

Investigating Data Imbalance and Its Influence on Machine Learning Model Reliability

Ethan Williams

Professor, American Institute of Machine Learning, USA

Abstract

Class imbalance represents an important methodological challenge in machine learning because many classification datasets contain substantially different numbers of observations across target categories. Such distributions frequently arise when the events of greatest analytical interest occur less often than routine observations. Conventional classification algorithms may consequently prioritize majority-class predictions, producing favorable aggregate accuracy while demonstrating limited sensitivity to the minority class. However, the relationship between class imbalance and predictive reliability is influenced by additional characteristics, including minority sample size, feature overlap, label quality, classification thresholds, and the costs associated with prediction errors. The present paper develops a hypothetical methodological framework for investigating the influence of class imbalance on machine learning model reliability. The proposed investigation compares four Random Forest classification configurations: untreated imbalanced learning, class-weighted learning, random minority oversampling, and the Synthetic Minority Over-sampling Technique. A fictional manufacturing quality dataset containing 18,000 product observations is introduced to demonstrate the intended research design. The dataset includes 17,100 nondefective products and 900 defective products, producing a 95:5 majority-to-minority class distribution.

The proposed methodology incorporates class distribution assessment, feature preparation, batch-independent data partitioning, controlled imbalance treatment, and comparative model evaluation. Precision, recall, F1-score, balanced accuracy, and probability calibration are considered relevant dimensions of classification reliability. Two illustrative analytical tables describe the hypothetical dataset distribution and the comparative performance of selected imbalance-handling strategies. All numerical results are manually constructed and do not represent completed computational experiments. The methodological discussion examines minority-class recognition, false-positive and false-negative trade-offs, sampling-related information loss, synthetic observation reliability, and the limitations of conventional accuracy measures. Particular attention is given to distinguishing improvements in classification metrics from demonstrated operational reliability. The paper concludes that class imbalance should be addressed through integrated data preparation, appropriate evaluation, and transparent model validation rather than through sampling procedures alone.

Keywords: Class Imbalance, Machine Learning Reliability, Imbalanced Classification, Minority-Class Detection, Synthetic Minority Over-sampling Technique, Cost-Sensitive Learning, Random Forest, Precision–Recall Analysis, Predictive Performance, Classification Evaluation.

1. Introduction

Machine learning classification has become an important analytical technique for identifying relationships within structured datasets and assigning observations to predefined categories. Applications frequently involve classifying customer behavior, identifying equipment conditions, detecting unusual transactions, assessing manufacturing quality, and recognizing operational events. The effectiveness of these systems depends on the statistical properties of their training data and the reliability of their predictive evaluation. One important difficulty arises when classification categories are represented by substantially different numbers of observations. This condition, commonly described as class imbalance, can influence model learning and complicate the interpretation of classification performance, particularly when the less frequently observed category has considerable practical importance.

Class imbalance occurs when one classification category contains substantially more observations than another. A manufacturing inspection dataset, for example, may contain many satisfactory products and comparatively few defective items. Similarly, information systems may record numerous normal transactions alongside a smaller number of unusual events. Such distributions do not automatically make classification impossible, but they can create difficulties when the minority class is poorly represented or difficult to distinguish from the majority class. Classification algorithms may identify patterns that reduce overall prediction errors while failing to recognize a substantial proportion of important minority observations. Consequently, favorable aggregate performance may conceal weaknesses relevant to the intended application.

The influence of class imbalance cannot be understood exclusively through the ratio between majority and minority observations. Minority sample size, feature overlap, measurement noise, and the complexity of class boundaries can also affect classification performance. A dataset containing strongly separable categories may be comparatively easy to classify despite unequal class frequencies. Conversely, a moderately imbalanced dataset can remain challenging when minority observations overlap substantially with the majority class. These considerations indicate that class imbalance should be examined as part of a broader learning problem rather than treated as an isolated numerical characteristic. Reliable classification requires an understanding of both the distribution and the informational quality of the available observations.

Several methodological approaches have been developed to address imbalanced classification. Data-level techniques modify the training distribution through oversampling, undersampling, or synthetic observation generation. Algorithm-level approaches adjust model learning through class weights, modified objectives, or cost-sensitive procedures. Ensemble methods can also incorporate resampling or weighting to improve the recognition of less frequently observed categories. However, each strategy introduces potential limitations. Oversampling may increase the influence of repeated observations, undersampling can discard useful majority-class information, and synthetic samples may not accurately represent genuine minority characteristics. Moreover, improved recall may occur alongside increased false-positive predictions or reduced probability calibration.

The present paper proposes a hypothetical comparative investigation of data imbalance and its influence on machine learning model reliability. The research design considers a fictional manufacturing quality classification dataset in which defective products represent 5% of observations. Four Random Forest configurations are examined to distinguish the effects of alternative imbalance-handling procedures while maintaining a common classification algorithm. The proposed evaluation emphasizes

precision, recall, F1-score, balanced accuracy, and probability reliability. Two illustrative tables demonstrate dataset organization and possible classification trade-offs. No actual manufacturing records were collected, and no machine learning experiments were executed. The study provides a methodological framework for future empirical evaluation.

2. Objectives of the Study

The primary objective of this study is to examine how unequal class distributions may influence machine learning classification reliability and how alternative imbalance-handling strategies can be evaluated. The investigation concentrates on minority-class recognition, prediction error trade-offs, and dependable model assessment. Three specific objectives establish the analytical direction of the proposed research framework.

2.1 Examination of Class Distribution and Minority Recognition

The first objective is to examine how an unequal distribution of classification categories may influence the recognition of minority observations. Machine learning models often optimize prediction objectives across all training examples, potentially giving substantial influence to the more frequently represented class. The proposed investigation considers how this condition could affect minority-class sensitivity and the interpretation of aggregate classification accuracy. Particular attention is given to identifying situations in which a model performs favorably on the majority class while failing to detect a meaningful proportion of minority observations.

The investigation further aims to distinguish the influence of class frequency from additional characteristics of the learning problem. Minority sample size, feature overlap, label consistency, and classification boundary complexity may contribute to differences in predictive reliability. Consequently, the proposed framework considers class imbalance as one component of a broader statistical challenge. This objective emphasizes the importance of evaluating the characteristics of minority observations instead of assuming that balancing the class distribution will automatically improve classification performance under every condition.

2.2 Comparative Evaluation of Imbalance-Handling Strategies

The second objective focuses on establishing a consistent framework for comparing alternative methods of handling class imbalance. Untreated Random Forest classification, class-weighted learning, random minority oversampling, and SMOTE are selected to represent different strategies. The intended comparison examines minority-class precision, recall, and F1-score using an independent dataset that preserves the original class distribution. Maintaining the same underlying classification algorithm allows the proposed investigation to concentrate on differences associated with imbalance treatment rather than changes in model architecture.

A further objective involves examining the trade-offs introduced by alternative resampling and weighting procedures. Increasing the influence of minority observations may improve recall while also producing additional false-positive predictions. Synthetic minority sampling may provide new training examples but can introduce unrealistic observations when the original feature space contains substantial class overlap. Consequently, the proposed framework emphasizes balanced performance interpretation rather than maximizing a single metric. This objective seeks to establish whether observed classification changes are consistent with the operational requirements of the prediction task.

2.3 Assessment of Prediction Reliability and Evaluation Integrity

The third objective is to identify appropriate measures for evaluating classifier reliability in the presence of imbalanced categories. Overall accuracy may be insufficient because the majority class can dominate the number of correct predictions. Precision, recall, F1-score, balanced accuracy, and precision–recall curves provide complementary information about classification behavior. The proposed study also considers probability calibration because a model may produce accurate class rankings without providing dependable numerical estimates of minority-event likelihood.

The investigation additionally aims to establish safeguards against methodological errors associated with resampling and model selection. Applying SMOTE or random oversampling before dividing observations into independent datasets can introduce information leakage and misleading performance estimates. Repeated optimization using test data may create further bias. Consequently, the proposed framework emphasizes training-only imbalance treatment, independent validation, and carefully documented evaluation procedures. This objective supports trustworthy model assessment rather than improvements produced through inappropriate experimental design.

3. Review of Literature

3.1 Early Investigations of Class Imbalance and Rare Events

King and Zeng (2001) investigated logistic regression in rare-event settings and examined difficulties associated with probability estimation when positive outcomes occur infrequently. Japkowicz and Stephen (2002) conducted a systematic investigation of class imbalance and considered its relationship with concept complexity, dataset size, and alternative learning procedures. Their research demonstrated that classification difficulties cannot be explained exclusively through class ratios. These contributions established important foundations for understanding the statistical and computational characteristics of learning from unequally represented categories.

Weiss (2004) examined the relationship between rare classes and rare observations within data mining. Batista et al. (2004) investigated alternative strategies for balancing training datasets and demonstrated that imbalance treatment does not necessarily improve learning under all conditions. Van Hulse et al. (2007) subsequently evaluated methods for learning from imbalanced data through comparative experiments. Together, these contributions indicate that effective classification depends on the interaction among class distribution, data characteristics, and model configuration. Their findings support careful experimental assessment rather than routine application of a single balancing procedure.

3.2 Resampling, Cost-Sensitive Learning, and Ensemble Approaches

Chawla et al. (2002) introduced the Synthetic Minority Over-sampling Technique as a method for creating artificial minority-class observations through interpolation. The approach provided an alternative to simply duplicating existing minority examples. Liu et al. (2009) examined exploratory undersampling strategies, including ensemble-based procedures designed to reduce information loss from majority-class observations. These studies established important data-level approaches for handling unequal class distributions. Their methodological relevance lies in demonstrating how changes to the training sample may influence the classifier's attention to less frequently represented categories.

Galar et al. (2012) reviewed ensemble approaches to imbalanced classification, including bagging, boosting, and hybrid strategies. López et al. (2013) investigated the importance of intrinsic data characteristics, including class overlap and noisy observations. Menardi and Torelli (2014) examined

classification assessment under imbalanced conditions and considered procedures for improving statistical reliability. Branco et al. (2016) provided a broader survey of predictive modeling involving imbalanced target distributions. Collectively, these studies emphasize that successful imbalance management depends on data characteristics, methodological assumptions, and the intended prediction objective.

3.3 Classification Metrics, Calibration, and Evaluation Reliability

Breiman (2001) established Random Forest as an ensemble learning method suitable for classification through multiple randomized decision trees. Davis and Goadrich (2006) examined the relationship between precision–recall and receiver operating characteristic curves, while Fawcett (2006) reviewed the interpretation of ROC analysis. He and Garcia (2009) subsequently surveyed learning from imbalanced data and highlighted evaluation difficulties associated with unequal class distributions. These contributions support the use of multiple classification metrics rather than reliance on aggregate accuracy alone.

Niculescu-Mizil and Caruana (2005) examined the relationship between classifier outputs and probability calibration. Zadrozny and Elkan (2002) investigated methods for transforming classifier scores into probability estimates. Hand (2009) examined limitations of conventional ROC-based performance assessment, while Saito and Rehmsmeier (2015) emphasized precision–recall analysis for imbalanced classification. Krawczyk (2016) reviewed continuing challenges in imbalanced learning. Together, these contributions establish that classifier reliability involves discrimination, probability estimation, and evaluation integrity rather than a single reported performance statistic.

4. Research Methodology

4.1 Research Design and Comparative Classification Framework

The present paper adopts a hypothetical comparative quantitative research design for investigating class imbalance in binary classification. A Random Forest classifier is selected as the common predictive algorithm because it provides a consistent foundation for evaluating different imbalance-handling procedures. Four configurations are considered: untreated learning, class-weighted learning, random minority oversampling, and SMOTE-based oversampling. The intended prediction task is the identification of defective manufacturing products from measurements available before final quality inspection. This framework has not been implemented experimentally.

The proposed analytical procedure consists of data quality assessment, feature preparation, batch-independent partitioning, imbalance treatment, model training, and independent classification evaluation. All configurations would use the same initial explanatory variables and a consistent hyperparameter search procedure. Class weighting and resampling would be applied exclusively during model development. The independent test set would retain the original distribution of nondefective and defective products. This arrangement is intended to support fair comparison without artificially changing the operational class prevalence during evaluation.

4.2 Hypothetical Dataset and Imbalance Treatment

For illustrative purposes, the study assumes a fictional manufacturing dataset containing 18,000 product observations distributed across 45 production batches. Each batch contributes 400 observations, including 380 nondefective and 20 defective products. The dataset therefore contains 17,100 majority-

class observations and 900 minority-class observations, representing a 95:5 class distribution. Thirty numerical explanatory variables are assumed to describe machine operating conditions, process-sensor characteristics, and material properties measured before final quality inspection. These records are hypothetical rather than collected or simulated empirical observations.

The proposed partition allocates 27 production batches to training, nine to validation, and nine to independent testing. Random minority oversampling would duplicate selected minority observations within the training data. SMOTE would construct synthetic minority examples using neighboring training observations in the standardized feature space. Class-weighted learning would assign greater importance to errors involving the minority class. All preprocessing parameters and synthetic-sample construction procedures would be estimated exclusively from training data, with appropriate nesting during cross-validation where applicable.

4.3 Model Reliability Measures and Evaluation Procedure

The proposed evaluation would calculate minority-class precision, recall, F1-score, balanced accuracy, and the area under the precision–recall curve. Precision represents the proportion of positive predictions that correspond to defective products, while recall indicates the proportion of actual defective products identified. F1-score summarizes the balance between these measures. Balanced accuracy gives equal weight to sensitivity and specificity. These metrics would be examined alongside the confusion matrix to distinguish changes in minority detection from increases in unnecessary positive predictions.

Probability calibration would be evaluated separately using reliability analysis and suitable scoring measures such as the Brier score. A model trained with altered class proportions may require careful calibration against observations representing the intended deployment prevalence. Hyperparameters and classification thresholds would be selected using development information, while independent test observations would remain unavailable until final evaluation. A completed investigation should report confidence intervals using batch-aware resampling and examine performance across different imbalance ratios. The present manuscript provides only constructed numerical examples.

5. Analysis

5.1 Illustrative Assessment of Class Distribution and Evaluation Bias

Class distribution represents a fundamental characteristic of supervised classification datasets because it influences the relative frequency of prediction outcomes. Table 1 presents the hypothetical allocation of nondefective and defective product observations across three batch-independent evaluation partitions. The original dataset contains 95% nondefective products and 5% defective products. This distribution demonstrates why aggregate accuracy can be misleading. A classifier predicting every observation as nondefective would achieve 95% accuracy while failing to identify any actual defective products.

The hypothetical distribution also demonstrates the importance of maintaining realistic class proportions during independent testing. Artificially balancing the test dataset would change the frequency of positive observations and could produce misleading interpretations of operational precision. The proposed framework therefore limits resampling to training data and preserves the original distribution during validation and testing. These assumptions provide a structured foundation for examining minority-class recognition. They do not describe actual manufacturing processes or empirically observed defect rates.

Table 1: Hypothetical Class Distribution Across Manufacturing Quality Evaluation Partitions

Classification Category	Training Observations	Validation Observations	Test Observations	Total
Nondefective Products (Majority)	10,260	3,420	3,420	17,100
Defective Products (Minority)	540	180	180	900
Total	10,800	3,600	3,600	18,000

The constructed test partition contains 180 defective and 3,420 nondefective observations. These counts establish the mathematical basis for illustrating minority-class classification metrics under different imbalance-handling procedures.

5.2 Illustrative Comparison of Imbalance-Handling Strategies

Classification reliability under class imbalance can be examined through the relationship between correctly identified minority observations and incorrectly generated positive predictions. Table 2 presents hypothetical classification outcomes for four Random Forest configurations. The fictional independent test set contains 180 defective products and 3,420 nondefective products. The results were manually constructed to illustrate possible differences in minority-class precision and recall. They do not represent predictions obtained from executed machine learning models.

The constructed comparison assigns the highest F1-score to SMOTE-based learning and the highest precision to untreated classification. Random oversampling produces higher illustrative recall than class weighting but lower precision. These patterns demonstrate potential trade-offs among imbalance-handling strategies. However, no configuration can be regarded as empirically superior because the experiments were not executed. A genuine investigation would require repeated training, independent testing, probability calibration, and uncertainty assessment before determining whether the observed differences are reproducible or operationally meaningful.

Table 2: Hypothetical Minority-Class Classification Performance Under Alternative Imbalance Strategies

Random Forest Configuration	True Positives	False Positives	Precision	Recall	F1-Score
Untreated Imbalanced Learning	90	30	0.750	0.500	0.600
Class-Weighted Learning	126	54	0.700	0.700	0.700
Random Minority Oversampling	135	90	0.600	0.750	0.667
SMOTE-Based Oversampling	144	72	0.667	0.800	0.727

The illustrative example shows that increasing minority-class recall can coincide with additional false-positive predictions. Therefore, the preferred imbalance-handling strategy cannot be determined from recall alone and should reflect the consequences of both error types.

6. Discussion

6.1 Majority-Class Dominance and Minority Prediction Reliability

Class imbalance can influence machine learning classification because the majority category contributes a larger proportion of training observations. When optimization procedures prioritize aggregate prediction accuracy, models may learn decision boundaries that classify common observations effectively while providing limited sensitivity to rare events. However, this effect depends on the learning algorithm, feature distributions, and complexity of the classification problem. Imbalance should therefore be examined alongside the separability and quality of minority observations rather than treated as an independent explanation for every classification error.

The hypothetical dataset illustrates the limitations of aggregate accuracy when positive events are uncommon. Predicting every test observation as nondefective would result in 95% accuracy despite identifying no defective products. The untreated Random Forest example assigns higher accuracy than this trivial classifier while still missing half of the defective observations. These constructed outcomes emphasize the importance of minority-sensitive evaluation. They do not establish how an actual Random Forest would behave when trained using the proposed manufacturing dataset.

6.2 Sampling Techniques and Classification Trade-Offs

Resampling procedures attempt to modify the influence of class distributions during model development. Random oversampling increases the frequency of existing minority observations, while SMOTE generates synthetic observations through interpolation. These approaches may provide additional emphasis on the minority class, but their effectiveness depends on the quality and structure of the available data. Repeated observations can increase the influence of noise, while synthetic examples may be unrealistic when minority patterns occupy disconnected or overlapping regions of the feature space.

The illustrative comparison demonstrates a possible situation in which resampling increases minority-class recall while influencing precision. Higher recall indicates that more defective products are identified, but additional false-positive predictions may require unnecessary quality inspections. The practical importance of these errors depends on the cost and consequences of each decision. Therefore, a balancing strategy should not be selected exclusively because it produces the highest sensitivity or F1-score. Its performance must be interpreted according to the intended operational objective and independently verified evidence.

6.3 Probability Calibration and Trustworthy Classification

Model reliability involves more than assigning correct class labels. In many applications, decision-makers also require dependable estimates of the probability that an observation belongs to the minority class. A classifier may rank observations effectively while producing probability values that do not correspond accurately to observed event frequencies. Resampling and class weighting can influence these estimates because they alter the training distribution or learning objective. Consequently, improved classification recall does not automatically imply better probability calibration.

Niculescu-Mizil and Caruana (2005) demonstrated the importance of examining probability estimates produced by supervised learning algorithms. Zadrozny and Elkan (2002) investigated methods for improving the relationship between model scores and class probabilities. These contributions support evaluating discrimination and calibration separately. Within the proposed framework, any probability correction would be developed using unresampled validation observations reflecting the intended operating distribution. A completed study should report calibration error and proper scoring measures alongside minority-class classification performance.

7. Challenges

7.1 Minority Sample Scarcity and Class Overlap

Minority-class sample scarcity presents a difficulty because rare observations may provide insufficient information for learning reliable classification boundaries. This problem becomes particularly important when the minority category contains multiple subgroups with different feature characteristics. A model trained on a small collection of positive examples may fail to recognize important forms of the minority class that are absent from the training dataset. Consequently, increasing the numerical representation of existing minority observations does not necessarily provide the missing information required for reliable classification.

Class overlap introduces an additional challenge when minority and majority observations occupy similar regions of the feature space. In such circumstances, even an appropriately balanced dataset may remain difficult to classify because the available measurements do not clearly distinguish the categories. Synthetic oversampling may intensify this difficulty when generated observations fall into ambiguous regions. The proposed framework therefore emphasizes examining feature distributions, minority subgroup representation, and label quality before selecting imbalance treatment. These considerations help distinguish distributional problems from limitations in the available predictive information.

7.2 Synthetic Sampling, Noise, and Information Leakage

Synthetic oversampling techniques may generate artificial observations that do not correspond to plausible examples within the underlying application. SMOTE assumes that interpolation between neighboring minority observations can create useful training examples, but this assumption may be inappropriate when features contain constraints, nonlinear boundaries, or disconnected minority subgroups. Synthetic observations constructed from noisy or incorrectly labeled examples can also reinforce undesirable patterns. Consequently, the suitability of oversampling depends on the characteristics and quality of the original feature space.

Information leakage creates another serious methodological challenge when resampling is conducted before the dataset is divided into training and independent testing partitions. Synthetic examples may incorporate information derived from observations that subsequently appear in the test set. This can produce overly favorable performance estimates and undermine the credibility of model comparisons. The proposed methodology therefore restricts resampling to development data and requires appropriate nesting during cross-validation. Independent evaluation must preserve the original observations and should not contain synthetic training examples.

7.3 Changing Class Prevalence and Operational Decision Reliability

Class distributions may change as operating conditions, data collection practices, and underlying processes evolve. A model developed using one minority-event prevalence may therefore encounter a different distribution during deployment. Such changes can influence precision even when sensitivity and specificity remain relatively stable. Consequently, evaluation results obtained from a fixed class ratio may not generalize directly to all operating environments. Reliable implementation requires monitoring both the frequency of minority events and changes in the statistical relationships among predictive features.

Operational decision reliability also depends on the consequences of false-positive and false-negative classifications. In manufacturing quality assessment, a missed defective product may create different costs from an unnecessary inspection of a satisfactory item. These costs are influenced by product criticality, inspection resources, and quality requirements. The proposed framework therefore emphasizes explicit decision thresholds and cost-sensitive assessment. A classification strategy should be selected through verified performance evidence and application-specific requirements rather than through a universal preference for higher recall or lower overall error.

8. Conclusion

Data imbalance represents an important methodological consideration in machine learning classification because unequal class frequencies can influence model training and complicate performance interpretation. The present paper developed a hypothetical comparative framework examining untreated Random Forest learning, class weighting, random minority oversampling, and SMOTE. Its principal focus was the relationship among minority-class representation, prediction error trade-offs, and model reliability. The proposed methodology demonstrates how alternative imbalance-handling strategies could be evaluated using consistent algorithms, independent test observations, and complementary performance measures.

The illustrative tables demonstrate possible methods for reporting class distributions and comparing minority-class precision, recall, and F1-score. However, all numerical values were manually constructed and do not represent actual manufacturing observations or completed classification experiments. Establishing reliable conclusions would require authentic or reproducibly simulated datasets, implemented models, appropriate validation, and statistical uncertainty assessment. Consequently, the hypothetical results cannot demonstrate that any particular imbalance-handling method provides superior predictive performance under genuine operating conditions.

Overall, the study identifies minority sample scarcity, class overlap, synthetic data reliability, probability calibration, information leakage, and changing class prevalence as important challenges in imbalanced classification. Future empirical investigations should examine these factors across different datasets, algorithms, and operational contexts. Data-level and algorithm-level interventions may improve minority recognition under suitable conditions, but their benefits must be established through independently verified evidence.

References

1. 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, 4165–4243. <https://doi.org/10.1007/s10994-023-06353-6>

2. Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20–29. <https://doi.org/10.1145/1007730.1007735>
3. Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, 49(2), Article 31. <https://doi.org/10.1145/2907070>
4. Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249–259. <https://doi.org/10.1016/j.neunet.2018.07.011>
5. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
6. 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, Article 137. <https://doi.org/10.1007/s10462-024-10759-6>
7. Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9268–9277). <https://doi.org/10.1109/CVPR.2019.00949>
8. Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240). Association for Computing Machinery. <https://doi.org/10.1145/1143844.1143874>
9. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
10. Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4), 463–484. <https://doi.org/10.1109/TSMCC.2011.2161285>
11. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyi, Y., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73, 220–239. <https://doi.org/10.1016/j.eswa.2016.12.035>
12. Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing* (pp. 878–887). Springer. https://doi.org/10.1007/11538059_91
13. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks* (pp. 1322–1328). IEEE. <https://doi.org/10.1109/IJCNN.2008.4633969>
14. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
15. Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5), 429–449. <https://doi.org/10.3233/IDA-2002-6504>
16. Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6, Article 27. <https://doi.org/10.1186/s40537-019-0192-5>
17. Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5, 221–232. <https://doi.org/10.1007/s13748-016-0094-0>

18. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2980–2988). <https://doi.org/10.1109/ICCV.2017.324>
19. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), Article e0118432. <https://doi.org/10.1371/journal.pone.0118432>
20. Exploratory analysis of methods, techniques, and metrics to handle class imbalance problem. (2024). Procedia Computer Science, 235, 863–877. <https://doi.org/10.1016/j.procs.2024.04.082>