NADOS: Reliability-Gated Difficulty-Aware Oversampling for Noisy Imbalanced Classification

Ednel Ashraff Misran*, Syahid Anuar, Adam Mohd Khairuddin
Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia

***Abstract*—Class imbalance becomes more challenging when minority underrepresentation is accompanied by label noise, because synthetic oversampling may amplify unreliable local structures if noisy minority samples are used as interpolation seeds. This study proposes Noise-Aware Adaptive-Difficulty Oversampling (NADOS), which separates the assessment of seed trustworthiness from the allocation of synthesis effort. NADOS evaluates each minority instance through two local criteria. A reliability score determines whether the instance is suitable for synthetic generation, while a difficulty score assigns synthesis priority among eligible instances. The method is evaluated on 22 binary imbalanced benchmark datasets, five controlled label-noise levels, four classifier families, and nine oversampling methods. Across the full benchmark, NADOS obtains average F1 = 0.7428, G-mean = 0.8511, balanced accuracy = 0.8552, and AUPRC = 0.7700. These results give NADOS the strongest average F1, G-mean, and balanced accuracy, while remaining competitive on AUPRC. Non-parametric statistical tests show significant differences among the compared methods and indicate that NADOS is consistently competitive under noisy imbalanced learning conditions. Separating seed reliability from synthesis difficulty provides a practical strategy for noisy imbalanced classification. The reliability gate limits unsafe seed usage, while the difficulty score preserves attention to difficult but trustworthy minority samples.**

***Keywords*—Imbalanced classification; label noise; oversampling; SMOTE; noisy data; minority learning**

I. Introduction

Class imbalance remains a persistent challenge in supervised learning because many classifiers favor the majority class when the training distribution is skewed. This issue is critical in tasks where the minority class is the main class of interest, such as medical screening, fault detection, fraud analysis, and anomaly identification. The problem becomes more difficult when the training labels are noisy, because the learner must recover a sparse minority concept from data that may contain incorrect or locally inconsistent labels [1, 2, 3].

Synthetic oversampling is widely used to address class imbalance. SMOTE and its variants increase minority representation by generating artificial minority samples instead of duplicating existing observations. However, synthetic generation is not always beneficial. If a seed instance is noisy, isolated, or located in an unstable overlap region, interpolation can create misleading samples and reinforce incorrect local structure. This risk is especially important in noisy imbalanced learning, where the minority class is sparse and more vulnerable to local corruption [4, 5, 6].

Recent oversampling methods reduce this risk through adaptive generation, boundary information, local density, cleaning mechanisms, and noise-aware weighting. These methods have improved synthetic sample generation, but a key design issue remains: difficult minority instances are not difficult for the same reason. Some difficult samples are informative boundary cases, while others are unreliable because they are mislabeled, isolated, or structurally inconsistent. Treating both cases alike can direct synthesis effort toward unsafe seeds. Conversely, focusing only on safe regions can neglect useful hard cases near the decision boundary [7, 8, 9, 10].

This study proposes Noise-Aware Adaptive-Difficulty Oversampling (NADOS) to address this distinction. NADOS separates minority-seed assessment into two decisions. Reliability acts as an eligibility gate: a minority instance must be sufficiently trustworthy before it can guide synthetic generation. Difficulty then acts as an allocation signal: among reliable seeds, harder instances receive greater synthesis priority. This design aims to avoid unsafe generation while retaining useful boundary information.

The main contributions are fourfold: a reliability-gated and difficulty-aware oversampling method for noisy imbalanced classification; a local scoring mechanism that separates seed trustworthiness from synthesis priority; a controlled benchmark evaluation across 22 binary imbalanced datasets, five label-noise levels, four classifier families, and nine oversampling methods; and non-parametric statistical validation against established SMOTE-family and noise-aware oversampling baselines.

The remainder of this study is organized as follows. Section II reviews related work. Section III presents the proposed NADOS method. Section IV describes the experimental design. Section V reports and discusses the results. Section VI concludes the study.

II. Related Work

This section reviews the main oversampling families that motivate NADOS. It first summarizes classical data-level methods, then discusses adaptive, boundary-oriented, cleaning-based, and noise-aware variants. The review emphasizes not only what each method does, but also what previous studies suggest about its behavior under overlap, sparsity, and label-noise conditions.

A. Classical Oversampling for Imbalanced Learning

Class imbalance has been widely studied because many standard learning algorithms optimize global error and

*Corresponding author.

therefore tend to favor the majority class when the class distribution is skewed [1, 2]. Data-level methods remain attractive because they can be applied before classifier training and do not require changes to the underlying model. Random oversampling is the simplest form of this strategy: minority instances are duplicated until the class distribution becomes less skewed. Although this can reduce majority-class dominance, it may also increase redundancy and overfitting because the minority distribution is enlarged by repetition rather than by new information.

SMOTE changed data-level imbalance learning by generating synthetic minority instances through interpolation between neighboring minority samples [4]. Instead of duplicating observations, SMOTE creates new samples along line segments between a minority seed and one of its minority nearest neighbors. This idea made synthetic oversampling a dominant baseline for imbalanced classification and motivated many later variants. Its main advantage is generality: it can be combined with many classifiers and applied to tabular datasets without requiring a problem-specific cost matrix.

However, SMOTE also introduced a structural risk. The method assumes that minority neighbors provide useful interpolation directions. When the minority seed is noisy, isolated, or located in a majority-dominated overlap region, interpolation can create synthetic samples that blur the class boundary or reinforce incorrect local structure. This problem is especially important when class imbalance and label noise occur together. In such cases, the learner is not only deprived of minority examples, but it may also receive synthetic minority points generated from unreliable seeds. Reviews of imbalanced learning repeatedly identify the structure of minority examples, local overlap, and noisy labels as factors that can make resampling unreliable [5, 1, 2].

Empirical studies have consistently shown that synthetic oversampling can improve minority recognition compared with simple resampling, but the gains depend strongly on data geometry and classifier behavior. The original SMOTE study reported improved ROC behavior relative to random over-sampling and under-sampling on several benchmark tasks [4]. Later comparative studies found that SMOTE combined with cleaning rules can reduce overlap effects in some settings, but no single resampling method is uniformly best across datasets [11, 5]. These outcomes support the need for methods that respond to local structure rather than applying a single global balancing rule.

B. Adaptive, Locality-Aware, and Noise-Aware Oversampling

One major extension of SMOTE is adaptive generation. ADASYN assigns more synthetic samples to minority instances that are harder to learn, using local class composition to estimate difficulty [7]. This is an important shift because it recognizes that not all minority samples should receive equal generation effort. The method attempts to move the decision boundary toward difficult minority examples by increasing their representation. The limitation is that local difficulty is not always equivalent to usefulness. A sample may be difficult because it lies near an informative boundary, but it may also be difficult because it is mislabeled, isolated, or locally inconsistent.

Boundary-oriented approaches address a related issue. Borderline-SMOTE focuses on minority samples near the class boundary, based on the assumption that boundary samples are more informative for discrimination than samples located deep inside safe minority regions [12]. SVM-SMOTE similarly uses support-vector information to guide generation near the decision boundary [13]. These methods are important competitors because they explicitly move away from uniform minority interpolation. Nevertheless, their emphasis on boundary regions can become problematic under label noise. Boundary and overlap regions are precisely where unreliable labels and ambiguous neighborhoods are more likely to appear.

Another line of work combines oversampling with data cleaning. SMOTE-Tomek and SMOTE-ENN use cleaning rules after synthetic generation to remove overlapping or locally inconsistent samples [11]. These hybrid approaches are useful because they acknowledge that oversampling can introduce ambiguous points. However, post-generation cleaning does not fully solve the seed-selection problem. If unreliable seeds guide interpolation, the method may generate poor candidates before the cleaning step is even applied, a concern also reflected in filtering-based approaches for noisy and borderline examples [6]. This motivates methods that improve synthetic generation earlier in the pipeline.

Noise-aware and density-aware methods attempt to control this risk more directly. WRND is particularly relevant because it uses relative neighborhood density to adapt oversampling for imbalanced noisy classification [8]. Other recent methods use clustering, safe-region estimation, local density, or filtering to improve robustness. These methods show that the field has moved beyond simple class balancing toward local structure-aware generation. However, many approaches still rely on a single local signal to influence both seed selection and synthesis intensity. This creates a design tension: the same evidence is often used to decide whether a sample should be trusted and how much it should be emphasized.

Prior experimental work also shows that adaptive and locality-aware methods can improve performance when the minority class contains informative boundary regions, but the same mechanisms may overemphasize unreliable points when label noise is present. ADASYN was reported to improve learning by concentrating synthesis on harder minority samples [7], while Borderline-SMOTE and related variants showed benefits from focusing on samples near decision boundaries [12, 13]. WRND extends this direction to noisy imbalanced classification by weighting generation with relative neighborhood density [8]. These outcomes suggest that local evidence is useful, but they also expose the central question addressed here: whether the evidence used to trust a seed should be separated from the evidence used to allocate synthesis effort.

C. Research Gap

This study addresses a gap in existing oversampling methods: the decision to trust a minority instance as a synthesis seed is often treated together with the decision about how much synthetic generation effort that instance should receive. These two decisions should be separated. Seed eligibility concerns whether a candidate minority instance is reliable

enough to guide interpolation, whereas synthesis allocation concerns how strongly a reliable seed should contribute to class balancing.

This distinction matters because difficulty can arise from different causes. A minority instance surrounded by majority samples may be an informative boundary case, but it may also be mislabeled or isolated. A method that increases generation near all difficult samples may amplify noise. In contrast, a method that selects only safe minority regions may miss boundary cases that are useful for discrimination. The methodological gap is therefore not simply the need for adaptive oversampling, but the need to separate seed eligibility from synthesis allocation.

NADOS addresses this gap by using reliability to determine seed eligibility and difficulty to determine synthesis priority. This differs from methods that treat local difficulty alone as a reason for more generation, and from methods that focus only on safe regions without explicitly preserving informative hard cases. The method first asks whether a minority instance should be trusted, and only then determines how much influence it should have during synthetic generation.

In relation to the preceding studies, NADOS treats seed eligibility and synthesis allocation as two separate decisions. Classical SMOTE mainly assumes that selected minority neighbors are trustworthy [4], ADASYN and Borderline-SMOTE emphasize difficult or boundary samples [12, 7], and cleaning hybrids remove some problematic samples only after generation [11, 14, 15]. NADOS instead screens candidate seeds before interpolation and then allocates generation pressure only among seeds that pass this screen (Fig. 1).

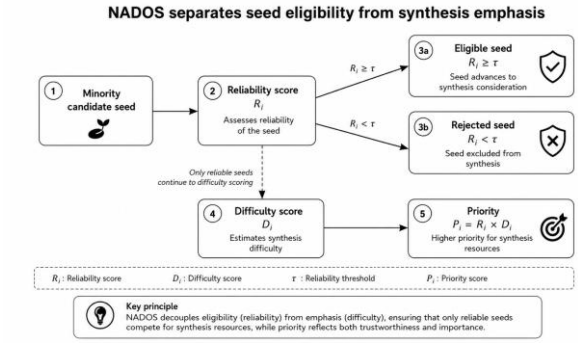

Fig. 1. NADOS decision logic separating seed eligibility from synthesis priority.

III. Proposed Method

This section defines the proposed NADOS method. It begins with the binary noisy imbalanced learning setting, introduces the local neighborhood evidence used by the method, defines the reliability and difficulty scores, and then describes how these scores are combined into a priority-based synthetic generation procedure.

A. Problem Setting

Let a binary training set be denoted as:

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

where, x_i is a feature vector, and y_i is the class label. The minority class is underrepresented relative to the majority class. The observed training data may also contain label noise. The goal is to generate synthetic minority samples that improve minority-sensitive classification performance without reinforcing unreliable local structure.

B. Local Neighborhood Evidence

For each minority instance x_i, NADOS constructs a k-nearest-neighbor neighborhood using the training fold. Let:

- m_i be the number of minority neighbors among the k-nearest neighbors.
- d_i^{min} be the average distance from x_i to minority neighbors.
- d_i^{maj} be the average distance from x_i to majority neighbors.

Neighborhood composition captures local minority support, while relative distance structure indicates whether the instance is closer to minority or majority regions.

C. Reliability Score

The reliability score estimates whether a minority instance is sufficiently trustworthy to be used as a synthesis seed. It combines minority-neighbor support with relative distance evidence:

$$R_i = \lambda\left(\frac{m_i}{k}\right) + (1-\lambda)\left(\frac{d_i^{\text{maj}}}{d_i^{\text{min}} + d_i^{\text{maj}}}\right)$$

where, λ controls the contribution of neighborhood composition. A higher R_i indicates stronger local minority support and greater distance from majority neighbors. A minority instance is eligible for synthetic generation only if:

$$R_i \geq \tau$$

where, τ is the reliability threshold.

The reliability score functions as a conservative screen rather than a full generation policy. The first term, m_i / k, measures same-class support in the local neighborhood. A minority instance with stronger minority support is less likely to be an isolated noisy point. The second term compares distances to majority and minority neighbors. If majority neighbors are relatively farther away, the instance is less likely to be embedded in a majority-dominated region. Combining these signals avoids relying only on neighbor counts, which can be unstable in sparse regions, or only on distance ratios, which can be sensitive to local scale.

D. Difficulty Score

The difficulty score estimates how much synthesis effort should be assigned to an eligible minority instance. NADOS combines a boundary component with a local sparsity component:

$$B_i = 1 - \left|2\left(\frac{m_i}{k}\right) - 1\right|$$

$$\rho_i = \frac{m_i}{k}$$

$$D_i = \eta B_i + (1 - \eta)(1 - \rho_i)$$

where, B_i is high near mixed neighborhoods and 1 - ρ_i reflects lower minority support. The parameter η controls the balance between boundary difficulty and sparsity difficulty.

The difficulty score is deliberately separated from reliability. A great difficulty score does not by itself make an instance suitable for interpolation; it only indicates that the instance may deserve more synthesis attention after passing the reliability gate. This prevents the method from treating unreliable difficult samples and informative difficult samples as equivalent. The boundary component emphasizes mixed neighborhoods, while the sparsity component gives attention to locally under-supported minority regions.

E. Synthesis Priority

NADOS assigns synthesis priority only to reliable seeds:

$$P_i = R_i \times D_i, if\ R_i \geq \tau;\ otherwise\ P_i = 0$$

The priority values are normalized across eligible minority seeds. Synthetic samples are then generated by selecting seed instances according to their normalized priorities and interpolating between the selected seed and one of its minority neighbors:

$$x_{new} = x_i + \alpha(x_j - x_i)$$

where, x_j is a minority neighbor of x_i and α is sampled uniformly from [0, 1].

The multiplicative priority formulation was selected because reliability and difficulty serve different roles. Reliability acts as a trust signal, whereas difficulty allocates synthesis effort among candidate seeds. A simple additive rule could allow a highly difficult but weakly reliable instance to retain substantial priority, which would conflict with the noise-aware motivation of NADOS. The product rule makes priority high only when both conditions are satisfied: the instance is sufficiently reliable and also informative enough to deserve synthesis emphasis. This provides a conservative first formulation for noisy imbalanced settings, where difficult minority instances may represent useful boundary cases but may also correspond to mislabeled or unstable samples.

F. Algorithm 1

Algorithm 1 NADOS oversampling procedure

1:	**input:** D, cmin, k, λ, η, τ, target ratio r	
2:	**initialize** D′ ← D and required synthetic count	
3:	***for*** each minority instance x_i **do**	
4:	compute R_i from support and distance evidence	▷ reliability
5:	compute D_i from boundary and sparsity components	▷ difficulty
6:	set $P_i = R_i \times D_i$ if $R_i \geq \tau$; else $P_i = 0$	▷ priority
7:	**end for**	
8:	***while*** target minority ratio r is not reached **do**	
9:	sample by normalized P_i; generate xnew = x_i + $\alpha(x_j - x_i)$; append to D′	▷ synthesis
10:	**output** resampled training set D′	

This procedure makes NADOS a reliability-gated extension of adaptive interpolation-based oversampling. The gate prevents unsafe seeds from driving generation, while the difficulty score preserves synthesis attention for reliable but challenging minority instances.

The computational cost of NADOS is dominated by nearest-neighbor search and synthetic sample generation. For each training fold, neighborhood evidence is computed for minority instances, after which generation proceeds similarly to SMOTE but with priority-weighted seed selection. Thus, the method adds local scoring overhead relative to plain SMOTE but avoids classifier-specific optimization or iterative model fitting inside the resampling step (Fig. 2).

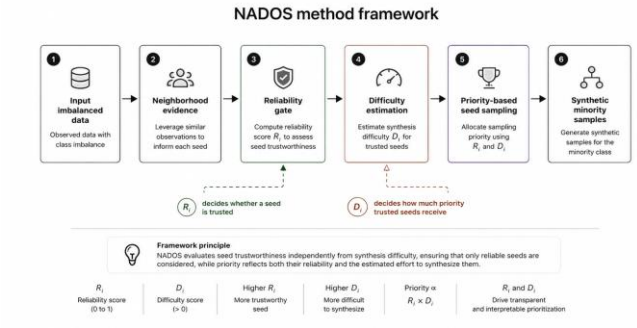

Fig. 2. NADOS method framework.

The main experiment used fixed settings across all datasets, classifiers, folds, and noise levels: k = 5, reliability weight λ = 0.50, difficulty weight η = 0.50, reliability threshold τ = 0.35, target minority ratio 1.00, product priority R_i x D_i, noise levels of 0%, 5%, 10%, 15%, and 20%, and random seed 42.

IV. Experimental Design

This section describes the benchmark datasets, controlled noise protocol, classifiers, baseline oversampling methods, evaluation metrics, statistical tests, and software environment used to evaluate NADOS. The experiments were implemented in Python using standard scientific-computing and machine-learning libraries, including scikit-learn and imbalanced-learn for model evaluation and baseline resampling support [16, 17].

A. Benchmark Datasets

The evaluation uses 22 binary imbalanced benchmark datasets: ecoli-0_vs_1, ecoli1, ecoli2, ecoli3, glass-0-1-2-3_vs_4-5-6, glass0, glass1, glass6, haberman, iris0, new-thyroid1, new-thyroid2, page-blocks0, pima, segment0, vehicle0, vehicle1, vehicle2, vehicle3, wisconsin, yeast1, and yeast3. These datasets are drawn from public imbalanced-learning benchmark collections derived from KEEL and UCI-style tabular datasets [18, 19]. They cover multiple sample sizes and imbalance ratios, giving a diverse benchmark for tabular imbalanced classification.

The benchmark contains 22 binary imbalanced tabular datasets with 150 to 5472 instances, 3 to 19 features, and imbalance ratios ranging from 1.82 to 8.79. The datasets include ecoli, glass, haberman, iris, new-thyroid, page-blocks, pima, segment, vehicle, wisconsin, and yeast variants.

B. Controlled Label Noise

Because the selected benchmark datasets do not provide a consistent natural-noise setting, controlled symmetric label noise is injected into the training folds. Noise levels of 0%, 5%, 10%, 15%, and 20% are used. Noise is applied only to training folds, while test folds remain unchanged. This design evaluates whether each oversampling method can learn from corrupted training data while preserving a clean evaluation reference.

This ordering is important for experimental validity. If noise injection or resampling were applied before the train-test split, synthetic samples or corrupted-label effects could leak into the evaluation fold. The protocol therefore follows a strict sequence: split the data, inject noise into the training fold only, apply resampling to the noisy training fold only, train the classifier, and evaluate on the untouched test fold. This makes the measured performance reflect robustness to noisy training data rather than contamination of the test reference (Fig. 3).

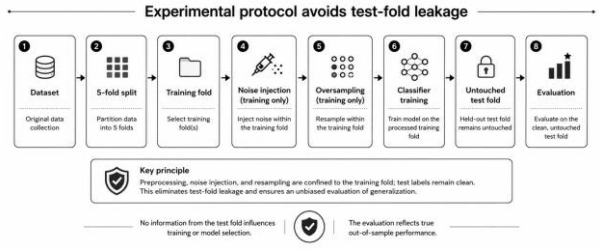

Fig. 3. Experimental protocol used to prevent test-fold leakage.

C. Classifiers and Baselines

Four classifiers are used: logistic regression, support vector machine, random forest, and gradient boosting. NADOS is compared against eight baselines: Random Oversampling, SMOTE, ADASYN, Borderline-SMOTE, SVM-SMOTE, SMOTE-ENN, SMOTE-Tomek, and WRND.

The classifier set is intentionally heterogeneous. Logistic regression provides a linear baseline, SVM captures margin-based behavior, random forest represents bagged tree ensembles, and gradient boosting represents sequential tree ensembles. This reduces the chance that the evaluation reflects compatibility with only one classifier family. The baseline set covers duplication, classical interpolation, adaptive difficulty-based oversampling, boundary-guided oversampling, hybrid cleaning, and a recent noise-aware method.

D. Metrics

The primary metrics are F1, G-mean, AUPRC, and balanced accuracy. Minority recall and runtime are also reported. These metrics are selected because standard accuracy is not sufficient for imbalanced classification and can hide poor minority-class recognition; precision-recall analysis is also recommended when the positive or minority class is rare [20, 21]. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Precision, recall, specificity, and the primary summary metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$G\text{-}mean = \sqrt{Recall \times Specificity}$$

$$Balanced\ accuracy = \frac{Recall + Specificity}{2}$$

$$AUPRC = \Sigma\,(Recall_t - Recall_{t-1})\,Precision_t$$

AUPRC summarizes the precision-recall curve and is useful when the minority class is the primary class of interest.

E. Statistical Testing

The comparison uses non-parametric statistical testing because multiple methods are evaluated across multiple datasets, folds, classifiers, and noise levels. Friedman tests are used to test whether overall differences exist among methods [22]. Pairwise Wilcoxon signed-rank tests with Holm correction are then used to compare NADOS against each baseline [23, 24].

Each dataset-fold-noise-classifier combination is treated as a comparison block. The Friedman test evaluates whether the methods differ in rank across these repeated benchmark conditions. Because the global test does not identify which method pairs differ, Wilcoxon signed-rank tests are then applied to paired NADOS-versus-baseline results. Holm correction is used to control the family-wise error rate across multiple pairwise comparisons. This combination of average performance, rank analysis, and corrected paired tests provides a more reliable interpretation than mean scores alone.

V. Results and Discussion

This section reports the empirical findings from three complementary perspectives: aggregate metric values, average rank analysis, and pairwise statistical comparisons against baselines. It then discusses robustness under increasing label noise and the runtime implications of the proposed scoring procedure.

A. Overall Performance

Overall performance is summarized across all datasets, folds, classifiers, and noise levels.

Across the full benchmark, NADOS obtains average F1 = 0.7428, G-mean = 0.8511, AUPRC = 0.7700, balanced accuracy = 0.8552, minority recall = 0.8445, and runtime = 0.1609 seconds. The closest competitor is SVM-SMOTE, with F1 = 0.7417, G-mean = 0.8506, AUPRC = 0.7746, and balanced accuracy = 0.8555. SMOTE-Tomek, SMOTE, Random Oversampling, and WRND remain competitive but are lower on the main balanced classification metrics.

NADOS obtains the highest average F1 and G-mean and is effectively tied with SVM-SMOTE on balanced accuracy. SVM-SMOTE obtains the highest average AUPRC. This suggests that NADOS is strongest for balanced minority-

sensitive classification performance, while AUPRC remains competitive but not dominant (Table I).

TABLE I. MEAN PERFORMANCE ACROSS ALL DATASETS, FOLDS, CLASSIFIERS, AND NOISE LEVELS

Method	F1	G-mean	Bal. Acc.	AUPRC
NADOS	0.7428	0.8511	0.8552	0.7700
SVM-SMOTE	0.7417	0.8506	0.8555	0.7746
SMOTE-Tomek	0.7356	0.8497	0.8536	0.7695
SMOTE	0.7329	0.8480	0.8519	0.7682
Random oversampling	0.7357	0.8468	0.8515	0.7663
WRND	0.7405	0.8379	0.8445	0.7595
SMOTE-ENN	0.7026	0.8376	0.8444	0.7513
ADASYN	0.6980	0.8342	0.8391	0.7517
Borderline-SMOTE	0.6955	0.8303	0.8355	0.7404

NADOS does not dominate every dataset-metric combination, but it has the strongest aggregate profile for F1, G-mean, and balanced accuracy. Noisy imbalanced benchmark performance is heterogeneous, and different resampling methods may be favored by different dataset geometries, classifiers, and noise levels. The contribution should therefore be interpreted through aggregate performance, average ranking, and statistically validated comparisons rather than as universal dominance.

The overall results are consistent with the method design. F1, G-mean, and balanced accuracy reward methods that improve minority recognition without allowing majority performance to collapse. NADOS performs well under these criteria because the reliability gate limits unsafe seed usage while the difficulty score still emphasizes challenging reliable regions. SMOTE-ENN achieves the highest average minority recall, but its F1 and G-mean are lower, indicating that recall gains alone are insufficient when precision or majority recognition deteriorates.

The AUPRC result should be interpreted cautiously. SVM-SMOTE achieves the highest AUPRC, suggesting that margin-guided boundary synthesis can be effective for ranking minority instances by predicted score. NADOS remains competitive but does not dominate this metric. This indicates that reliability-gated generation improves classification balance more clearly than probability ranking quality in the current configuration.

B. Average Rank Analysis

The Friedman ranking results show that NADOS achieves the best average rank for F1, G-mean, and balanced accuracy. For AUPRC, SVM-SMOTE obtains the best average rank, while NADOS ranks third but remains close to SMOTE-Tomek and SMOTE.

The Friedman and average-rank results support the aggregate performance findings.

The Friedman tests are significant for all primary metrics ($p < 0.001$), indicating systematic differences among methods across benchmark blocks. NADOS has the best average rank for F1 (4.2786), G-mean (4.3755), and balanced accuracy (4.3873). For AUPRC, SVM-SMOTE obtains the best average rank (4.2130), followed by SMOTE-Tomek and SMOTE, while NADOS remains close with rank 4.6200 and average AUPRC = 0.7700.

The significant Friedman tests indicate that the performance differences among methods are not random across the benchmark conditions. The rank profile supports the main claim that NADOS is a strong general-purpose oversampling method under noisy imbalanced settings, particularly for metrics that directly reflect balanced class recognition.

The rank results are useful because average scores can be influenced by a small number of large datasets or easier benchmark conditions. Ranking each method within repeated comparison blocks provides a distribution-free view of consistency. NADOS obtains the best average rank for three of the four primary metrics, which supports its robustness across datasets, classifiers, folds, and noise levels. Its third-place AUPRC rank reinforces the need to position the method as balanced and robust rather than universally superior (Fig. 4).

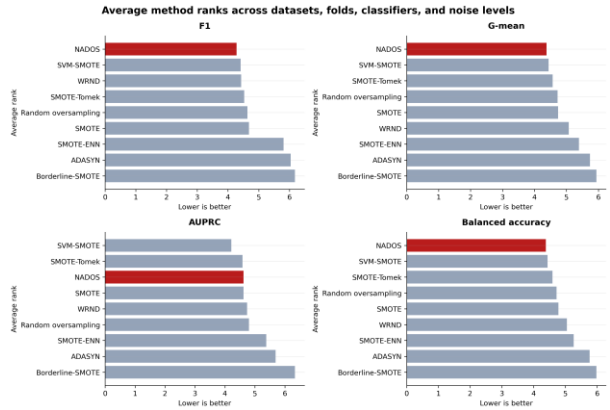

Fig. 4. Average method ranks across the primary evaluation metrics.

C. Pairwise Comparison Against Baselines

Wilcoxon-Holm pairwise tests show that NADOS significantly improves F1 over Borderline-SMOTE, ADASYN, SMOTE-ENN, SMOTE, Random Oversampling, and SMOTE-Tomek. The F1 differences relative to SVM-SMOTE and WRND are not statistically significant after Holm correction, suggesting that these methods remain close competitors.

For G-mean and balanced accuracy, NADOS significantly improves over most baselines, including WRND, SMOTE-ENN, ADASYN, Borderline-SMOTE, SMOTE, Random Oversampling, and SMOTE-Tomek. Differences relative to SVM-SMOTE are not significant. For AUPRC, NADOS improves over several weaker baselines but does not outperform SVM-SMOTE.

These findings should be interpreted cautiously. NADOS should not be framed as universally dominant, but as a method that consistently strengthens balanced minority-sensitive performance while remaining competitive with the strongest SMOTE-family baseline.

The pairwise results clarify the practical meaning of the average ranks. NADOS provides statistically significant gains over ADASYN and Borderline-SMOTE across F1, G-mean, and balanced accuracy. Since these methods also emphasize difficult or boundary samples, the result supports the argument that difficulty should not be used alone as a generation signal. NADOS also improves over WRND for G-mean and balanced accuracy, suggesting that reliability-gated synthesis can compete with a recent density-aware noise-robust strategy. The non-significant differences against SVM-SMOTE indicate that margin-guided synthesis remains a strong competitor.

The closest Wilcoxon-Holm comparisons reinforce this interpretation.

The Wilcoxon-Holm tests show significant F1 gains for NADOS over Borderline-SMOTE (mean difference = 0.0476, $p < 0.001$), ADASYN (0.0441, $p < 0.001$), SMOTE (0.0100, $p < 0.001$), and SMOTE-Tomek (0.0075, $p < 0.001$). The F1 differences against SVM-SMOTE (0.0012, $p = 0.3224$) and WRND (0.0023, $p = 0.9411$) are not significant. For G-mean and balanced accuracy, NADOS significantly improves over WRND and most SMOTE-family baselines, while differences against SVM-SMOTE remain non-significant.

D. Component Ablation and Parameter Sensitivity

To isolate the contributions of the two NADOS components, two ablated variants were evaluated using the same benchmark protocol. The reliability-only variant applies the reliability gate but distributes synthesis uniformly among eligible seeds. The difficulty-only variant removes the reliability gate and allocates synthesis according to the difficulty score alone. Full NADOS performs best across all four reported metrics, indicating that the reliability gate and difficulty-weighted allocation contribute complementary effects rather than one component fully subsuming the other (Table II).

TABLE II. ABLATION COMPARISON FOR FULL NADOS AND COMPONENT VARIANTS

Variant	F1	G-mean	Bal. Acc.	AUPRC
Full NADOS	0.7428	0.8511	0.8552	0.7700
Reliability-only	0.7354	0.8478	0.8523	0.7657
Difficulty-only	0.7184	0.8438	0.8480	0.7597

TABLE III. SELECTED PARAMETER SENSITIVITY SUMMARY ON REPRESENTATIVE DATASETS

Setting	Parameter	Mean F1	Mean G-mean	Mean AUPRC
rel30/diff70	lambda/eta	0.5706	0.7620	0.6173
rel50/diff50	lambda/eta	0.5748	0.7595	0.6227
rel70/diff30	lambda/eta	0.5726	0.7566	0.6237
tau=0.20	threshold	0.5287	0.7862	0.6014
tau=0.35	threshold	0.5467	0.7806	0.6065
tau=0.50	threshold	0.5748	0.7595	0.6227

A limited parameter sensitivity check evaluates the reliability threshold and the relative reliability-difficulty weighting on representative datasets. The main configuration uses $k = 5$, lambda = 0.50, eta = 0.50, and tau = 0.35. The value $k = 5$ was selected as a local-neighborhood setting commonly used in SMOTE-family experiments and was fixed to avoid dataset-specific tuning. The threshold and weighting checks show that performance varies by dataset, which supports the decision to report a fixed reproducible configuration rather than tune parameters separately for each benchmark condition (Table III).

E. Robustness Under Label Noise

The controlled noise protocol evaluates whether each method remains useful as training labels become increasingly corrupted. NADOS is designed for this setting because unreliable seeds should be filtered before synthetic generation. The reliability gate is expected to matter most as noise increases, since noisy minority instances are more likely to produce harmful synthetic samples when used as interpolation seeds.

The noise-level results show a clear robustness trend. In the no-noise setting, SMOTE-Tomek, SMOTE, and SVM-SMOTE are slightly ahead of NADOS on G-mean. Once label noise is introduced, NADOS becomes increasingly competitive and obtains the strongest G-mean at 10% and 15% noise. At 20% noise, SVM-SMOTE remains slightly ahead, but NADOS remains stronger than SMOTE, SMOTE-Tomek, and WRND on G-mean and balanced accuracy. This supports the intended role of the reliability gate: NADOS is most useful when the training data contain noisy or unstable minority candidates.

This trend also explains why the method should be evaluated under noise rather than only on clean imbalanced data. In the no-noise setting, classical SMOTE-family methods remain highly competitive because many minority seeds are usable. As noise increases, unsafe seed selection becomes more costly. NADOS does not prevent all performance degradation, but it delays the point at which noisy local structure damages balanced performance. The clearest advantage appears at moderate noise levels, especially 10% and 15%, where reliability screening is valuable, but neighborhood evidence remains informative.

Overall, the G-mean trend shows that NADOS is most competitive in moderate-noise settings. At the highest noise level, SVM-SMOTE is slightly stronger, but NADOS remains close and stronger than SMOTE, SMOTE-Tomek, and WRND on the main balanced metrics (Fig. 5).

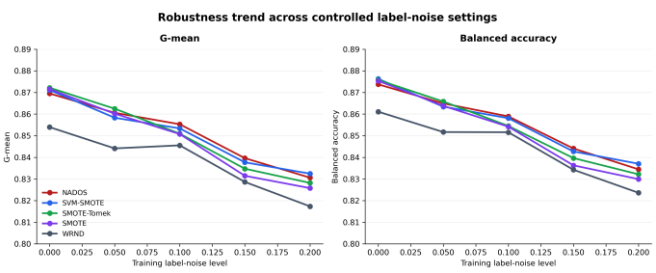

Fig. 5. Robustness trend across controlled label-noise settings.

F. Runtime and Computational Complexity

The runtime results indicate that NADOS remains practical at the benchmark scale used in this study. As shown in Table IV, NADOS records an average runtime of 0.1609 seconds and an average runtime rank of 4.80 across the benchmark comparisons. It is not the fastest method: SMOTE-ENN, Random Oversampling, WRND, and SMOTE-Tomek have better average runtime ranks. However, NADOS remains faster on average than SVM-SMOTE, SMOTE, Borderline-SMOTE, and ADASYN, while providing stronger aggregate F1, G-mean, and balanced accuracy.

TABLE IV. RUNTIME COMPARISON ACROSS ALL BENCHMARK SETTINGS

Method	Mean runtime (s)	Runtime rank
SMOTE-ENN	0.1042	1.51
Random oversampling	0.1497	3.94
SMOTE-Tomek	0.1661	4.35
WRND	0.1499	4.53
NADOS	0.1609	4.80
SVM-SMOTE	0.1683	5.84
SMOTE	0.1739	6.05
Borderline-SMOTE	0.1809	6.92
ADASYN	0.1868	7.04

The additional cost comes from computing local neighborhood evidence before generating synthetic samples. For n training instances, n_min minority instances, d features, k nearest neighbors, and G generated synthetic samples, a direct implementation computes distances between each minority instance and the training set. This requires O(n_min n d) distance operations. After the neighbor lists are available, synthetic interpolation requires O(Gd) operations because each generated sample is constructed across d features. The remaining reliability, difficulty, and priority calculations are linear in the number of minority instances, O(n_min), and are therefore smaller than the neighborhood-search cost.

In practical terms, NADOS has a similar computational profile to SMOTE-family methods that also depend on nearest-neighbor search. Its extra scoring stage adds overhead, but it does not require repeated classifier fitting, iterative optimization, or ensemble training inside the resampling process. For small and medium tabular benchmarks, this cost is acceptable. For very large or high-dimensional datasets, approximate nearest-neighbor search, feature normalization, dimensionality reduction, or batch-wise implementation may be needed.

G. Discussion

The results also identify boundary conditions where NADOS should not be described as universally superior. SVM-SMOTE is the closest competitor, obtains a slightly higher AUPRC and balanced accuracy in the aggregate table, and is not significantly different from NADOS on F1 after Holm correction. This suggests that margin-guided boundary synthesis can be highly competitive when the classifier score ranking is important. NADOS is therefore best characterized as having a strong and balanced performance profile under the evaluated noisy imbalanced benchmark, rather than as dominating every metric or dataset condition.

The results support the central methodological argument: separating seed reliability from synthesis difficulty is beneficial in noisy imbalanced classification. NADOS does not simply assign more synthetic samples to difficult regions. Instead, it first examines whether a minority instance is trustworthy enough to guide generation. This distinction helps avoid the main weakness of difficulty-only adaptive oversampling, where noisy or unstable points may receive excessive synthetic emphasis.

The method is not uniformly superior on every metric. SVM-SMOTE remains highly competitive and obtains stronger AUPRC. This limitation should be acknowledged directly. The value of NADOS lies in its balanced performance profile, explicit noise-aware decision logic, and strong average ranking on F1, G-mean, and balanced accuracy.

The results support the main hypothesis: a useful oversampling method for noisy imbalanced data should not treat all difficult minority samples as equally valuable. The reliability gate and difficulty score serve different purposes. The gate reduces generation from unstable seeds, while the difficulty score preserves attention to hard but useful samples. This division is the main conceptual contribution of NADOS and explains why it is competitive with both classical SMOTE-family methods and a recent noise-aware baseline.

VI. THREATS TO VALIDITY

This study uses controlled symmetric label noise, which supports reproducibility but does not capture all forms of real-world label corruption. The empirical claims should therefore be scoped to the symmetric training-label noise setting evaluated here. In practice, label noise may be class-dependent, feature-dependent, annotator-dependent, or concentrated in specific subregions of the feature space. Future work should test NADOS under class-dependent and feature-dependent corruption models before generalizing the conclusions to all noisy-data settings.

The benchmark is limited to binary tabular imbalanced datasets. This is appropriate for the first evaluation of NADOS, but conclusions should not be generalized automatically to multiclass, image, text, temporal, or streaming data. The method depends on k-nearest-neighbor evidence, so its behavior can be affected by the distance metric, feature scaling, mixed numerical-categorical attributes, missing values, and high-dimensional distance concentration. In the present implementation, continuous features are normalized within the training pipeline, and Euclidean neighborhoods are used; alternative metrics, metric learning, or categorical-aware distance functions may be needed for heterogeneous datasets.

The parameter settings are fixed for the main evaluation. This supports reproducibility and fair comparison, but it may not be optimal for every dataset. Adaptive parameter selection, threshold tuning, and alternative priority functions are left for future work. Finally, the implementation evaluates NADOS against a focused set of reproducible baselines. Additional recent oversampling methods may provide further comparison

opportunities if stable implementations and consistent benchmark protocols are available.

VII. Conclusion

This study proposed NADOS, a reliability-gated and difficulty-aware oversampling method for noisy imbalanced classification. The method separates minority-seed trustworthiness from synthesis priority by using reliability as an eligibility gate and difficulty as an allocation signal among eligible seeds. Evaluation across 22 benchmark datasets, five label-noise levels, four classifier families, and eight oversampling baselines shows that NADOS achieves the best average F1, G-mean, and balanced accuracy while remaining competitive on AUPRC. These findings support reliability-gated adaptive oversampling as a practical strategy for noisy imbalanced learning.

Future work should extend NADOS to adaptive thresholding, broader external datasets, multiclass settings, and more detailed component-level ablation. It should also examine alternative priority-combination rules, including additive weighting, soft-gated products, probabilistic seed selection, and learned weighting functions, to determine whether adaptive priority functions improve performance under different noise regimes, imbalance ratios, and feature-space conditions.

VIII. Declarations

A. Declaration on Generative AI

The authors used generative AI-assisted tools for language refinement and editorial checking. All scientific ideas, experimental design, implementation, results, interpretations, and final manuscript content were reviewed and verified by the authors.

B. Data Availability

The datasets used in this study are publicly available benchmark imbalanced classification datasets in KEEL-style format and related public tabular repositories [18, 19]. The dataset names and benchmark characteristics are summarized in the manuscript.

C. Code Availability

The implementation code, benchmark scripts, parameter settings, preprocessing pipeline, and random seeds are retained by the authors and may be made available upon reasonable request, subject to thesis completion, institutional requirements, and repository preparation.

D. Funding

This work was supported/funded by the Universiti Teknologi Malaysia under UTM NEXUS Translasi (PY/2025/00988).

E. Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263-1284. https://doi.org/10.1109/TKDE.2008.239

[2] Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5:221-232. https://doi.org/10.1007/s13748-016-0094-0

[3] Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell Data Anal. 2002;6(5):429-449. https://doi.org/10.3233/IDA-2002-6504

[4] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357. https://doi.org/10.1613/jair.953

[5] Fernandez A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res. 2018;61:863-905. https://doi.org/10.1613/jair.1.11192

[6] Saez JA, Luengo J, Stefanowski J, Herrera F. SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Inf Sci. 2015;291:184-203. https://doi.org/10.1016/j.ins.2014.08.051

[7] He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks. IEEE; 2008. pp. 1322-1328. https://doi.org/10.1109/IJCNN.2008.4633969

[8] Li M, Zhou H, Liu Q, Gong X, Wang G. WRND: a weighted oversampling framework with relative neighborhood density for imbalanced noisy classification. Expert Syst Appl. 2024;241:122593. https://doi.org/10.1016/j.eswa.2023.122593

[9] Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-Level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: Advances in Knowledge Discovery and Data Mining. Berlin: Springer; 2009. pp. 475-482. https://doi.org/10.1007/978-3-642-01307-2_43

[10] Barua S, Islam MM, Yao X, Murase K. MWMOTE: Majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans Knowl Data Eng. 2014;26(2):405-425. https://doi.org/10.1109/TKDE.2012.232

[11] Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newsl. 2004;6(1):20-29. https://doi.org/10.1145/1007730.1007735

[12] Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Advances in Intelligent Computing. Lecture Notes in Computer Science, vol 3644. Berlin: Springer; 2005. pp. 878-887. https://doi.org/10.1007/11538059_91

[13] Nguyen HM, Cooper EW, Kamei K. Borderline over-sampling for imbalanced data classification. Int J Knowl Eng Soft Data Paradigms. 2011;3(1):4-21. https://doi.org/10.1504/IJKESDP.2011.039875

[14] Tomek I. Two modifications of CNN. IEEE Trans Syst Man Cybern. 1976;SMC-6(11):769-772. https://doi.org/10.1109/TSMC.1976.4309452

[15] Wilson DL. Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans Syst Man Cybern. 1972;SMC-2(3):408-421. https://doi.org/10.1109/TSMC.1972.4309137

[16] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825-2830.

[17] Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res. 2017;18(17):1-5.

[18] Dua D, Graff C. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science; 2019.

[19] Alcala-Fdez J, Fernandez A, Luengo J, Derrac J, Garcia S, Sanchez L, Herrera F. KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. J Mult-Valued Logic Soft Comput. 2011;17(2-3):255-287.

[20] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. ACM; 2006. pp. 233-240. https://doi.org/10.1145/1143844.1143874

[21] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432

[22] Demsar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1-30.

[23] Wilcoxon F. Individual comparisons by ranking methods. Biom Bull. 1945;1(6):80-83. https://doi.org/10.2307/3001968

[24] Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6(2):65-70.