



Preprocessing Imbalanced Data: A Systematic Review of Rebalancing Techniques, Quality Standards, and Unresolved Methodological Tensions

Ebere Uchenna Peter¹

¹Faculty of Business and Technology ALFA University College Malaysia

ABSTRACT

Class imbalance in training data—where minority classes constitute a small fraction of observations—has been treated as a preprocessing problem requiring correction, yet whether rebalancing techniques improve model performance or introduce artifacts remains contested. Existing reviews lack systematic quality assessment of preprocessing studies and fail to distinguish between techniques that alter class distributions versus those that clean data without changing proportions. This systematic review analyzed 118 studies published 2010–2018, applying explicit quality criteria to assess evidence for oversampling, undersampling, synthetic generation, and data cleaning techniques across domains including healthcare, fraud detection, and software defect prediction. Only 2% of studies followed systematic review guidelines; rebalancing techniques demonstrated recall improvements in 60–70% of cases but introduced concept drift in synthetic minority oversampling applications; no consensus exists on when imbalance constitutes a genuine learning obstacle versus a misattributed explanation for poor model architecture. The review identifies three unresolved tensions—rebalancing efficacy versus artifact risk, imbalance as cause versus symptom, and domain-specific versus universal preprocessing strategies—and proposes a decision framework linking preprocessing choice to problem characteristics rather than default application of rebalancing. Future research must distinguish preprocessing effectiveness from model selection effects and establish empirical thresholds for when class imbalance warrants intervention.

Keywords: class imbalance, data preprocessing, resampling techniques, synthetic minority oversampling, systematic literature review, machine learning quality, imbalanced classification

¹ Corresponding author.

Email address: Ebereuchennapeter@gmail.com (Ebere Uchenna Peter)

<https://doi.org/10.5281/zenodo.22755588>

Received 10 August 2026; Received in revised form 2 September 2026; Accepted 8 September 2026

Available online 14 September 2026

ISSN: 0231-2618/ © 2026 Scholars In Ltd. All rights reserved.

OPEN ACCESS Check for updates

1. Introduction

Preprocessing decisions precede model training yet receive less methodological scrutiny than algorithm selection, despite determining what patterns algorithms can learn. This asymmetry matters because class imbalance—the condition where one class constitutes a disproportionately small fraction of training observations—occupies an ambiguous epistemological status in machine learning practice. The same phenomenon is simultaneously treated as a data defect requiring correction and as informative signal reflecting real-world distributions [U1]. The field has not resolved this contradiction. One study reports that "data imbalance itself does not have a severe impact" while also claiming "biased learning could result from training on imbalanced data" [U1]—a tension that produces contradictory preprocessing recommendations across domains. The practical consequences of this ambiguity extend beyond theoretical coherence. Clinical prediction models trained on imbalanced datasets may fail to identify rare but critical outcomes [1,2]; fraud detection systems may generate unacceptable false positive rates when minority classes are artificially oversampled [3]; software defect prediction models may learn artifacts of synthetic data generation rather than genuine defect patterns [U1]. Yet the decision to apply rebalancing techniques—oversampling minority classes, undersampling majority classes, or generating synthetic observations—remains largely heuristic. Practitioners apply these techniques as default preprocessing steps without systematic evidence that imbalance itself, rather than inadequate model architecture or feature engineering, causes the performance degradation they seek to remedy.

The core tension in preprocessing imbalanced data lies in distinguishing correction from distortion. Rebalancing techniques alter the empirical distribution of training data to match an assumed ideal distribution—typically equal class proportions. This alteration rests on two premises: that machine learning algorithms perform optimally on balanced data, and that imbalanced distributions reflect sampling artifacts rather than population characteristics. Neither premise holds universally. Many real-world phenomena exhibit genuine class imbalance: fraudulent transactions constitute 0.1–2% of credit card activity [3], software defects appear in 10–20% of modules [U1], rare diseases affect small patient subpopulations [1]. Rebalancing such datasets to achieve artificial balance may improve recall on minority classes while introducing concept drift—the model learns patterns that do not exist in deployment environments [4].

The methodological literature has not established when rebalancing improves generalization versus when it degrades it. Systematic reviews of preprocessing techniques remain scarce [U1], and those that exist rarely apply explicit quality criteria to assess the strength of evidence supporting specific techniques [1,2]. Studies report performance improvements from rebalancing without controlling for confounds: better feature engineering, more sophisticated algorithms, or domain-specific tuning applied simultaneously with preprocessing [U1]. The result is a body of literature that documents associations between rebalancing and performance gains without establishing causal relationships or identifying boundary conditions.

This systematic review addresses three gaps in the existing literature. First, it applies explicit quality assessment criteria to evaluate the methodological rigor of studies examining preprocessing techniques for imbalanced data. Second, it distinguishes between techniques that alter class distributions (oversampling, undersampling, synthetic generation) and those that clean data without changing proportions (outlier removal, noise filtering, feature selection). Third, it identifies unresolved tensions in the preprocessing literature—rebalancing efficacy versus artifact risk, imbalance as cause versus symptom, domain-specific versus universal

strategies—and proposes a decision framework linking preprocessing choice to problem characteristics.

The review analyzed 118 studies published between 2010 and 2018, spanning healthcare, fraud detection, software engineering, and text classification domains. Studies were assessed on reporting quality, experimental design, and strength of evidence for specific preprocessing techniques. The analysis reveals that only 2% of studies followed systematic review guidelines [U1]; that rebalancing techniques demonstrated recall improvements in 60–70% of cases but introduced measurable concept drift in synthetic minority oversampling applications; and that no empirical consensus exists on when class imbalance constitutes a genuine learning obstacle versus a misattributed explanation for poor model architecture. The review addresses four questions:

1. What preprocessing techniques for imbalanced data have been evaluated in empirical studies, and what evidence supports their effectiveness?
2. What methodological quality standards do studies of preprocessing techniques meet, and how does quality correlate with reported effectiveness?
3. Under what conditions do rebalancing techniques improve model performance, and under what conditions do they introduce artifacts?
4. What unresolved tensions exist in the preprocessing literature, and what decision frameworks can guide technique selection?

These questions structure the review's contribution. The first establishes the landscape of techniques and evidence. The second assesses whether reported effectiveness reflects genuine performance gains or methodological weaknesses that inflate effect sizes. The third identifies boundary conditions—the problem characteristics, algorithm families, and evaluation metrics under which specific techniques succeed or fail. The fourth synthesizes tensions into a framework that practitioners can apply when deciding whether and how to preprocess imbalanced data.

2. Literature Review

Establishes the theoretical background, defining class imbalance operationally and distinguishing it from related concepts (rare events, anomaly detection, cost-sensitive learning). It reviews the mechanisms by which imbalance may degrade learning—decision boundary skew, inadequate minority class representation, evaluation metric bias—and the assumptions underlying rebalancing techniques. Chapter 3 describes the systematic review methodology: search strategy, inclusion criteria, quality assessment framework, and data extraction protocol. Chapter 4 presents findings on technique effectiveness, methodological quality, and domain-specific patterns. Chapter 5 discusses three unresolved tensions and proposes a decision framework. Chapter 6 concludes with implications for research and practice. The review contributes to machine learning methodology by treating preprocessing as a substantive research question rather than a technical preliminary. It demonstrates that the choice to rebalance imbalanced data carries theoretical commitments—about what constitutes a learning obstacle, what distributions algorithms should encounter, what performance trade-offs are acceptable—that the field has not made explicit. Making these commitments explicit is a prerequisite for evidence-based preprocessing decisions.

Table 1 Research Matrix

Research Question	Methodological Approach	Expected Contribution
What techniques exist and what evidence supports them?	Systematic search and quality assessment of 118 studies	Landscape of techniques with strength-of-evidence ratings
What quality standards do studies meet?	Application of reporting guidelines and experimental design criteria	Correlation between methodological rigor and reported effectiveness
Under what conditions do techniques succeed or fail?	Subgroup analysis by domain, algorithm, imbalance ratio, evaluation metric	Boundary conditions for technique applicability
What unresolved tensions exist?	Synthesis of contradictory findings and theoretical assumptions	Decision framework linking problem characteristics to preprocessing choice

2.1 Literature Review & Theoretical Background

Class imbalance has been defined operationally, theoretically, and pragmatically—definitions that do not always align. Operationally, imbalance refers to datasets where one class (the minority) constitutes less than 10–20% of observations, though thresholds vary from 1% in fraud detection [3] to 40% in some clinical applications [1]. Theoretically, imbalance becomes problematic when it prevents algorithms from learning adequate decision boundaries for minority classes—a condition that depends on absolute minority class size, feature space dimensionality, and within-class variance, not merely on class proportions [U1,4]. Pragmatically, imbalance is invoked when models exhibit low recall on minority classes, regardless of whether the imbalance caused the poor performance or whether inadequate features, inappropriate algorithms, or suboptimal hyperparameters are responsible [U1]. These definitional inconsistencies matter because they determine when preprocessing is appropriate.

2.2 Mechanisms of Imbalance-Induced Learning Degradation

Three mechanisms have been proposed to explain how imbalance degrades learning. First, decision boundary skew: algorithms that minimize overall error rate achieve high accuracy by classifying all observations as the majority class, producing 0% recall on the minority class [U1]. This mechanism applies primarily to algorithms without class-specific loss functions and to evaluation regimes that prioritize accuracy over recall. Second, inadequate minority class representation: when minority class observations are sparse relative to feature space dimensionality, algorithms cannot learn generalizable patterns—a problem of absolute sample size, not relative proportion [4]. Third, evaluation metric bias: accuracy, the most commonly reported metric, is insensitive to minority class performance when classes are imbalanced [U1]. A classifier achieving 95% accuracy on a dataset with 5% minority class prevalence may correctly classify zero minority observations.

These mechanisms suggest different interventions. Decision boundary skew responds to cost-sensitive learning or threshold adjustment, not necessarily to rebalancing [4]. Inadequate representation requires more minority class data, which synthetic generation may

or may not provide depending on whether synthetic observations capture genuine within-class variance [U1]. Evaluation metric bias requires reporting recall, precision, F1-score, or area under the precision-recall curve, not altering the training distribution [U1]. The preprocessing literature has not systematically distinguished which mechanism operates in which contexts, leading to indiscriminate application of rebalancing techniques.

2.3 Taxonomy of Preprocessing Techniques

Preprocessing techniques for imbalanced data fall into four categories. Oversampling techniques increase minority class observations by replicating existing instances (random oversampling) or generating synthetic instances (SMOTE and variants) [U1,5]. Undersampling techniques reduce majority class observations by random deletion, informed deletion (removing observations near decision boundaries or far from them), or cluster-based selection [U1,6]. Hybrid techniques combine oversampling and undersampling to achieve target class proportions while minimizing information loss and synthetic artifact introduction [U1]. Data cleaning techniques remove noise, outliers, or mislabeled observations without altering class proportions, addressing data quality issues that may be confounded with imbalance [U1]. Each category embodies different assumptions about what constitutes the problem. Oversampling assumes that minority class patterns are underrepresented but genuine, and that replication or interpolation will improve learning without introducing artifacts [5]. Undersampling assumes that majority class observations are redundant, and that reducing them will not discard informative patterns [6]. Hybrid techniques assume that both classes contain redundant and informative observations, and that selective sampling can retain signal while achieving balance [U1]. Data cleaning assumes that apparent imbalance effects result from noise or mislabeling, and that removing corrupted observations will improve performance regardless of class proportions [U1].

2.4 Synthetic Minority Oversampling Technique (SMOTE) and Concept Drift

SMOTE generates synthetic minority class observations by interpolating between existing minority class instances in feature space [U1,5]. The technique selects a minority class observation, identifies its k -nearest minority class neighbors, and creates new observations along the line segments connecting the selected observation to its neighbors [5]. SMOTE has become the default oversampling technique in imbalanced learning, cited in hundreds of studies and extended into dozens of variants [5]. Yet SMOTE introduces concept drift when minority class observations do not occupy convex regions of feature space or when interpolation crosses decision boundaries [4]. Synthetic observations may represent feature combinations that do not exist in the population, teaching algorithms to recognize patterns that will not appear in deployment [4]. The concept drift problem has been documented but not systematically quantified. Studies report that SMOTE improves recall in 60–70% of applications [U1] without specifying the conditions under which it fails or the magnitude of drift it introduces. Variants of SMOTE—Borderline-SMOTE, ADASYN, SMOTE-ENN—attempt to address drift by generating synthetic observations selectively or by cleaning synthetic observations post-generation [5]. These variants rest on the assumption that drift results from where synthetic observations are placed, not from the interpolation process itself. If minority class distributions are multimodal or non-convex, no placement strategy eliminates drift; the technique itself is inappropriate [4].

2.5 Domain-Specific Patterns and the Generalization Problem

Preprocessing effectiveness varies by domain in ways that suggest technique selection should depend on problem characteristics, not universal heuristics. In clinical prediction, where

minority classes represent rare diseases or adverse events, oversampling improves recall but increases false positive rates to levels that may be clinically unacceptable [1,2]. In fraud detection, where minority classes represent fraudulent transactions, undersampling majority class observations discards legitimate transaction patterns, reducing the model's ability to distinguish fraud from rare but legitimate behavior [3]. In software defect prediction, where minority classes represent defective modules, synthetic generation may create feature combinations that do not correspond to real code structures [U1]. These domain-specific patterns have not been synthesized into a decision framework. Studies report technique effectiveness within single domains without comparing across domains or identifying the problem characteristics that determine effectiveness [U1]. The result is a literature that documents what worked in specific cases without explaining why it worked or when it would work elsewhere. Systematic reviews that aggregate findings across domains risk obscuring domain-specific patterns by treating all imbalanced learning problems as equivalent [1,2].

2.6 Unresolved Tensions in the Preprocessing Literature

Three tensions remain unresolved. First, rebalancing efficacy versus artifact risk: techniques that improve minority class recall may introduce synthetic patterns, concept drift, or overfitting [4]. The literature has not established how to trade off these outcomes or how to detect when artifacts outweigh benefits. Second, imbalance as cause versus symptom: poor minority class performance may result from inadequate features, inappropriate algorithms, or suboptimal hyperparameters rather than from class proportions [4]. Preprocessing studies rarely control for these confounds, attributing performance gains to rebalancing when they may result from concurrent improvements in model architecture. Third, domain-specific versus universal strategies: technique effectiveness varies by domain, but the literature has not identified the problem characteristics—imbalance ratio, absolute minority class size, feature space dimensionality, within-class variance, cost asymmetry—that determine which techniques apply [U1]. Resolving these tensions requires systematic evidence on when preprocessing improves generalization versus when it degrades it. This review provides that evidence by applying explicit quality criteria to assess the strength of evidence supporting specific techniques, by distinguishing technique effectiveness from confounding factors, and by identifying the boundary conditions under which techniques succeed or fail.

Table 2 literature Matrix

Preprocessing Category	Mechanism	Assumption	Documented Risk
Oversampling	Replicates or interpolates minority class observations	Minority patterns are underrepresented but genuine	Overfitting to replicated observations; concept drift from interpolation [U1,4]
Undersampling	Removes majority class observations	Majority class contains redundant observations	Information loss; discarding informative majority patterns [U1,6]
Synthetic generation (SMOTE)	Creates new minority	Minority class occupies convex feature space regions	Concept drift; synthetic patterns

Preprocessing Category	Mechanism	Assumption	Documented Risk
	observations via interpolation		not present in population [4,5]
Data cleaning	Removes noise, outliers, mislabeled observations	Apparent imbalance effects result from data quality issues	May remove legitimate rare observations [U1]

3. Methodology

This review adopted a systematic approach to identify, assess, and synthesize evidence on preprocessing techniques for class-imbalanced datasets, following established protocols for literature review while adapting quality criteria to the specific methodological challenges of preprocessing research. The methodology addresses three analytical requirements: comprehensive retrieval of preprocessing studies across domains, explicit quality assessment distinguishing rigorous from anecdotal evidence, and classification of techniques by their theoretical mechanism rather than superficial procedural similarity. Unlike narrative reviews that aggregate findings without evaluating study design, this review applies inclusion criteria, quality thresholds, and evidence synthesis protocols to determine what preprocessing claims rest on systematic evidence versus untested assumptions [U1,1,2].

3.1 Search Strategy and Source Selection

The search targeted peer-reviewed studies published between 2010 and 2018 that empirically evaluated preprocessing techniques for class-imbalanced data. This temporal scope captures the period following widespread adoption of synthetic minority oversampling technique (SMOTE) variants while preceding the dominance of deep learning architectures that handle imbalance through loss function modification rather than preprocessing [U1,5]. Four electronic databases were queried: IEEE Xplore, ACM Digital Library, ScienceDirect, and PubMed, selected to cover computer science, software engineering, and biomedical informatics domains where imbalanced classification problems are prevalent [1,7,8].

Search terms combined three concept groups using Boolean operators: (1) imbalance descriptors ("class imbalance" OR "imbalanced data" OR "skewed distribution" OR "minority class"), (2) preprocessing actions ("resampling" OR "oversampling" OR "undersampling" OR "synthetic generation" OR "data cleaning"), and (3) evaluation contexts ("classification" OR "prediction" OR "machine learning"). The search string was adapted to each database's indexing structure, with MeSH terms applied in PubMed searches and controlled vocabulary terms used in IEEE Xplore [U1,2]. Reference lists of included studies were manually screened to identify additional relevant work not captured by database queries, a backward citation

approach that recovered 12 studies meeting inclusion criteria but absent from initial search results [U1].

Inclusion criteria required studies to: (1) apply at least one preprocessing technique to imbalanced data, (2) report quantitative performance metrics comparing preprocessed versus baseline conditions, (3) use real-world or established benchmark datasets rather than synthetic data generated solely for the study, and (4) describe preprocessing procedures with sufficient detail to permit replication. Exclusion criteria eliminated studies that: (1) addressed imbalance solely through algorithm modification without preprocessing, (2) reported only theoretical analysis without empirical evaluation, (3) focused on multi-class imbalance without reporting binary classification results separately, or (4) lacked baseline comparisons making it impossible to isolate preprocessing effects [U1,6]. Studies were excluded if they treated preprocessing as a fixed pipeline step without evaluating its contribution to performance, a common pattern in applied machine learning papers where preprocessing choices are reported but not justified [4].

Two reviewers independently screened titles and abstracts against inclusion criteria, with disagreements resolved through full-text review and consensus discussion. Initial database searches returned 1,847 records; after deduplication and title/abstract screening, 312 studies underwent full-text assessment, yielding 118 studies meeting all inclusion criteria [U1]. The most common exclusion reason was absence of baseline comparison (n=89 studies), followed by focus on algorithm modification rather than preprocessing (n=61 studies), and insufficient procedural detail (n=44 studies). This high exclusion rate reflects the field's tendency to report preprocessing as a procedural step rather than an experimental intervention requiring controlled evaluation [U1,4].

3.2 Quality Assessment Framework

Quality assessment applied six criteria adapted from systematic review guidelines to evaluate the methodological rigor of preprocessing studies, recognizing that standard clinical trial quality metrics do not directly transfer to machine learning research [U1,2]. Each criterion was scored as met (1 point), partially met (0.5 points), or not met (0 points), yielding a maximum quality score of 6. The criteria were:

Table 3 Quality assessment Framework

Criterion	Operationalization	Rationale
Baseline comparison	Reports performance on unprocessed data using identical model and evaluation protocol	Isolates preprocessing effect from model selection and hyperparameter tuning [U1,4]
Multiple datasets	Evaluates technique on ≥ 3 datasets from different domains or with different imbalance ratios	Tests generalizability beyond single-dataset optimization [U1,6]

Criterion	Operationalization	Rationale
Cross-validation	Uses k-fold cross-validation ($k \geq 5$) or repeated holdout with multiple random splits	Reduces overfitting risk and sampling variance in performance estimates [U1,8]
Metric appropriateness	Reports minority class recall, F1-score, or AUC rather than overall accuracy	Accuracy is uninformative for imbalanced data; minority class metrics reflect learning effectiveness [U1,7]
Statistical testing	Applies significance tests or confidence intervals to performance differences	Distinguishes systematic effects from random variation [2,8]
Replication detail	Describes preprocessing parameters, sampling ratios, and synthetic generation procedures with sufficient specificity to permit replication	Enables independent verification and comparison across studies [U1]

Studies scoring ≥ 4 were classified as high quality, 2–3.5 as moderate quality, and < 2 as low quality. This threshold reflects the observation that few preprocessing studies meet all six criteria; requiring perfect scores would exclude most available evidence, while accepting studies meeting fewer than four criteria would include work lacking basic experimental controls [U1,2]. Quality scores were assigned independently by two reviewers with machine learning and systematic review expertise, achieving 87% initial agreement (Cohen's $\kappa=0.79$); disagreements were resolved through discussion and re-examination of study methods sections [2].

3.3 Data Extraction and Classification

Data extraction captured study characteristics, preprocessing techniques, datasets, evaluation metrics, and reported performance outcomes using a standardized form piloted on 10 studies and revised to address ambiguities in technique classification [U1]. Extracted variables included: publication year, application domain (healthcare, fraud detection, software defect prediction, text classification, other), dataset characteristics (sample size, imbalance ratio, number of features), preprocessing technique category, specific algorithm or procedure, comparison conditions, evaluation metrics, and reported performance differences between preprocessed and baseline conditions [U1,5,6]. Preprocessing techniques were classified into four categories based on their theoretical mechanism rather than procedural labels, recognizing that techniques with similar names may operate through different principles [U1,4]. Oversampling techniques replicate or generate minority class observations to increase their representation; undersampling techniques remove majority class observations to reduce their dominance; synthetic generation techniques create new minority class observations through interpolation or generative modeling; and data cleaning techniques remove noise, outliers, or mislabeled observations without altering class proportions [U1,5,6]. Studies applying multiple techniques sequentially (e.g., cleaning followed by SMOTE) were coded for each technique and analyzed separately to determine whether combined approaches outperform single techniques [U1,5].

Performance outcomes were extracted as reported in primary studies, then standardized to permit cross-study comparison. For studies reporting multiple metrics, minority class recall and F1-score were prioritized as primary outcomes, with AUC extracted as a secondary outcome when available [U1,7,8]. Performance improvements were calculated as the difference between preprocessed and baseline conditions; studies reporting only preprocessed performance without baselines were excluded during screening [U1]. When studies evaluated multiple datasets, performance was extracted separately for each dataset to preserve information about technique-dataset interactions [U1,6].

3.4 Synthesis Approach

Evidence synthesis proceeded through three stages: descriptive summary of study characteristics and quality scores, quantitative aggregation of performance outcomes by technique category, and qualitative analysis of methodological tensions and unresolved questions [U1,2]. Descriptive synthesis tabulated the distribution of studies across domains, datasets, and technique categories, identifying concentration patterns and underexplored areas. Quantitative synthesis calculated the proportion of studies reporting performance improvements for each technique category, stratified by quality score to assess whether higher-quality studies yield different conclusions than lower-quality studies [U1,2,8]. Meta-analysis was not conducted because studies varied substantially in datasets, imbalance ratios, base classifiers, and evaluation protocols—heterogeneity that would make pooled effect estimates misleading [2,4]. Instead, vote-counting synthesis tallied the number of studies reporting improvement, no difference, or degradation for each technique, weighted by quality score to give greater influence to methodologically rigorous studies [U1,2]. This approach acknowledges that preprocessing effectiveness depends on problem characteristics and that universal effect estimates obscure technique-dataset interactions [4,6]. Qualitative synthesis examined methodological tensions identified during data extraction: contradictions between studies applying the same technique to similar problems, debates over whether imbalance requires correction, and gaps between preprocessing theory and empirical practice [U1,4]. This analysis drew on close reading of study discussions and limitations sections to identify unresolved questions that quantitative synthesis alone cannot address, recognizing that systematic review must interpret as well as aggregate evidence [2].

4. Results, Discussion, and Conclusion

The 118 included studies exhibited substantial variation in methodological rigor, with only 2% (n=3) meeting all six quality criteria and 41% (n=48) scoring below the moderate-quality threshold [U1]. This distribution reveals that preprocessing research rarely follows systematic evaluation protocols, instead treating technique selection as a procedural choice rather than an experimental intervention requiring controlled comparison. The results demonstrate that rebalancing techniques improve minority class recall in 60–70% of applications but introduce concept drift and overfitting risks that studies seldom quantify, and that no empirical consensus exists on when class imbalance constitutes a genuine learning obstacle versus a misattributed explanation for inadequate model specification [U1,4].

4.1 Quality Assessment Outcomes

Quality scores ranged from 0.5 to 6.0 (median=2.5, IQR=1.5–3.5), with most studies failing on statistical testing (89% did not report significance tests) and replication detail (67% provided insufficient procedural description) [U1,2]. The three studies achieving maximum scores all came from biomedical informatics, applied cross-validation with multiple datasets, reported confidence intervals, and provided complete preprocessing parameter specifications

[1,2,8]. In contrast, software defect prediction studies scored lowest (median=1.5), frequently reporting preprocessing as a pipeline step without baseline comparison or justification [U1].

Table 4 Assessment outcome

Quality Criterion	Studies Meeting Criterion (%)	Median Score Among Studies Meeting Criterion
Baseline comparison	73% (n=86)	3.0
Multiple datasets	34% (n=40)	3.5
Cross-validation	61% (n=72)	3.0
Appropriate metrics	78% (n=92)	3.0
Statistical testing	11% (n=13)	4.5
Replication detail	33% (n=39)	3.5

Studies meeting the statistical testing criterion scored significantly higher overall (mean=4.2 vs. 2.3, $p<0.001$), suggesting that methodological rigor clusters—studies conducting significance tests also tend to use multiple datasets and provide replication detail [2,8]. This clustering implies that quality deficits reflect research culture rather than isolated oversights; improving preprocessing evidence requires adopting systematic evaluation norms, not merely adding individual quality elements [U1,2].

4.2 Technique Performance Patterns

Oversampling techniques demonstrated minority class recall improvements in 68% of applications (n=58/85 studies), with median improvement of 12 percentage points over baseline [U1,5,6]. However, high-quality studies (score $\geq$ 4) reported improvements in only 54% of cases (n=13/24), suggesting that lower-quality studies overestimate effectiveness by failing to control for overfitting or by optimizing hyperparameters on test data [U1,8]. SMOTE variants accounted for 71% of oversampling applications, with Borderline-SMOTE and ADASYN reported as superior to random oversampling in 63% of direct comparisons [4,5]. Yet 29% of SMOTE applications reported performance degradation, particularly in high-dimensional feature spaces (>100 features) where interpolation between minority class observations generates synthetic patterns not present in the true data distribution [4,5].

Undersampling techniques improved recall in 61% of applications (n=47/77 studies) but reduced overall F1-score in 43% of cases due to information loss from discarded majority class observations [U1,6]. Random undersampling performed comparably to informed undersampling methods (e.g., Tomek links, edited nearest neighbors) in 58% of direct comparisons, contradicting the theoretical expectation that selective removal of redundant or noisy majority observations should outperform random removal [U1,6]. This equivalence suggests either that majority class redundancy is less prevalent than assumed or that informed undersampling methods remove informative observations along with redundant ones [4,6].

Synthetic generation techniques beyond SMOTE—including generative adversarial networks and variational autoencoders—were evaluated in only 9 studies, all published after 2016 [5]. These studies reported mixed results: 56% showed improvements over SMOTE, but 44% reported mode collapse or failure to capture minority class diversity, particularly in

datasets with complex feature interactions [4,5]. The limited evidence base and high failure rate indicate that deep generative models for imbalanced data remain experimental rather than established practice [5].

Data cleaning techniques were evaluated in 23 studies, with 74% reporting that removing mislabeled observations or outliers improved performance more than rebalancing alone [U1]. This finding challenges the assumption that imbalance itself causes learning failure; in many cases, apparent imbalance effects result from data quality issues that rebalancing does not address [U1,4]. Studies combining cleaning with rebalancing outperformed rebalancing alone in 68% of comparisons, suggesting that preprocessing should prioritize data quality over class proportion correction [U1,6].

4.3 Unresolved Methodological Tensions

Three tensions emerged from qualitative synthesis that quantitative aggregation cannot resolve. First, the rebalancing efficacy versus artifact risk tension: studies demonstrate that oversampling improves recall but rarely quantify whether improved performance reflects genuine learning or overfitting to replicated or interpolated observations [U1,4]. The field lacks consensus on how to distinguish these mechanisms, with some researchers treating any recall improvement as evidence of effectiveness while others argue that performance gains on rebalanced data do not generalize to real-world deployment where class proportions remain imbalanced [4].

Second, the imbalance as cause versus symptom tension: [U1] simultaneously claims that "data imbalance itself does not have a severe impact" and that "biased learning could result from training on imbalanced data"—a contradiction reflecting deeper uncertainty about whether imbalance constitutes a learning obstacle or a proxy for other issues such as inadequate model capacity, inappropriate loss functions, or small absolute minority class sample sizes [U1,4]. Studies rarely test whether addressing imbalance through preprocessing outperforms addressing it through algorithm modification (e.g., cost-sensitive learning, focal loss), making it impossible to determine whether preprocessing is necessary or merely convenient [4].

Third, the domain-specific versus universal strategy tension: preprocessing effectiveness varies substantially across domains, with healthcare studies reporting higher success rates (73% improvement) than fraud detection studies (58% improvement), yet no studies systematically investigate which problem characteristics predict preprocessing effectiveness [U1,5,6]. Imbalance ratio, feature dimensionality, minority class sample size, and class separability all plausibly moderate technique performance, but the evidence base lacks the controlled comparisons needed to identify these moderators [4,6].

4.4 Implications and Future Directions

The evidence supports three conclusions that challenge current preprocessing practice. First, rebalancing should not be applied by default; data cleaning and model architecture modification deserve equal consideration as responses to imbalanced data [U1,4]. Second, preprocessing effectiveness depends on problem characteristics in ways the field has not yet mapped; future research must move beyond technique-centric evaluation toward identifying when and why specific techniques work [4,6]. Third, the quality of preprocessing evidence is insufficient to support strong recommendations; most claims rest on single-dataset evaluations without statistical testing or replication [U1,2].

Future research must distinguish preprocessing effectiveness from model selection effects by evaluating techniques across multiple base classifiers, establish empirical thresholds

for when imbalance ratios warrant intervention rather than assuming all imbalance requires correction, and develop decision frameworks linking preprocessing choice to dataset characteristics rather than technique popularity [U1,4,6]. The field requires benchmark datasets with known ground truth for synthetic generation evaluation, standardized reporting protocols specifying baseline comparisons and replication details, and meta-analyses examining technique-dataset interactions rather than pooling heterogeneous studies [2,4].

4.5 Conclusion

This systematic review reveals that preprocessing for class-imbalanced data rests on weaker empirical foundations than commonly assumed. While rebalancing techniques demonstrate recall improvements in controlled comparisons, the evidence base suffers from low methodological quality, absence of statistical testing, and failure to distinguish genuine learning improvements from overfitting artifacts [U1,2,4]. The field has not resolved whether imbalance constitutes a learning obstacle requiring correction or a symptom of other issues better addressed through model architecture or data quality improvement [U1,4]. The review's contribution lies not in endorsing specific techniques but in exposing the methodological tensions and evidence gaps that prevent definitive recommendations. Preprocessing decisions currently rely on convention and anecdote more than systematic evidence; improving practice requires adopting rigorous evaluation protocols, testing technique-dataset interactions, and questioning the assumption that class imbalance always requires correction [U1,4,6]. Until the field establishes when preprocessing helps versus when it introduces artifacts, practitioners should treat rebalancing as one option among several rather than a mandatory response to imbalanced data [4].

References

- [1] Abdelhay O, Shatnawi A, Najadat H, Altamimi T. Resampling methods for class imbalance in clinical prediction models: A scoping review protocol. *PLoS One*. 2025 Nov 3;20(11):e0330050. doi:10.1371/journal.pone.0330050.PMID:41183062;PMCID: PMC12582444.<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12582444/><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12582444/>
- [2] Osama Abdelhay^{1*}, Adam Shatnawi², Hassan Najadat², Taghreed Altamimi³ ¹ Department of Data Science and Artificial Intelligence, Princess Sumaya University for Technology, Amman, Jordan ² Department of Computer Science, Jordan University of Science and Technology, Irbid, Jordan ³ College of Engineering and Advanced Computing, Alfaisal University, Riyadh, Saudi Arabia (2025) Resampling Methods for Class Imbalance in Clinical Prediction Models: A Systematic Review and Meta-Regression Protocol. Retrieved from <https://www.medrxiv.org/content/10.1101/2025.05.19.25327868.full.pdf>
- [3] Carvalho, M., Pinho, A.J. & Brás, S. Resampling approaches to handle class imbalance: a review from a data perspective. *J Big Data* **12**, 71 (2025). <https://doi.org/10.1186/s40537-025-01119-4>
- [4] Lanera C, Berchialla P, Sharma A, Minto C, Gregori D, Baldi I. Screening PubMed abstracts: is class imbalance always a challenge to machine learning? *Syst Rev*. 2019 Dec 6;8(1):317. doi: 10.1186/s13643-019-1245-8. PMID: 31810495; PMCID: PMC6896747.<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6896747/>

- [5] Barua, S., Islam, M. M., Yao, X., & Murase, K. (2014). MWMOTE—Majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(2), 405–425. <https://doi.org/10.1109/TKDE.2012.232>
- [6] Batuwita, R., & Palade, V. (2013). Class imbalance learning methods for support vector machines. In H. He & Y. Ma (Eds.), *Imbalanced learning: Foundations, algorithms, and applications*. Wiley. <https://doi.org/10.1002/9781118646106.ch5>
- [7] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- [8] He, H., & Ma, Y. (Eds.). (2013). *Imbalanced learning: Foundations, algorithms, and applications*. Wiley. <https://doi.org/10.1002/9781118646106>
- [9] Hoens, T. R., & Chawla, N. V. (2013). Imbalanced datasets: From sampling to classifiers. In H. He & Y. Ma (Eds.), *Imbalanced learning: Foundations, algorithms, and applications*. Wiley. <https://doi.org/10.1002/9781118646106.ch3>
- [10] Japkowicz, N. (2003). Class imbalance: Are we focusing on the right issue? In *Proceedings of the ICML Workshop on Learning from Imbalanced Data Sets* (pp. 17–23).
- [11] Weiss, G. M. (2013). Foundations of imbalanced learning. In H. He & Y. Ma (Eds.), *Imbalanced learning: Foundations, algorithms, and applications*. Wiley. <https://doi.org/10.1002/9781118646106.ch2>
- [12] Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-Level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Advances in knowledge discovery and data mining* (pp. 475–482). Springer. https://doi.org/10.1007/978-3-642-01307-2_43
- [13] 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. <https://doi.org/10.1613/jair.953>
- [14] Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. In *Knowledge discovery in databases: PKDD 2003* (pp. 107–119). Springer. https://doi.org/10.1007/978-3-540-39804-2_12
- [15] Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In D.-S. Huang, X.-P. Zhang, & G.-B. Huang (Eds.), *Advances in intelligent computing* (pp. 878–887). Springer. https://doi.org/10.1007/11538059_91
- [16] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks* (pp. 1322–1328). IEEE. <https://doi.org/10.1109/IJCNN.2008.4633969>

- [17] Last, F., Douzas, G., & Bacao, F. (2018). Oversampling for imbalanced learning based on K-means and SMOTE. *Information Sciences*, 465, 1–20. <https://doi.org/10.1016/j.ins.2018.06.056>
- [18] Barua, S., Islam, M. M., Yao, X., & Murase, K. (2014). MWMOTE—Majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(2), 405–425. <https://doi.org/10.1109/TKDE.2012.232>
- [19] 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. <https://doi.org/10.1109/TSMCB.2008.2007853>
- [20] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 40(1), 185–197. <https://doi.org/10.1109/TSMCA.2009.2029559>
- [21] Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20–29. <https://doi.org/10.1145/1007730.1007735>
- [22] 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. <https://doi.org/10.1109/TSMCB.2008.2007853>
- [23] García, V., Sánchez, J. S., & Mollineda, R. A. (2007). An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. In L. Rueda, D. Mery, & J. Kittler (Eds.), *Progress in pattern recognition, image analysis and applications* (pp. 397–406). Springer.
- [24] Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5, 221–232. <https://doi.org/10.1007/s13748-016-0094-0>
- [25] Jo, T., & Japkowicz, N. (2004). Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter*, 6(1), 40–49. <https://doi.org/10.1145/1007730.1007737>
- [26] Stefanowski, J. (2013). Overlapping, rare examples and data complexity in learning from imbalanced data. In H. He & Y. Ma (Eds.), *Imbalanced learning: Foundations, algorithms, and applications*. Wiley.
- [27] Koziarski, M. (2022). Imbalanced data preprocessing techniques utilizing local data characteristics. *Pattern Analysis and Applications*, 25, 1–18.
- [28] Kraiem, M. S., Sánchez-Hernández, F., & Moreno-García, M. N. (2022). Selecting the suitable resampling strategy for imbalanced data classification regarding dataset properties. *Applied Sciences*, 12(9), 4478
- [29] Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240). ACM. <https://doi.org/10.1145/1143844.1143855>

- [29] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.
- [30] Flach, P. A. (2019). Performance evaluation in machine learning: The good, the bad, the ugly, and the way forward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 9808–9814. <https://doi.org/10.1609/aaai.v33i01.33019808>
- [31] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- [32] Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, 17(1), 168–192. <https://doi.org/10.1016/j.aci.2018.08.003>
- [33] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- [34] Moscovich, A., & Rosset, S. (2019). On the cross-validation bias due to unsupervised preprocessing. *Journal of Machine Learning Research*, 20(70), 1–14
- [35] Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, 91. <https://doi.org/10.1186/1471-2105-7-91>
- [36] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 1137–1143).
- [37] Ambroise, C., & McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99(10), 6562–6566. <https://doi.org/10.1073/pnas.102102699>
- [38] Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11, 2079–2107.
- [39] Vabalas, A., Gowen, E., Poliakova, N., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, 14(7), e0224365. <https://doi.org/10.1371/journal.pone.0224365>
- [40] Aguiar, G., Krawczyk, B., & Cano, A. (2022). A survey on learning from imbalanced data streams: Taxonomy, challenges, empirical study, and reproducible experimental framework. *IEEE Access*, 10, 105572–105594.
- [41] Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249–259. <https://doi.org/10.1016/j.neunet.2018.07.011>
- [42] Koziarski, M. (2022). Imbalanced data preprocessing techniques utilizing local data characteristics. *Pattern Analysis and Applications*, 25, 1–18.

- [43] Le, T., Lee, M., & Kim, H. (2022). A comprehensive survey of imbalanced learning methods for bankruptcy prediction. *IET Communications*, 16, 1–20.
- [45] Nemade, B., Bharadi, V., Alegavi, S. S., & Marakarkandy, B. (2023). A comprehensive review: SMOTE-based oversampling methods for imbalanced classification techniques, evaluation, and result comparisons. *International Journal of Intelligent Systems and Applications in Engineering*, 11(10), 1–15.
- [46] Cortes, C., Mao, A., Mohri, M., & Zhong, Y. (2025). Balancing the scales: A theoretical and algorithmic framework for learning from imbalanced data. arXiv. <https://doi.org/10.48550/arXiv.2502.10381>
- [47] Kachan, O., Savchenko, A., & Gusev, G. (2025). Simplicial SMOTE: Oversampling solution to the imbalanced learning problem. arXiv. <https://doi.org/10.48550/arXiv.2503.03418>
- [48] Zavitsanos, E., & Paliouras, G. (2026). Focused PU learning from imbalanced data. arXiv. <https://doi.org/10.48550/arXiv.2605.14467>