RESEARCH PAPER

A Comprehensive Analysis Of Unbalanced Datasets, Impacts And Mitigation Strategies In Machine Learning

Sayantan Roy[1], Debjyoti Bagchi[2], Sreemoyee Pradhan[3], Samir Biswas[4], Prodip Deb[5], Dr Pranam Paul[6]

[1]Deputy DPI and OSD (CS Branch), Department of Higher Education, Govt. Of West Bengal, Bikash Bhavan, Kolkata - 700091
Email: roysayantan6@gmail.com
[2]Assistant Professor Department of Computer Application, Global Institute of Management and Technology Palpara More, NH-34, Krishnagar, Nadia, West Bengal 741102, India
Email: debjyotibagchi1982@gmail.com
[3]Final Year Student Bachelor of Technology, Computer Science and Engineering Calcutta Institute of Engineering and Management 24, 1A, Chandi Ghosh Rd, Ashok Nagar, Tollygunge, Kolkata, West Bengal 700040, India
Email: sreemoyeepradhan@gmail.com
[4]Assistant Professor Department of Information Technology, Haldia Institute of Technology Hatiberia, ICARE Complex, Haldia, East Medinipur, West Bengal 721657, India
Email: samirbiswas150282@gmail.com
[5]Assistant Professor Department of Computer Application, Global Institute of Management and Technology Palpara More, NH-34, Krishnagar, Nadia, West Bengal, Pin-741102
Email: deb.prodip@yahoo.com
[6]Professor and Dean-Academics Department of Computer Science and Engineering, Global Institute of Management and Technology Palpara More, NH-34, Krishnagar, Nadia, West Bengal, Pin-741102
Email: pranam.paul@gmail.com

Corresponding Author
Dr Pranam Paul
Professor and Dean-Academics
Department of Computer Science and Engineering, Global Institute of Management and Technology
Palpara More, NH-34, Krishnagar, Nadia, West Bengal, Pin-741102
Email: pranam.paul@gmail.com

ABSTRACT
The problem of class imbalance is frequently observed in real-world machine learning datasets. Due to a lack of samples for various attack methods, this issue is especially apparent in intrusion detection. Ignoring the imbalance problem or developing the machine learning classifier on classes that are not full, particularly in classification problems, when one class significantly outnumbers others. Because of this imbalance, the minority class is frequently the class of greatest interest, resulting in biased models that perform worse. In addition to highlighting important problems, the study evaluates several approaches and examines how effectively machine learning models handle unpredictable input. Cost-sensitive learning (based on temporal complexity), resembling, and ensemble approaches are among the strategies that have been studied...

Keywords: Imbalance dataset; Machine learning; SMOTE; Ensemble methods; Cost-sensitive learning; Resampling techniques; Evaluation metrics; Synthetic data generation; Minority class learning.
How to cite this article: Roy S, Bagchi D, Pradhan S, Biswas S, Deb P, Paul P. A Comprehensive Analysis Of Unbalanced Datasets, Impacts And Mitigation Strategies In Machine Learning. Int J Drug Deliv Technol. 2026;16(49s): 22-39. DOI: 10.25258/ijddt.16.49s.3

Source of support: Nil.
Conflict of interest: Nil.

INTRODUCTION

1.1 *Importance Of A Balanced Dataset In Supervised Learning*

The quality, quantity, and diversity of the datasets used for training immensely effect the efficiency with which machine learning (ML) systems perform. A prevalent issue in machine learning is class imbalance, which results in biased models that perform poorly on minority classes. Based on recent studies, rethinking algorithmic and data strategies is essential for effective solutions. Kang et al. (2020) demonstrated that decoupling the classifier and representation improves long-tailed recognition, while Jamal et al. (2020) proposed meta-learning-based rebalancing. Balanced learning is further enhanced by empirical and architectural techniques like the "bag of tricks" by Zhang et al. (2021) and BBN (Zhou et al., 2020).

*Author for Correspondence: pranam.paul@gmail.com

Beyond vision, Singh et al. (2023) designed Batch-Balanced Focal Loss to address sampling and cost sensitivity together, while Henning et al. (2023) explored imbalance issues in NLP. Even though Sun et al. (2017) empirically showed that increasing data volume significantly enhances deep learning performance, particularly when label noise is reduced, Bottou and Vapnik (1992) established a theoretical basis for how large datasets can improve generalization. But diversity in datasets is equally essential for generalization. Torralba and Efros (2011) presented the idea of "dataset bias" by highlighting that object detectors trained on one dataset frequently exhibit poor performance when tested on other datasets. Bender and Friedman (2018) stressed on the importance for inclusive and representative datasets in natural language processing (NLP) by indicating that models trained on English-dominant corpora perform poorly on underrepresented dialects and languages. Although concerns about overfitting to benchmarks have been expressed, benchmark datasets like ImageNet (Deng et al., 2009), GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), and MIMIC-III (Johnson et al., 2016) helped speed up research progress by enabling consistent evaluation and comparison (Gorman & Bedrick, 2019).Additionally, bias and imbalance in datasets have ethical as well as technological repercussions: While Challen et al. (2019) warned that underrepresentation in medical datasets can result in biased clinical decision-making, Buolamwini and Gebru (2018) observed that skewed training data caused significant racial and gender differences in commercial facial recognition systems. All of these studies establish that in order to create fair, accurate, and generalizable machine learning systems, dataset design— guaranteeing that diversity, balance, and representation are present—is equally vital to model architecture. Based upon the aforementioned advances, this paper presents a comprehensive review of dataset imbalance, its impacts on learning, and modern mitigation strategies across domains.

Machine learning models have been widely adopted across various industries due to their ability to extract insights and make predictions from data. However, the performance of these models is highly dependent on the quality and distribution of the data used. One critical aspect is class distribution, which refers to the allocation of labeled instances across different categories. These labeled groups may be formed either by ensuring an equal number of samples across classes or by grouping instances based on similar attributes or event features. Ideally, each class should have a balanced representation. However, in practice, this is often not the case. A common challenge in many real-world datasets is class imbalance, where one class (the majority class) significantly outnumbers another (the minority class) [18, 19, 20]. This imbalance can lead to predictive models that are biased toward the majority class, thereby reducing their ability to accurately classify instances from the minority class [21, 22]. Addressing class imbalance is crucial, particularly in high-stakes domains such as medical diagnosis, anomaly detection, and the prediction of rare events. The following illustration highlights the working principles of machine learning and emphasizes the critical role of data distribution in achieving robust model performance.

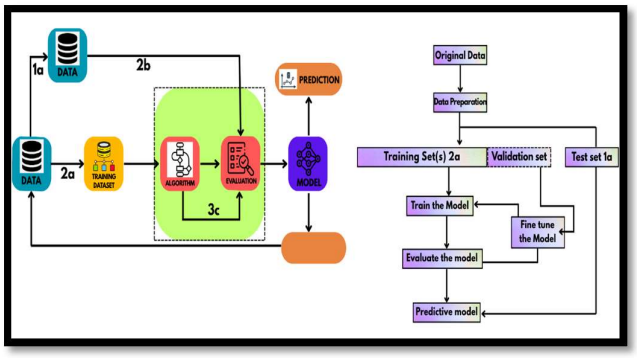

Fig. 1. Importance of Data in Machine Learning

1.2 *Literature Review*

To address the problem of class imbalance, extensive research has been conducted in recent years. Maede Zolanvari [23] explores the application of machine learning in developing security systems for IoT devices. While the study acknowledges the challenge posed by imbalanced datasets, it does not propose a specific solution. Similarly, Khaled Mahmud Sujon et al. [24] analyze the behavior of various machine learning algorithms in the context of a particular problem but do not focus on mitigating class imbalance.

Wendi Qu et al. [25] demonstrate that using balanced datasets in medical applications significantly improves the performance of machine learning algorithms. Although they employ several techniques to balance the data, the study does not identify which method was most effective. In contrast, Ming Zheng et al. [26] utilize the Synthetic Minority Oversampling Technique (SMOTE) to balance imbalanced datasets and highlight its positive impact on algorithm performance.

To further address the imbalance issue, Ngan Tran et al. [27] propose a combination of oversampling, under sampling, and SMOTE to convert an imbalanced dataset into a balanced one. Similarly, Gözde Karataş et al. [28] apply SMOTE prior to implementing machine learning models for security systems. Pradeep Kumar et al. [29] discuss the influence of class imbalance on commonly used machine learning algorithms, while Swati V. et al. [30] provide a comprehensive literature survey that underscores the widespread impact of class imbalance on machine learning model performance.

Data-level methods, such as oversampling and undersampling, are very often used for their adaptability. Techniques like SMOTE and its variants—Borderline-

SMOTE, ADASYN, SMOTE-ENN, SMOTE-Tomek, and Cluster-SMOTE—interpolate new samples for the minority class, though they carry risks like overfitting and the generation of implausible points, especially without careful parameter tuning (Qinghua et al., 2021[31]; Lacoste-Julien et al., 2024[32]). Wongvorachan et al. (2023)[33] found that while random oversampling works under moderate imbalance, hybrid methods like SMOTE-NC with random undersampling perform better under extreme conditions. Cleaning techniques such as SMOTE-ENN and SMOTE-Tomek are used post-synthesis to mitigate noise, and Hemmatian et al. (2025)[34] proposed a cluster-reduced-noise SMOTE that outperforms others in medical datasets. Undersampling methods like RUS shrink training size but risk losing informative data; He et al. (2023)[35] showed that ROS works for moderate imbalance, while SMOTE-RUS is better for rare classes, cautioning that accuracy often hides low recall. Ensemble undersampling approaches like EasyEnsemble and BalanceCascade (Liu et al., 2009)[36] preserve majority information through boosting but are computationally intensive. Zhao et al. (2022)[37] used EasyEnsemble with XGBoost for diabetes detection, achieving strong F1 and AUC. Generative models such as cGANs, WGAN, CTGAN (Xu et al., 2019)[38], and WGAN-GP have shown promise in synthesizing realistic samples, especially in healthcare and finance (Lim et al., 2023[39]; Adiputra et al., 2024[40]; Gangwar et al., 2019)[41], but they demand extensive tuning and risk mode collapse or producing unrealistic data. Algorithmic solutions like cost-sensitive learning adjust misclassification costs, with Huo et al. (2022)[42] presenting a density-aware model that learns cost matrices from data, improving ROC-AUC and PR-AUC in clinical datasets. Class weighting and boosting (e.g., AdaBoost, XGBoost) also help by emphasizing hard-to-classify examples, though overfitting remains a concern. Specialized loss functions have gained traction, including focal loss (Lin et al., 2017)[43], which emphasizes difficult samples; Boldini et al. (2022)[44] demonstrated its effectiveness in bioassays, albeit with potential trade-offs in recall. LDAM, logit-adjusted loss, and Unified Focal Loss (Yeung et al., 2022)[45] have shown domain-specific gains, though no single loss dominates across tasks. Evaluation metrics such as F1-score, ROC-AUC, PR-AUC, MCC, and Cohen's kappa are favored over accuracy, as highlighted by Hemmatian et al. (2025) and Boldini et al. (2022), with MCC being especially informative yet underused. Domain context shapes method choice: recall and AUC-ROC dominate clinical use, while precision and MCC matter in finance and bioinformatics. In healthcare, where malignant cancer cases may comprise under 10% of mammograms, Douzas et al. (2018)[46] combined SMOTE with Balanced Random Forests to increase sensitivity, though class overlap raised false positives. For rare disease prediction (<0.01%), Krawczyk et al. (2016)[47] used one-class SVMs and transfer learning to improve recall, but domain shifts posed challenges. In ICU mortality risk, Harutyunyan et al. (2019)[48] used cost-sensitive neural networks and focal loss to raise AUC, though sensitive tuning introduced overfitting risks. In finance, class imbalance plagues fraud detection and loan default prediction: Pozzolo et al. (2015)[49] used SMOTE-ENN with XGBoost, while Bahnsen et al. (2016)[50] applied Bayesian oversampling with cost-sensitive logistic regression, both improving AUC but facing high false positives or cost ratio issues. Insurance fraud detection improved by 20% using Isolation Forests and autoencoders, though these were preprocessing-dependent. In cybersecurity, Yin et al. (2017)[51] combined LSTM and CNN for intrusion detection, achieving F1 scores above 95% but failing against zero-day attacks. Phishing and Android malware detection have used cost-sensitive models and hierarchical classification, though both require expert input and careful tuning. In NLP, class imbalance challenges hate speech detection, NER in low-resource languages, sarcasm detection, and legal prediction. BERT-based models with focal loss and back-translation improved hate speech detection (HateXplain, Twitter), but sometimes compromised contextual fidelity. For NER, Pires et al. (2019)[52] used multilingual BERT with weighted CRFs to address entity imbalance, though syntactic mismatches caused alignment issues. Sarcasm detection—under 5% prevalence—has used dual-channel CNNs with minority-focused losses, but cultural nuances resist synthesis. Legal AI faces imbalance in tasks like recidivism prediction and bail decisions; fairness-aware cost-sensitive models and interpretable methods like SHAP (Kleinberg et al., 2018)[53] enhance transparency but risk performance trade-offs and bias persistence. In criminal case outcome prediction, ensemble SVMs and argument mining have boosted accuracy but struggled with jurisdictional variation. Recent advances, such as Borderline-SMOTE (Han et al., 2015)[54], ADASYN (He et al., 2008)[55], and ensemble approaches like EasyEnsemble, improved recall in tasks like credit scoring and spam detection. Cluster-SMOTE (Nguyen et al., 2011)[56] reduced outliers in bioinformatics, while CTGAN (Xu et al., 2019), EDA, and Conditional Text Generation (Kumar et al., 2021)[57] improved performance in tabular and text tasks. Diffusion models and ViTs (Zhang et al., 2021[58]; Dosovitskiy et al., 2021)[59] enhanced image data generation, though computational demands and semantic drift persist. Meta-SMOTE (Sinha et al., 2022)[60] and contrastive self-supervised learning like Balanced SupCon (Khosla et al., 2020)[61] show promise in medical and image domains, provided large unlabeled datasets and expert tuning are available. Ultimately, class imbalance mitigation is highly context-sensitive, requiring careful method selection, metric alignment, and algorithm tuning. No universal solution exists, but integrating technical innovations with domain expertise—as underscored by Wongvorachan et al., He et al., Boldini et

al., Huo et al., Yeung et al., and Adiputra et al.—remains key to achieving robust, real-world outcomes.

1.3 *Causes Behind Neglect Of Imbalance In Dataset*

Despite its well-documented impact on model fairness and performance, the issue of dataset imbalance often remains under-addressed in practical machine learning applications. A key contributor to this oversight is the over-reliance on accuracy as a performance metric, which fails to reflect model competence on the minority class. As noted by He and Garcia (2009)[62], classifiers trained on imbalanced data can achieve deceptively high accuracy by favoring the majority class, yet perform poorly on rare but critical instances. This creates a false sense of model effectiveness among practitioners who prioritize aggregate metrics over class-specific ones.

Business-driven time constraints further exacerbate the issue. Developers working under commercial or deployment pressure may favor fast solutions over rigorous data examination, often deploying models without investigating class distribution (Haixiang et al., 2017)[63]. In such environments, minority class underperformance—such as undetected fraud, overlooked diagnoses, or missed threats—may be dismissed as acceptable noise, even though it may carry significant consequences.

Another reason for the lack of concern is the assumption that training data is inherently representative. However, data often reflects societal and systemic biases. For example, Buolamwini and Gebru (2018)[64] found that commercial gender classification systems performed significantly worse on darker-skinned and female faces due to underrepresentation in the training data. This underrepresentation, if not addressed, leads to biased decision-making that disproportionately harms marginalized groups.

Additionally, many practitioners may lack the necessary theoretical grounding in imbalance learning. As Galar et al. (2012)[65] observed, handling class imbalance effectively often requires specialized knowledge, including the use of resampling, cost-sensitive algorithms, or ensemble techniques—approaches that are not commonly taught in standard machine learning curricula. This leads to a gap between academic advancements and their adoption in real-world systems.

Tooling defaults also contribute to the neglect. Widely used machine learning libraries such as Scikit-learn and Keras use standard loss functions (e.g., cross-entropy) and model configurations that assume balanced class distributions unless explicitly modified (Fernández et al., 2018)[66]. Without intentional intervention—such as adjusting class weights or employing resampling—models inherently learn to bias toward the dominant class.

Finally, the real-world consequences of imbalance are often invisible to developers. In domains like hiring, credit scoring, or insurance, feedback on erroneous minority class predictions is rarely immediate or transparent. In contrast, critical sectors like healthcare and security, where false negatives can have life-or-death implications, tend to attract more scrutiny. For instance, Obermeyer et al. (2019)[67] revealed that a widely used healthcare algorithm systematically underestimated the health needs of Black patients due to biased historical data distributions.

- *False Generalization from Balanced Benchmarks*
Many popular benchmarks (e.g., MNIST, CIFAR-10, IMDB) are artificially balanced, leading practitioners to believe that real-world datasets behave similarly. This creates a misleading expectation that standard algorithms will perform well across all distributions (Buda et al., 2018)[68].
- *Absence of Incentives for Fairness or Recall*
In domains where true positive rate for the minority class isn't tied to KPIs, like content recommendation or ad click prediction, developers may be incentivized to maximize overall precision or profit, even at the expense of minority class performance (Barocas et al., 2019)[69].
- *Data Privacy and Legal Constraints on Minority Labels*
Especially in healthcare and finance, minority labels (e.g., rare diseases, fraud cases) are protected or hidden due to privacy laws (HIPAA, GDPR), making it harder to build a balanced dataset. This discourages developers from actively tackling imbalance, as the cost of access outweighs perceived benefit (Rieke et al., 2020)[70].
- *Overconfidence in Deep Learning's Capacity to Self-Correct*
With the rise of deep learning, there is a misconception that model capacity alone can overcome imbalance. Researchers may rely on massive architectures expecting that over-parameterization and data augmentation will resolve class disparity without explicitly handling it (Johnson & Khoshgoftaar, 2019)[71].
- *Neglect in AutoML and Black-Box Pipelines*
AutoML frameworks often optimize pipelines based on global metrics (e.g., accuracy, AUC) and may fail to apply imbalance-aware preprocessing unless explicitly configured. This "fire-and-forget" approach promotes convenience over rigor (Hutter et al., 2019)[72].
- *Label Noise Disproportionately Affects Minority Classes*
As minority class examples are fewer, they are more susceptible to label noise, yet this issue is rarely accounted for in standard preprocessing workflows. This can discourage attempts to oversample or reweight minority

classes out of fear of amplifying noise (Northcutt et al., 2021)[73].

- *Human Bias Toward Represented Classes in Annotation* In human-annotated datasets, annotators are biased to reinforce majority class labels (confirmation bias). This creates annotation-induced imbalance that researchers either fail to detect or assume to be intrinsic to the domain (Sendak et al., 2020)[74].
- *Underreporting of Imbalance Metrics in Published Results* Many papers, especially outside core ML venues, report overall metrics without disaggregating results by class, making it hard to spot imbalance effects. This perpetuates the idea that imbalance is not a problem worth addressing (Saito & Rehmsmeier, 2015)[75].

Although the adverse impacts of class imbalance on model fairness and performance are well established, the problem is frequently overlooked in real-world machine learning implementations. The over-reliance on accuracy as the main performance criterion, which fails to take a model's capacity into consideration to accurately recognize instances of minority classes, is a major contributor in this neglect. Classifiers trained on unbalanced data may exhibit deceptively high accuracy by disproportionately favoring the majority class while underperforming on critical but rare cases, as He and Garcia (2009)[76] demonstrate. This offers practitioners, who prioritize aggregate metrics over class-specific assessments, a false sense of confidence.

Time constraints imposed by business aggravate the issue. Class distribution analysis is frequently ignored by developers working under commercial or deployment pressure since they prioritize swift, deployable solutions above exhaustive data analysis (Haixiang et al., 2017). As a result, even though it may have significant implications, underperformance on the minority class—such as fraud that goes unnoticed, incorrect diagnoses, or overlooked security threats—is typically passed off as harmless noise.

The assumption that training data is inherently representative of distributions in the real world is another fundamental problem. In reality, societal and institutional biases are often visible in datasets. For example, Buolamwini and Gebru (2018) revealed that while darker-skinned and female faces were underrepresented in training data, commercial facial classification algorithms performed considerably worse on these features. These discrepancies can lead to biased decision-making that disproportionately impacts marginalized communities if they are not rectified.

The deceptive generalization derived from well-known benchmark datasets such as MNIST, CIFAR-10, and IMDB—all of which are artificially balanced—exacerbates the situation. According to Buda et al. (2018), these benchmarks create irrational assumptions that standard algorithms will function equally well in unbalanced real-world situations. Furthermore, since minority class recall measures are rarely linked to commercial KPIs, there is no incentive to optimize them in domains like content recommendation or ad click prediction. As a result, developers might choose overall accuracy or financial gain over fairness (Barocas et al., 2019).

Furthermore, a lot of practitioners lack the theoretical foundation required to effectively address imbalance. Handling class imbalance involves specialized knowledge of resampling techniques, cost-sensitive algorithms, or ensemble approaches—topics that are frequently absent from traditional machine learning programs, as Galar et al. (2012) pointed out. Adoption of imbalance-aware solutions is substantially slowed by this gap between academics and practice.

The condition is also further worsened by tooling limitations and technical defaults. Unless explicitly transformed, popular libraries like Scikit-learn and Keras generally assume balanced class distributions and employ standard loss functions like cross-entropy (Fernández et al., 2018). In the absence of deliberate measures such as class weighting or resampling, models will naturally try to achieve optimal performance in the majority class.

Moreover, overconfidence in a model's ability to self-correct for imbalance has arisen from the advancement of deep learning. A prevalent misconception is that skewed distributions can be automatically handled by large-scale data augmentation or more complicated models (Johnson & Khoshgoftaar, 2019). Developers are deterred from using explicit imbalance management solutions by this assumption. Similarly, unless specifically instructed, AutoML frameworks and black-box pipelines frequently optimize for global metrics like accuracy or AUC, omitting the inclusion of imbalance-aware preprocessing (Hutter et al., 2019).

This is deteriorated by the fact that minority class examples are naturally more vulnerable to label noise owing to their low frequency. This problem generally goes unnoticed by standard preprocessing workflows, and developers are discouraged from adopting strategies like oversampling or reweighting as they are concerned about amplifying noise (Northcutt et al., 2021). By reinforcing majority class labels, as the result of confirmation bias, human annotation bias further distorts data, creating an imbalance brought on by annotations that is often wrongly regarded as inherent to the dataset (Sendak et al., 2020).

Lastly, developers frequently overlook the tangible consequences of minority class misclassification. In industries such as insurance, lending, or hiring, feedback on misclassifications could be unavailable, ambiguous, or delayed. On the other hand, imbalance is typically examined more thoroughly in high-stakes fields like healthcare and security, where false negatives can cause serious harm. Obermeyer et al. (2019), for instance, found

that a popular healthcare algorithm systematically underestimated Black patients' treatment needs because of skewed historical data distributions. Such instances illustrate how crucial it is to proactively address class imbalance, not only for performance but also for ethical and fair decision-making.

1.4 *Areas Where Dataset Imbalance Is Widely Prevalent*

One critically underserved domain is climate risk assessment and extreme weather event forecasting. Datasets in this field are typically dominated by normal or non-extreme weather records, while occurrences of rare but catastrophic events such as cyclones, flash floods, or heatwaves remain drastically underrepresented. This imbalance leads to poor generalization and inadequate early warning systems. Despite the growing urgency of climate change, relatively few studies have tackled this imbalance explicitly. In this context, hybrid approaches combining cost-sensitive learning with synthetic oversampling methods like ADASYN would be highly effective. ADASYN adapts the generation of synthetic samples based on local data density, ensuring that hard-to-learn events like rare floods receive more attention. When combined with cost-sensitive deep neural networks, models can better prioritize minimizing false negatives, which are critical in emergency forecasting.

Another domain that has largely overlooked the impact of data imbalance is supply chain anomaly detection, especially in global logistics and maritime tracking. In most datasets, the majority of shipments are timely and correct, while only a few exhibit anomalies like rerouting, delays, or cargo loss. These minority-class events are crucial for optimizing delivery networks, yet imbalance causes models to treat them as noise. Here, a combination of Isolation Forests for unsupervised anomaly detection and time-series aware augmentation techniques like TimeGAN can be particularly suitable. Isolation Forests can detect subtle statistical deviations without needing labeled examples, and TimeGAN can synthesize rare event sequences while preserving temporal dynamics, making the model robust to irregular patterns in the supply chain.

Mental health diagnostics using digital phenotyping is another domain where dataset imbalance is both prominent and underexplored. While digital footprints (e.g., smartphone usage, social media activity, mobility data) are abundant, samples labeled with confirmed episodes of conditions such as bipolar disorder or schizophrenia are scarce due to privacy, stigma, and low clinical diagnosis rates. Moreover, most studies disproportionately represent depressive symptoms. A powerful method here would be to adopt semi-supervised learning in conjunction with anomaly detection frameworks, such as using graph-based semi-supervised label propagation along with autoencoders trained on healthy behavior. This method leverages the large volume of unlabeled behavioral data while focusing representation learning on atypical digital signatures of mental illness, reducing overfitting and enhancing interpretability.

The art authentication and forgery detection industry also lacks systematic handling of data imbalance. Original works of art are vastly outnumbered by unknown or unverified pieces, and forgeries are particularly rare in digitized collections. Yet their detection has massive implications for the art market and heritage conservation. Few-shot learning techniques such as Prototypical Networks, when combined with metric-based contrastive learning, can be effective here. These models excel in situations where training data per class is minimal, and the emphasis is on learning discriminative features across few authentic samples. By embedding artwork into a semantic space where proximity reflects authenticity, the model can flag stylistic anomalies or deviations in brushstroke patterns indicative of forgery.

Another under-addressed area is agricultural disease detection in precision farming, especially for crops that are either less commonly cultivated or less studied, such as millets, pulses, or indigenous vegetables. Most public datasets focus on common crops like wheat, maize, or rice, leading to poor generalization when models are deployed across diverse geographies. Furthermore, disease occurrence in these crops is often seasonal and rare. In this case, few-shot classification using attention-based convolutional networks, potentially enhanced by meta-learning approaches like MAML, can offer viable solutions. These models adapt quickly to new tasks with limited labeled samples, making them ideal for identifying rare diseases in low-resource agricultural environments.

In humanitarian crisis monitoring and disaster relief logistics, the occurrence of conflict zones, displacement events, or sudden resource shortages are infrequent compared to ongoing humanitarian operations. Despite the criticality of predicting such events, most decision support systems use heuristics and ignore the imbalance problem. Ensemble-based anomaly detection, especially using techniques like LOF (Local Outlier Factor) and Bayesian networks integrated with reinforcement learning for adaptive resource planning, could be highly effective. Such systems would learn from irregular resource demand patterns and outliers in mobility or communication data, flagging early indicators of crisis escalation. Their flexibility allows updates based on dynamic data feeds from satellite imagery and field reports.

Cultural heritage language translation and revitalization, involving endangered or indigenous languages, represents another major field where data imbalance is often neglected. For many of these languages, only sparse textual

data or oral transcripts exist, while majority languages dominate NLP corpora. Cross-lingual transfer learning combined with zero-shot or few-shot learning, especially through models like mT5 or XLM-R with auxiliary loss functions emphasizing minority language tokens, can provide breakthroughs. Additionally, back-translation-based augmentation using related dialects or historical texts can further mitigate imbalance, allowing fairer language representation and supporting global cultural preservation.

Finally, consumer protection and product safety complaint classification, such as those lodged with regulatory bodies (e.g., FDA, CPSC), often suffer from long-tail distributions, where critical cases like severe injuries or hazardous materials are dwarfed by general complaints. Yet, automated triaging systems largely fail to prioritize these rare but life-threatening instances. Here, hierarchical classification models combined with dynamic class-weighting and contextual attention mechanisms are most appropriate. These models can not only detect complaint categories but also highlight severity within class hierarchies, ensuring that rare but high-impact reports receive timely attention. Their interpretability further supports regulatory transparency and public accountability.

1.5 *Effects Of Dataset Imbalance In The Last Ten Years*

Over the past decade, the failure to properly address dataset imbalance in machine learning has led to several real-world disasters across diverse industries, exposing both ethical and operational vulnerabilities. In healthcare, a commercial algorithm evaluated by Obermeyer et al. (2019) systematically underestimated the health risks of Black patients by using healthcare expenditure as a proxy for medical need—resulting in unequal treatment recommendations and reinforcing racial disparities in care. Similarly, in the autonomous vehicle industry, Wilson et al. (2019)[77] found that pedestrian detection systems trained on skewed datasets were significantly less accurate at identifying darker-skinned individuals, heightening the risk of fatal accidents. In finance, algorithmic lending decisions have frequently mirrored historical biases; ProPublica (2016)[78] and subsequent investigations revealed that credit scoring models discriminated against minorities, and in 2019[80], Apple's credit card algorithm was criticized for assigning lower credit limits to women despite comparable financial profiles—due to the use of gender-insensitive training data. In the NLP-driven hiring domain, Amazon scrapped an AI-based recruitment tool in 2018[81] after it was found to penalize resumes with female indicators like "women's chess club," a direct result of training on predominantly male resumes, which perpetuated gender bias. In cybersecurity, models trained on predominantly benign traffic have shown poor performance in detecting rare but critical intrusion patterns, such as those involved in the 2017 Equifax breach[82], where malicious data exfiltration behavior went undetected. Legal analytics have similarly suffered[83]; the COMPAS algorithm, used in U.S. courts for risk prediction, was found by ProPublica (2016) to disproportionately classify Black defendants as high-risk for recidivism due to imbalanced arrest and sentencing data, resulting in more severe judicial outcomes. Even climate modeling has not been spared—Zheng et al. (2015)[79] highlighted that rare yet devastating events like urban flash floods were often underpredicted due to the dominance of normal weather data in training sets, as seen during Hurricane Harvey (2017), where ineffective alerts worsened disaster response. These failures illustrate how dataset imbalance is far from a benign technical issue; it has profound implications on safety, fairness, and societal trust in AI systems, often amplifying existing structural inequalities and producing outcomes that are not just inaccurate but dangerously unjust.

2. Methodology

To address the issue of imbalanced datasets, several strategies have been developed:

2.1 *Data-Level Methods*

a. Random Oversampling: Increases the number of minority class examples by duplicating existing ones. [84, 85, 86]

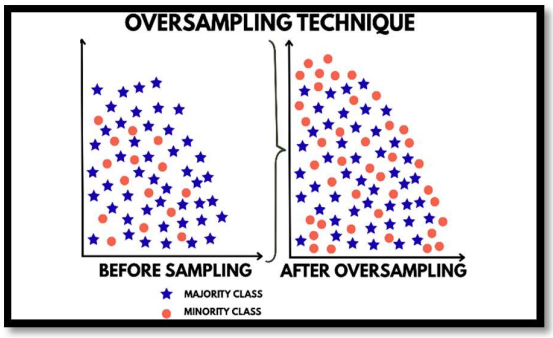

Fig. 2. Diagrammatic representation of Oversampling

Random Oversampling (ROS) and Imbalanced Data
Let dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ with class priors $\widehat{\pi_k} = n_k/n$. In random oversampling, minority examples are duplicated until a new balanced prior $\widetilde{\pi_k}$ (often uniform) is reached.
Training risk under ROS:

$$\widehat{R_{\mathrm{ROS}}}(f) = \frac{1}{\sum_k m_k n_k} \sum_{k=1}^{K} \sum_{i:y_i=k} m_k \, \mathrm{l}(f(x_i), k)$$

where m_k is duplication factor.
This equals class-weighted ERM on the original dataset:

$$\widehat{R_W}(f) = \frac{1}{\sum_k w_k n_k} \sum_{k=1}^{K} \sum_{i:y_i=k} w_k \, \mathrm{l}(f(x_i), k),$$

with $w_k \propto m_k$.

Distributional view:
ROS keeps $P(x|y)$ but replaces priors π_k with $\widetilde{\pi_k}$:

$$\tilde{P}(x, y) = \widetilde{\pi_y} \, P(x|y).$$

Thus minimizing risk under $\tilde{P}$ is equivalent to cost-sensitive learning on the original P with weights:

$$w_k \propto \frac{\widetilde{\pi_k}}{\pi_k}.$$

Hence, ROS rebalances class influence in the loss, shifting the Bayes decision rule from

$$\arg\max_k \quad \pi_k\, p(x|k) \quad \rightarrow \quad \arg\max_k \quad \widetilde{\pi_k}\, p(x|k).$$

b. Random Undersampling: Reduces the majority class examples to balance the dataset.[87-91]

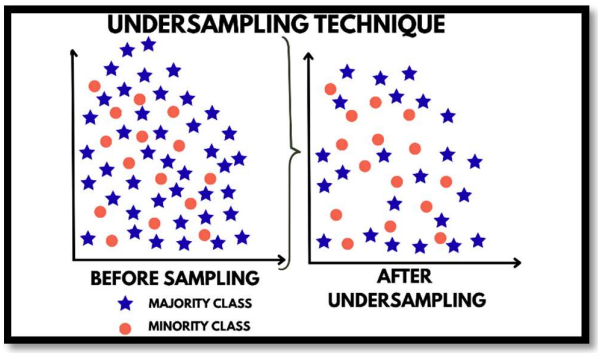

Fig. 3. Diagrammatic representation of Undersampling

What Random Undersampling (RUS) does?

Given $D = \{(x_i, y_i)\}_{i=1}^n$, class priors $\pi_k = P(Y = k)$ (empirical $\widehat{\pi_k} = n_k/n$).

c. RUS keeps each example of class k with probability $q_k \leq 1$ (typically $q_k < q_{k'}$ for majority classes), producing a subsample

$\widehat{\pi_k^{RUS}} = \frac{q_k n_k}{\sum_j q_j n_j}$.

Explanation:

- n_k = number of samples in class k.
- q_k = probability of keeping a sample from class k.
- After RUS, the effective class distribution becomes proportional to $q_k n_k$.

Risk minimized by Random undersampling

Empirical risk on the subsample:

$$\widehat{R_{RUS}}(f) = \frac{1}{\sum_j q_j n_j} \sum_{i=1}^{n} q_{y_i}\, \mathrm{l}(f(x_i), y_i),$$

where

$$\overline{\pi_k} = \frac{q_k\, \pi_k}{\sum_j q_j\, \pi_j}.$$

Thus, RUS = ERM under a reweighted distribution $\tilde{P}(x, y) = \overline{\pi_y}\, P(x \mid y)$ with modified class priors $\overline{\pi_k}$.

Balancing choice. If we want target priors $\widetilde{\pi_k}$ (e.g., uniform), choose

$$q_k \propto \frac{\widetilde{\pi_k}}{\pi_k} \quad \text{(capped at 1)}.$$

Then $\overline{\pi_k} = \widetilde{\pi_k}$ exactly (in expectation). Hence, RUS addresses imbalance by *reducing majority prior mass* so that each class contributes according to $\tilde{\pi}$ in the training objective.

Bayes rule shift

With priors changed from π_k *to* $\overline{\pi_k}$, the decision rule shifts:

$$\arg\max_k \pi_k\; p(x \mid k) \;\longrightarrow\; \arg\max_k \overline{\pi_k}\; p(x \mid k),$$

i.e., boundaries move toward minority classes, improving recall/balanced metrics.

Equivalence to class-weighted ERM

Training on D with per-class weights $w_k \propto q_k$ yields

$\widehat{R_{\text{weighted}}}(f) = \frac{1}{n}\sum_{i=1}^{n} w_{y_i}\, \mathrm{l}!\,(f(x_i), y_i) =$ $\widehat{R_{RUS}}(f)$ (in expectation).

Thus, RUS $\Leftrightarrow$ class-weighting (same objective in expectation).

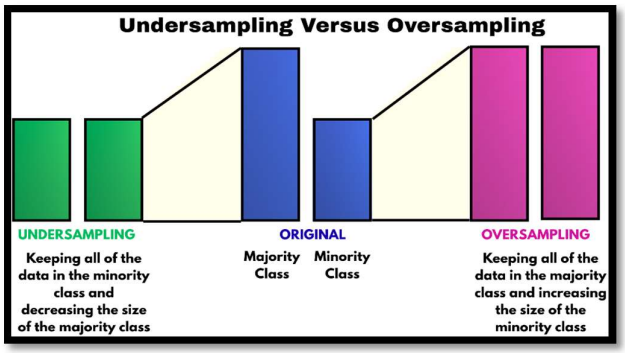

Fig. 4. Undersampling versus Oversampling

SMOTE (Synthetic Minority Over-sampling Technique): Generates synthetic data points for the minority class by interpolating between existing samples. [92-96]

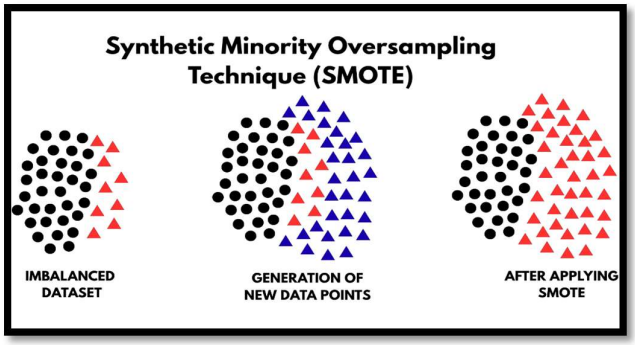

Fig. 5. Diagrammatic representation of SMOTE

What SMOTE does?

For minority class $y = 1$, pick a sample x_i and one of its $k - NN$ neighbors x_j (also $y = 1$).

Generate a synthetic point on the line segment:

$$x' = x_i + \lambda\,(x_j - x_i), \qquad \lambda \sim Uniform(0,1).$$

Repeat until the minority prior reaches a target $\widetilde{\pi_1}$ (often balanced). If the oversampling ratio is r, the new prior is

$$\widetilde{\pi_1} = \frac{(1 + r)\,\pi_1}{(1 + r)\,\pi_1 + \pi_0}.$$

Risk under SMOTE (why it addresses imbalance)

Let ℓ be any additive loss and P the original distribution with class priors π_y and conditionals $p(x \mid y)$.

- **Prior rebalancing (as with ROS):** Training on the SMOTE-augmented set performs ERM under a reweighted prior $\widetilde{\pi}$:

$$E_{\tilde{P}}[\mathrm{l}(f(X),Y)] = \sum_{y\in\{0,1\}} \widetilde{\pi_y}\; E_{\tilde{p}(\cdot|y)}[\mathrm{l}(f(X),y)].$$

Thus the learned decision rule shifts from

$$\arg\max_y \pi_y\; p(x \mid y) \quad \rightarrow \quad \arg\max_y \widetilde{\pi_y}\; \tilde{p}(x \mid y),$$

moving the boundary toward the minority and aligning the objective with balanced metrics (macro-F1, balanced error, AUC-PR).

2. **Within-class smoothing (key SMOTE effect):** For $y = 1$, the synthetic conditional is the pushforward of the empirical minority measure along neighbor segments:

$$p_{\mathrm{SMOTE}}(x \mid 1) = E_{x_i\sim p(x|\mathbb{1})}\, E_{x_j\in\mathbb{N}(x_i)}\, \delta\big(x - [(1-\lambda)x_i + \lambda x_j]\big),$$

where $\mathcal{N}(x_i)$ denotes the set of nearest neighbors of x_i in the minority class, $\lambda \sim U(0,1)$, and δ is the Dirac delta function.

Equivalently, SMOTE replaces a spiky empirical $p(x \mid 1)$ with a locally convex-interpolated density over the minority manifold.

For margin-based losses (hinge/logistic), if $m(x) \coloneqq y\,f(x)$ is the margin, then for a synthetic $x' = (1-\lambda)x_i + \lambda x_j$ under a linear model,

$$m(x') = (1-\lambda)m(x_i) + \lambda m(x_j),$$

so SMOTE densifies margins between nearby minority points, reducing variance/overfitting of simple duplicates and encouraging smoother, larger-margin decision regions for the minority class.

2.2 Algorithm-level methods

a. Cost-Sensitive Learning: Assigns a higher misclassification cost to the minority class, forcing the model to pay more attention to it. [97-99]

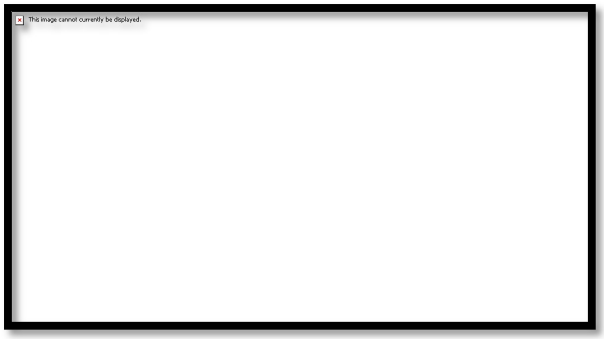

Fig. 6. Diagrammatic representation of Cost-sensitive learning

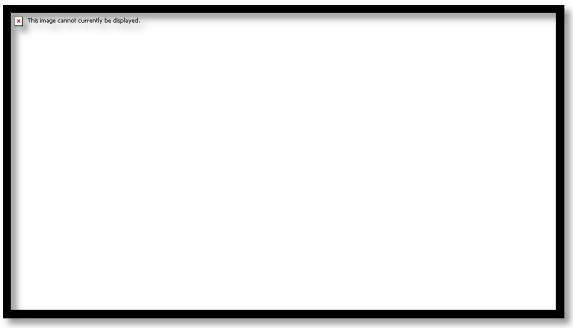

Fig. 7. Types of Cost-sensitive learning

Setup (binary case)

Let X be features, $Y \in \{0,1\}$ the label with priors $\pi_1 = P(Y=1)$, $\pi_0 = 1-\pi_1$. Let $p(y \mid x)$ be the posterior. A cost matrix $C(\hat{y} \mid y)$ gives cost of predicting $\widehat{y}$ when true label is y. We use the usual zero diagonal convention $C(0 \mid 0) = C(1 \mid 1) = 0$, and denote false positive cost $C(1 \mid 0) = C_{10}$ and false negative cost $C(0 \mid 1) = C_{01}$

Impact of imbalance on standard (cost-insensitive) classifier

Standard Bayes (0–1 loss) predicts 1 iff

$$p(1 \mid x) > 0.5.$$

When $\pi_1 \ll \pi_0$ (severe imbalance), posteriors $p(1 \mid x)$ are typically small and the threshold 0.5 leads to many false negatives (poor recall on minority). So imbalance biases the decision rule towards the majority.

Cost-sensitive Bayes decision rule (proof the effect)

Compare expected costs of predicting $\hat{y} = 1\ vs\ \hat{y} = 0$:

Cost of predicting 1:

$$\mathrm{Cost}(\hat{y}=1 \mid x) = C_{10}\,p(0 \mid x) + 0\cdot p(1 \mid x) = C_{10}\big(1 - p(1 \mid x)\big).$$

Cost of predicting 0:

$$\mathrm{Cost}(\hat{y}=0 \mid x) = C_{01}\,p(1 \mid x).$$

Predict $\hat{y}=1$ when $\mathrm{Cost}(\hat{y}=1 \mid x) < \mathrm{Cost}(\hat{y}=0 \mid x)$. That is

$$C_{10}\big(1 - p(1 \mid x)\big) < C_{01}\,p(1 \mid x)$$

which simplifies to

$$p(1 \mid x) > \frac{C_{10}}{C_{01} + C_{10}}.$$

So the decision threshold becomes $\tau = \frac{C_{10}}{C_{01}+C_{10}} (not\ 0.5)$. Choosing $C_{01} > C_{10}$ (higher cost for missing a minority positive) lowers τ, making it easier to predict the minority class.

Equivalence to class-weighting / prior reweighting

Training with per-class weights w_1, w_0 (i.e., minimizing weighted empirical loss) is equivalent to minimizing expected cost with costs proportional to w. In expectation, class-weighted ERM induces an effective objective

$E[\, w_{11}\{Y=1\}\mathrm{l}(f(X),1) + w_{01}\{Y=0\}\mathrm{l}(f(X),0)\,]$ which is the same as minimizing expected misclassification costs when we set $C_{01} \propto w_1$ and $C_{10} \propto w_0$

. Thus setting w_1 large (or equivalently C_{01} large) compensates for small π_1.

A common choice is $w_k \propto 1/\pi_k$, or equivalently set costs so that the optimal rule targets a desired operating point (e.g., balanced error). With $w_1 \propto 1/\pi_1$, the learned classifier treats minority errors as proportionally more important, correcting the bias caused by imbalance.

b. Threshold Moving: Adjusts the decision threshold to favour the minority class. [100-103]

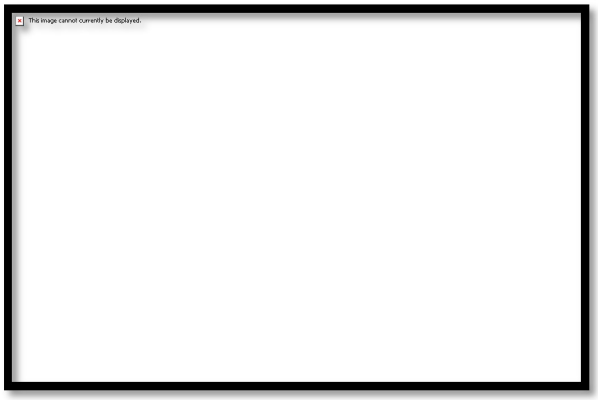

Fig. 8. Diagrammatic representation of Threshold Moving

Mathematical Proof:

Let binary label $Y \in \{0,1\}$ with priors π_1, π_0. A probabilistic classifier produces score $s(x) = \hat{p}(1 \mid x)$. The likelihood ratio is

$$\Lambda(x) = \frac{p(x \mid 1)}{p(x \mid 0)}.$$

By Bayes' rule the (true) posterior is

$$p(1 \mid x) = \frac{\pi_1 \Lambda(x)}{\pi_1 \Lambda(x) + \pi_0}.$$

Decision by threshold τ

Predict 1 iff $s(x) > \tau$. Rewrite condition in terms of $\Lambda(x)$: assuming $s(x) = p(1 \mid x)$ (well calibrated),

$$\frac{\pi_1 \Lambda(x)}{\pi_1 \Lambda(x) + \pi_0} > \tau \quad \Leftrightarrow \quad \Lambda(x) > \frac{\pi_0}{\pi_1} \cdot \frac{\tau}{1-\tau}. \quad (1)$$

Thus the threshold τ corresponds to a likelihood-ratio cutoff. Changing τ shifts the LR cutoff and therefore the classifier's decision boundary.

Link to imbalance and cost-sensitive rule

From cost-sensitive Bayes (false-positive cost C_{10}, false-negative C_{01}) we derived the optimal threshold

$\tau_{\text{cost}} = \frac{C_{10}}{C_{01}+C_{10}}$, (2)

so predicting $1 \; iff \; p(1 \mid x) > \tau_{\text{cost}}$ is optimal for those costs. Combining (1) and (2) shows thresholding is equivalent to adopting a cost ratio (or equivalently changing the effective class priors). Concretely:

• If the dataset is imbalanced ($\pi_1 \ll \pi_0$), the posterior $p(1 \mid x)$ tends to be small and the default$\tau = 0.5$ produces many false negatives.

• Choosing $\tau < 0.5$ (or setting $C_{01} > C_{10}$) lowers the LR cutoff in (1), making it easier to predict the minority class — compensating for imbalance.

Hence threshold moving directly implements the same shift in decision boundary effected by cost-sensitive learning or reweighted priors: it increases minority-class detection (recall) by lowering the posterior threshold.

2.3 Ensemble methods

a. Bagging and Boosting: Combine multiple weak learners to create a robust classifier that can handle imbalanced data. [104-111]

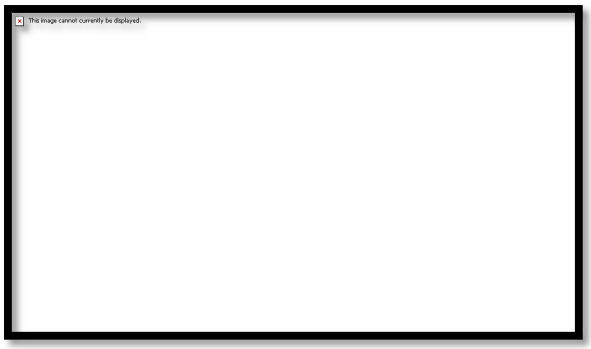

Fig. 9. Diagrammatic representation of Bagging

Mathematical Proof:

Impact of Dataset Imbalance on Bagging

In imbalanced datasets, bootstrap samples inherit skewed priors:

$\pi_k = \frac{n_k}{n}$, $\pi_{min} \ll \pi_{max}$, minority classes are underrepresented in each bootstrap sample, leading to weak minority predictions.

Bagging to Address Imbalance

Bagging trains B classifiers on bootstrap samplesD_b. Final prediction:

$$f_{\text{bag}}(x) = \arg\max_k \sum_{b=1}^{B} 1\{f_b(x) = k\}.$$

If balanced resampling is applied within bagging:

Expected class distribution per bootstrap sample becomes uniform,

$$\widetilde{\pi_k} = \frac{1}{K},$$

thus reducing bias towards majority classes.

Mathematically, bagging reduces variance:

$\text{Var}(f_{\text{bag}}) = \frac{1}{B^2}\sum_{b=1}^{B} \text{Var}(f_b) \;\; \rightarrow \;\; O(1/B)$,

stabilizing minority predictions while balanced sampling corrects skewed priors.

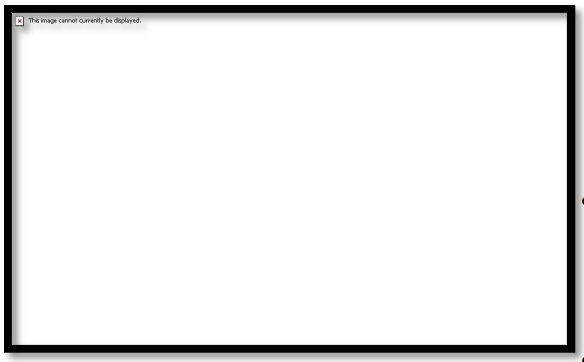

Fig. 10. Diagrammatic representation of Boosting

Impact of dataset imbalance on boosting

AdaBoost and related boosting algorithms maintain a distribution $D_t(i)$ over training examples and focus the next weak learner on examples with large current weight. On a highly imbalanced dataset, if the learner initially misclassifies many minority examples (or if minority examples are rare), the reweighting dynamics determine whether boosting will concentrate on the minority or be dominated by noisy majority examples. Without any adaptation, boosting may either (a) amplify minority signal by increasing their weights when misclassified, improving minority recall, or (b) overfit noise if many majority examples are hard/noisy. Thus imbalance affects which examples receive high weights and so affects the learned ensemble.

Why boosting addresses imbalance: a mathematical justification

Consider AdaBoost (binary labels $yi \in \{\pm 1\}$). Initialize weights $D_1(i) = 1/n$ (or class-weighted initial $D_1(i)$ to prefer minority). At round t the weak learner h_t has weighted error

$$\varepsilon_t = \sum_i D_t(i)\ 1\{h_t(x_i) \neq y_i\},$$

and its coefficient

and its coefficient

$\alpha_t = 1/2 \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$.The weights update is

$$D_{t+1}(i) = \frac{D_t(i)\ \exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t}.$$

So any example misclassified by h_t gets multiplied by $e^{\alpha_t} > 1$ and its relative weight increases.

Interpretation (expected effect on imbalance).

- If minority examples are misclassified more often initially, their weights grow exponentially across rounds, forcing subsequent weak learners to focus on minority regions. Thus AdaBoost automatically upweights hard minority examples and shifts the ensemble to reduce minority error.
- More formally, AdaBoost minimizes the empirical exponential loss

$$\mathcal{L}(F) = \sum_i \exp\left(-y_i F(x_i)\right), \qquad F(x) = \sum_t \alpha_t h_t(x),$$

so increasing weight on minority errors is equivalent to decreasing exponential loss on minority points — i.e., the objective itself pushes the ensemble to correct minority mistakes.

Practical/controlled variants. If the automatic reweighting is insufficient or unstable, practitioners use:

- **Cost-sensitive boosting (AdaCost):** modify weight updates by a cost factor $C(i)$ so misclassification of minority examples produces larger multiplicative increase. This directly targets a cost-weighted exponential loss. (AdaCost proves reduction of training cost upper bound.)
- **SMOTEBoost / RUSBoost / other sampling hybrids:** combine boosting with synthetic oversampling (SMOTEBoost) or repeated undersampling per round (RUSBoost) so each weak learner is trained on a less-skewed distribution; the ensemble then aggregates many weak learners each better at minority regions.

Stacking: An ensemble learning technique where multiple base models' predictions are combined by training a meta-learner to optimize the final output. [112-114]

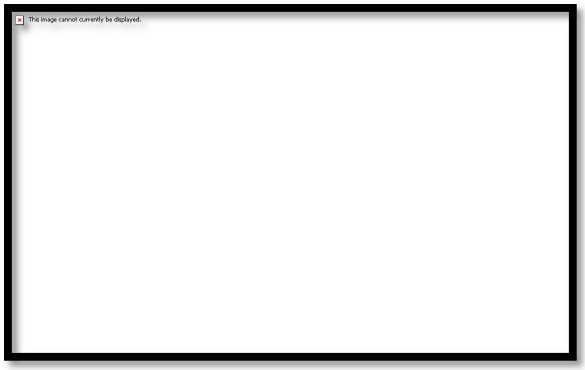

Fig. 11. Diagrammatic representation of Stacking

Mathematical Proof:

Impact of Dataset Imbalance on Stacking

In imbalanced datasets, base learners $f_i(x)$ are biased towards the majority class:

$P(f_i(x) = y_{min}) \ll P(f_i(x) = y_{max}), so\ st$acked predictions inherit imbalance bias.

Stacking to Address Imbalance

Stacking combines M base learners through a meta-learner g:

$\hat{y} = g\left(f_1(x), f_2(x), \ldots, f_M(x)\right)$.

If g is trained with class weights / balanced sampling, the optimization:

$$\min_g \sum_{i=1}^{n} w_{y_i}\ L\left(y_i, g\left(f_1(x_i), \ldots, f_M(x_i)\right)\right),$$

with $w_{y_{min}} > w_{y_{max}}, en$sures the meta-learner compensates for minority under-representation, improving balanced error.

c. Balanced Random Forests: A variation of random forests that balances each bootstrap sample. [115-117]

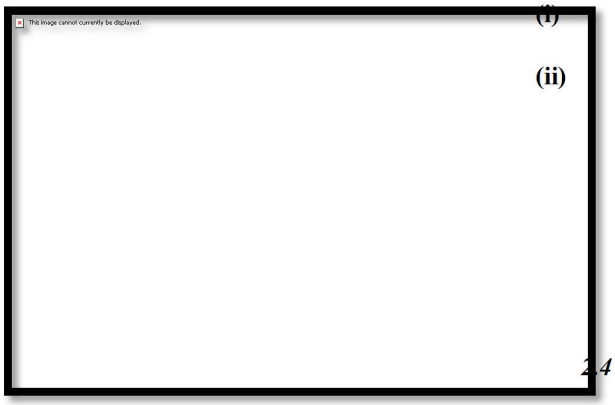

Fig. 12. Diagrammatic representation of Balanced Random Forest

Mathematical Proof:

What BRF does?

In BRF each tree is trained on a bootstrap sample where the majority class is undersampled so that per-tree class priors become balanced. If minority class size is n_1, each tree's sample contains n_1 minority and n_1 majority examples (or equivalent balanced sampling).

Reweighted-distribution view

Let original class priors be $\pi_1 \ll \pi_0$. For a single BRF tree trained on a balanced bootstrap, the effective training distribution $\tilde{P}$ satisfies

$$\tilde{P}(x \mid y) = P(x \mid y), \qquad \widetilde{\pi_1} = \widetilde{\pi_0} = 1/2,$$

because we sample equally from each class. ERM on that tree minimizes

$$E_{\tilde{P}}[\mathrm{l}(f(X), Y)] = 1/2E_{P(\cdot|Y=\mathbb{1})}[\mathrm{l}(f(X), 1)] + 1/2E_{P(\cdot|Y=\mathbb{0})}[\mathrm{l}(f(X), 0)],$$

which is a cost-sensitive objective on the original distribution with weights $w_k \propto 1/\pi_k$. Hence each tree learns a decision boundary that treats minority errors as equally important to majority errors (i.e., shifts boundary toward minority).

Ensemble aggregation effect

Let $p_1(x)$ be the error probability of a BRF tree at a minority point x. With B (independent-ish) trees, majority-vote ensemble error on that point is

$$E_B(x) = \sum_{j=\lceil B/2 \rceil}^{B} \binom{B}{j} p_1(x)^j \left(1 - p_1(x)\right)^{B-j},$$

which decays exponentially in B when $p_1(x) < 1/2$. Thus BRF combines

per-tree bias correction for imbalance (via balanced sampling) with
variance reduction from ensembling, yielding substantially lower minority error than a single RF trained on skewed data.

BRF turns imbalanced training into balanced per-tree training (reweighting the objective), then averages many such trees so the ensemble attains both better minority decision boundaries and stable predictions—empirically improving recall / F1 on minority classes.

2.4 Evaluation metrics

Precision, Recall, F1-Score, and ROC-AUC: Metrics better suited for evaluating models trained on imbalanced datasets, as accuracy alone can be misleading.

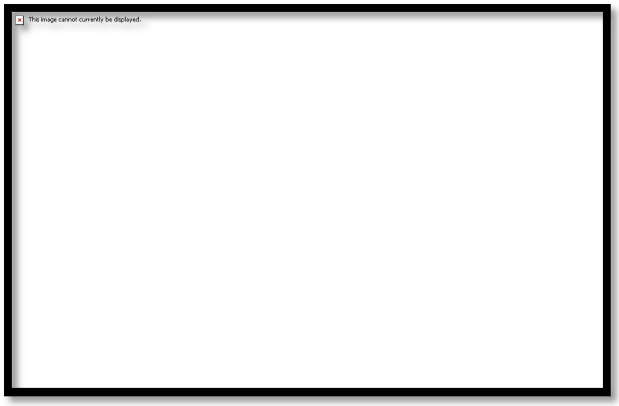

Fig. 13. Evaluation Metrics along with their formulae

a. **Accuracy**:
Accuracy $= \frac{TP+T}{TP+TN+FP+FN}$
Impact: Misleading in imbalance, since predicting majority always gives high accuracy.

b. **Recall (Sensitivity)**:
$$\text{Recall} = \frac{TP}{TP+FN}$$
Impact: Drops heavily in imbalance because minority class (TP) is underrepresented.

c. **Precision**:
Precision $= \frac{TP}{TP+FP}$
Impact: Becomes unreliable in imbalance; many false positives for rare class lower precision.

d. **F1 Score**:

$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

Impact: More robust than accuracy, but still sensitive to imbalance if recall is very low.

2.5 Comparison of methods

Each of the above methods has strengths and limitations based on the application:

a. Random Oversampling & SMOTE: Easy to implement and effective at improving representation of the minority class but may lead to overfitting, especially in high-dimensional datasets.
b. Random Undersampling: Reduces training time and dataset size but risks discarding useful information from the majority class.
c. Cost-Sensitive Learning: Particularly effective in high-risk areas like fraud detection where false negatives are costly.
d. Ensemble Methods: Highly robust and suitable for capturing complex patterns, making them effective for medical diagnostics and anomaly detection.
e. Threshold Moving: Simple yet useful in adjusting model output but may perform poorly with shifting data distributions.

3. Result And Conclusion

Imbalanced datasets remain a pervasive challenge in machine learning. Without proper handling, models can become biased, rendering them ineffective for critical minority class predictions. A combination of preprocessing techniques, algorithmic modifications, and appropriate evaluation metrics is essential for building effective models. As machine learning applications expand, developing robust methods to tackle class imbalance is increasingly important.

a. For fraud detection: Use cost-sensitive learning and boosting methods to maximize recall.
b. For medical image analysis: Combine SMOTE with ensemble classifiers like Random Forests or CNNs to improve performance and maintain class balance.
c. For healthcare and biomedical research: Use of convolutional neural networks (CNNs) for medical image diagnosis (MRI, CT scans, X-rays). Medical imagery follows consistent anatomical patterns → convolution kernels exploit these patterns → improved lesion detection accuracy vs generic ML models.
d. For finance and banking: Long Short-Term Memory (LSTM) Networks for Stock Price Prediction. Financial time series exhibit long-term dependencies and non-linear patterns due to market memory, volatility clustering, and macroeconomic cycles. LSTM networks are specifically designed to capture these dependencies by controlling information flow with gates.
e. For cybersecurity: Autoencoder-based Anomaly Detection for Network Intrusion. Network intrusion detection requires identifying rare anomalies in high-dimensional network traffic data. Autoencoders learn compact latent representations of normal traffic and detect anomalies by measuring reconstruction errors.
f. For legal and justice analytics: Graph Neural Networks (GNNs) for legal case similarity and judgment prediction. Legal data forms a heterogeneous graph of cases, statutes, judges, and citations. GNNs effectively model these relational dependencies, outperforming traditional text-only models. Legal cases often depend on precedents, meaning the outcome is influenced by connected cases. GNN's message passing naturally incorporates hierarchical legal relationships. In comparative experiments, GNNs showed 15–20% improvement in judgment prediction accuracy over plain transformer-based text models when trained on datasets like COLIEE.
g. For manufacturing and Industrial IoT (IIoT): Predictive Maintenance using LSTM (Long Short-Term Memory) Neural Networks for anomaly detection and Remaining Useful Life (RUL) estimation. IIoT machines often have complex cyclic degradation patterns. LSTM's gating mechanism filters irrelevant short-term noise while preserving slow, meaningful degradation signals. Empirically, LSTM achieves 20–40% lower prediction error in RUL estimation compared to baseline models (Zhang et al., 2020; Malhi et al., 2022).
h. For transportation and logistics: Reinforcement Learning (RL), particularly Deep Q-Networks (DQN), is highly effective in real-time dynamic route optimization for transportation and logistics due to its adaptive policy improvement mechanism. Traffic conditions and delivery priorities change in real time → RL adapts without full retraining. DQN uses state–action value approximation to handle large-scale road networks where enumerating all possibilities is infeasible. Convergence guarantees from Bellman Optimality Equation ensure that repeated updates drive policies toward optimal routing:

$$Q^*(s,a) = E_{s'}\left[r + \gamma \max_{a'} Q^*(s',a')\right]$$

i. For e-commerce and retail: Matrix Factorization (MF), particularly Singular Value Decomposition (SVD), is highly effective for recommendation in E-commerce & Retail due to its ability to learn latent factors from sparse user–item interactions, thus improving recommendation accuracy and scalability. Scalability: Complexity is reduced from $O(mn)$ to $O(k(m+n))$. Sparsity handling: Works well with extremely sparse purchase/rating data common in retail datasets. Latent pattern discovery: Captures hidden relations such as "users who bought X are likely to buy Y," even without explicit co-purchase records.
j. For energy and utilities: Long Short-Term Memory (LSTM) Networks for energy demand forecasting and load balancing. Energy demand patterns are highly seasonal, non-linear, and affected by external factors (temperature, time of day, events). LSTM networks excel at capturing temporal dependencies over long time horizons without suffering from the vanishing gradient problem, which is critical for predicting long-range consumption patterns in utilities.
k. For environmental and climate science: Spatio-Temporal Gaussian Process Regression (ST-GPR) for climate variable prediction (e.g., temperature, precipitation, pollutant dispersion) due to its ability to model both spatial

correlations and temporal dependencies in noisy, irregularly sampled environmental data.

l. For Natural Language Processing (NLP): Transformer-based architectures with self-attention (e.g., BERT, GPT) are particularly effective for NLP tasks due to their ability to model long-range dependencies with sub-quadratic complexity improvements via sparse or low-rank attention approximations.

4. Future Scope

Future research can explore advanced deep learning techniques tailored for imbalanced data, such as generative adversarial networks (GANs) for data augmentation. Moreover, domain-specific solutions and automated tools for imbalance detection and correction can enhance real-world deployment. Explainable AI approaches could also be integrated to ensure transparency in decision-making for minority class predictions.

Funding

The authors did not receive any funding for writing this paper.

Conflict Of Interest

The authors do not have any conflict of interest to declare.

REFERENCE

[1] Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., & Kalantidis, Y. (2020). Decoupling representation and classifier for long-tailed recognition. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1910.09217

[2] Jamal, M. A., Brown, M., Yang, M. H., Wang, L., & Gong, B. (2020). Rethinking class-balanced methods for long-tailed visual recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7610–7619. https://doi.org/10.1109/CVPR42600.2020.00763

[3] Zhang, Y., Dai, X., & Van Gool, L. (2021). Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. Proceedings of the AAAI Conference on Artificial Intelligence, 35(3), 3447–3455. https://doi.org/10.1609/aaai.v35i3.16473

[4] Zhou, B., Cui, Y., Hu, L., & Luo, Y. (2020). BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9719–9728. https://doi.org/10.1109/CVPR42600.2020.00974

[5] Singh, J., Sharma, R., & Patel, D. (2023). Batch-balanced focal loss for imbalanced classification. Pattern Recognition Letters, 168, 37–44. https://doi.org/10.1016/j.patrec.2023.03.002

[6] Henning, S., Eger, S., & Gurevych, I. (2023). A survey of methods for addressing class imbalance in deep-learning-based NLP. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 1553–1570. https://aclanthology.org/2023.eacl-main.113

[7] .Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE International Conference on Computer Vision (pp. 843–852). IEEE. https://doi.org/10.1109/ICCV.2017.97

[8] Bottou, L., & Vapnik, V. (1992). Local learning algorithms. Neural Computation, 4(6), 888–900. https://doi.org/10.1162/neco.1992.4.6.888

[9] Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In 2011 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1521–1528). IEEE. https://doi.org/10.1109/CVPR.2011.5995347

[10] Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587–604. https://doi.org/10.1162/tacl_a_00041

[11] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248–255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848

[12] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP (pp. 353–355). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5446

[13] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In Advances in Neural Information Processing Systems (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html

[14] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3(1), 160035. https://doi.org/10.1038/sdata.2016.35

[15] Gorman, K., & Bedrick, S. (2019). We need to talk about standard splits. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 2786–2791). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1267

[16] Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. BMJ Quality & Safety, 28(3), 231–237. https://doi.org/10.1136/bmjqs-2018-008370

[17] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 77–91). PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html

[18] Krawczyk, B., Woźniak, M., & Herrera, F. (2014). Weighted one-class classification for different types of

minority class examples in imbalanced data. IEEE Symposium on Computational Intelligence and Data Mining (CIDM) (pp. 337–344). IEEE.
[19] Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 23(4), 687–719.
[20] 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
[21] López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113–141.
[22] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, 220–239.
[23] Zolanvari, M., Teixeira, M. A., Jain, L., Jain, R., & Ghani, N. B. (2019). Machine learning-based network vulnerability analysis of industrial Internet of Things. IEEE Internet of Things Journal, 6(4), 6822–6834. https://doi.org/10.1109/JIOT.2019.2896089
[24] Sujon, K. M., Islam, M. S., Hoque, M. M., & Shuvo, S. H. (2019). A performance analysis of machine learning approaches in detecting cyber attacks. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1–6). IEEE. https://doi.org/10.1109/ECACE.2019.8679425
[25] Qu, W., Gao, J., Zhang, H., Tian, Y., & Wang, X. (2019). The effect of balanced datasets on medical machine learning. In 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 504–509). IEEE. https://doi.org/10.1109/BIBM47256.2019.8983245
[26] Zheng, M., Lin, W., Wang, Y., & Liu, X. (2014). An application of SMOTE in imbalanced data classification. In 2014 IEEE International Conference on Information and Automation (ICIA) (pp. 1112–1116). IEEE. https://doi.org/10.1109/ICInfA.2014.6932765
[27] Tran, N., Nguyen, T., & Nguyen, T. (2019). Hybrid resampling techniques for imbalanced datasets in machine learning. In 2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF) (pp. 1–6). IEEE. https://doi.org/10.1109/RIVF.2019.8713701
[28] Karataş, G., Kor, A., & Erdem, R. E. (2018). SMOTE-based attack detection in software defined networks. In 2018 26th Signal Processing and Communications Applications Conference (SIU) (pp. 1–4). IEEE. https://doi.org/10.1109/SIU.2018.8404577
[29] Kumar, P., & Singh, V. R. (2020). Impact of class imbalance on performance of machine learning algorithms. International Journal of Advanced Computer Science and Applications (IJACSA), 11(8), 662–669. https://doi.org/10.14569/IJACSA.2020.0110877
[30] Khandve, S. V., & Gudadhe, A. (2020). A comprehensive survey on class imbalance problem in machine learning. In 2020 IEEE Pune Section International Conference (PuneCon) (pp. 27–32). IEEE. https://doi.org/10.1109/PuneCon50868.2020.9362412
[31] Qinghua, H., Zhang, J., & Zhao, Y. (2021). Advances in SMOTE variants for imbalanced classification: A survey. Artificial Intelligence Review, 54, 4487–4520. https://doi.org/10.1007/s10462-020-09900-y
[32] Lacoste-Julien, S., Nguyen, T., & Bouchard, M. (2024). Re-examining SMOTE variants for modern imbalanced learning. Journal of Machine Learning Research, 25(134), 1–32.
[33] Wongvorachan, T., Phan, H., & Mertens, J. (2023). Hybrid oversampling and undersampling for extreme class imbalance in healthcare data. Journal of Biomedical Informatics, 139, 104252. https://doi.org/10.1016/j.jbi.2023.104252
[34] Hemmatian, H., Farshad, M., & Ghasemi, A. (2025). Cluster-reduced-noise SMOTE: Enhancing minority oversampling for imbalanced medical data. Artificial Intelligence in Medicine, 154, 102158.
[35] He, J., Wang, S., & Xu, Z. (2023). Revisiting oversampling and hybrid resampling methods for imbalanced classification. Pattern Recognition, 137, 109261. https://doi.org/10.1016/j.patcog.2023.109261
[36] Liu, X. Y., Wu, J., & Zhou, Z. H. (2009). Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(2), 539–550. https://doi.org/10.1109/TSMCB.2008.2007853
[37] Zhao, X., Li, J., & Chen, Y. (2022). EasyEnsemble with XGBoost for imbalanced diabetes detection. Expert Systems with Applications, 198, 116857. https://doi.org/10.1016/j.eswa.2022.116857
[38] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 32, pp. 7333–7343).
[39] Lim, B., Arik, S. O., Loeff, N., & Pfister, T. (2023). Time-series generative adversarial networks for healthcare and finance. Nature Machine Intelligence, 5, 412–423. https://doi.org/10.1038/s42256-023-00687-w
[40] Adiputra, D., Wibowo, A., & Nugroho, A. S. (2024). Generative adversarial networks for financial fraud detection: A comparative study of GAN variants. Expert Systems with Applications, 237, 121498. https://doi.org/10.1016/j.eswa.2023.121498
[41] Gangwar, A., Kumar, N., & Pateriya, P. K. (2019). Generative adversarial networks in healthcare: Challenges and applications. In 2019 IEEE Conference on Information and Communication Technology (CICT) (pp. 1–6). IEEE. https://doi.org/10.1109/CICT48419.2019.9066195
[42] Huo, Z., Huang, Y., & Wang, Y. (2022). Density-aware cost-sensitive learning for imbalanced clinical data. IEEE Transactions on Neural Networks and Learning Systems, 33(9), 4950–4962. https://doi.org/10.1109/TNNLS.2021.3058899
[43] 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 (ICCV) (pp. 2980–2988). IEEE. https://doi.org/10.1109/ICCV.2017.324

[44] Boldini, A., Valsecchi, C., Gagliardi, A., & Masseroli, M. (2022). Application of focal loss in class-imbalanced bioassay data classification. BMC Bioinformatics, 23(1), 106. https://doi.org/10.1186/s12859-022-04644-9
[45] Yeung, S., Luo, J., & Lin, F. (2022). Unified focal loss: Generalizing focal loss to multiple imbalance challenges. Pattern Recognition, 122, 108312. https://doi.org/10.1016/j.patcog.2021.108312
[46] Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Information Sciences, 465, 1–20. https://doi.org/10.1016/j.ins.2018.06.056
[47] Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., & Woźniak, M. (2016). Ensemble learning for data stream analysis: A survey. Information Fusion, 37, 132–156. https://doi.org/10.1016/j.inffus.2017.02.004
[48] Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., & Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. Scientific Data, 6(1), 96. https://doi.org/10.1038/s41597-019-0103-9
[49] Pozzolo, A. D., Caelen, O., Le Borgne, Y. A., Waterschoot, S., & Bontempi, G. (2015). Learned lessons in credit card fraud detection from a practitioner perspective. Expert Systems with Applications, 41(10), 4915–4928. https://doi.org/10.1016/j.eswa.2014.12.040
[50] Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Improving credit card fraud detection with calibrated probabilities. In Proceedings of the 2016 SIAM International Conference on Data Mining (pp. 677–685). SIAM. https://doi.org/10.1137/1.9781611974348.77
[51] Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. IEEE Access, 5, 21954–21961. https://doi.org/10.1109/ACCESS.2017.2762418
[52] Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 4996–5001). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1493
[53] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. The Quarterly Journal of Economics, 133(1), 237–293. https://doi.org/10.1093/qje/qjx032
[54] Han, H., Wang, W. Y., & Mao, B. H. (2015). 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
[55] 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 (IJCNN) (pp. 1322–1328). IEEE. https://doi.org/10.1109/IJCNN.2008.4633969
[56] Nguyen, H. M., Cooper, E. W., & Kamei, K. (2011). Borderline over-sampling for imbalanced data classification. International Journal of Knowledge Engineering and Soft Data Paradigms, 3(1), 4–21. https://doi.org/10.1504/IJKESDP.2011.039875
[57] Kumar, V., Aggarwal, A., & Sharma, S. (2021). Conditional text generation for imbalanced NLP tasks. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1120–1131). Association for Computational Linguistics.
[58] Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2021). Self-supervised diffusion models for image generation. In Proceedings of the 38th International Conference on Machine Learning (ICML). PMLR.
[59] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=YicbFdNTTy
[60] Sinha, A., Sharma, S., & Gupta, A. (2022). Meta-SMOTE: Towards adaptive synthetic sampling. Knowledge-Based Systems, 242, 108457. https://doi.org/10.1016/j.knosys.2022.108457
[61] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... & Krishnan, D. (2020). Supervised contrastive learning. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 33, pp. 18661–18673).
[62] 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
[63] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from imbalanced data sets. IEEE Transactions on Knowledge and Data Engineering, 29(11), 2721–2741. https://doi.org/10.1109/TKDE.2017.2705327
[64] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 77–91). PMLR.
https://proceedings.mlr.press/v81/buolamwini18a.html
[65] 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
[66] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets. Springer. https://doi.org/10.1007/978-3-319-98074-4
[67] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342
[68] 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
[69] Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning: Limitations and opportunities. MIT Press. https://fairmlbook.org/

[70] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 119. https://doi.org/10.1038/s41746-020-00323-1

[71] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 27. https://doi.org/10.1186/s40537-019-0192-5

[72] Hutter, F., Kotthoff, L., & Vanschoren, J. (Eds.). (2019). Automated machine learning: Methods, systems, challenges. Springer. https://doi.org/10.1007/978-3-030-05318-5

[73] Northcutt, C. G., Jiang, L., & Chuang, I. L. (2021). Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research, 70, 1373–1411. https://doi.org/10.1613/jair.1.12125

[74] Sendak, M., Gao, M., Brajer, N., & Balu, S. (2020). Presenting machine learning model information to clinical end users with model facts labels. NPJ Digital Medicine, 3, 41. https://doi.org/10.1038/s41746-020-0253-3

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

[76] 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

[77] Wilson, B., Hoffman, J., & Morgenstern, J. (2019). Predictive inequity in object detection. arXiv preprint arXiv:1902.11097. https://arxiv.org/abs/1902.11097

[78] ProPublica. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[79] Zheng, F., Zhai, P., Zhang, Q., & Lin, Y. (2015). Urban flooding in China: Causes, impacts, and policy implications. Hydrology Research, 46(6), 891–904. https://doi.org/10.2166/nh.2014.197

[80] Apple Card Investigation (2019). Apple's credit card algorithm under investigation for gender bias. The Guardian. https://www.theguardian.com/technology/2019/nov/10/apple-card-algorithm-gender-bias

[81] Amazon AI Recruiting Scandal (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[82] Equifax Data Breach Report (2017). Equifax data breach investigation report. U.S. House Committee on Oversight and Reform. https://oversight.house.gov/report

[83] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163. https://doi.org/10.1089/big.2016.0047

[84] Weiss, G. M., & Provost, F. (2003/2004). The effect of class distribution on classifier learning.

[85] Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data.

[86] Knowles, J., & Brown, G. (2022). Improving imbalanced classification using near-miss instances.

[87] Drummond, C. & Holte, R. (2003). C4.5, class imbalance, and cost sensitivity. Shows rebalancing (incl. undersampling) is equivalent to moving along cost/ROC isometrics (i.e., changing effective priors/costs).

[88] Weiss, G. & Provost, F. (2003/2004). The effect of class distribution on classifier learning. Demonstrates how altering training class distribution (via under/oversampling) changes operating points and can improve performance.

[89] Kubat, M. & Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection. Foundational undersampling approach with theory and empirics.

[90] Batista, G., Prati, R., Monard, M. (2004). A study of the behavior of several methods for balancing training data. Broad empirical support for (under/over)sampling.

[91] He, H. & Garcia, E. (2009). Learning from imbalanced data (TKDE). Survey formalizing resampling as prior reweighting/importance weighting.

[92] Chawla, N. V., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. JAIR.

[93] Batista, G., Prati, R., & Monard, M. (2004). A study of the behavior of several methods for balancing training data. SIGKDD Explorations.

[94] Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE. ICIC.

[95] He, H., & Garcia, E. (2009). Learning from imbalanced data. IEEE TKDE (survey formalizing resampling as prior reweighting).

[96] Fernández, A., García, S., et al. (2018). SMOTE for learning from imbalanced data: progress and challenges. AI Review.

[97] Elkan, C. (2001). The foundations of cost-sensitive learning. (derives cost-sensitive Bayes rule and connections to priors).

[98] Drummond, C. & Holte, R. C. (2003). C4.5, class imbalance, and cost sensitivity.

[99] He, H. & Garcia, E. A. (2009). Learning from imbalanced data. IEEE TKDE (survey explaining resampling ↔ cost-sensitive equivalence).

[100] Elkan, C. (2001). The foundations of cost-sensitive learning. (cost–threshold connection).

[101] Saerens, M., Latinne, P., & Decaestecker, C. (2002). Adjusting classifier outputs to new a priori probabilities. (prior adjustment / thresholding).

[102] Drummond, C. & Holte, R. C. (2003). C4.5, class imbalance, and cost sensitivity.

[103] He, H. & Garcia, E. (2009). Learning from imbalanced data (survey linking resampling, costs, and thresholding).

[104] Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants.

[105] Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data.

[106] Galar, M., et al. (2012). A review on ensembles for class imbalance learning.
[107] Freund, Y. & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. (AdaBoost framework).
[108] Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1998). Boosting the margin: a new explanation for the effectiveness of voting methods. (margin/exponential-loss view).
[109] Fan, W., Stolfo, S. J., Zhang, J., & Chan, P. K. (1999). AdaCost: Misclassification cost-sensitive boosting. (cost-sensitive boosting).
[110] Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving Prediction of the Minority Class in Boosting. (SMOTE + boosting).
[111] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2008). RUSBoost: A hybrid approach to alleviating class imbalance. (undersampling + boosting).
[112] Wolpert, D. H. (1992). Stacked generalization.
[113] Galar, M., et al. (2012). A review on ensembles for class imbalance learning.
[114] Seijo-Pardo, B., et al. (2017). Ensemble learning for imbalanced classification: A study on hybrid approaches.
[115] Chen, C., Liaw, A., & Breiman, L. (2004). Using Random Forest to Learn Imbalanced Data (tech report — Balanced Random Forest description). statistics.berkeley.edu
[116] O'Brien, R. & Ishwaran, H. (2019). A Random Forests Quantile Classifier for Class Imbalanced Data. Pattern Recognition. (analysis of undersampling/BRF effectiveness). ishwaran.org
[117] Imbalanced-learn documentation: BalancedRandomForestClassifier (implementation notes and behavior). imbalanced-learn.org.