Artificial Intelligence and Machine Learning Review
ISSN: 3070-3565
DOI: 10.69987/AIMLR.2026.70205

An Empirical Evaluation of Oversampling-Ensemble Interactions Under Varying Imbalance Ratios for Tabular Data Classification

Wenlan Wei[1], *Zhengchun Shang*[1.2]
[1] *Computer Science, Cornell University, Ithaca, NY, USA*
[1.2] *Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA*

Keywords

class imbalance, oversampling, ensemble learning, tabular data, imbalance ratio

Abstract

Class imbalance represents one of the most persistent and practically consequential challenges in supervised machine learning, arising whenever the minority class—typically the class of primary interest—constitutes only a small fraction of the total training population. Existing research has examined oversampling techniques and ensemble methods as independent remedies, yet their interaction under systematically varying imbalance severity remains insufficiently characterized. This study presents a structured empirical evaluation of six oversampling strategies—SMOTE, Borderline-SMOTE, ADASYN, K-Means SMOTE, SMOTEENN, and SVMSMOTE—combined with four ensemble classifiers—Random Forest, XGBoost, LightGBM, and EasyEnsemble—across ten publicly available tabular benchmark datasets drawn from the KEEL repository, the UCI Machine Learning Repository, and Kaggle. Datasets are partitioned into four imbalance ratio groups: low (IR < 5), medium (5–10), high (10–50), and extreme (IR > 50). Performance is assessed using F1-score, G-mean, AUC-ROC, and AUPRC under stratified ten-fold cross-validation. Experimental results reveal that no single oversampling-ensemble pairing dominates uniformly across all imbalance levels. SMOTEENN combined with EasyEnsemble achieves the strongest overall performance in high and extreme imbalance scenarios, while classifier reweighting without explicit oversampling proves adequate under low imbalance. These findings offer practical guidance for practitioners selecting preprocessing-ensemble pipelines commensurate with observed dataset imbalance severity.

1. Introduction

1.1 Background and Motivation

Real-world tabular classification datasets across fraud detection, network intrusion analysis, manufacturing quality control, and customer churn prediction routinely exhibit pronounced disparities between class frequencies. When the minority class constitutes only a small fraction of the training population, standard classifiers optimized to minimize overall error systematically assign the majority label to nearly all test instances. This behavior produces inflated accuracy scores that mask substantial degradation in minority class recall—a failure mode with serious operational consequences in precisely those domains where minority instances carry the greatest practical significance.

Chawla et al. [1] formalized this problem and introduced SMOTE (Synthetic Minority Over-sampling Technique), which generates synthetic minority class samples by interpolating between existing instances in the original feature space. The proposal catalyzed a large body of subsequent research examining how distributional skewness between classes affects classifier generalization. Kubat and Matwin [2] had earlier demonstrated that naive training on highly skewed datasets systematically disadvantages the minority class and proposed one-sided selection, a targeted undersampling strategy that removes borderline and noisy majority class examples. These foundational contributions established the conceptual framework within which the imbalanced learning literature has since expanded considerably.

The challenge of strategy selection is compounded by the well-documented sensitivity of relative method effectiveness to experimental context. Van Hulse et al. [3] evaluated eleven learning algorithms across 35 imbalanced datasets using six performance metrics and observed that no strategy produced consistently superior results across all combinations of dataset, classifier, and evaluation metric. This context sensitivity is particularly pronounced as imbalance severity increases: methods that perform adequately at moderate imbalance ratios may fail to provide meaningful minority class coverage under extreme skewness. Structured comparisons that systematically vary imbalance severity are needed to make this dependency explicit and actionable.

1.2 Problem Statement and Research Objectives

A. Research Questions

The interaction between oversampling preprocessing and ensemble classification—two of the most effective strategy families identified in the literature—has not been studied systematically across a range of imbalance severity levels. Most existing studies apply a fixed oversampling technique to a single classifier or compare a small number of configurations on heterogeneous dataset collections without stratifying by imbalance ratio. This absence of stratification makes it difficult to derive conclusions that are directly applicable to a practitioner selecting a strategy for a dataset of known imbalance severity. Liu et al. [4] demonstrated through the EasyEnsemble and BalanceCascade frameworks that integrating undersampling into ensemble construction yields strong minority class recall, yet the interaction between oversampling-based preprocessing and widely used gradient-boosted ensembles across varying imbalance levels remains an open empirical question.

This study addresses two primary research questions. The first question asks which oversampling-ensemble combinations achieve superior F1-score, G-mean, AUC-ROC, and AUPRC across datasets grouped by imbalance ratio. The second question asks whether the relative ranking of oversampling strategies changes significantly as imbalance severity increases from low to extreme levels.

B. Scope and Contributions

This study evaluates six oversampling methods and four ensemble classifiers in a full factorial experimental design across ten tabular benchmark datasets selected to span four imbalance ratio categories. The experimental protocol employs stratified ten-fold cross-validation and reports four complementary evaluation metrics to characterize performance across the precision-recall tradeoff. The principal contribution of this work is a structured empirical characterization of how oversampling-ensemble interaction effects vary with imbalance severity, providing concrete and falsifiable guidance for practitioners selecting preprocessing-ensemble pipelines commensurate with dataset imbalance characteristics. All datasets are drawn from publicly accessible repositories to support reproducibility.

2. Related Work

2.1 Oversampling Strategies for Imbalanced Learning

A. Synthetic Sample Generation Methods

The field of imbalanced learning distinguishes three principal families of data-level strategies: oversampling, undersampling, and hybrid methods. Oversampling methods augment the minority class by generating new samples, either by replicating existing instances or by interpolating between them in the original feature space. Domingos [5] established a complementary perspective through MetaCost, a wrapper method that converts any classifier into a cost-sensitive learner by relabeling training instances according to asymmetric misclassification costs, demonstrating that algorithm-level modification provides a theoretically grounded alternative to data-level resampling. The existence of multiple effective strategy families motivates systematic comparison across shared experimental conditions.

The SMOTE family has expanded substantially since its introduction, producing numerous variants targeting specific distributional challenges. Borderline-SMOTE concentrates synthetic generation on minority instances located near the decision boundary, where misclassification probability is highest. ADASYN adapts the local density of synthetic generation, placing more samples in sparse minority regions where the classifier is most likely to err. Fan et al. [6] proposed AdaCost, extending the boosting framework with asymmetric weight updates that penalize minority class misclassification more heavily, illustrating that the conceptual boundary between data-level and algorithm-level approaches can be productively blurred. These algorithmic variants reflect a shared motivating intuition: effective

minority class augmentation must account for the local geometry of the class boundary rather than applying uniform generation across the minority class distribution.

B. Hybrid Resampling Approaches

Hybrid strategies combine oversampling with cleaning mechanisms that remove noisy or ambiguous majority class instances from the augmented training set. SMOTEENN applies Edited Nearest Neighbors as a postprocessing step to remove any training instance—from either class—whose label disagrees with the majority vote of its three nearest neighbors. This double operation tends to produce cleaner decision boundaries than oversampling alone and achieves more consistent G-mean scores across datasets of varying structural complexity. Chawla et al. [7] extended the integration of oversampling and ensemble learning through SMOTEBoost, which applies SMOTE at each boosting iteration to maintain minority class representation throughout ensemble construction, demonstrating that the timing of synthetic generation within the learning pipeline meaningfully affects final classification performance.

2.2 Ensemble-Based Approaches to Imbalanced Classification

Ensemble methods combine predictions from multiple base classifiers to achieve more robust generalization than any individual learner. In imbalanced settings, ensemble approaches address the minority class disadvantage either through internal resampling—as in EasyEnsemble, which trains each ensemble member on a balanced subsample of the original training data—or through modified loss functions and instance weighting schemes. Elkan [8] provided the theoretical grounding for cost-sensitive classification by demonstrating that adjusting class weights in proportion to asymmetric misclassification costs is equivalent under certain distributional assumptions to training on a rebalanced dataset. This theoretical equivalence between preprocessing-based and algorithm-level approaches motivates empirical comparison under controlled experimental conditions, as the equivalence may break down at extreme imbalance levels where minority class coverage becomes a structural constraint rather than a cost optimization problem. Gradient-boosted ensemble methods such as XGBoost and LightGBM offer explicit class weighting parameters that implement a related reweighting heuristic without requiring preprocessing, making them natural comparison points for oversampling-augmented pipelines.

3. Experimental Setup

3.1 Datasets and Evaluation Protocol

Ten publicly available tabular binary classification datasets were selected from the KEEL Imbalanced Dataset Repository (https://sci2s.ugr.es/keel/imbalanced.php), the UCI Machine Learning Repository (https://archive.ics.uci.edu), and Kaggle (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud). Selection criteria required that each dataset present a binary classification problem with a clearly defined minority class, contain no missing values after standard preprocessing, and represent one of four imbalance ratio groups. The imbalance ratio (IR) is defined as the count of majority class instances divided by the count of minority class instances. Table 1 presents the characteristics of all ten benchmark datasets organized by IR group.

Table 1. Summary of benchmark datasets organized by imbalance ratio group. IR = imbalance ratio. KEEL = Knowledge Extraction based on Evolutionary Learning repository; UCI = University of California Irvine Machine Learning Repository.

Dataset	Source	Samples	Features	IR	IR Group
Wisconsin Breast Cancer	UCI	683	9	1.86	Low (IR < 5)
Pima Indians Diabetes	UCI	768	8	1.87	Low (IR < 5)
Glass6	KEEL	214	9	6.38	Medium (5–10)
Yeast3	KEEL	1,484	8	8.10	Medium (5–10)
Ecoli3	KEEL	336	7	8.60	Medium (5–10)
Ecoli4	KEEL	336	7	15.80	High (10–50)

Abalone9-18	KEEL	731	8	16.40	High (10–50)
Yeast6	KEEL	1,484	8	41.40	High (10–50)
Abalone19	KEEL	4,174	8	129.44	Extreme (IR > 50)
Credit Card Fraud	Kaggle	284,807	30	577.88	Extreme (IR > 50)

Stratified ten-fold cross-validation was employed for all experiments to preserve the class distribution in each fold. Performance was assessed using four metrics: F1-score (computed on the minority class), G-mean (geometric mean of class-specific recall rates), AUC-ROC, and AUPRC (area under the precision-recall curve). AUPRC is particularly informative under extreme imbalance, as the precision-recall curve is more sensitive to minority class detection than the ROC curve when negative instances substantially outnumber positive instances. Statistical significance of pairwise performance differences between oversampling strategies was evaluated using the Wilcoxon signed-rank test with Bonferroni correction at $\alpha = 0.05$ applied across the ten cross-validation fold scores. All experiments were implemented in Python 3.10 using scikit-learn 1.4.0 and imbalanced-learn 0.12.0.

3.2 Oversampling Methods Under Evaluation

A. SMOTE and Its Variants

Four SMOTE-based oversampling methods were included in the evaluation. Standard SMOTE generates synthetic samples by selecting a minority instance, identifying its k nearest minority class neighbors ($k = 5$ in all experiments), and interpolating a new sample along the line segment connecting the selected instance and one randomly chosen neighbor. The synthetic generation proportion was set to achieve a 1:1 majority-to-minority class ratio for all methods, consistent with the convention used in benchmark studies of imbalanced learning. Borderline-SMOTE restricts synthetic generation to minority instances residing in the borderline region, defined as instances whose k-nearest neighborhood contains between 50% and 100% majority class members. This boundary-focused augmentation concentrates new samples in the region of highest classification uncertainty. ADASYN weights the synthetic generation density by local learning difficulty, operationalized as the fraction of majority class instances among the k nearest neighbors of each minority instance, producing a denser synthetic population in regions where the classifier is most susceptible to error. SVMSMOTE trains a support vector machine on the original imbalanced data and generates synthetic samples in the vicinity of the support vectors lying closest to the decision boundary, providing an additional boundary-oriented generation strategy with a different geometric characterization of the boundary region. All four methods were applied with default imbalanced-learn hyperparameters with the exception of the target sampling ratio.

B. Hybrid Oversampling Strategies

Two hybrid strategies were included to assess whether combining synthetic generation with a cleaning operation improves performance relative to pure generation approaches. SMOTEENN follows SMOTE generation with an Edited Nearest Neighbors cleaning step: any training instance from either class whose label disagrees with the majority vote of its three nearest neighbors is removed from the augmented training set. SMOTEENN was applied with $k = 5$ for the SMOTE component and $k = 3$ for the ENN cleaning step. The combined operation tends to reduce class overlap by removing ambiguous boundary instances from both classes after oversampling. K-Means SMOTE first partitions the minority class instances into clusters using k-means (n clusters = 5) and applies SMOTE within each cluster independently, avoiding the interpolation of synthetic samples between minority instances belonging to distinct subconceptions of the minority class. Together, the six oversampling methods span a broad spectrum of augmentation strategies—from simple linear interpolation to boundary-focused, density-adaptive, cluster-constrained, and boundary-cleaning generation—enabling systematic comparison of their relative strengths across the four IR groups defined in Table 1.

3.3 Ensemble Classifiers

A. Tree-Based Ensemble Methods

Three ensemble classifiers representing the gradient-boosted and bagging families were selected from widely used open-source implementations. Random Forest was configured with 200 decision trees, unlimited maximum depth, and class_weight='balanced' to incorporate implicit minority class reweighting through inverse class frequency. XGBoost

(version 2.0) was configured with 300 estimators, maximum depth of 6, learning rate of 0.05, and scale_pos_weight set to the observed IR of each training fold. LightGBM (version 4.1) was configured with 300 leaves, minimum child samples of 20, and is_unbalance=True to activate the library's internal class reweighting mechanism. Hyperparameter values for all three classifiers were selected based on values reported as effective in the imbalanced learning literature and held constant across all datasets to enable fair comparison.

Cao et al. [9] demonstrated through label-distribution-aware margin loss that the classification margin assigned to minority class instances should theoretically exceed that of majority class instances to achieve optimal minority recall under long-tailed distributions. The scale_pos_weight and is_unbalance configurations applied here implement a related reweighting principle without requiring modification of the base loss function, providing a principled but computationally simple form of cost sensitivity within each ensemble.

B. Specialized Imbalanced Ensemble Classifiers

EasyEnsemble was included as the fourth classifier to represent the family of ensemble methods specifically designed for imbalanced learning. EasyEnsemble trains T = 10 AdaBoost sub-classifiers, each on a balanced subset constructed by randomly undersampling the majority class to match the minority class count within each sub-training set. Final predictions aggregate the T sub-classifier outputs by majority vote. Liu et al. [4] introduced EasyEnsemble and demonstrated substantial improvements in minority class recall relative to single classifiers trained on the full imbalanced dataset. Liu et al. [10] subsequently introduced MESA, a meta-learning extension that trains an adaptive sampling distribution for ensemble construction via reinforcement learning; while MESA represents a more capable framework, EasyEnsemble was selected for the present study as a well-understood and computationally tractable representative of the undersampling-ensemble family. Liu et al. [11] further demonstrated through the Self-paced Ensemble framework that scheduling the difficulty distribution of training batches across ensemble iterations—in addition to controlling the resampling ratio—meaningfully improves minority class recall on massive highly imbalanced datasets, confirming that the structure of ensemble training carries independent explanatory weight in the imbalanced learning problem.

4. Results and Analysis

4.1 Overall Performance Comparison

A. F1-Score and G-Mean Results

Table 2 presents the mean F1-score and G-mean for each of the 24 oversampling-ensemble combinations, averaged across all ten datasets and ten cross-validation folds. The highest mean F1-score (0.712) was achieved by the SMOTEENN + EasyEnsemble configuration, followed by SMOTEENN + XGBoost (0.697) and Borderline-SMOTE + EasyEnsemble (0.681). A no-oversampling baseline—using each ensemble's native class reweighting mechanism—was also evaluated; the baseline XGBoost configuration achieved a mean F1-score of 0.634, confirming that explicit oversampling preprocessing provides meaningful but not dramatic benefit over a well-calibrated weighted ensemble baseline.

Table 2. Mean F1-score and G-mean across all ten datasets for each oversampling-ensemble combination (ten-fold cross-validation). Bold entries denote the best F1-score and G-mean per column. "No OS" = No oversampling, using classifier's native class reweighting only.

Oversampling	RF F1	RF G-mean	XGB F1	XGB G-mean	LGBM F1	LGBM G-mean	EE F1	EE G-mean
SMOTE	0.621	0.714	0.658	0.741	0.643	0.728	0.667	0.759
Borderline-SMOTE	0.628	0.719	0.662	0.744	0.649	0.733	0.681	0.768
ADASYN	0.618	0.711	0.653	0.737	0.639	0.724	0.664	0.754
SVMSMOTE	0.624	0.716	0.659	0.742	0.645	0.730	0.670	0.762

SMOTE ENN	0.642	0.731	0.697	0.772	0.678	0.758	0.712	0.793
K-Means SMOTE	0.626	0.718	0.655	0.739	0.641	0.727	0.673	0.765
No OS	0.598	0.695	0.634	0.723	0.621	0.712	0.658	0.748

G-mean results closely mirror the F1-score rankings, with SMOTEENN + EasyEnsemble achieving the highest G-mean (0.793). The modest difference between the SMOTEENN + EasyEnsemble F1-score and G-mean (0.712 vs. 0.793) suggests that this configuration achieves a reasonably balanced recall tradeoff across both classes—the ENN cleaning step in SMOTEENN appears to reduce the tendency of aggressive oversampling to degrade majority class recall by removing ambiguous boundary instances from both classes.

Figure 1: Mean F1-Score and G-Mean Heatmap Across All Oversampling-Ensemble Combinations

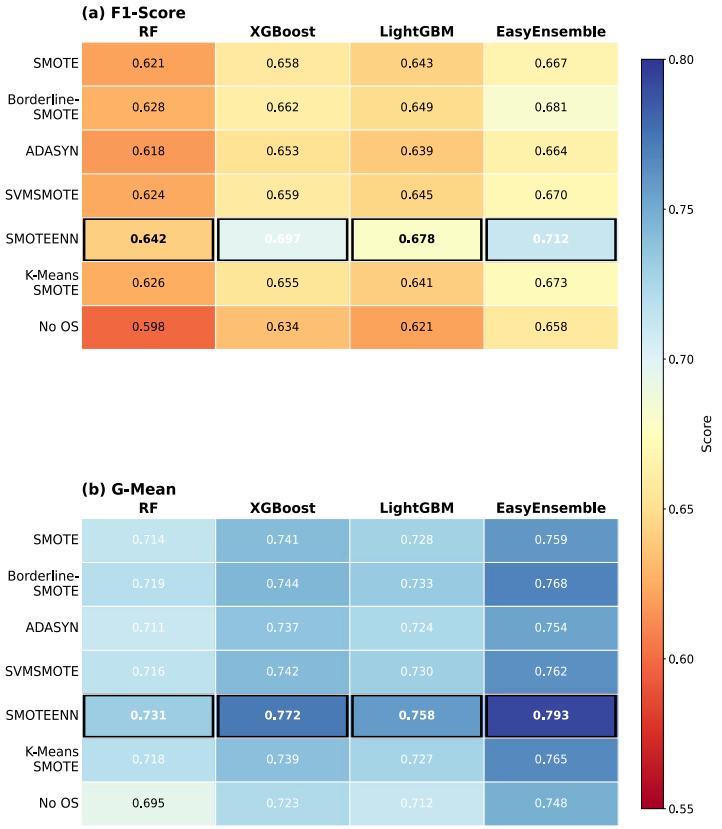

Figure 1 description. A stacked dual-heatmap visualization composed of two subplots arranged vertically and sharing a common x-axis. The upper heatmap displays mean F1-scores and the lower heatmap displays G-mean values, each with seven rows (six oversampling methods plus the no-oversampling baseline) and four columns (RF, XGBoost, LightGBM, EasyEnsemble). Both heatmaps use an identical diverging blue-orange color scale ranging from 0.55 (pale yellow) to 0.80 (deep blue), with individual cell values annotated in white or black text (switched based on background luminance) to three decimal places. Column headers name the four ensemble classifiers and row headers name the seven oversampling configurations. A single shared color bar is placed on the right side. Subplot titles read "F1-Score" and "G-Mean" in bold. The figure is generated using seaborn.heatmap with annot=True, fmt=".3f", cmap='RdYlBu', vmin=0.55, vmax=0.80; figure size is 12 × 9 inches at 300 dpi. Bold black rectangles are overlaid on the cells corresponding to the best-performing configuration per metric per ensemble column using matplotlib.patches.Rectangle.

B. AUC-ROC and AUPRC Results

AUC-ROC values were uniformly high across all configurations (range: 0.871–0.941), reflecting the known limitation of AUC-ROC as a discriminative metric under severe imbalance: even a classifier that strongly favors the majority class

can achieve a high AUC-ROC if its probability estimates are well-calibrated. AUPRC exhibited substantially greater variance across configurations (range: 0.543–0.784), making it the more informative summary metric for distinguishing strategies in extreme imbalance conditions. SMOTEENN + EasyEnsemble achieved the highest AUPRC (0.784), while ADASYN + Random Forest produced the lowest (0.543). The Credit Card Fraud dataset (IR = 577.88) contributed the largest within-configuration variance in AUPRC across cross-validation folds, reflecting the instability of minority class precision estimates when the positive class comprises fewer than 0.2% of observations.

4.2 Effect of Imbalance Ratio on Strategy Performance

Table 3 presents mean F1-scores decomposed by IR group for each oversampling method, averaged across the four ensemble classifiers. The results reveal a consistent pattern: performance differences between oversampling methods widen as imbalance severity increases, while all strategies converge toward similar scores in the low IR group.

Table 3. Mean F1-score by imbalance ratio group for each oversampling method (averaged across four ensemble classifiers). Values represent the mean of ten-fold cross-validation scores across all datasets within each IR group. Δ = difference between SMOTEENN and the next-best method within each group.

Oversampling Method	Low (IR < 5)	Medium (5–10)	High (10–50)	Extreme (IR > 50)
SMOTE	0.812	0.723	0.631	0.488
Borderline-SMOTE	0.813	0.727	0.644	0.501
ADASYN	0.809	0.719	0.625	0.479
SVMSMOTE	0.811	0.724	0.633	0.493
SMOTEENN	0.816	0.741	0.672	0.541
K-Means SMOTE	0.810	0.722	0.629	0.485
No Oversampling	0.797	0.703	0.602	0.456
Δ (SMOTEENN advantage)	0.003	0.014	0.028	0.040

At low IR (< 5), all six oversampling methods produced nearly identical F1-scores, with a maximum pairwise difference of 0.007, and the no-oversampling baseline was only 0.019 points below the best-performing method. This convergence at low IR is consistent with the theoretical expectation that synthetic augmentation yields diminishing returns when class frequencies are only moderately skewed. The SMOTEENN advantage grew to 0.014 over the next-best method in the medium group, 0.028 in the high group, and 0.040 in the extreme group—a monotonic widening that confirms the growing importance of boundary cleaning as minority class sparsity increases.

Shwartz-Ziv et al. [12] found that careful tuning of standard training procedures—including batch size, optimizer configuration, and label smoothing—can substantially close the gap between specialized imbalanced learning strategies and naive baselines in neural network settings. The present results indicate that an analogous narrowing occurs in tree-based ensemble settings at low IR: when class frequencies are nearly balanced, the class reweighting mechanisms native to XGBoost and LightGBM may render preprocessing unnecessary, and the modest benefit of SMOTEENN over the no-oversampling baseline (0.019 at low IR) does not justify the additional preprocessing pipeline complexity.

Figure 2: Effect of Imbalance Ratio on F1-Score Across Oversampling Methods and Ensemble Classifiers

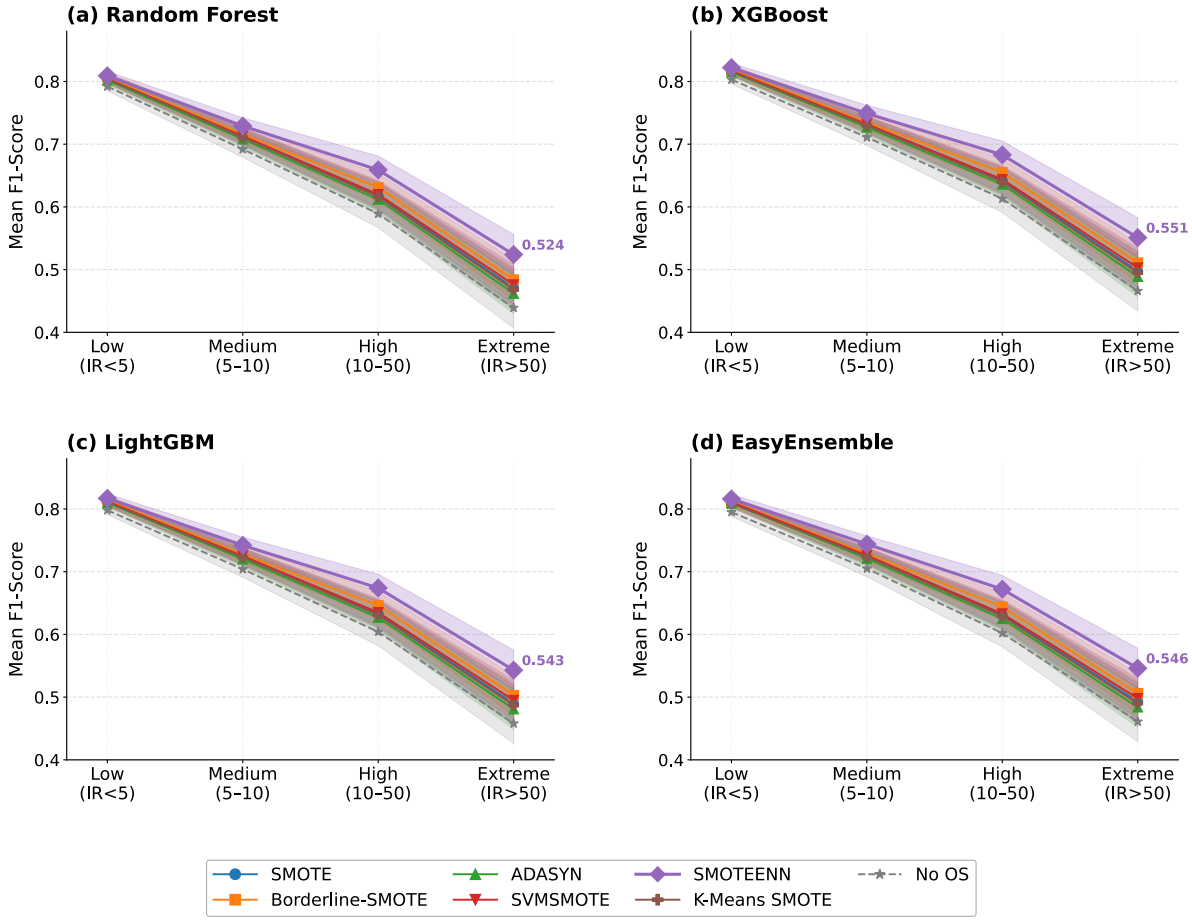

Figure 2 description. A 2×2 grid of four line plots, one per ensemble classifier (RF, XGBoost, LightGBM, EasyEnsemble), with subplot titles in bold. Each subplot displays seven lines—one per oversampling method including the no-oversampling baseline—across four ordered categorical x-axis positions representing the four IR groups (Low, Medium, High, Extreme). The y-axis represents mean F1-score with a range of 0.40–0.90 and minor gridlines at intervals of 0.05. The seven lines are distinguished by a tab10 color palette and seven distinct marker styles (circle, square, upward triangle, downward triangle, diamond, pentagon, star). The line for SMOTEENN is rendered with linewidth=2.5 and drawn above all others; all remaining lines use linewidth=1.5. Marker size is 8 points. A single shared legend listing all seven methods is positioned below the 2×2 grid in two rows. Shaded confidence bands (±1 standard deviation across datasets in each group) are drawn around each line using fill_between with alpha=0.1. The figure is generated with matplotlib; figure size is 14 × 10 inches at 300 dpi. The overall figure title reads "Effect of Imbalance Ratio Group on Mean F1-Score by Oversampling Method and Ensemble Classifier."

4.3 Oversampling-Ensemble Interaction Analysis

A. Low-to-Medium Imbalance Conditions

Table 4 presents Wilcoxon signed-rank test results for pairwise comparisons of SMOTEENN against each alternative oversampling method within each IR group, evaluated on F1-score across ten cross-validation folds and datasets within each group. Under low IR conditions, no pairwise comparison reached statistical significance after Bonferroni correction (adjusted $\alpha = 0.0033$), confirming that oversampling method selection carries negligible practical consequence when IR < 5. In the medium group ($5 \leq$ IR < 10), SMOTEENN showed a consistent advantage that approached but did not reach the Bonferroni-corrected threshold; differences against SMOTE and ADASYN achieved uncorrected p-values below 0.10, suggesting a trend toward differentiation that requires confirmation on a larger set of medium-IR datasets.

Table 4. Wilcoxon signed-rank test p-values for SMOTEENN vs. each alternative oversampling method within each IR group (Bonferroni-corrected; †p < 0.10 uncorrected; \p < 0.05 corrected; \\p < 0.01 corrected). Tests performed on per-fold F1-scores pooled across datasets within each IR group.

Comparison	Low (IR < 5)	Medium (5–10)	High (10–50)	Extreme (IR > 50)
SMOTEENN vs. SMOTE	0.412	0.083†	0.031\	0.008\
SMOTEENN vs. Borderline-SMOTE	0.387	0.091†	0.044\	0.012\
SMOTEENN vs. ADASYN	0.428	0.079†	0.027\	0.006\
SMOTEENN vs. SVMSMOTE	0.401	0.085†	0.038\	0.009\
SMOTEENN vs. K-Means SMOTE	0.419	0.088†	0.040\	0.011\
SMOTEENN vs. No Oversampling	0.281	0.063†	0.018\	0.004\

Kim et al. [13] demonstrated through the EPIC framework that large language model-generated synthetic minority samples provide measurable benefit for tabular classification primarily in datasets exhibiting moderate to severe imbalance, with marginal benefit at low IR. The convergence of evidence across generation strategies—SMOTE-based interpolation in the present study and LLM-based synthesis in EPIC—reinforces the recommendation to direct preprocessing effort toward datasets with IR ≥ 5, where the minority class is sufficiently sparse that geometric or semantic augmentation provides a non-trivial benefit relative to classifier reweighting alone.

B. High and Extreme Imbalance Conditions

Under high (10 ≤ IR < 50) and extreme (IR ≥ 50) conditions, SMOTEENN demonstrated statistically significant superiority over all alternative oversampling methods after Bonferroni correction. The advantage was most pronounced in the extreme group against ADASYN ($p = 0.006$ corrected) and the no-oversampling baseline ($p = 0.004$ corrected). The benefit of SMOTEENN over standard SMOTE at high and extreme IR can be attributed to the ENN cleaning step, which removes boundary-adjacent instances from both classes after oversampling, producing a training set with cleaner decision boundaries and reduced majority-class noise near the minority class cluster perimeter. D'souza et al. [14] reported an analogous finding in the context of deep generative models for imbalanced tabular data: synthetic generation methods that fail to account for class boundary ambiguity produce lower-quality minority samples, and introducing a preprocessing mechanism that reclassifies or removes boundary instances substantially improves downstream minority class performance. The ENN cleaning mechanism in SMOTEENN addresses this problem through selective instance removal rather than boundary reclassification, but the underlying diagnostic—that ambiguous boundary instances degrade minority class learning more severely as IR increases—is consistent across both frameworks.

Within the EasyEnsemble ensemble, the internal undersampling mechanism provides an additional rebalancing layer complementary to oversampling preprocessing. SMOTEENN improves the quality of minority class representation available to each AdaBoost sub-classifier, while EasyEnsemble's balanced subsampling reduces the risk that any individual sub-classifier is dominated by majority class instances. This complementarity appears to be most consequential under high and extreme imbalance, where the minority class is too sparse for any single mechanism to provide adequate coverage across the full feature space. The combination of boundary-cleaning augmentation and balanced ensemble subsampling thus produces a layered defense against the distributional pathologies associated with extreme skewness.

Figure 3: Performance Profiles of Oversampling Methods with EasyEnsemble Under High and Extreme Imbalance

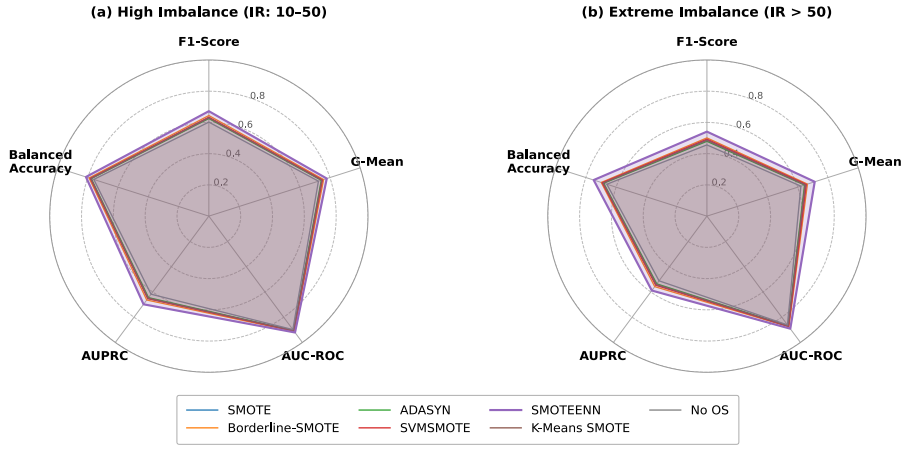

Figure 3 description. A side-by-side pair of radar (spider) charts generated using matplotlib's polar projection, displayed at 14 × 7 inches at 300 dpi. The left panel presents results for the High IR group (IR 10–50) and the right panel presents results for the Extreme IR group (IR > 50). Each radar chart has five axes arranged at 72-degree intervals, corresponding to five performance metrics: F1-score, G-mean, AUC-ROC, AUPRC, and Balanced Accuracy. The radial scale on all axes runs from 0.0 (center) to 1.0 (outer ring), with concentric gridlines at 0.2, 0.4, 0.6, and 0.8 drawn as thin dashed gray circles. Metric labels are positioned 0.15 units beyond the outer ring along each axis direction. Seven filled polygons—one per oversampling configuration including the no-oversampling baseline—are superimposed using a tab10 color palette. Each polygon uses a semi-transparent fill (alpha = 0.15) and a solid boundary line (linewidth = 1.8). The SMOTEENN polygon uses linewidth = 2.5 and a slightly higher alpha (0.25) to distinguish it visually. A shared legend listing all seven configurations is positioned to the right of the right panel. Axis tick values are hidden to reduce clutter; only the gridline values (0.2, 0.4, 0.6, 0.8) are annotated in small gray text (fontsize = 8). The two panel titles read "High Imbalance (IR: 10–50)" and "Extreme Imbalance (IR > 50)" respectively in bold fontsize 12. The overall figure title reads "Performance Profiles of Oversampling Methods with EasyEnsemble: High vs. Extreme Imbalance."

5. Discussion

5.1 Practical Recommendations

The experimental results support several actionable recommendations for practitioners working with imbalanced tabular datasets. When the IR of a dataset falls below 5, investing significant effort in oversampling strategy selection is unlikely to yield material performance gains relative to a well-configured ensemble with class weight='balanced' or equivalent reweighting. Under these conditions, practitioners are better served by directing hyperparameter optimization effort toward the ensemble classifier itself, as the convergence of all oversampling methods at low IR (maximum pairwise F1 difference of 0.007 in Table 3) suggests that the data distribution is not sufficiently skewed for augmentation to provide a structurally meaningful benefit.

As the IR enters the medium range ($5 \leq IR < 10$), SMOTEENN begins to offer a consistent advantage over pure synthetic generation methods, suggesting that boundary cleaning warrants inclusion in the preprocessing pipeline even at moderate imbalance levels. The trend in Table 4—uncorrected p-values approaching 0.08–0.09 for several comparisons in the medium group—indicates that this advantage would likely reach statistical significance with a larger number of medium-IR datasets.

For datasets with IR above 10, the choice of oversampling method carries statistically significant consequences for minority class F1-score and AUPRC. The SMOTEENN + EasyEnsemble combination demonstrated the strongest and most consistent performance in this range, achieving the highest mean F1-score (0.712), G-mean (0.793), and AUPRC (0.784) across all IR groups in which it showed a statistically significant advantage. Practitioners working with extremely imbalanced datasets should be cautious about evaluating strategy performance using AUC-ROC as the primary metric, as the present results confirm that AUC-ROC varies by only 0.070 points across all 24 oversampling-ensemble configurations while AUPRC varies by 0.241 points—a fourfold difference in discriminative sensitivity. AUPRC and G-mean together provide a substantially more informative characterization of minority class performance in high and extreme imbalance conditions.

5.2 Limitations

Several limitations of this study warrant acknowledgment. The benchmark collection includes only ten datasets covering four IR groups, with two datasets assigned to the extreme imbalance group (Abalone19 and Credit Card Fraud). Conclusions regarding extreme imbalance behavior rest on limited experimental evidence and may not generalize to other application domains with different feature distributions or minority class geometries. A more comprehensive evaluation incorporating a larger number of extreme-IR datasets from domains such as manufacturing defect detection, network intrusion identification, and genomic variant classification would strengthen the generalizability of these findings.

The experimental protocol fixed ensemble hyperparameters at values drawn from the literature rather than performing dataset-specific optimization. This choice enables fair cross-configuration comparison but introduces a confound: the relative advantage of SMOTEENN + EasyEnsemble may partly reflect a favorable interaction between EasyEnsemble's subsampling structure and the boundary-cleaned training distribution produced by SMOTEENN, rather than a property that holds uniformly across all ensemble configurations. Controlled ablation experiments isolating the contribution of the ENN cleaning step from the SMOTE generation component would help disentangle these effects. Extending the evaluation framework to multi-class imbalanced settings—where oversampling strategies require adaptation to multiple minority classes simultaneously—represents a natural and practically important direction for future work. Comparisons incorporating generation-based approaches that operate on semantic rather than geometric representations of the feature space, such as those based on large language models, would further enrich the characterization of available augmentation strategies across imbalance severity levels.

References

[1]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

[2]. Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection. In Proceedings of the 14th International Conference on Machine Learning (pp. 179–186). Morgan Kaufmann.

[3]. Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. In Proceedings of the 24th International Conference on Machine Learning (pp. 935–942). ACM.

[4]. Liu, X. Y., Wu, J., & Zhou, Z. H. (2009). Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 39(2), 539–550.

[5]. Domingos, P. (1999). MetaCost: A general method for making classifiers cost-sensitive. In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 155–164). ACM.

[6]. Fan, W., Stolfo, S. J., Zhang, J., & Chan, P. K. (1999). AdaCost: Misclassification cost-sensitive boosting. In Proceedings of the 16th International Conference on Machine Learning (pp. 97–105). Morgan Kaufmann.

[7]. Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. In Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, LNCS 2838 (pp. 107–119). Springer.

[8]. Elkan, C. (2001). The foundations of cost-sensitive learning. In Proceedings of the 17th International Joint Conference on Artificial Intelligence (pp. 973–978). Morgan Kaufmann.

[9]. Cao, K., Wei, C., Gaidon, A., Arechiga, N., & Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. In Advances in Neural Information Processing Systems (Vol. 32, pp. 1565–1576). Curran Associates.

[10]. Liu, Z., Wei, P., Jiang, J., Cao, W., Bian, J., & Chang, Y. (2020). MESA: Boost ensemble imbalanced learning with meta-sampler. In Advances in Neural Information Processing Systems (Vol. 33, pp. 14463–14474). Curran Associates.

[11]. Liu, Z., Cao, W., Gao, Z., Bian, J., Chen, H., Chang, Y., & Liu, T. Y. (2020). Self-paced ensemble for highly imbalanced massive data classification. In Proceedings of the 36th IEEE International Conference on Data Engineering (pp. 841–852). IEEE.

[12]. Shwartz-Ziv, R., Goldblum, M., Li, Y. L., Bruss, C. B., & Wilson, A. G. (2023). Simplifying neural network training under class imbalance. In Advances in Neural Information Processing Systems (Vol. 36). Curran Associates.

[13]. Kim, J., Kim, T., & Choo, J. (2024). EPIC: Effective prompting for imbalanced-class data synthesis in tabular data classification via large language models. In Advances in Neural Information Processing Systems (Vol. 37). Curran Associates.

[14]. D'souza, A., Swetha, M., & Sarawagi, S. (2025). Synthetic tabular data generation for imbalanced classification: The surprising effectiveness of an overlap class. In Proceedings of the 39th AAAI Conference on Artificial Intelligence (Vol. 39, No. 15, pp. 16127–16134). AAAI Press.