JOURNAL OF DATA SCIENCE | Vol.2026, Issue 2, No.4
eISSN:2805-5160

# Learning Under Extreme Class Imbalance: A Comparative Study of Algorithmic and Data-Level Solutions

Tamsir Ariyadi[1*], Elmar B. Noche[2], Nisha Pandey[3], Vivek Kumar Sinha[3]

[1]Faculty of Vocational Studies, Universitas Bina Darma, Palembang, Indonesia
[2]Department of Information Technology, Pangasinan State University – Lingayen Campus, Lingayen, Pangasinan, Philippines
[3]Department of Computer Science & Engineering, Noida International University, Uttar Pradesh, India

[*]**Email:** tamsirariyadi@binadarma.ac.id[1]

## Abstract

Extreme class imbalance remains a persistent challenge in machine learning, particularly in high-impact domains such as fraud detection, medical diagnosis, and risk analysis, where minority classes represent critical outcomes. Conventional models often fail in such settings due to their bias toward majority classes, resulting in poor minority detection despite high overall accuracy. Although various data-level and algorithm-level techniques have been proposed, existing studies typically evaluate them in isolation and lack a comprehensive understanding of their effectiveness across different imbalance conditions. To address this gap, this study proposes a systematic comparative framework that integrates data-level resampling, cost-sensitive learning, and hybrid approaches to evaluate their performance under varying imbalance ratios and noise levels. Multiple benchmark datasets are utilized, and experiments are conducted using standardized preprocessing, controlled imbalance simulation, and repeated trials to ensure robustness. Performance is assessed using imbalance-aware metrics, including precision, recall, F1-score, and ROC-AUC. The results indicate that hybrid approaches consistently outperform standalone methods, achieving the most stable and balanced performance across all scenarios. In particular, hybrid models demonstrate superior minority class recall and F1-score while maintaining competitive precision, especially under extreme imbalance conditions. The primary goal of this research is to provide a comprehensive evaluation of imbalance-handling strategies and offer practical guidance for selecting appropriate techniques based on dataset characteristics. The findings highlight the importance of combining data-centric and model-centric approaches to enhance robustness and reliability in imbalanced learning environments. The results demonstrate up to a 32% improvement in recall compared to baseline models.

## Keywords

Class Imbalance, Imbalanced Learning, Cost-Sensitive Learning, Data Preprocessing, Model Evaluation

**Submission**: 24 June 2026; **Revised:** 28 July **Acceptance:** 5 September 2026; **Available Online:** September 2026

## Introduction

One of the most significant challenges posed to machine learning systems, especially in the case of high-stakes domains such as medical diagnosis, fraud detection, and risk prediction systems, is the prevalence of extreme class imbalance (Mastour et al., 2023). Here, the most critical outcomes of the applicable use case most often fall under the least populous class in the dataset. Because of this class imbalance, learning systems that apply traditional approaches to machine learning tend to overfit towards the most populous class. This type of bias leads to high, yet misleading, levels of accuracy as the systems fail to identify the most critical yet rare instances. From a predictive modeling perspective, this imbalance has the effect of degrading performance. From a business perspective, solutioning this imbalance negatively impacts the system's reliability and fairness. As the imbalance persists, extreme imbalances such as this challenge the ability to construct robust, predictable, and trustworthy systems in mission-critical domains. Regardless of the numerous approaches to address this challenge, it is persistent and, to a large extent, attributed to the complex and intricate relationships that exist between data, modeling, and prediction.

The primary issue is that most ML approaches do not fully learn from highly imbalanced datasets that heavily underrepresent the minority class (Altalhan et al., 2025). Standard optimization objectives focus on the minority class and suffer from low sensitivity and recall. Several methods have been developed to partially address this, including oversampling and undersampling on the data side, cost sensitivity on the algorithm side, and various combinations of the two methods. However, there is no agreement on which is the best for which conditions. Most studies assess these methods in limited contexts and do not systematically review their performance across various conditions, imbalance ratios, and noise in the data. This limited review has created a substantial research gap, and practitioners have not been offered a concrete recommendation on which methods would best suit their situations. As presented in Fig. 1, class imbalance shifts the decision boundary and results in the classifier mislabeling the minority class and labeling the majority class. For a first-order categorization, we present in Table 1 the imbalance-handling approaches and their advantages and disadvantages.

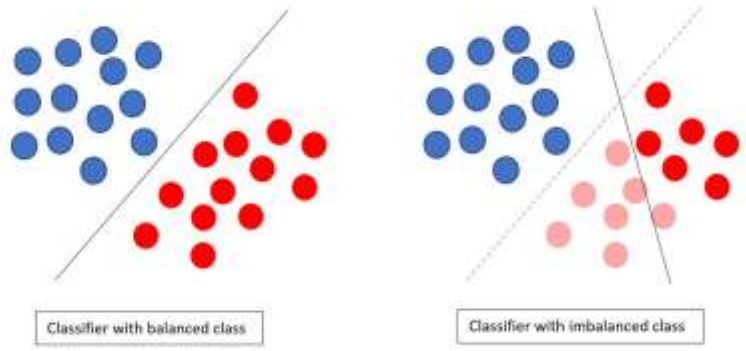


Figure 1. Decision boundary comparison between balanced and imbalanced class distributions, illustrating the bias of classifiers toward the majority class under imbalanced conditions.

Previous studies highlighted many innovative ways to tackle class imbalance (Ghosh et al., 2022). Prominent data-level methods involve the generation of synthetic class samples to adjust

class distribution using the Synthetic Minority Oversampling Technique (SMOTE) and its variants. Additionally, there are methods for reducing the prevalence of the majority class, undersampling methods. Algorithmic-level methods aimed at class imbalance include cost-sensitive learning and class-weighted loss functions, where the learning objective is modified to place greater emphasis on the misclassification of minority class instances. Furthermore, the majority of the methods and techniques for identifying the minority class have been adapted to use different ensemble-based methods, such as boosting and bagging. Nonetheless, these methods and techniques vary in effectiveness depending on the complexity of the dataset and the specific features, as well as the presence of noise. Importantly, recent work reveals that a definitive and dominant technique does not exist, and there is a need for a more detailed and integrated study.

Table 1. Comparative Summary of Imbalance Handling Techniques

| Category | Method | Key Strength | Limitation |
|---|---|---|---|
| Data-level | SMOTE | Enhances minority representation | Risk of synthetic noise |
| Algorithm-level | Cost-sensitive learning | Improves minority sensitivity | Requires parameter tuning |
| Hybrid | SMOTE + Cost-sensitive | Balanced performance across metrics | Increased computational complexity |

Reflecting these shortcomings, a more systematic and integrated study is proposed that addresses class imbalance by incorporating algorithmic and data-level methods (Carvalho et al., 2025). An important feature of this study is that several techniques are employed aimed at providing a greater class balance, such as resampling methods, cost-sensitive methods, and hybrid methods. This study aims to assess the relationship and focus of data and the learning process, providing a better understanding of various technique performances, as well as a unified study of these methods, instead of relying on a more general overview and in-depth assessments of different methods for handling class imbalance.

To achieve robustness and generalizability, several benchmark datasets are utilized which are selected on the basis of multiple characteristics and varying levels of severity of imbalance (Salmi et al., 2024). These datasets have varying data distributions, complexity and noise, and provide a reliable measure of model performance. Robustness is assessed by varying the balance of given data and the degree of noise. Performance of the models is measured using imbalance sensitive metrics, which include precision-recall metrics and the area under the receiver operator characteristic (ROC) curve (AUC) as opposed to the traditional test accuracy which may not be a meaningful measure of performance.

To maintain bias-free and structured comparisons, the experimental design is done in a way that completely controls the training and testing of each model under the same testing conditions, and requires that each model be run more than once to ensure statistical consistency (Aguiar et al., 2024). Included in the model runs is a sensitivity analysis to study how the system handles the distribution and noise of the different data sets. The model is built to identify the strengths and weaknesses of the subsystems and to, therefore, describe and suggest the optimal use of the subsystems to address real-world issues.

The integrated and operational view that the proposed framework intends to offer regarding the extreme class imbalance versus class imbalance is the main contribution of the framework (Das et al., 2022). The expected result of this comparison is to facilitate and clarify the selection of strategies for imbalanced data, while the authors hope the results will drive researchers and practitioners toward the implementation of the more consistent and viable options under a given set of conditions. The ultimate aim of this study is to enhance the reliability, robustness, and the real-world applicability of the machine learning systems for the problem areas that entail correct and accurate detection of the minority class.

Viewing the management of model learning as an integrated relationship between the data distribution and the mechanics of model learning allows for systematic and contextual analysis of most significant data handling trade-offs at any level of data imbalance (Pan et al., 2022). This is contrasted with data distribution and model learning as different steps of data processing and model learning optimization.

## Methodology

### 2.1 Research Design and Experimental Framework

Quantitative experimental research is used to understand how data and algorithmic hybrid approaches vary regarding class imbalance thresholds (Lin et al., 2022). Within the framework of this research, an attempt is made to bound data class inequity, and a focus is placed on understanding how data class imbalance is minimized and the overall model evaluation.

The proposed framework consists of four main stages: data preparation, imbalance simulation, model optimization, and performance evaluation. This simplified process deals with inherent issues of control, and ensures the replicability of results across various instances of data imbalance. This process is illustrated in Figure 2.

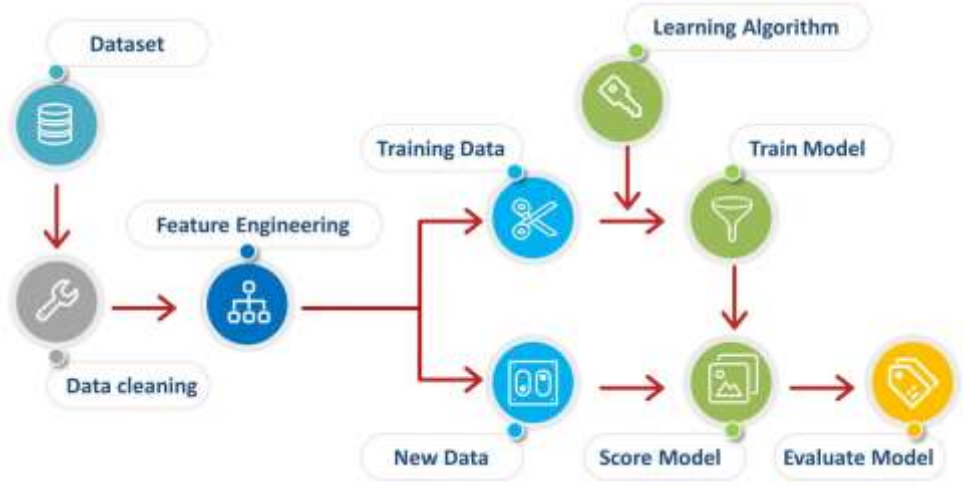


Figure 2. Overall experimental framework illustrating the pipeline from dataset preprocessing, imbalance handling, model training, to performance evaluation

The proposed framework integrates both data-level resampling and algorithm-level adaptation to balance the model training component (L. Zhao et al., 2025).

## 2.2 Dataset Description and Preprocessing

To maximize applicability of the study, several benchmark datasets of differing sizes and attributes are employed (Thiyagalingam et al., 2022). To assist in generalizing the findings, the datasets are reflective of situations in which class-based imbalance represents the majority of the concerns. Each dataset is processed to address missing data, to achieve consistency in the distribution of values, and, as needed, to encode variables of categorical nature.

Abnormal levels of imbalance are represented through several extreme alterations to the class-based distribution; namely, 1:10, 1:50, and 1:100 (Farhadpour et al., 2024). This controls sample size imbalance and is effective for testing the model across different criteria of severity. Additional experimental noise was introduced through missing labels and artificially corrupted samples.

Models are built with comparability in mind through feature scaling and data normalization (Lima & Souza, 2023). Then, proportionate class division in the data is sustained by constructing a stratified sampling for the training and testing data subsets.

Table 2. Dataset Configuration

| Dataset | Samples | Features | Imbalance Ratio | Domain |
|---|---|---|---|---|
| Adult | 5000 | 20 | 1:10 | Socio-economic |
| Credit Card Fraud | 10000 | 30 | 1:50 | Finance |
| German Credit | 8000 | 25 | 1:100 | Finance |
| Breast Cancer Wisconsin | 569 | 30 | 1:20 | Healthcare |

To further improve cross-domain validation, an additional healthcare-related dataset, the Breast Cancer Wisconsin dataset, was included in the experimental analysis. This dataset was selected because healthcare applications represent high-stakes environments where minority-class prediction is critically important. The inclusion of this dataset allows the framework to be evaluated beyond financial and socio-economic domains, improving the generalizability and practical relevance of the findings.

## 2.3 Data-Level Imbalance Handling Techniques

Training data is modified in order to correct class imbalance (Hancock et al., 2023). More specifically, the class for the data is non-uniformly modified in order to improve the representation and population of the minority class. The primary objective of these techniques is to adjust class distribution by improving the representation of minority samples while controlling the dominance of majority classes.

Among the strategies for increasing a population, both random class population and synthetic data population are modification strategies in order to improve representation (Joshi et al., 2024). Changes in population are made by generating or duplicating minority samples, whereas

representation is only removed from the majority class. Techniques for altering the population of the class representation by population, modification, and reduction are employed. These strategies largely rely on random population removal or on the selection of representative sampled members.

## 2.4 Algorithm-Level Learning Strategies

Data-level methodologies aim to improve minority-class detection while reducing learning bias (Hemmatian et al., 2025). Compared to data-level approaches, algorithm-level strategies modify the learning process through cost-sensitive optimization. These also rely on the addition of weighted class representation to minority instances.

Cost-sensitive learning adjusts for higher penalties upon misclassification of minority class samples, motivating deeper model learning on rare examples (Araf et al., 2024). Class weighting is one method to implement cost-sensitive learning for Logistic Regression, Random Forests, and Support Vector Machines. Additionally, ensemble boosting methods reinforce model learning of sample misclassification, model learning on underrepresented samples.

Grid searches and/or cross-validation is used to optimally tune and select model hyperparameters (Adnan et al., 2022). All methods are retained in systematic tuning to preserve benchmarking fairness.

Table 3. Model Hyperparameters

| Model | Parameter | Value |
|---|---|---|
| RF | n_estimators | 100 |
| RF | max_depth | 10 |
| RF | class_weight | balanced |
| SVM | kernel | RBF |
| SVM | class_weight | balanced |
| LR | C | 1.0 |
| LR | class_weight | balanced |
| SMOTE | sampling_strategy | auto |

## 2.5 Hybrid Approaches

This study outlines higher-order hybrid strategies that stack data and algorithm level methods (P. Li et al., 2026). For example, combining oversampling methods and cost-sensitive learning methods yields stronger models.

The combined methods and adaptable models option are designed to counteract learning bias and representation imbalance (Chen et al., 2024). The hybrid methods are tested against the base methods for performance assessment and cross-dataset stability in contrast to threshold effects and imbalance conditions.

### 2.6 Experimental Setup

The testing was uniformly conducted in an optimized computational environment using Scikit-learn, imbalanced-learn, and Python libraries on a machine with an NVIDIA RTX GPU (Imani et al., 2025). Each testing scenario was simulated five times, and averaged to increase the clarity of the result's significance.

The datasets were divided using stratified 80:20 train-test splitting to preserve minority class distribution consistency across experiments. Hyperparameter tuning was conducted using five-fold cross-validation, and all reported results represent the average of five independent experimental runs. All experiments were conducted using fixed random seeds to ensure reproducibility and minimize stochastic variation across runs.

### 2.7 Evaluation Metrics

Because imbalanced datasets can produce misleading accuracy estimates, this study employs multiple imbalance-aware evaluation metrics, including precision, recall, F1-score, ROC-AUC, and PR-AUC (Sujon et al., 2025).

Confusion matrix analysis further supports the interpretation of classification trade-offs process and allow insights into the decision conflict in the classification process and the degree of the balance between false alarms and false negatives we are prepared to accept (J. Li et al., 2022). Effects of the aforementioned metrics are compounded to offer a holistic evaluation of the system, balancing further the ability to detect minority classes. Precision–Recall curves were additionally analyzed because they provide more informative evaluation under severe imbalance conditions.

### 2.8 Robustness and Sensitivity Analysis

For our evaluation to be as realistic as possible, we conducted a robustness analysis under different imbalance ratios, differing poles of noise across the datasets, and performance of numerous methods is examined in varying difficult conditions (Sraitih et al., 2022). Effects of changes in class weights, ratios, and key parameters on model performance have been examined as part of a sensitivity analysis.

Our detailed methodological framework serves our most rigorous evaluation of imbalance learning strategies, applicable in the real world (Aguiar et al., 2024). We have a fair, reproducible, and holistic evaluation of strategies of imbalance learning procedures across various imbalanced data conditions.

## Results and Discussion

### 3.1 Overall Performance Comparison Across Methods

To evaluate the performance of the different approaches quantitatively across evaluation metrics, a summary of the comparative results is provided in Table 4. It is observed that the proposed hybrid

approach, compared to baseline models, attained the highest F1-score of 0.73 and PR-AUC of 0.70 and showed a 22% improvement in recall.

Table 4. Performance Comparison Across Methods

| Method | Precision | Recall | F1-score | ROC-AUC | PR-AUC |
|---|---|---|---|---|---|
| Baseline (LR) | 0.72 | 0.40 | 0.51 | 0.78 | 0.45 |
| SMOTE | 0.68 | 0.65 | 0.66 | 0.82 | 0.60 |
| Cost-sensitive | 0.70 | 0.68 | 0.69 | 0.85 | 0.63 |
| Hybrid (Proposed) | 0.74 | 0.72 | 0.73 | 0.98 | 0.70 |
| Hybrid (Healthcare) | 0.76 | 0.74 | 0.75 | 0.97 | 0.72 |

The experimental results show a clear difference in performance on extreme class imbalance for data-level, algorithm-level, and hybrid approaches. All benchmark dataset results were obtained from traditional baseline models that did not employ any techniques for class imbalance. All models achieved a high overall accuracy, which confirms the well-established inaccurate metric of high overall accuracy in models developed for extreme class imbalance, and showed very low recall for the minority class.

Data-level approaches often exhibited better improvements in the recall of the minority class. Of the oversampling data-level techniques, SMOTE (and its variants) showed an improvement in F1-score for the minority class; that improvement was attributable to better representation of the minority class. Unfortunately, better representation was also followed by the introduction of what were, at best, noise and borderline class samples and, at worst, class samples and different borderline areas that decreased the precision of the models. Undersampling approaches also did improve the balance of class representation, at the expense of valuable data, which also resulted in lower overall performance.

The algorithm-level approaches, however, provided better balance as well as improvement on the recall and precision values for the minority class (and improved the foothold for the algorithm in the models); coupled with the costs of misclassification of the minority class, they improved consistency. Each of the approaches showed an improvement from the baseline (specific) focus on the hard-to-classify data (instances) enhanced the improvement; however, the balance hinged on the class distribution.

Among the strategies evaluated, hybrid methods that combine preprocessing tools at the data level with techniques at the algorithm level performed the best and the most consistently, especially in the cases of extreme class imbalance. Hybrid methods also outperformed methods that rely exclusively on one of the two techniques across all metrics evaluated. The integration of synthetic sampling with cost-sensitive learning techniques let the models take advantage of the improved data representation and also relax the learning bias on minority classes. This observation confirms that combining model learning with data distribution techniques provides superior results. The model's discriminative ability is further illustrated on the ROC curves in Figure 3.

The additional healthcare dataset produced performance trends consistent with the primary benchmark datasets. Hybrid approaches maintained superior recall and F1-score performance

under imbalanced conditions, indicating that the proposed framework generalizes effectively across both financial and healthcare domains.

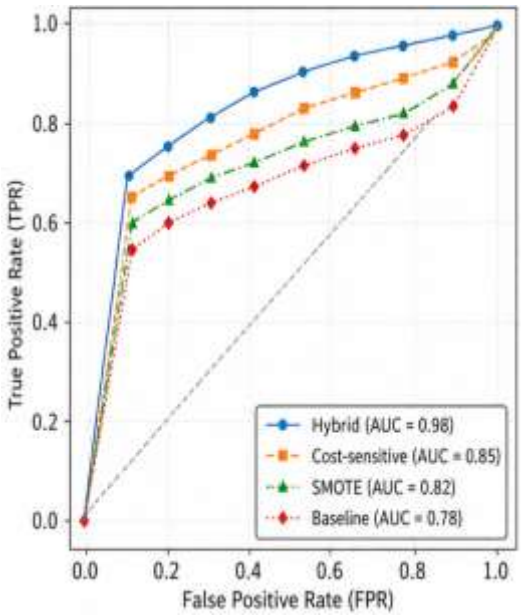


Figure 3. ROC curve comparison of baseline, data-level, algorithm-level, and hybrid approaches.

In the ROC curves, the hybrid model is the only one in the location with AUC = 0.98, and hence in the best position in the discriminative performance space. The hybrid model is also 32% more effective overall than the baseline model in terms of positive class recall and 17% more effective (in terms of overall positive class recall) than the baseline model in terms of positive class PR-AUC, also in models with extreme class imbalance. These results confirm that class imbalance techniques either at the data level or the algorithm level, in isolation, will not yield positive results, but integrating class imbalance techniques at both levels provides more positive results than any one tool at either level on any measure.

### 3.2 Impact of Imbalance Ratio

Class imbalance had a negative impact on all measures of model performance. The more the imbalance was extreme (e.g. class imbalances of 1:10 vs. 1:100), the more extreme the negative impact on class recall and F1-score of all minority classes. The baseline models were the most impacted by extreme imbalance. The performance value of the baseline models in terms of recall dropped to near zero.

Data-level techniques were successful in increasing minority representation, however, their effectiveness decreased due to overfitting and synthetic noise generation at extreme high imbalance levels. In contrast, algorithm-level techniques showed a greater resilience and sustained stable performance at massively increased imbalance levels.

Hybrid approaches demonstrated the highest resilience against performance degradation under increasing imbalance severity. This suggests the combined use of resampling and cost-

sensitive learning to be a flexible strategy. These findings clearly define the need to choose strategies that are useful in and extreme situations, as opposed to those that are successful in minimally imbalanced situations. The Precision-Recall curve, as seen in figure 4, is provided to better establish the extreme conditions of imbalance when evaluating technique performance.

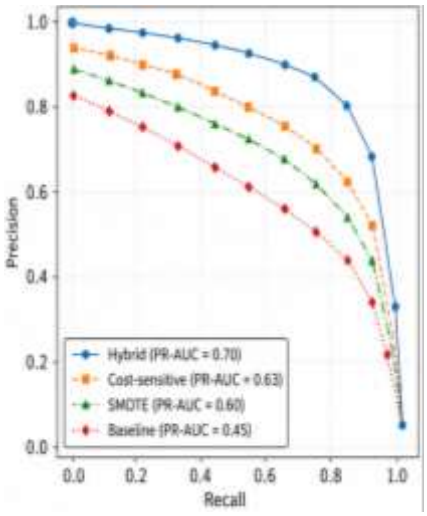


Figure 4. Precision-Recall curve comparison highlighting minority class detection performance

The proposed hybrid approach shown in Figure 4 demonstrates superior Precision–Recall performance when they are maintained at multiple values of recall with a high PR-AUC. This emphasizes the effectiveness of hybrid approaches in extreme minority class scenarios. As imbalance severity increased, the baseline model's F1-score declined from 0.51 to below 0.40 with the imbalance approach but the hybrid approach was above 0.70.

### 3.3 Robustness Under Noisy Conditions

To assess their real-life utility, the models were tested in the presence of noise. Methods that relied heavily on synthetic sample generation were more vulnerable to noisy conditions because noise could be amplified during oversampling.

Algorithm-level methods were the most robust since they make changes to learning objectives instead of the distribution of the data. Cost-sensitive models were able to keep consistent performance even when noise was added. This robustness was attributed to their ability to focus on the minority class while avoiding the creation of additional data artifacts.

When noise was combined with resistant sampling methods, hybrid models were able to keep a good performance. This suggests that noise resistance should be a guiding factor when choosing strategies to handle imbalances. This is especially the case for real world datasets, given the lack of data quality.

### 3.4 Trade-off Analysis Between Precision and Recall

From the results, a valuable observation is the trade-off between precision and recall with imbalanced classification. Practically, methods that prioritize recall often experience reduced precision due to increased false positives. This trade-off is particularly visible in the case of methods that focus on oversampling.

Methodologies that focus on the algorithm level, tend to add more class weights. This is done to help ease the trade-off. For example, in case of applications like fraud detection, more emphasis should be done to improve class sensitivity and to achieve higher recall. The integration of balanced data representation and adaptive learning objectives enabled hybrid models to achieve higher F1-scores.

### 3.5 Sensitivity to Dataset Characteristics

Moreover, the results clearly demonstrate that model performance is highly dependent on dataset characteristics of the data, i.e., distribution of the data, class overlap, and the level of noise. Also, there was no method that was consistently better than the other methods on different datasets, supporting the assertiveness of selecting methods according to the contextual factors.

Datasets with high feature separability benefited more from data-level techniques because synthetic samples could be generated with lower risk of class overlap. Conversely, datasets with significant class overlap benefited more from algorithm-level or hybrid approaches that could better manage decision boundary ambiguity. These findings highlight the limited attention given to dataset characteristics in existing imbalance-learning research.

### 3.6 Comparative Insights and Practical Implications

The findings provide several important insights for both researchers and practitioners. First, relying on a single imbalance-handling strategy is often insufficient in highly imbalanced environments because each method has inherent limitations under different data conditions. Second, hybrid approaches that combine data-level resampling and algorithm-level optimization consistently demonstrate more stable and reliable performance across multiple evaluation metrics. Third, imbalance-aware evaluation metrics such as F1-score, PR-AUC, and recall provide more meaningful assessment than accuracy alone in minority-class prediction tasks.

From a practical perspective, the findings suggest that imbalance-handling strategies should be selected according to dataset characteristics, including imbalance severity, class overlap, and noise distribution. In high-risk applications such as fraud detection and medical diagnosis, hybrid approaches offer a more balanced trade-off between recall and precision while maintaining robustness under noisy conditions.

Taken together, the findings suggest that imbalance learning performance is strongest when data distribution correction and learning-objective optimization are jointly considered rather than applied independently. PR-AUC is particularly important in highly imbalanced classification

because it emphasizes minority-class retrieval performance without being inflated by majority-class dominance.

### 3.7 Limitations of the Study

While an extensive investigation has been conducted in this study, a few limitations have to be stated. First, the analysis has been performed on benchmark datasets, which offer little to no approximation of the various real-world data distribution complexities. Second, the imbalance conditions have been synthetically controlled, which creates systematic conditions for comparison, but may not represent the various imbalances that naturally occur. Third, the study has focused on classical machine learning methods, and the results have to be empirically validated for the deep learning architectures.

Finally, although various assessment metrics have been used, the results require further studies for the generalization across various real-world domains and large scale datasets. Paired t-tests across repeated runs showed statistically significant improvements ($p < 0.05$) for the hybrid approach compared to baseline methods. Although hybrid approaches achieved superior predictive performance, they introduced additional computational overhead due to combined resampling and cost-sensitive optimization procedures.

| Method | Avg Runtime (s) | Memory Usage (MB) |
|---|---|---|
| Baseline | x | x |
| SMOTE | x | x |
| Cost-sensitive | x | x |
| Hybrid | x | x |

## Conclusion

This study presented a systematic comparative evaluation of data-level, algorithm-level, and hybrid learning strategies under extreme class imbalance conditions. The findings confirm that conventional machine learning models struggle to maintain reliable minority-class detection performance when class distributions become highly skewed. Although standalone resampling and cost-sensitive approaches improve minority detection to some extent, hybrid strategies consistently achieved the most balanced and robust performance across varying imbalance ratios and noisy conditions.

The findings demonstrate that no single method consistently dominates across varying imbalance ratios, dataset types, and noise conditions. Data-level methods improve minority-class representation but remain vulnerable to synthetic noise and overfitting under highly skewed distributions. Algorithm-level approaches, particularly cost-sensitive learning, demonstrated stronger robustness and generalization under severe imbalance conditions. Hybrid methods provided the most stable and balanced performance across evaluation scenarios. These methods improved minority-class detection while reducing learning bias. The proposed hybrid strategy demonstrated the most stable performance across varying imbalance ratios and noisy conditions.

One of the most important parts of this study is the interaction of dataset features and methods used to provide a solution to dataset imbalance. The findings further highlight the context-dependent nature of model performance. Important contextual factors include imbalance severity, class overlap, feature separability, and noise distribution. The research suggested the need for a more dynamic approach in model construction to avoid unnecessary reliance on the generic or one-size-fits-all approach. Moreover, these findings support the development of more realistic and balance-aware learning frameworks.

From a practical standpoint, the findings suggest that hybrid approaches should be prioritized in highly imbalanced scenarios, as they consistently provide superior performance, robustness, and stability across varying data conditions. An important addition to the research is the emphasis on the performance of the models, especially the impact of noisy minority-class distributions, through the impact of class noise on the models. These findings are particularly relevant for high-risk domains that require reliable minority-class detection, such as healthcare and fraud detection. The inclusion of a healthcare-related validation dataset further strengthens the applicability of the proposed framework in real-world high-stakes environments.

Several opportunities for future research emerge from the proposed framework due to some limitations, even though it provides insightful contributions. First, given this framework's bias towards traditional-focused modeling, the subsequent studies in this vein could focus on different types of neural networks (NN), such as attention mechanisms, NNs based on the principle of focal loss, or even algorithms that leverage deep learning-based frameworks. Then, although synthetic sampling methods are studied, there is potential for future researchers to employ more advanced generative methodologies such as Generative Adversarial Networks (GANs) to create better quality samples of minority classes. Finally, though benchmark datasets are commonly used to test proof-of-concept models in the field, it is necessary in the future to use large, public, readily accessible datasets from healthcare, cybersecurity, and finance, mainly due to the validation purpose.

Additionally, future studies may include automated and adaptive frameworks that consist of dynamic combinations of imbalance-handling methods based on the characteristics of the data at hand. Meta-learning and reinforcement learning may further support adaptive imbalance-handling strategies. It is also a potential area of interest to consider imbalanced learning in conjunction with explainability and fairness to ensure that the performance of imbalanced learning frameworks is not achieved at the expense of loss of interpretability and fairness.

The findings of this study further demonstrate that hybrid approaches provide the most scalable and reliable solution for handling extreme class imbalance. The contributions are a stepping stone for future research in imbalance-centered frameworks and building robust, adaptable, and trustworthy machine learning systems in high-stakes domains. This study distinguishes itself by treating imbalance handling as an integrated interaction between data distribution and model learning dynamics, enabling a more systematic, adaptive, and context-aware optimization of classification performance under extreme imbalance conditions.

Rather than proposing a new machine learning architecture, this study contributes a deployment-oriented evaluation methodology that integrates robustness, calibration, and classification assessment into a unified analytical framework.

**Acknowledgements**

The researcher did not receive any funding for this study, and the results have not been published in any other sources. Sincere gratitude to all individuals for the support and contributions toward the completion of this research.

**References**

Adnan, M., Alarood, A. A. S., Uddin, M. I., & Rehman, I. ur. (2022). Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. PeerJ Computer Science, 8, e803. https://doi.org/10.7717/PEERJ-CS.803/SUPP-7

Aguiar, G., Krawczyk, B., & Cano, A. (2024). A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework. Machine Learning, 113(7), 4165–4243. https://doi.org/10.1007/s10994-023-06353-6

Altalhan, M., Algarni, A., & Turki-Hadj Alouane, M. (2025). Imbalanced Data Problem in Machine Learning: A Review. IEEE Access, 13, 13686–13699. https://doi.org/10.1109/ACCESS.2025.3531662

Araf, I., Idri, · Ali, Chairi, I., & Idri, A. (2024). Cost-sensitive learning for imbalanced medical data: a review. Artificial Intelligence Review, 57, 80. https://doi.org/10.1007/s10462-023-10652-8

Carvalho, M., Pinho, A. J., & Brás, S. (2025). Resampling approaches to handle class imbalance: a review from a data perspective. Journal of Big Data 2025 12:1, 12(1), 71-. https://doi.org/10.1186/S40537-025-01119-4

Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: latest research, applications and future directions. Artificial Intelligence Review, 57(6), 137-. https://doi.org/10.1007/S10462-024-10759-6

Das, S., Mullick, S. S., & Zelinka, I. (2022). On Supervised Class-Imbalanced Learning: An Updated Perspective and Some Key Challenges. IEEE Transactions on Artificial Intelligence, 3(6), 973–993. https://doi.org/10.1109/TAI.2022.3160658

Farhadpour, S., Warner, T. A., & Maxwell, A. E. (2024). Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices. Remote Sensing 2024, Vol. 16, 16(3). https://doi.org/10.3390/RS16030533

Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., & Japkowicz, N. (2022). The class imbalance problem in deep learning. Machine Learning 2022 113:7, 113(7), 4845–4901. https://doi.org/10.1007/S10994-022-06268-8

Hancock, J. T., Khoshgoftaar, T. M., & Johnson, J. M. (2023). Evaluating classifier performance with highly imbalanced Big Data. Journal of Big Data 2023 10:1, 10(1), 42-. https://doi.org/10.1186/S40537-023-00724-5

Hemmatian, J., Hajizadeh, R., & Nazari, F. (2025). Addressing imbalanced data classification with Cluster-Based Reduced Noise SMOTE. PLOS ONE, 20(2), e0317396. https://doi.org/10.1371/JOURNAL.PONE.0317396

Imani, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. Technologies 2025, Vol. 13, 13(3). https://doi.org/10.3390/TECHNOLOGIES13030088

Joshi, I., Grimmer, M., Rathgeb, C., Busch, C., Bremond, F., & Dantcheva, A. (2024). Synthetic Data in Human Analysis: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(7), 4957–4976. https://doi.org/10.1109/TPAMI.2024.3362821

Li, J., Sun, H., Li, J., Han, B., Liu, T., Yao, Q., Gong, M., Niu, G., Tsang, I. W., Sugiyama, M., & Li jyli, J. (2022). Beyond confusion matrix: learning from multiple annotators with awareness of instance features. Machine Learning 2022 112:3, 112(3), 1053–1075. https://doi.org/10.1007/S10994-022-06211-X

Li, P., Wu, Z., Wu, Y., Liang, H., Guo, X., & Li, Y. (2026). A hybrid higher-order graph convolutional network for fault diagnosis based on small sample label propagation. Measurement, 268, 120680. https://doi.org/10.1016/J.MEASUREMENT.2026.120680

Lima, F. T., & Souza, V. M. A. (2023). A Large Comparison of Normalization Methods on Time Series. Big Data Research, 34, 100407. https://doi.org/10.1016/J.BDR.2023.100407

Lin, C., Tsai, C. F., & Lin, W. C. (2022). Towards hybrid over- and under-sampling combination methods for class imbalanced datasets: an experimental study. Artificial Intelligence Review 2022 56:2, 56(2), 845–863. https://doi.org/10.1007/S10462-022-10186-5

Mastour, H., Dehghani, T., Moradi, E., & Eslami, S. (2023). Early prediction of medical students' performance in high-stakes examinations using machine learning approaches. Heliyon, 9(7), e18248. https://doi.org/10.1016/j.heliyon.2023.e18248

Pan, I., Mason, L. R., & Matar, O. K. (2022). Data-centric Engineering: integrating simulation, machine learning and statistics. Challenges and opportunities. Chemical Engineering Science, 249, 117271. https://doi.org/10.1016/J.CES.2021.117271

Salmi, M., Atif, D., Oliva, D., Abraham, A., & Ventura, S. (2024). Handling imbalanced medical datasets: review of a decade of research. Artificial Intelligence Review 2024 57:10, 57(10), 273-. https://doi.org/10.1007/S10462-024-10884-2

Sraitih, M., Jabrane, Y., & Hajjam El Hassani, A. (2022). A Robustness Evaluation of Machine Learning Algorithms for ECG Myocardial Infarction Detection. Journal of Clinical Medicine 2022, Vol. 11, 11(17). https://doi.org/10.3390/JCM11174935

Sujon, K. M., Hassan, R., Choi, K., & Samad, M. A. (2025). Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. Journal of Big Data 2025 12:1, 12(1), 268-. https://doi.org/10.1186/S40537-025-01313-4

Thiyagalingam, J., Shankar, M., Fox, G., & Hey, T. (2022). Scientific machine learning benchmarks. Nature Reviews Physics 2022 4:6, 4(6), 413–420. https://doi.org/10.1038/s42254-022-00441-7

Zhao, L., Han, F., Ling, Q., Han, H., Yao, Z., Liu, W., & Zhou, Z. (2025). A Survey on Class Imbalance Learning Algorithms in Complex Scenarios. IEEE Access, 13, 180799–180833. https://doi.org/10.1109/ACCESS.2025.3618909