$. Features $x_i = \left[x_i^1, \ldots, x_i^d\right]$ are defined over a d-dimensional feature space $X \in \mathcal{X} = \mathbb{R}^d$, and labels are binary $y_i \in \{0, 1\}$ (our method can be extended to multiclass via one-vs-rest). The observed dataset consists of n^- observations from the $Y = 0$ negative or *majority* class, and n^+ observations from the $Y = 1$ positive or *minority* class. In many real-world applications, $n^+ \ll n^-$, creating class imbalance.

The observed dataset $\mathcal{D}^{\text{obs}}$ can be viewed as a finite sample from its true (generally unknown) underlying distribution $\mathcal{P}^{\text{tru}} := p^{\text{tru}}(x, y)$, collected through an *observation mechanism* governed by O. In particular, let $O_i \in \{0, 1\}$ indicate whether a population instance $(X_i, Y_i) \sim \mathcal{P}^{\text{tru}}$ is observed (and part of $\mathcal{D}^{\text{obs}}$), and define the class-conditional *observation propensity*

$$\pi_y(x) = P(O = 1 \mid X = x, Y = y), \quad y \in \{0, 1\}.$$

The resulting *observed distribution* of $\mathcal{D}^{\text{obs}}$ is $\mathcal{P}^{\text{obs}} := p^{\text{obs}}(x, y) = P(X = x, Y = y \mid O = 1)$ which satisfies $p^{\text{obs}}(x, y) \propto p^{\text{tru}}(x, y)\pi_y(x)$.

Class-count imbalance occurs when $p^{\text{obs}}(x, 1) \ll p^{\text{obs}}(x, 0)$, with $\pi_0(x) = c \cdot \pi_1(x)$ for some constant c. In other words, the probability of observing any instance is *uniformly* (by a factor of c) higher for class $Y = 0$ than $Y = 1$. The major challenge in class-count imbalance is thus the uniform shift of classification probabilities caused by observing an uneven number of instances across classes.

In contrast, LMI occurs when $p^{\text{obs}}(x, 1) \ll p^{\text{obs}}(x, 0)$ and there does not exist any constant $c : \pi_0(x) = c \cdot \pi_1(x), \forall x \in X$. Under LMI, minority instances from particular regions of $\mathcal{X}$ are systematically less-observed than other regions, with $\pi_1(x) \neq \pi_1(x')$ for some $x, x' \in X$. The observed dataset $\mathcal{D}^{\text{obs}}$ thus exhibits both a marginal imbalance $P(Y = 1 \mid O = 1)$ and a feature-dependently distorted minority distribution $p^{\text{obs}}(x \mid y = 1)$ relative to $p^{\text{tru}}(x \mid y = 1)$. Importantly, this distortion is distinct from natural heterogeneity in $\mathcal{P}^{\text{tru}}$, as minority instances may be naturally rare in particular regions with small $p^{\text{tru}}(x \mid y = 1)$. However, LMI refers to an additional class-and-feature-dependent suppression through the observation mechanism $\pi_y(x)$ that affects minority representation in $\mathcal{D}^{\text{obs}}$ despite their natural probability of occurrence $P(X = x \mid Y = 1)$.

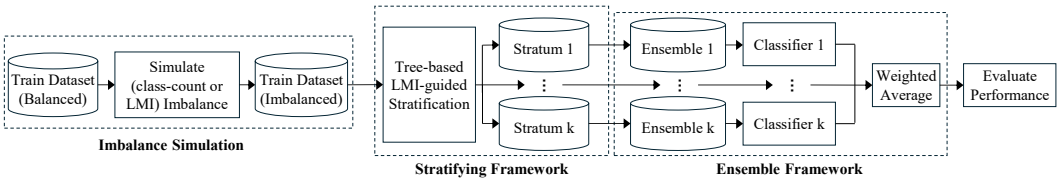


Figure 2: Overall stratified ensembling (SE) framework and evaluation protocol.

Under LMI, learning from $\mathcal{D}^{\text{obs}}$ faces two coupled challenges: the minority class is rare with $p^{\text{obs}}(x,1) \ll p^{\text{obs}}(x,0)$, *and* the observed minority is a biased subset of its true distribution $\mathcal{P}^{\text{tru}}$. Therefore, correcting only class-count imbalance—the main objective of most imbalance mitigation methods—may overfit to $\mathcal{P}^{\text{obs}}$ and generalize poorly to $\mathcal{P}^{\text{tru}}$. This presents a fundamental flaw in imbalanced learning for real-life settings affected by LMI, potentially generating misleading conclusions over the relationship between X and outcomes Y.

4 Proposed Methodology

We give the theoretical groundwork of our proposed Stratified Ensembling (SE) algorithm in §4.1. SE exhibits a nuanced bias-variance tradeoff which we discuss and address in §4.2. Building on these theoretical results, we develop SE in §4.3, and establish our proposed evaluation protocol in §4.4. The entire proposed framework for SE is depicted in Figure 2.

4.1 Theoretical Motivation

The central challenge under LMI is not only that minority instances are scarce, but that the observed relationship between class Y and features X in $\mathcal{D}^{\text{obs}}$ can be systematically distorted by the observation mechanism O. As such, models trained on $\mathcal{D}^{\text{obs}}$ may learn patterns that are specific to the observed distribution $\mathcal{P}^{\text{obs}}$ rather than the true underlying distribution $\mathcal{P}^{\text{tru}}$.

To address this, we first characterize LMI distortion in logit form, as it cleanly separates natural prevalence densities $p^{\text{tru}}(x,y)$ from observation effects O. Lemma 1 shows how class-conditional observation propensities $\pi_y(x)$ distort the posterior $P(Y=1 \mid X=x)$ into its observed counterpart $P(Y=1 \mid X=x, O=1)$ on the logit (lgt) scale.

Lemma 1. *For any feature value x, the posterior class probability under the observed distribution $\mathcal{P}^{\text{obs}}$ satisfies* $\text{lgt}\left(P\left(Y=1 \mid X=x, O=1\right)\right) = \text{lgt}\left(P\left(Y=1 \mid X=x\right)\right) + \gamma(x)$, $\gamma(x) = \log\left(\frac{\pi_1(x)}{\pi_0(x)}\right)$.

By Lemma 1, the observed odds are tilted by the ratio of observation propensities

$$\frac{P(Y=1 \mid X=x, O=1)}{P(Y=0 \mid X=x, O=1)} = \frac{\pi_1(x)}{\pi_0(x)} \cdot \frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)}.$$

Moreover, the observed class-conditional feature distribution is obtained by reweighting the true underlying distribution by the observation propensity

$$p^{\text{obs}}(x \mid Y=y) \propto p^{\text{tru}}(x \mid Y=y)\pi_y(x), y \in \{0,1\}.$$

When $\pi_0(x) = c \cdot \pi_1(x)$—as in class-count imbalance—the selection term collapses into the constant c, and the observed posterior is uniformly shifted compared to the true population posterior. As such, class-count imbalance induces a constant shift in posterior log-odds, which can be addressed, for example, by reweighting the training sample. Conversely, under LMI, the feature-dependent shift $\gamma(x)$ distorts learning non-uniformly across different regions of the feature space.

The ideal approach to address LMI is to directly learn $\gamma(x)$ and mitigate it. However, as established in Lemma 2, these shifts cannot be separated from true posterior $P(Y=1 \mid X=x)$ without using prior knowledge (which we do not assume access to).

Lemma 2. *Given only $\mathcal{D}^{\text{obs}}$, $P(Y=1 \mid X=x)$ and $\gamma(x)$ cannot be separately identified.*

Despite this restriction, we propose to stratify the feature space $\mathcal{X}$ into strata $\mathcal{S} = \{S_1, \ldots, S_{|\mathcal{S}|}\}$ where $\gamma(x)$ is *approximately* constant within any stratum S_k. As each stratum $S_k \subseteq \mathcal{X}$ is a relatively

LMI-unbiased sample of its region and closer to class-count imbalance, it allows us to effectively address its imbalance via existing mitigation techniques. Theorem 3 establishes the best-case scenario in which such strata exactly address LMI distortion.

Theorem 3. *Assume there exist strata $\mathcal{S}$ such that for any stratum $S_k \in \mathcal{S}$, $\gamma^k(x) = c^k$ for some stratum-specific constant c^k. Then for all $x \in S_k$, the Bayes-optimal ranking of instances within S_k is identical under $\mathcal{P}^{\text{obs}}$ and $\mathcal{P}^{\text{tru}}$.*

Theorem 3 shows that it is possible to convert the complex, feature-dependent distortion caused by LMI into a simpler stratum-dependent offset that is easier to accommodate for standard learners and existing imbalance mitigation techniques. However, the assumption of Theorem 3 is unlikely to be achieved in practice without prior knowledge on the observation mechanism O. In such cases, Proposition 4 shows that even if $\gamma^k(x)$ is only approximately constant within a stratum S^k, the Bayes-optimal rankings are preserved for pairs of instances that are not near-ties in their true posterior log-odds.

Proposition 4. *Suppose that for a stratum $S_k \in \mathcal{S}$, we have $\gamma^k(x) = c^k + \delta^k(x)$, where $\delta^k(x)$ is the within-stratum deviation from the stratum-specific distortion constant c^k. The Bayes-optimal ranking of instances within S_k is preserved for any pair of instances whose true posterior log-odds differ by more than $2\max_{x \in S^k} |\delta^k(x)|$.*

Proposition 4 shows that the exact condition of Theorem 3 is not necessary for stratification to be useful in mitigating LMI. Accordingly, SE does not require or assume that its learned strata have exactly constant LMI distortion. The goal is instead to reduce within-stratum variation in $\gamma(x)$ sufficiently to limit LMI-induced ranking errors. The important implication is the magnitude of the remaining within-stratum deviation $\delta^k(x)$. When deviation from the stratum-specific distortion constant c^k is small, the observed distribution can alter only the ranking of instances whose true posterior log-odds are already close. Thus, reducing the within-stratum variation of $\gamma^k(x)$ limits the set of pairs for which LMI can induce ranking errors. This provides the theoretical motivation for our stratifying procedure: the goal is not to recover perfectly unbiased strata, but to construct strata in which the residual LMI distortion is sufficiently small.

4.2 Bias-Variance Tradeoff in Stratification

Although stratification can reduce LMI-induced distortion, it also increases variance as each stratum contains fewer instances. This brings a nuanced bias-variance tradeoff into the procedure, which we address in two ways.

First, instead of restricting each stratum S_k to its few minority instances $(x_i \mid y_i = 1) \in S_k$, we combine all observed minority instances $(x_i \mid y_i = 1) \in \mathcal{D}^{\text{obs}}$ with the majority instances $(x_i \mid y_i = 0) \in S_k$ of that stratum. Using all minority points in each stratum lowers variance for the minority class, while still preserving the core benefit of stratifying. In particular, the majority instances paired with minority points are drawn from the specific region of a stratum S_k where the LMI distortion for $Y = 1$ is approximately constant. Therefore, the decision boundary is less distorted by majority points of regions with a different, and potentially much more severe, LMI.

Importantly, the primary benefit of stratifying under LMI bias is not due to minority points from other strata, and rather the disproportionate influence of majority observations in regions where minority instances are under-observed. Such majority points appear as abundant evidence for class $Y = 0$ in parts of the space where $Y = 1$ actually occurs but is selectively unobserved. Therefore, including such majority points systematically distorts the classifier toward predicting $\hat{y}_i = 0$. By restricting the majority class to S_k, we exclude precisely these harmful majority instances from mismatched strata, while the pooled minority set mainly acts to stabilize estimation of the decision boundary. As we later show in our ablation studies in Appendix C.5, classifiers trained on such strata have better predictive performance compared to classifiers trained only on minority points within their associated stratum.

As our second approach for an effective bias-variance tradeoff, we ensemble the base classifier f trained on strata $\mathcal{S} = \{S_1, \ldots, S_k\}$, and use a performance-weighted average of probabilities

$$\hat{P}(x) = \frac{\sum_{S_k \in \mathcal{S}} \alpha_k P_{f(S_k)}(x)}{\sum_{S_k \in \mathcal{S}} \alpha_k}$$

over the entire training set, and not just stratum-specific instances. Here, α_k is an internal validation-based reliability weight, computed using a metric such as MCC, F1, or AUC. These weights are used only to reduce the influence of unstable base learners arising from differences in stratum size or local separability. They should not be, however, interpreted as unbiased estimates of target-population performance under P^{tru}. Values $P_{f(S_k)}(x)$ are classification probabilities of the base classifier f trained on stratum S_k. Ensembling is well-known to reduce variance, and hence, the resulting predictions are expected to be more robust than any single classifier $f(S_k)$.

4.3 LMI-guided Stratified Ensembling Algorithm

Our SE algorithm is an approximate procedure motivated by Proposition 4, that aims to create strata with reduced LMI. To this end, we develop a custom objective function that minimizes an approximate measure of within-node LMI. Using this objective function, our stratifying tree initializes with all instances in the root node and recursively partitions the feature space $\mathcal{X}$ along features $X^j \in \mathcal{X}$. The procedure terminates when further splitting a parent node is not estimated to reduce LMI distortion in its children. The resulting leaves define strata $\mathcal{S} = \{S_1, \ldots, S_{|\mathcal{S}|}\}$.

As our first step, we generate a one-dimensional ranking score $\eta(x)$ that orders minority probabilities according to $P(Y = 1|X = x, O = 1)$. The only goal of scores $\eta(x)$ is to create inexpensive and statistically stable partitions of instances $(x_i, y_i) \in \mathcal{D}^{\text{obs}}$ that can be used to compare observed minority prevalence across comparable regions of the feature space. Without reducing $\mathcal{X} \in \mathbb{R}^d$ to a single ordering $\eta(x)$, partitioning the feature space becomes expensive and typically unstable in high dimensions. Given this role, the ranking scores do not require calibration, do not directly determine the feature-space partitions in SE, and are not used by the final base classifiers. In our primary implementation, we use an inexpensive closed-form difference-in-means score for $\eta(x)$, while they can also be set, for example, as the predicted sample confidence scores of any auxiliary classifier fit on $\mathcal{D}^{\text{obs}}$. As we show in Appendix C.5, SE is robust to substantially different choices for $\eta(x)$, including an uninformative random ordering.

Next, at any node m, we define $\mathcal{B}_m = \{b_1, \ldots, b_{|\mathcal{B}_m|}\}$ as quantile bins of $\eta(x)$. For each bin b, the observed within-bin minority probability is given by

$$q_m(b) \coloneqq P(Y = 1 \mid X \in \mathcal{X}_m, \eta(X) \in b, O = 1).$$

These probabilities can be approximated, for example, via a simple fraction of minority counts $n_b^+/(n_b^+ + n_b^-)$ within each bin, similar to general decision tree classifiers. The heterogeneity of the observed minority prevalence at node m can then be measured via a weighted variance of $q_m(b)$ on the logit scale:

$$\sigma^2_{\mathcal{B}_m}(\text{lgt}(q_m(b))) = \frac{\sum\limits_{b \in \mathcal{B}_m} |b| \left(\text{lgt}\left(q_m(b)\right) - \frac{\sum\limits_{b' \in \mathcal{B}_m} |b'| \text{lgt}\left(q_m(b')\right)}{\sum\limits_{b' \in \mathcal{B}_m} |b'|} \right)^2}{\sum\limits_{b \in \mathcal{B}_m} |b|}.$$

Larger values of $\sigma^2_{\mathcal{B}_m}(q)$ indicate higher variation of observed minority prevalence across ranking regions. However, as previously discussed, this can be due to feature-dependent LMI distortion caused by $\gamma(x)$, or natural heterogeneity of the minority distribution $p^{\text{tru}}(x \mid y = 1)$ at that region. While it is not possible to exactly separate these components without prior knowledge on the observation mechanism O, we propose a compromise that takes advantage of the additive log-odds structure established in Lemma 1.

Instead of directly using values $\sigma^2_{\mathcal{B}_m}(q)$ to measure LMI, we use a parent-child relationship of our stratifying tree to approximately isolate LMI distortion from natural minority prevalence. In particular, we first enforce the child and parent nodes of our stratifying tree to use the same binning partitions. That is, after splitting a parent node m into left and right child nodes v_l, v_r, each child inherits the parent bins corresponding to its specific region. This shared-binning rule ensures that any parent-child comparison is performed over similar neighborhoods, rather than on differences induced by changing bin definitions.

For any parent node m, child node v, and shared bin $b \in \mathcal{B}_m \cap \mathcal{B}_v = \mathcal{B}_v$, define the residual

$$\Delta_{m \to v}(b) = \text{lgt}\left(q_m(b)\right) - \text{lgt}\left(q_v(b)\right).$$

Residuals $\Delta_{m\to v}(b)$ can be theoretically decomposed into a parent-child drift of natural prevalence and LMI distortion as given in Theorem 5.

Theorem 5. *Residuals $\Delta_{m\to v}(b)$ are composed of a sum of natural-prevalence drift $\mathrm{lgt}\left(P(Y=1 \mid x\in\mathcal{X}_m, \eta(x)\in b)\right) - \mathrm{lgt}\left(P(Y=1 \mid x\in\mathcal{X}_v, \eta(x)\in b)\right)$ and LMI-distortion drift $\gamma_m(b) - \gamma_v(b)$.*

Given this decomposition, Corollary 6 shows that when parent and child nodes have similar natural heterogeneity on the logit scale and within a bin $b \in \mathcal{B}_v$, residuals $\Delta_{m\to v}(b)$ primarily reflect differences in LMI-distortion.

Corollary 6. *If the within-bin probabilities $q_m(b)$ of a bin $b \in \mathcal{B}_v$ shared among parent m and child v do not substantially change the logit of the true posterior $\mathrm{lgt}\left(P(Y=1 \mid x\in\mathcal{X}_m, \eta(x)\in b)\right)$ then residuals $\Delta_{m\to v}(b)$ primarily reflect LMI-distortion drift.*

As $\gamma(x)$ is not identifiable from $\mathcal{D}^{\text{obs}}$, this decomposition does not allow us to certify that a learned stratum has approximately constant LMI distortion. Rather, Corollary 6 provides a conditional link between the observable residual and the target distortion. When natural-prevalence drift within shared bins is small, variation in the residual primarily reflects variation in LMI distortion. Our SE algorithm thus uses this residual as an observed-data surrogate rather than as an exact estimator of $\gamma(x)$.

Specifically, we use $\Delta_{m\to v}(b)$ to define a node-level objective function that favors partitions with more homogeneous parent-child prevalence changes across shared bins, which may help identify regions with reduced LMI distortion under the conditions discussed above. At a node m, a candidate split is specified by a feature X^j and a threshold among the ones defined by bin set $\mathcal{B}_m$. For the candidate split (feature and threshold) we compute the weighted score

$$\frac{|v_l|}{|m|}\,\sigma^2_{\mathcal{B}_{v_l}}(\Delta_{m\to v_l}) + \frac{|v_r|}{|m|}\,\sigma^2_{\mathcal{B}_{v_r}}(\Delta_{m\to v_r}),$$

and choose the split that minimizes it (equivalently maximizing the gain from the parent), subject to a stopping criterion based on the parent node's observed heterogeneity. The recursion terminates when no feasible split yields a gain.

In the case that Corollary 6 is not entirely satisfied, the stratifying algorithm may select a feature that reduces natural-prevalence drift more than LMI-distortion drift. Given our bias-variance tradeoff approach in §4.2, such a split corresponds to the general undersampling of the majority class in the resulting children. While such splits may not reduce LMI-distortion, they also do not deteriorate the condition of Proposition 4, as shown in Proposition 7.

Proposition 7. *For any pair $x, x' \in v$, the ranking-stability condition in Proposition 4 is weakly no harder to satisfy in the child v than in the parent m.*

In practice, over-partitioning is further constrained as a split is accepted only when it improves the SE objective and satisfies the minimum leaf-size requirements. Moreover, each resulting classifier reuses the full observed minority set while ensembling reduces the variance introduced by partitioning, as discussed in §4.2. Appendix C.5 further shows that performance remains stable when the minimum leaf sizes and score-bin count are varied.

4.4 Gold-Standard Evaluation Protocol

An important challenge in LMI class imbalance is that sampling test sets from datasets that are potentially LMI-imbalanced biases true model performance. To highlight and address this potential bias, we propose to measure imbalance performance using gold-standard datasets that are originally near-balanced and unbiased, and simulate LMI or class-count imbalance on those datasets. That is, we first partition the dataset into training $\mathcal{D}^{\text{train}}$ and test $\mathcal{D}^{\text{test}}$ sets using stratified folds. We then simulate imbalance via controlled instance deletions—based on features X and class $Y=1$ for LMI, and randomly for class-count imbalance. The simulation algorithms are given in Appendix B.3.

All methods are applied to the resulting imbalanced training set, while performance is evaluated on the gold-standard test set. This design ensures that comparisons across methods reflect their ability to recover the predictive power over the true underlying distribution $\mathcal{P}^{\text{tru}}$, rather than the biased $\mathcal{P}^{\text{obs}}$.

Our simulation-based evaluation is necessary as true-LMI distorted datasets are difficult to obtain, particularly as benchmarks for imbalance learning. Therefore, we emphasize that our protocol is not

meant to replace current evaluation practices. Rather, we seek to highlight the risks of benchmarking methods solely on test sets drawn from potentially LMI datasets. In practice, gold-standard test sets can be specifically generated by controlling the data collection procedure, for example, via targeted labeling, external cohorts, delayed-outcome follow-ups, or auditing data collection.

5 Numerical Evaluation

Our experiments evaluate (1) the performance of our approach versus benchmarks under LMI; and (2) how standard hold-out evaluation can yield misleading results when applied to LMI data. We provide further numerical results in Appendix C, including time and memory usage, varying degrees of LMI, class-specific and additional evaluation metrics, detailed non-averaged results, ablation and hyperparameter sensitivity studies, alternative LMI mechanisms, and performance under general class-count imbalance.

5.1 Evaluation Framework

We construct 5-fold cross-validation hold-out samples taken from benchmark datasets of the KEEL repository (Alcalá-Fdez et al., 2011), detailed in Appendix B.1. Then, on training data $\mathcal{D}^{\text{train}}$, we simulate 95% deletion or equivalently a 1:20 IR (corresponding to mildly severe imbalance) with either class-count or LMI simulated according to the procedure detailed in Appendix B.3. Following prior work (Rezvani & Wang, 2023; Altalhan et al., 2025), we report MCC, and F1 as our main criteria, but also report other metrics in Appendix C.3.

We evaluate on five base classifiers and a host of popular and state-of-the-art imbalance methods detailed in Appendix B.2. All algorithms and classifiers are tuned via a grid search and 5-fold cross-validation performed on the training set. We also report two reference empirical baselines. The gold standard baseline (GLD): performance on the original balanced datasets, and the damaged baseline (DMG): performance on imbalanced datasets without any imbalance correction technique. These baselines contextualize the gains achieved by different methods.

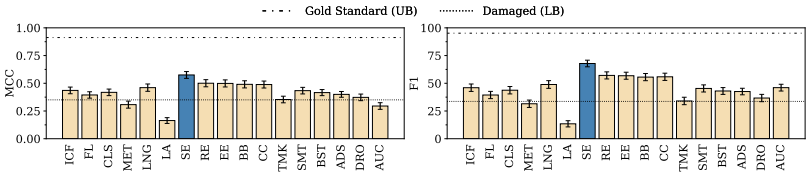
All algorithms are coded in Python, occasionally with a C++ backend (all code is provided in the supplementary materials) and executed on a PC with an Intel Xeon processor W-2255, 264GB of RAM, and Ubuntu 20.04.5 operating system. The SE code is open source.

5.2 Performance under LMI

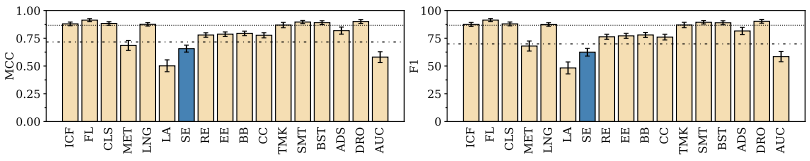
Figure 3 gives the MCC and F1 under LMI. All classifiers show notable performance drops when trained on damaged data (DMG lower bound in Figure 3a), confirming the adverse effect of LMI. SE achieves the strongest overall performance on the undamaged test population, driven particularly by improved generalization to high-LMI regions (Figure 3c). On the other hand, other methods overfit the majority class in under-observed regions of the feature space, giving misleadingly high performance in low-LMI regions (Figure 3b) but considerably low performance on high-LMI regions.

The misleading advantage of competing methods is itself evidence of overfitting to the biased LMI observation pattern. In particular, DMG—trained directly on the damaged data without any imbalance correction—substantially exceeds GLD in these regions, even though GLD is trained on the original unbiased data. At the same time, DMG performs considerably worse on both the full population and in high-LMI regions. Therefore, its higher score in Figure 3b does not reflect a superior target-population decision rule. Rather, it reflects overfitting to the regions overrepresented by the LMI observation mechanism. By contrast, SE reduces overfitting and remains close to GLD in low-LMI regions, substantially improves performance in high-LMI, and achieves the highest overall MCC and F1 on the full undamaged test data.

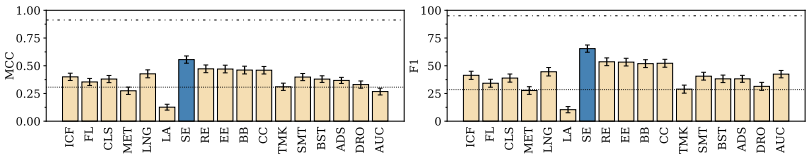
Lastly, SE also substantially outperforms the other ensembling baseline methods (RE, EE, BB), suggesting that its feature-guided stratification strategy provides benefits beyond those obtained by conventional random undersampling and ensembling.



(a) Overall performance on unbiased hold-out sets



(b) Performance on low-LMI regions



(c) Performance on high-LMI regions

Figure 3: Average performance under LMI across datasets and base classifiers.

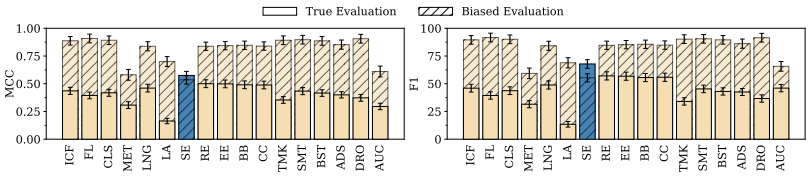


Figure 4: True versus biased evaluation by conventional sampling of LMI-biased hold-out sets

5.3 Biased Evaluation by Sampling LMI Data

Standard evaluation protocols will mislead model performance if they sample the test sets from the same LMI-effected data. To reveal this flaw, we simulate 85% LMI (to retain sufficient instances for subsequent hold-out samples), and evaluate models on hold-out sets drawn solely from the damaged dataset (mimicking conventional practice) to end at 95% LMI. This setup mirrors real-world conditions where the test data shares the same structural bias as training, but it fails to reflect the true minority population under LMI.

Figure 4 reports performance average results. Under LMI, evaluating on data drawn from the same biased distribution used for training can overstate performance and obscure generalization. In particular, this biased evaluation misleadingly favors models that overfit to the observed minority subset while penalizing those that generalize to the entire population. These results reinforce the importance of accounting for the nature of imbalance when evaluating mitigation methods.

6 Conclusion

This paper studies LMI class imbalance, where minority observability depends jointly on class labels and predictive features. Under LMI, popular imbalance methods fail to generalize and instead amplify biases in the observed data. To address this, we develop a stratifying procedure that uses an observed-data surrogate to seek subsets with reduced within-stratum LMI distortion. These strata are then used to create an ensemble of classifiers that are trained on the entire set of observed minority instances plus the majority instances of a generated stratum. Our method outperforms popular and state-of-the-art techniques under LMI on multiple criteria. We also introduce an evaluation protocol that reveals how standard validation practices can misrepresent performance under LMI.

References

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., and Herrera, F. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17(2–3):255–287, 2011.

Altalhan, M., Algarni, A., and Alouane, M. T.-H. Imbalanced data problem in machine learning: A review. *IEEE Access*, 2025.

Bhat, S. D., More, A., Soni, M., and Agrawal, S. Prior2posterior: Model prior correction for long-tailed learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1289–1298. IEEE, 2025.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Diochnos, D. I. and Trafalis, T. B. Learning reliable rules under class imbalance. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 28–36. SIAM, 2021.

Eban, E., Schain, M., Mackey, A., Gordon, A., Rifkin, R., and Elidan, G. Scalable learning of non-decomposable objectives. In *Artificial intelligence and statistics*, pp. 832–840. PMLR, 2017.

Fu, Y., Huang, S., Feng, Z., and Ma, Y. Rdam: Domain adaptation under small and class-imbalanced samples. *Knowledge-Based Systems*, pp. 114909, 2025.

Han, H., Wang, W.-Y., and Mao, B.-H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pp. 878–887. Springer, 2005.

He, H., Bai, Y., Garcia, E. A., and Li, S. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 1322–1328. Ieee, 2008.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Khan, A. A., Chaudhari, O., and Chandra, R. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications*, 244:122778, 2024.

Lee, J., Ko, J. U., Kim, T., Kim, Y. C., Jung, J. H., and Youn, B. D. Domain adaptation with label-aligned sampling (dalas) for cross-domain fault diagnosis of rotating machinery under class imbalance. *Expert Systems with Applications*, 243:122910, 2024.

Lemaitre, G., Nogueira, F., and Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of machine learning research*, 18(17):1–5, 2017.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Liu, X.-Y., Wu, J., and Zhou, Z.-H. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008.

López, V., Fernández, A., and Herrera, F. On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed. *Information Sciences*, 257:1–13, 2014.

McDermott, M. B., Zhang, H., Hansen, L. H., Angelotti, G., and Gallifant, J. A closer look at auroc and auprc under class imbalance. *Advances in Neural Information Processing Systems*, 37: 44102–44163, 2024.

Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., and Kumar, S. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

Narasimhan, H. Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 1646–1654. PMLR, 2018.

Palli, A. S., Jaafar, J., Gilal, A. R., Alsughayyir, A., Gomes, H. M., Alshanqiti, A., and Omar, M. Online machine learning from non-stationary data streams in the presence of concept drift and class imbalance: A systematic review. *Journal of Information and Communication Technology*, 23 (1):105–139, 2024.

Pradipta, G. A., Wardoyo, R., Musdholifah, A., Sanjaya, I. N. H., and Ismail, M. Smote for handling imbalanced data problem: A review. In *2021 sixth international conference on informatics and computing (ICIC)*, pp. 1–8. IEEE, 2021.

Qi, Q., Luo, Y., Xu, Z., Ji, S., and Yang, T. Stochastic optimization of areas under precision-recall curves with provable convergence. *Advances in neural information processing systems*, 34: 1752–1765, 2021.

Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020.

Ren, M., Zeng, W., Yang, B., and Urtasun, R. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pp. 4334–4343. PMLR, 2018.

Rezvani, S. and Wang, X. A broad review on class imbalance learning techniques. *Applied Soft Computing*, 143:110415, 2023.

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

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., and Meng, D. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32, 2019.

Tan, S., Peng, X., and Saenko, K. Class-imbalanced domain adaptation: An empirical odyssey. In *European Conference on Computer Vision*, pp. 585–602. Springer, 2020.

Wang, X., Lian, L., Miao, Z., Liu, Z., and Yu, S. X. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020.

Yen, S.-J. and Lee, Y.-S. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3):5718–5727, 2009.

Yu, T., Wang, J., Wang, H., Lin, M., Liu, H., Yee, N. S., and Ma, F. Towards collaborative fairness in federated learning under imbalanced covariate shift. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 3645–3656, 2025.

Zhu, B., Zhao, K., Cui, J., Sun, Q., Zhou, Y., Yang, X., and Zhang, H. Reducing class-wise performance disparity via margin regularization. *arXiv preprint arXiv:2602.00205*, 2026.

A Proofs

Proof of Lemma 1

By Bayes' rule, for $y \in \{0, 1\}$ we have

$$P(Y = y \mid X = x, O = 1) = \frac{P(O = 1 \mid X = x, Y = y)P(Y = y \mid X = x)}{P(O = 1 \mid X = x)}.$$

Taking the odds ratio under the observed distribution yields

$$\frac{P(Y = 1 \mid X = x, O = 1)}{P(Y = 0 \mid X = x, O = 1)} = \frac{P(O = 1 \mid X = x, Y = 1)P(Y = 1 \mid X = x)}{P(O = 1 \mid X = x, Y = 0)P(Y = 0 \mid X = x)}$$

$$= \frac{\pi_1(x)}{\pi_0(x)} \cdot \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)}.$$

Applying lgt to both sides gives

$$\mathrm{lgt}\left(\frac{P(Y = 1 \mid X = x, O = 1)}{P(Y = 0 \mid X = x, O = 1)}\right) = \mathrm{lgt}\left(\frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)}\right) + \mathrm{lgt}\left(\frac{\pi_1(x)}{\pi_0(x)}\right).$$

Given $\mathrm{lgt}(p) = \log\left(\frac{p}{1-p}\right)$, and by defining $\gamma(x) \coloneqq \log\left(\frac{\pi_1(x)}{\pi_0(x)}\right)$ proves the claim.

Proof of Lemma 2

For any function $h(x)$, we can write $\ell^{\text{obs}}(x) = \ell(x) + \gamma(x) = (\ell(x) + h(x)) + (\gamma(x) - h(x))$. Thus the same observed posterior log-odds are consistent with infinitely many decompositions into a natural log-odds component and an observation-distortion component. Therefore, $\gamma(x)$ cannot be recovered from $\ell^{\text{obs}}(x)$ alone without additional structure.

Proof of Theorem 3

Fix any stratum $S_k \in \mathcal{S}$, and let

$$\beta^{\text{tru}}(x) \coloneqq P(Y = 1 \mid X = x), \quad \beta^{\text{obs}}(x) \coloneqq P(Y = 1 \mid X = x, O = 1),$$

denote the true and observed posteriors, respectively. By Lemma 1, for any $x \in S_k$ we have,

$$\mathrm{lgt}\big(\beta^{\text{obs}}(x)\big) = \mathrm{lgt}\big(\beta^{\text{tru}}(x)\big) + \gamma(x), \qquad \gamma(x) = \log\left(\frac{\pi_1(x)}{\pi_0(x)}\right).$$

By the statement, $\forall x \in S_k$ we have $\gamma(x) = c^k$, and hence

$$\beta^{\text{obs}}(x) = \mathrm{lgt}^{-1}\left(\mathrm{lgt}\big(\beta^{\text{tru}}(x)\big) + c^k\right),$$

where $\mathrm{lgt}^{-1}(t) = 1/(1 + e^{-t})$ is the inverse-logit function.

Since both $\mathrm{lgt}(\cdot)$ and $\mathrm{lgt}^{-1}(\cdot)$ are strictly increasing, the mapping

$$f_k(u) \coloneqq \mathrm{lgt}^{-1}\left(\mathrm{lgt}(u) + c^k\right), \qquad u \in (0, 1),$$

is strictly increasing in u. Therefore, for any two points $x, x' \in S_k$ that

$$\beta^{\text{tru}}(x) > \beta^{\text{tru}}(x') \Longrightarrow f_k\big(\beta^{\text{tru}}(x)\big) > f_k\big(\beta^{\text{tru}}(x')\big) \Longrightarrow \beta^{\text{obs}}(x) > \beta^{\text{obs}}(x'),$$

Since the Bayes-optimal ranking under either distribution is obtained by sorting instances by their posterior probabilities, the Bayes-optimal ranking within S_k is identical under $\mathcal{P}^{\text{obs}}$ and $\mathcal{P}^{\text{tru}}$.

Proof of Proposition 4

Fix any stratum $S_k \in \mathcal{S}$, and let

$$\beta^{\text{tru}}(x) \coloneqq P(Y=1 \mid X=x), \quad \beta^{\text{obs}}(x) \coloneqq P(Y=1 \mid X=x, O=1),$$

By Lemma 1 and the statement we have

$$\operatorname{lgt}\left(\beta^{\text{obs}}(x)\right) = \operatorname{lgt}\left(\beta^{\text{tru}}(x)\right) + c^k + \delta^k(x).$$

For any $x, x' \in S_k$ we thus have

$$\operatorname{lgt}\left(\beta^{\text{obs}}(x)\right) - \operatorname{lgt}\left(\beta^{\text{obs}}(x')\right) = \operatorname{lgt}\left(\beta^{\text{tru}}(x)\right) - \operatorname{lgt}\left(\beta^{\text{tru}}(x')\right) + \delta^k(x) - \delta^k(x').$$

Since $|\delta^k(x)| \leq \max_{x \in S^k} |\delta^k(x)|$ and $|\delta^k(x')| \leq \max_{x' \in S^k} |\delta^k(x')|$, we have $|\delta^k(x) - \delta^k(x')| \leq 2\max_{x \in S^k} |\delta^k(x)|$. Therefore, if

$$\left|\operatorname{lgt}\left(\beta^{\text{tru}}(x)\right) - \operatorname{lgt}\left(\beta^{\text{tru}}(x')\right)\right| > 2\max_{x \in S^k} |\delta^k(x)|,$$

the residual term cannot change the sign of the true log-odds difference, and hence, the observed and true rankings agree.

Proof of Theorem 5

By definition, we have

$$q_m(b) = P(Y=1 \mid X \in \mathcal{X}_m,\ \eta(X) \in b,\ O=1),$$
$$q_v(b) = P(Y=1 \mid X \in \mathcal{X}_v,\ \eta(X) \in b,\ O=1).$$

By applying Lemma 1,

$$\operatorname{lgt}\big(q_m(b)\big) = \operatorname{lgt}\big(P(Y=1 \mid X \in \mathcal{X}_m,\ \eta(X) \in b)\big) + \gamma^m(b),$$
$$\operatorname{lgt}\big(q_v(b)\big) = \operatorname{lgt}\big(P(Y=1 \mid X \in \mathcal{X}_v,\ \eta(X) \in b)\big) + \gamma^v(b).$$

Subtracting the terms gives

$$\Delta_{m \to v}(b) = \Big[\operatorname{lgt}\big(P(Y=1 \mid X \in \mathcal{X}_m,\ \eta(X) \in b)\big) - \operatorname{lgt}\big(P(Y=1 \mid X \in \mathcal{X}_v,\ \eta(X) \in b)\big)\Big] + \big[\gamma^m(b) - \gamma^v(b)\big].$$

The first term corresponds to natural-prevalence drift, and the second term to LMI-distortion drift.

Proof of Corollary 6

By the assumption, for each shared bin b the true log-odds are approximately invariant when conditioning from $\mathcal{X}_m$ to $\mathcal{X}_v$, i.e.,

$$\operatorname{lgt}\big(P(Y=1 \mid X \in \mathcal{X}_m,\ \eta(X) \in b)\big) \approx \operatorname{lgt}\big(P(Y=1 \mid X \in \mathcal{X}_v,\ \eta(X) \in b)\big).$$

Therefore, the natural-prevalence drift of $\Delta_{m \to v}(b)$ is approximately zero, and we can conclude

$$\Delta_{m \to v}(b) = \operatorname{lgt}\big(q_m(b)\big) - \operatorname{lgt}\big(q_v(b)\big) \approx \gamma^m(b) - \gamma^v(b).$$

In the special case where natural-prevalence drift cancels exactly, the approximation becomes equality.

Proof of Proposition 7

Let $c^m = \arg\min_c \sup_{x \in m} |\gamma(x) - c|$ and $c^v = \arg\min_c \sup_{x \in v} |\gamma(x) - c|$. Since $v \subset m$, we have

$$\sup_{x \in v} |\gamma(x) - c^v| \leq \sup_{x \in v} |\gamma(x) - c^m| \leq \sup_{x \in m} |\gamma(x) - c^m|.$$

Thus the best achievable within-child deviation from a constant distortion is no larger than the best achievable within-parent deviation.

Table 1: Datasets, their number of instances, and dimension of X.

Dataset	N	d	Dataset	N	d
page-blocks0	7860	10	segment0	3166	19
abalone19	6626	8	shuttle-c0-vs-c4	2728	9
yeast3	2114	8	yeast4	2294	8
yeast5	2304	8	yeast6	2318	8

B Evaluation Details

We detail our benchmark datasets in Appendix B.1, introduce baseline imbalance mitigation methods and their hyperparameters in Appendix B.2, and give our overall class-count and LMI simulation procedure in Appendix B.3. The pseudocode of our proposed SE is given in Appendix B.4.

B.1 Datasets

We considered all balanced benchmark datasets from the KEEL repository (Alcalá-Fdez et al., 2011) with more than 2,000 instances. This was to give us enough instances to simulate class-count or LMI. The datasets and their properties are given in Table 1.

B.2 Algorithms

We evaluate on five base classifiers: logistic regression, support vector machines, random forests, multi-layer perceptrons, and CatBoost, covering linear, kernel-based, ensemble, neural, and boosting models with linear and non-linear decision boundaries.

We consider popular and state-of-the-art imbalance methods in six categories detailed below. For fairness across algorithms, we use standard implementations (scikit-learn, imbalanced-learn, or PyTorch) and widely adopted default hyperparameters and tune more complex algorithms where stated. The implementation code of all algorithms is provided in the supplementary materials.

B.2.1 Cost-based Methods

Inverse Class Frequency (ICF): Each class receives a weight inversely proportional to its training-set frequency. In particular, the minority class weight is set to $\frac{N}{2n^+}$ and the majority to $\frac{N}{2n^-}$.

Focal Loss (FL): Extends cross-entropy by down-weighting easy examples (Lin et al., 2017). The loss for a true class with predicted probability P_t is $-\alpha_t(1-P_t)^\gamma \log(P_t)$. We use the canonical defaults $\alpha = 0.25$ (minority class weight) and $\gamma = 2$ (focusing parameter), and tune them via 5-fold cross-validation.

Meta Reweighting (MET) : Iteratively learns per-training-point weights so that training instances whose gradients align with (i.e., most reduce) the meta loss get higher weight while harmful updates get down-weighted. We tuned base parameters using 5-fold cross-validation.

Long-tail rebalancing with tuned priors (LNG): Trains a base classifier, and then post-hoc adjusts its predicted probability of the minority group by subtracting a log-odds shift equal to $\tau \log\left(\frac{p}{1-p}\right)$, where p is the observed minority rate, to compensate for long-tail class priors. We tuned base parameters using 5-fold cross-validation.

Class-Balanced Loss (CLS): Weights classes using their "effective number of samples" $E(n_y) = (1-\beta^{n_y})/(1-\beta)$ (Cui et al., 2019). Class weight is proportional to $\frac{1}{E(n_y)}$. Following Cui et al. (2019), we set $\beta = \frac{(n-1)}{n}$, the standard default and tune it using 5-fold cross-validation.

B.2.2 Debiasing Methods

Logit Adjustment (LA): Corrects for skewed class priors by adding $\tau \log \pi_y$ to each class logit before cross-entropy, where π_y is the empirical class prior (Menon et al., 2020). We use the standard setting $\tau = 1$ and tune it using 5-fold cross-validation.

B.2.3 Ensemble-based Methods

Easy Ensemble (EE): Constructs multiple balanced subsets obtained through majority undersampling of the dataset and trains AdaBoost learners on each set (Liu et al., 2008). We use the `EasyEnsembleClassifier` Python library from imbalanced-learn with default hyperparameters.

Balanced Bagging (BB): Implements bagging with balanced bootstrap samples. We use the `BalancedBaggingClassifier` Python library from imbalanced-learn (Lemaitre et al., 2017) with default hyperparameters.

Random Ensembling (RE): Constructs a stratified ensemble using random subsets of the dataset. Instead of using our proposed objective function to guide stratification and the stopping criteria, RE randomly undersamples the majority class until each stratum is class-count balanced.

B.2.4 Undersampling Methods

Cluster Centroids (CC): Replaces the majority class with its k cluster centroids, reducing its size to match the minority count (Yen & Lee, 2009). We use the `ClusterCentroids` Python library from imbalanced-learn (Lemaitre et al., 2017) with default hyperparameters.

Tomek Links (TMK): Undersamples the majority set by removing borderline instances that form Tomek pairs (nearest neighbors of opposite class). We use the `TomekLinks` Python library from imbalanced-learn (Lemaitre et al., 2017) with default hyperparameters.

B.2.5 Oversampling Methods

SMOTE (SMT): Generates synthetic minority samples by interpolating between each minority point and its nearest neighbors (Chawla et al., 2002). We use the `SMOTE` Python library from imbalanced-learn (Lemaitre et al., 2017) with default hyperparameters.

Borderline SMOTE (BST): A variant of SMOTE that oversamples only minority instances near the decision boundary (with many majority neighbors) (Han et al., 2005). We use the `BorderlineSMOTE` Python library from imbalanced-learn (Lemaitre et al., 2017) with default hyperparameters.

Adaptive Synthetic oversampling (ADS): Generates more synthetic points in regions where the minority is sparsely represented relative to the majority (He et al., 2008). We use the `ADASYN` Python library from imbalanced-learn (Lemaitre et al., 2017) with default hyperparameters.

B.2.6 Optimization-based Methods

AUC-PR directive objectives (AUC): iteratively refits the classifier while updating sample weights to directly target AUC-PR: for a set of recall anchors, it finds the score threshold on positives that achieves each recall, then up-weights (i) positives that fall below that threshold (misses) and (ii) negatives that exceed it (false positives), with a larger penalty on false positives when recall is high. The weights are clipped/floored, normalized, and EMA-blended each iteration for stability. We tuned base parameters using 5-fold cross-validation.

Distributionally Robust Optimization (DRO): alternates between fitting the classifier with sample weights $w = q.n$ and updating the weight distribution q to put more mass on high-loss examples (via an exponential/multiplicative update), then projects q back onto a capped simplex so no single point can dominate (cap $= \frac{1}{\alpha n}$). This implements a DRO worst-case reweighting over an α-sized effective subset, with optional smoothing back toward uniform to keep all weights positive. We tuned base parameters using 5-fold cross-validation.

B.3 LMI and Class-count Imbalance Simulation

Algorithm 1 gives the pseudocode of our damaging procedure to simulate LMI or class-count imbalance. The damaging procedure simulates LMI via a soft deletion mechanism that incrementally removes minority instances based on a specified ordering O^{dmg} of minority instances in the training data. This ordering is an intentional design, such that removing instances with lower order simulates a non-random deletion mechanism for LMI. Without such ordering, deletion becomes approximately random and closer to general class-count imbalance.

We consider six different LMI mechanisms, spanning monotonic vs non-monotonic, single vs multiple feature driven, observed vs unobserved feature driven, minority vs majority based, and linear vs categorical deletion.

- Monotonic LMI: The default deletion mechanism that orders instances by a randomly selected feature $X \in \mathcal{X}$. Instances with lower values of X receive the highest deletion priority.
- Non-monotonic LMI: Instances are ordered by the absolute value of their percentile rank difference to the median $|(\text{rank}(X) - 0.5)|$ of a randomly selected feature $X \in \mathcal{X}$. Here, $\text{rank}(X)$ determines the rank of an instance by ordered values of X in $\mathcal{D}^{\text{train}}$. Therefore, instances with higher or lower values of X (tails) receive the highest deletion priority.
- Interaction LMI: Instances are ordered by the geometric mean of the percentile ranks $\sqrt{\text{rank}(X) \times \text{rank}(X')}$ of two randomly selected features $X, X' \in \mathcal{X}$. Instances with the lower values of this nonlinear combination of percentile ranks receive the highest deletion priority.
- Unobserved LMI: Instances are ordered by values of an unobserved feature. Feature X is randomly selected from $\mathcal{X}$ and used only for ordering. The training and test datasets are both updated to $\mathcal{X} \setminus X$, and therefore, all mitigation procedures and base classifiers never observe X. Instances with lower values of X receive the highest deletion priority.
- Majority-bias LMI: Instances are ordered by their average distance to the nearest majority points. Instances nearest to majority points have higher deletion probability.
- Categorical LMI: Instances are ordered by their quantile bin assignment of a randomly selected feature $X \in \mathcal{X}$. Instances in lower quintiles receive the highest deletion priority.

Feature interactions LMI makes observation propensity depend on a nonlinear interaction between two observed features, testing whether SE can recover localized structure that is not attributable to any single covariate. Unobserved-driver LMI introduces feature-dependent selection using a variable that is omitted from the observed feature set. Majority-bias LMI determines minority deletion propensity using local majority concentration, producing imbalance that depends on the surrounding class distribution rather than directly on a single minority feature. Non-monotonic LMI deletes minority instances from both tails of a driving feature, creating a distortion pattern that cannot be represented by a one-directional monotonic mechanism. Lastly, categorical LMI induces under-observation according to discretized quantile groups, testing whether SE remains effective when the observation mechanism operates over discrete rather than continuous feature regions.

When applicable, each deletion mechanism is simulated five times using a distinct, randomly selected feature(s) $X \in \mathcal{X}$, and repeated five times based on a random seed that determines the soft deletion likelihood.

Deletion likelihoods vary smoothly with the ordering O^{dmg} to ensure the possibility that some minority points remain observable across all regions. This preserves partial support of the original distribution $\mathcal{P}^{\text{tru}}$ while inducing systematic distortions in higher deletion-probability regions, modeling settings where imbalance arises from bias that skews but possibly does not fully censor subpopulations. Furthermore, by applying deletions progressively at increasing thresholds (65%, 75%, 85%, 95%), we generate a controlled family of datasets that allow us to measure the effects of varying degrees of LMI in Appendix C.2.

B.4 Stratifying Ensembling (SE) Pseudocode

Algorithm 2 gives the pseudocode of our proposed approach.

Algorithm 1 Imbalance Simulation Algorithm

Require: Minority instances $\mathcal{D}^{\text{min}} = \{i \in \mathcal{D}^{\text{train}} : y_i = 1\}$, damage level θ^{tar}, damage type $\texttt{how} \in \{\text{Class-count}, \text{LMI}\}$, ordering O^{dmg}.

Initialize set of damaged indices $\mathcal{I}^{\text{dam}} \leftarrow \emptyset$
for $\theta \in \{0.65, 0.75, 0.85, 0.95\}$ **do**
  Set $n^{\text{tar}} \leftarrow \theta \times |\mathcal{D}^{\text{min}}|$
  if $\texttt{how} = \texttt{class-count}$ **then**
    Sample $n^{\text{tar}} - \left|\mathcal{I}^{\text{dam}}\right|$ points at random: $\mathcal{I}^{\text{dam}} \leftarrow \mathcal{I}^{\text{dam}} \cup \text{Sample}\left(\{i \in \mathcal{D}^{\text{min}}\},\, n^{\text{tar}}\right)$
  else
    $\lambda \leftarrow 0.9 - \dfrac{0.1 \times \left|\mathcal{I}^{\text{dam}}\right|}{n^{\text{tar}}}$
    for Instances $i \in \mathcal{D}^{\text{min}}$ according to ordering O^{dmg} **do**
      if $i \notin \mathcal{I}^{\text{dam}}$ **and** $\text{Uniform}(0,1) \leq \lambda$ **then**
        $\mathcal{I}^{\text{dam}} \leftarrow \mathcal{I}^{\text{dam}} \cup \{i\}$
        $\lambda \leftarrow \lambda - 0.1/n^{\text{tar}}$
      end if
      if $\left|\mathcal{I}^{\text{dam}}\right| = n^{\text{tar}}$ **then**
        break
      end if
    end for
  end if
  if $\theta = \theta^{\text{tar}}$ **then**
    break
  end if
end for
return $\mathcal{D}^{\text{train}} \setminus \mathcal{I}^{\text{dam}}$

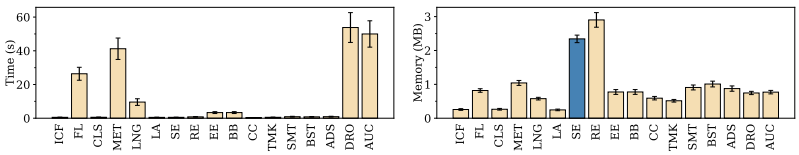


Figure 5: Average computational time and peak memory usage.

C Additional Numerical Results

We provide time and peak memory usage in C.1, performance under varying levels of LMI in C.2, class-specific performance and additional performance metrics in C.3, disaggregated results by classifier and dataset in C.4, ablation tests in C.5, hyperparameter sensitivity in C.6, performance under alternative LMI mechanisms in C.7, and performance under general class-count imbalance in C.8.

C.1 Time and Peak Memory Usage

Figure 5 gives the average computational time and peak memory usage. Our algorithm is considerably faster than or competitive with existing methods, while using on average 1.5 MB higher memory than most other algorithms. These results provide good evidence that our method is scalable to larger datasets.

C.2 Varying Degrees of LMI

Figure 6 shows how MCC and F1 change with increasing LMI for the top-performing methods of Figure 3, averaged across datasets and classifiers. SE performs best at higher LMI levels (beyond 80% or a 1:5 IR), with its advantage growing as imbalance increases. This is because lower LMI levels retain more of the underrepresented minority instances, and are closer to class-count imbalance

Algorithm 2 Stratified Ensembling (SE) Algorithm

Require: Training set $\mathcal{D}^{\text{train}} = \{(x_i, y_i)\}_{i=1}^{n}$, $\mathcal{D}^{\text{maj}} = \{i : y_i = 0\}$, $\mathcal{D}^{\text{min}} = \{i : y_i = 1\}$; base learner f; Number of root bins $|\mathcal{B}_0|$; min leaf size $n_{\min}$; Validation metric $\text{Perf}(\cdot)$

Compute proxy scores $\eta(x_i)$ for all i
Form global $|\mathcal{B}_m|$ quantile bins $\mathcal{B} = \{b_1, \ldots, b_{|\mathcal{B}_0|}\}$
Grow a stratifying tree on X to obtain leaf strata $\mathcal{S} = \{S_1, \ldots, S_K\}$ as follows:
for node m (top-down), $|m| \geq n_{\min}$ **do**
  For each $b \in \mathcal{B}_m$, estimate (smoothed) $q_m(b) \approx \frac{n_{m,b}^{(1)} + 0.5}{n_{m,b}^{(0)} + n_{m,b}^{(1)} + 1}$
  for candidate split (j, τ) **do**
    Split $m \to (v_l, v_r)$ using feature X^j at threshold τ
    for $v \in \{v_l, v_r\}$ **do**
      $\Delta_{m \to v}(b) \leftarrow \text{lgt}(q_m(b)) - \text{lgt}(q_v(b))$ for shared bins $b \in \mathcal{B}_v$
      $\sigma^2_{\mathcal{B}_v}(\Delta_{m \to v}(b)) \leftarrow \frac{\sum_{b \in \mathcal{B}_v} |b| \left(\Delta_{m \to v}(b) - \frac{\sum_{b' \in \mathcal{B}_v} |b'| \Delta_{m \to v}(b')}{\sum_{b' \in \mathcal{B}_v} |b'|} \right)^2}{\sum_{b \in \mathcal{B}_v} |b|}$
    end for
    $\mathcal{L}(X^j, \tau; m) \leftarrow \frac{|v_l|}{|m|} \sigma^2_{\mathcal{B}_{v_l}}(\Delta_{m \to v_l}(b)) + \frac{|v_r|}{|m|} \sigma^2_{\mathcal{B}_{v_r}}(\Delta_{m \to v_r}(b))$
  end for
  Choose $(j^\star, \tau^\star) \in \arg\min_{(j,\tau)} \mathcal{L}(X^j, \tau; m)$
  if $\mathcal{L}(X^{j^\star}, \tau^\star; m) \leq 0.999 \sigma^2_{\mathcal{B}_m}(q)$ (0.1% improvement) **then**
    Split m on $X^{j^\star}$ at $\tau^\star$
  else
    Stop at leaf node m
  end if
end for
for $k = 1, \ldots, K$ **do**
  $\mathcal{D}_k^{\text{train}} \leftarrow \{(x_i, 0) : i \in \mathcal{D}^{\text{maj}},\ x_i \in S_k\} \cup \{(x_i, 1) : i \in \mathcal{D}^{\text{min}}\}$
  Fit $f_k \leftarrow \text{TRAIN}(f, \mathcal{D}_k^{\text{train}})$
  Weight $\alpha_k \leftarrow \text{Perf}(f_k)$ on validation
end for
return $\hat{p}(x) = \frac{\sum_{k=1}^{K} \alpha_k P_{f_k}(Y=1 | X=x)}{\sum_{k=1}^{K} \alpha_k}$

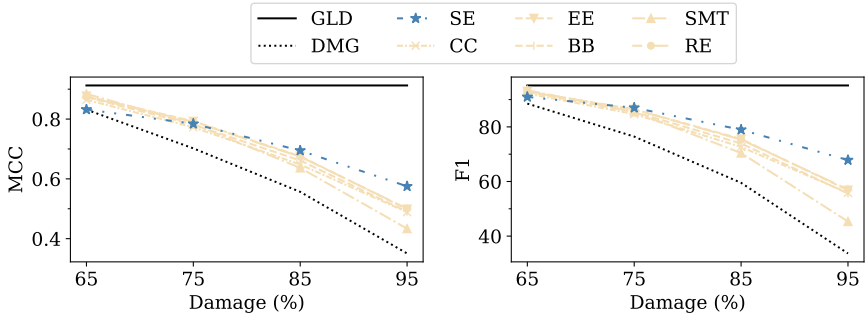


Figure 6: MCC and F1 performance under varying levels of LMI.

scenarios. Note that at higher deletion rates, some feature regions may lose nearly all observed minority instances, making it impossible for any algorithm to recover performance.

C.3 Class-specific and Additional Metrics

To better evaluate the performance of SE, we report precision, recall, and F1 separately for both classes in Figure 7. SE achieves higher precision and F1 for the majority class, and higher recall and F1 for the minority class compared to all other competing methods. The results therefore indicate

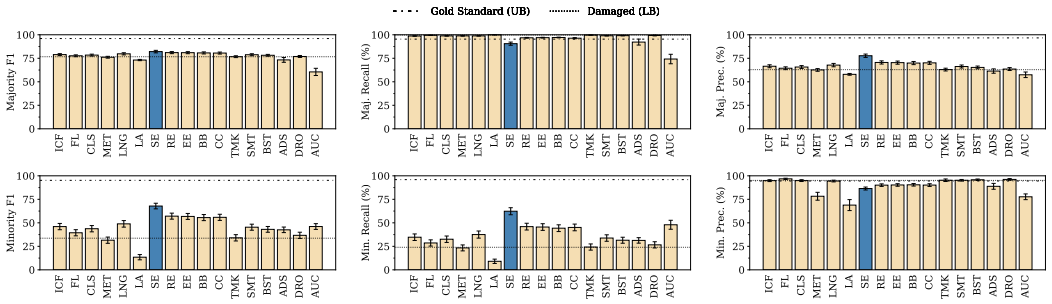


Figure 7: Class-specific F1, Precision, and Recall under LMI.

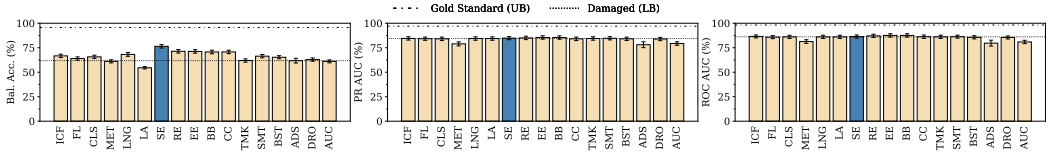


Figure 8: Balanced accuracy, PR AUC, and ROC AUC performance under LMI.

that SE achieves a more balanced allocation of predictive performance across classes under LMI, rather than improving minority detection at the cost of majority-class detection.

Figure 8 gives the balanced accuracy, PR AUC, and ROC AUC performance, averaged over all datasets and classifiers. SE remains the best-performing method under balanced accuracy, but all methods perform similarly under AUC measures. This shows that AUC metrics are much less informative under LMI, and confirms recent literature that has cautioned against treating AUC measures as automatically preferable under class imbalance (McDermott et al., 2024; Richardson et al., 2024).

C.4 Disaggregated Results by Classifier and Dataset

Table 2 gives the average MCC and F1 performance by classifier. SE achieves higher average MCC and F1 than the benchmarks for each of the five base classifiers.

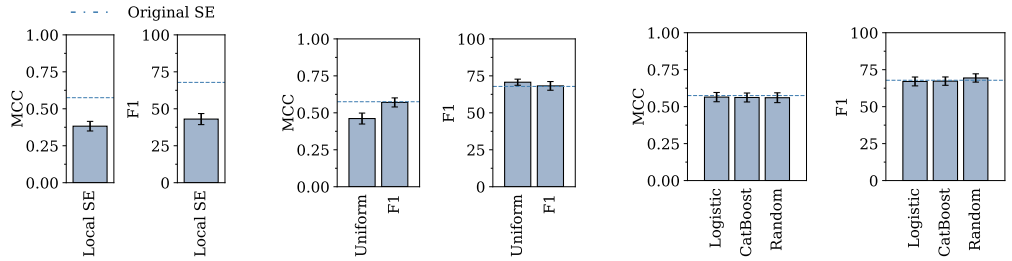
Table 3 gives the average MCC and F1 performance by dataset. Our proposed SE approach gives higher F1 performance in five, and higher MCC performance in six out of eight datasets. Even when SE does not achieve the highest performance, it remains close to the highest-performing competing method.

Table 2: Detailed Performance by Base Classifier.

Metric	Classifier	UB	Cost-based					Debias	Ensemble-based				Under-sampling		Over-sampling			Optimization		LB
		GLD	ICF	FL	CLS	MET	LNG	LA	SE	RE	EE	BB	CC	TMK	SMT	BST	ADS	DRO	AUC	DMG
MCC	LR	0.81	0.54	0.45	0.51	0.28	0.49	0.12	**0.57**	0.52	0.52	0.52	0.53	0.34	0.52	0.49	0.51	0.38	0.42	0.34
	SVM	0.90	0.38	0.37	0.36	0.33	0.42	0.22	**0.53**	0.47	0.47	0.46	0.45	0.36	0.42	0.40	0.39	0.36	0.33	0.35
	RF	0.98	0.34	0.35	0.34	0.35	0.47	0.20	**0.59**	0.46	0.45	0.43	0.46	0.36	0.38	0.37	0.34	0.36	0.35	0.35
	NN	0.91	0.51	0.42	0.49	0.28	0.50	0.08	**0.59**	0.56	0.56	0.55	0.54	0.37	0.46	0.44	0.41	0.39	0.30	0.36
	CAT	0.96	0.41	0.38	0.39	0.29	0.42	0.19	**0.59**	0.49	0.49	0.49	0.46	0.34	0.39	0.38	0.36	0.37	0.07	0.34
	Avg	0.91	0.44	0.39	0.42	0.31	0.46	0.16	**0.58**	0.50	0.50	0.49	0.49	0.35	0.43	0.42	0.40	0.37	0.29	0.35
F1	LR	89.6	61.5	48.3	58.2	32.6	54.2	10.4	**70.2**	61.5	61.6	60.7	61.6	34.2	59.3	55.2	59.1	40.3	56.0	33.8
	SVM	94.4	38.2	36.1	35.6	32.2	44.2	19.7	**62.5**	53.8	53.7	52.6	52.3	34.3	43.9	41.4	41.3	34.6	32.4	33.9
	RF	98.8	31.2	32.4	31.4	32.4	48.9	14.7	**66.1**	50.6	49.8	46.8	51.2	32.9	36.0	34.6	32.1	32.9	32.3	32.6
	NN	95.3	56.9	43.8	54.4	30.6	55.0	6.9	**71.3**	64.0	63.8	62.7	61.2	36.7	48.6	46.5	44.2	40.0	49.2	36.1
	CAT	97.9	42.2	36.5	39.3	29.9	42.2	15.5	**69.1**	55.6	55.1	55.1	52.8	32.2	38.9	37.6	36.1	35.6	60.4	31.8
	Avg	95.2	46.0	39.4	43.7	31.5	48.9	13.4	**67.8**	57.1	56.8	55.6	55.8	34.0	45.4	43.0	42.5	36.7	46.1	33.6

Table 3: Detailed Performance by Dataset.

Metric	Dataset	UB	Cost-based					Debias	Ensemble-based				Under-sampling		Over-sampling			Optimization		LB
		GLD	ICF	FL	CLS	MET	LNG	LA	SE	RE	EE	BB	CC	TMK	SMT	BST	ADS	DRO	AUC	DMG
MCC	page-blocks0	0.89	0.35	0.32	0.33	0.29	0.36	0.20	**0.40**	0.35	0.35	0.35	0.32	0.30	0.35	0.34	0.38	0.31	0.29	0.29
	segment0	0.99	0.60	0.60	0.56	0.46	**0.67**	0.26	0.60	0.56	0.56	0.54	0.54	0.55	0.61	0.56	0.63	0.60	0.48	0.55
	abalone19	0.82	0.25	0.24	0.24	0.17	0.23	0.08	0.24	0.21	0.22	0.21	0.21	0.20	0.28	0.27	**0.32**	0.23	0.21	0.20
	shuttle-c0-vs-c4	1.00	0.82	0.78	0.80	0.76	**0.92**	0.51	**0.92**	0.89	0.90	0.90	0.87	0.77	0.85	0.78	0.42	0.77	0.41	0.77
	yeast3	0.87	0.34	0.29	0.32	0.10	0.36	0.03	**0.63**	0.43	0.42	0.40	0.50	0.21	0.32	0.32	0.33	0.22	0.22	0.21
	yeast4	0.85	0.31	0.27	0.30	0.21	0.31	0.06	**0.48**	0.38	0.39	0.38	0.35	0.22	0.31	0.30	0.33	0.25	0.23	0.22
	yeast5	0.97	0.43	0.32	0.41	0.28	0.41	0.11	**0.72**	0.64	0.62	0.61	0.58	0.29	0.39	0.37	0.41	0.30	0.27	0.28
	yeast6	0.91	0.39	0.33	0.38	0.19	0.42	0.06	**0.62**	0.55	0.54	0.54	0.54	0.29	0.37	0.38	0.38	0.30	0.26	0.29
	Avg	0.91	0.44	0.39	0.42	0.31	0.46	0.16	**0.58**	0.50	0.50	0.49	0.49	0.35	0.43	0.42	0.40	0.37	0.29	0.35
F1	page-blocks0	94.0	34.3	30.9	32.4	26.1	36.4	13.2	42.6	36.1	36.7	35.4	30.2	26.8	35.1	33.5	38.9	28.6	**42.9**	26.2
	segment0	99.4	65.1	65.5	61.1	49.1	**72.4**	23.8	71.2	62.7	62.5	61.0	63.4	59.1	66.1	61.1	67.8	65.0	63.2	59.1
	abalone19	89.9	26.0	21.3	24.1	17.0	22.6	4.6	**37.0**	28.4	28.8	28.1	26.1	16.1	28.9	27.5	35.2	18.8	29.3	15.8
	shuttle-c0-vs-c4	99.9	84.2	80.9	82.5	78.8	**93.6**	52.7	93.5	90.9	91.8	91.7	89.3	79.2	86.4	80.1	42.3	79.5	71.0	79.2
	yeast3	93.0	37.2	29.2	34.6	12.2	41.4	1.3	**79.5**	55.6	54.1	51.0	62.7	19.3	34.9	34.8	36.5	21.4	36.3	18.5
	yeast4	91.8	32.8	24.4	30.6	22.2	32.1	2.6	**64.1**	46.8	47.3	45.3	43.2	17.6	31.9	29.6	35.3	21.8	38.1	17.3
	yeast5	98.3	45.4	30.4	43.5	26.7	45.5	6.1	**79.7**	71.6	69.4	68.6	66.3	26.5	40.6	38.3	43.5	29.3	42.1	25.7
	yeast6	95.1	43.1	32.8	41.1	20.2	47.3	3.2	**75.2**	64.6	63.8	63.5	65.5	27.7	38.9	39.6	40.9	29.2	45.5	27.4
	Avg	95.2	46.0	39.4	43.7	31.5	48.9	13.4	**67.8**	57.1	56.8	55.6	55.8	34.0	45.4	43.0	42.5	36.7	46.1	33.6



(a) Local vs Global (original setting) pairing of minority instances in SE.

(b) Uniform and F1-based vs MCC-based (original setting) weighting.

(c) Alternative logistic regression (LR), CatBoost (CatB), and Random (RND) calculation of ranking scores $\eta(x)$, vs difference-in-means (original setting).

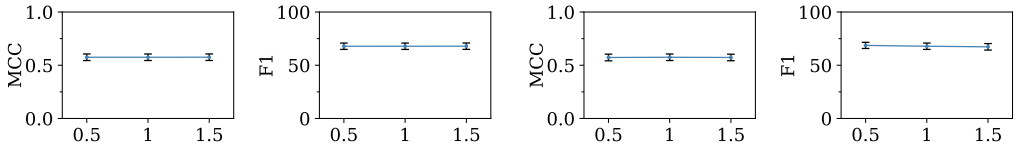
Figure 9: Ablation Studies.

C.5 Ablation Tests

We provide three sets of ablation tests in Figure 9. First, we examine our bias-variance tradeoff strategy of combining all minority instances with the majority instances of each stratum $S_k \in \mathcal{S}$. We compare this global minority pooling strategy with a local alternative that combines only the minority and majority instances within each stratum. As shown in Figure 9a, global pooling achieves superior performance, supporting its role in mitigating the sparsity of minority observations within individual strata, and ultimately providing a better bias-variance tradeoff.

Second, we examine the contribution of MCC-based ensemble weighting by comparing it with uniform and F1-based averaging. As shown in Figure 9(b), uniform averaging produces approximately six percentage points lower MCC, while improving F1 by approximately two points. F1-based averaging results in almost identical results to MCC-based averaging. These results suggest that performance-based weighting substantially improves results, despite relying on validation scores measured under the biased observation distribution.

Finally, we examine the sensitivity of SE to the choice of ranking proxy $\eta(x)$ used to construct the strata. Our original SE algorithm employs an inexpensive, model-agnostic difference-in-means direction, setting $\eta(x) \propto w^\top x$, where $w_j = (\mu_{1j} - \mu_{0j})/\sqrt{(\sigma_{1j}^2 + \sigma_{0j}^2)/2}$, and μ_{yj} and σ_{yj}^2 denote the class-conditional mean and variance of feature $X^j \in X$ for class $Y = y$. We replace this proxy with confidence scores from logistic regression and CatBoost classifiers trained on $\mathcal{D}^{\text{obs}}$, as well as a completely random ordering containing no predictive information about class membership. Results in Figure 9c demonstrate robustness to model-based proxies, while random ordering preserves approximately the same F1 and reduces average MCC by approximately 0.03.



(a) Sensitivity to bin count of root node $|\mathcal{B}_0|$. (b) Sensitivity to minimum leaf size $n_{\min}$.

Figure 10: Sensitivity Results.

Table 4: Performance on Alternative LMI Mechanisms.

Metric	LMI Mechanism	UB	Cost-based					Debias	Ensemble-based				Under-sampling		Over-sampling			Optimization		LB
		GLD	ICF	FL	CLS	MET	LNG	LA	SE	RE	EE	BB	CC	TMK	SMT	BST	ADS	DRO	AUC	DMG
MCC	Monotonic	0.91	0.44	0.39	0.42	0.31	0.46	0.16	**0.58**	0.50	0.50	0.49	0.49	0.35	0.43	0.42	0.40	0.37	0.29	0.35
	Non-monotonic	0.91	0.49	0.52	0.47	0.38	0.62	0.16	**0.62**	0.58	0.58	0.57	0.57	0.38	0.49	0.45	0.44	0.48	0.36	0.38
	Interaction	0.91	0.47	0.41	0.45	0.32	0.49	0.16	**0.58**	0.54	0.54	0.53	0.53	0.37	0.46	0.43	0.43	0.39	0.31	0.36
	Unobserved	0.90	0.53	0.46	0.52	0.34	0.55	0.17	**0.62**	0.59	0.60	0.59	0.60	0.40	0.55	0.51	0.51	0.41	0.34	0.40
	Majority-biased	0.91	0.47	0.42	0.45	0.34	0.47	0.17	**0.58**	0.51	0.51	0.50	0.54	0.38	0.48	0.46	0.50	0.41	0.33	0.38
	Categorical	0.91	0.52	0.53	0.50	0.37	0.60	0.18	**0.62**	0.60	0.59	0.58	0.57	0.40	0.53	0.51	0.52	0.50	0.39	0.39
	Avg	0.91	0.49	0.46	0.47	0.34	0.53	0.17	**0.60**	0.55	0.55	0.54	0.55	0.38	0.49	0.46	0.47	0.43	0.34	0.38
F1	Monotonic	95.2	46.0	39.4	43.7	31.5	48.9	13.4	**67.8**	57.1	56.8	55.6	55.8	34.1	45.3	43.0	42.6	36.7	46.1	33.6
	Non-monotonic	95.2	51.8	54.0	49.2	42.6	67.4	13.0	**72.5**	65.4	65.3	64.0	64.1	37.1	52.0	48.0	48.0	49.8	52.3	36.8
	Interaction	95.2	49.1	41.4	46.9	33.2	51.8	13.1	**68.2**	61.2	60.9	59.7	59.3	35.8	47.7	44.3	45.0	38.5	47.5	35.4
	Unobserved	94.2	57.7	47.5	55.5	37.2	59.5	14.4	**73.6**	70.3	70.8	70.0	67.9	40.9	59.6	55.8	56.5	42.7	50.3	40.3
	Majority-biased	95.2	49.5	40.8	46.5	33.8	48.8	13.3	**66.7**	57.3	56.6	55.0	58.9	36.2	50.0	47.1	52.4	39.4	49.0	36.0
	Categorical	95.2	56.3	56.4	54.0	42.5	65.0	15.3	**74.3**	67.9	67.5	66.3	65.6	40.4	57.2	55.1	56.6	52.9	55.0	39.6
	Avg	95.0	51.7	46.6	49.3	36.8	56.9	13.8	**70.5**	63.2	63.0	61.8	61.9	37.4	52.0	48.9	50.2	43.3	50.0	37.0

Together, these findings indicate that SE benefits from global minority pooling and performance-based aggregation while remaining relatively insensitive to the specific ranking proxy used to construct its strata.

C.6 Hyperparameter Sensitivity

We evaluate sensitivity to parameters of SE by varying the number of root node bins $|\mathcal{B}_0|$ from 50 to 150 (with 100 being the default), and the minimum leaf-size constraints from $0.5\times$ to $1.5\times$ their default values. Across these settings, average MCC and F1 vary by less than one percentage point, including on the smaller datasets. These results indicate that SE is not overly sensitive to precise choices of $|\mathcal{B}_0|$ or the minimum leaf sizes.

C.7 Alternative LMI Mechanisms

Our primary experiments use a monotonic LMI mechanism driven by a single observed feature. We adopt this setting deliberately as a controlled benchmark rather than as a claim that real-world LMI is generally monotonic or uni-variate. A single-feature monotonic mechanism allows the source and severity of observation distortion to be varied transparently and consistently across datasets, while holding other aspects of the data-generating process fixed, making it possible to isolate the effect of increasing LMI severity and to compare methods under a common, interpretable distortion process. To assess whether the performance of SE is limited to this simplifying benchmark, we additionally evaluate five qualitatively different LMI mechanisms, including feature interactions, an unobserved driver, majority-concentration bias, non-monotonic deletion, and categorical or grouped deletion (see Appendix B.3 for detailed definitions and overall simulation algorithm).

Table 4 gives the average results for each mechanism.

SE remains the best-performing method under all six mechanisms. As expected, the advantage is smallest when the LMI-driving variable is unobserved. Let X denote the observed features and Z the unobserved driver. The effective observation propensity available through the observed data is

$$\tilde{\pi}_1(x) = \mathbb{E}[\pi_1(X, Z) \mid X = x] .$$

When Z is associated with observed features, SE can exploit the localized distortion expressed through those proxies. If the omitted driver is independent of X and no observed feature carries

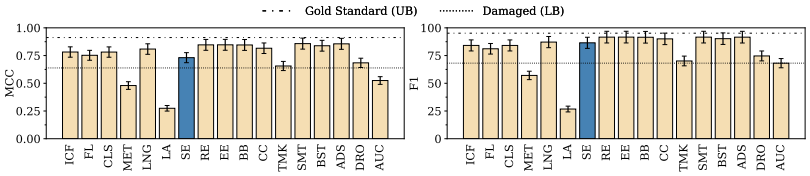


Figure 11: Performance under class-count imbalance.

information about the selection mechanism, then $\tilde{\pi}_1(x)$ becomes constant and the problem reduces to class-count imbalance with respect to the observed feature space. In that limiting case, no method using only X can exploit localized structure. The positive gains in the omitted-driver setting therefore show that SE remains useful when the true driver is available only indirectly through correlated observed variables.

C.8 Class-count Imbalance

Figure 11 gives the average performance of all methods under the general class-count imbalance problem. Since the minority class retains sufficient representation under class-count imbalance, existing methods excel in mitigating imbalance. On the other hand, SE's reliance on majority strata increases variance leading to slightly lower performance.

Accordingly, SE should not be viewed as a universal replacement for standard class-imbalance methods. When no less-biased validation sample or external information about the observation process is available, the presence of LMI cannot be certified from $\mathcal{D}^{\text{obs}}$ alone. In such settings, SE is most appropriately used as a robustness analysis when substantive knowledge makes localized observation bias plausible.

D Limitations

Our study has several limitations. First, our method is developed and evaluated primarily for tabular classification. Extensions to non-tabular data may be possible by applying stratification to learned representations, although these adaptations are left for future work as they are best addressed using specialized algorithms and classifiers. Second, our split objective is not an exact identifier of the observation distortion $\gamma(x)$. As we prove in §4.1, this is impossible without external knowledge. Therefore, SE is a practical theory-motivated approximation.

Finally, without additional assumptions or information, minority behavior cannot generally be identified in regions with no observed minority support. Likewise, LMI cannot be addressed from observed data alone when it is driven entirely by an unobserved feature that is uncorrelated with all observed features. This is because the observation distortion leaves no detectable localized structure in the observed feature space of such scenarios. Our approach is therefore most appropriate when LMI suppresses but does not entirely eliminate minority support in affected regions, and when the drivers of LMI are observed directly or are at least reflected through correlated observed features.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer:Yes

   Justification: Please see results in main body §5 and appendix C.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: Yes.

   Justification: Please see limitation section in Appendix D

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: Yes.

   Justification: Please see results and proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: Yes.

   Justification: All information is given in the paper, along with the source code given in additional material.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: Yes.

   Justification: All data are well-known benchmark imbalance sets from KEEL.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: Yes.

   Justification: Please see results section.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: Yes.

   Justification: All error bars are given as 95% confidence intervals.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: Yes.

   Justification: Please see Evaluation Setting section.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: Yes

Justification: It complies.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: NA

    Justification: foundational research and not tied to particular applications.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: NA

    Justification: no such risk.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: Yes.

    Justification: all benchmarks are cited.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: NA

    Justification: no specific asset used

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: NA

    Justification: no crowdsourcing

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: NA

    Justification: Does not apply

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: NA

    Justification: LMMs were used only for writing and formatting.