

Review Article

A Comprehensive Review of Data Balancing Techniques in Machine Learning: Methods, Challenges, and Future Directions

Ahmed Adil Nafea^{1,*}, ¹ Department of Artificial Intelligence, College of Computer Science and IT, University of Anbar, Ramadi, Iraq

ARTICLE INFO

Article History

Received 16 Jul 2026
Revised 31 Aug 2026
Accepted 25 Sep 2026
Published 31 Oct 2026

Keywords

Class imbalance,
Data balancing,
SMOTE,
Resampling techniques,
Cost-sensitive learning.

ABSTRACT

The problem of class imbalance is an important problem in machine learning (ML), leading to bias in the trained models and poor accuracy, especially for the minority classes. This review performs a thorough characterization of Under-sampling, over-sampling, algorithm-level, and their hybrid techniques. It evaluates popular techniques such as Random Under-sampling (RUS), Synthetic Minority Over-sampling Technique (SMOTE), and Cost-Sensitive Learning (CSL), and explores advanced hybrid strategies like SMOTE-ENN and SMOTE+Tomek Links. The effectiveness of these approaches is compared across a wide range of datasets and ML techniques, highlighting their advantages, trade-offs, and practical limitations, including risks of overfitting, information loss, and computational complexity. This paper also discusses challenges in multi-class imbalance scenarios and high-dimensional data. Finally, it discusses future research directions including balancing with Deep Learning (DL), AutoML, and fairness-aware modeling. It provides a reliable overview of the state of the art by presenting a summary of literature and details gaps, serving as a blueprint for researchers and potential users who are interested in accurate strategies to solve the class imbalance problem in various real-world applications.

1. INTRODUCTION

The increasing volume and complexity of data has led to the necessity of extracting knowledge from large datasets. Data Mining (DM) addresses this deficiency, drawing on foundations from statistics, probability, Artificial Intelligence (AI), and Machine Learning (ML), and other related fields [1]. ML is a significant branch of AI that draws on statistics, probability, and optimization. It allows systems to learn from data and make informed decisions or predictions [2]. ML techniques are classified into three primary paradigms, namely supervised, unsupervised, and reinforcement learning. Widely used supervised classifiers include Decision Tree (DT), Classification And Regression Tree (CART), Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Multilayer Perceptron (MLP), Artificial Neural Network (ANN), Adaptive Boosting (AdaBoost), Gradient Boosting, eXtreme Gradient Boosting (XGBoost), Light Gradient-Boosting Machine (LightGBM), Categorical Boosting (CatBoost), and Extra Trees (ET), while deep architectures such as convolutional networks (e.g., ResNet50) and graph convolutional networks (e.g., ESA-GCN) are applied to image and graph data [3]. Clustering and classification are two basics employed to extract knowledge from data. The capability of these techniques to identify hidden patterns and develop models from data is what gives them their efficiency strength [4]. The classification techniques aim to predict categorical classes, such as fault detection, sentiment analysis, image classification, plant species recognition, and disease diagnosis, based on a trained model [5],[6]. In fact, classification is the process of accurately predicting the class labels, while clustering is an unsupervised technique for knowledge extraction that groups data by their similarities, maximizing within-cluster similarity and between-cluster separation [7]. Learning from imbalanced data is a significant subarea of ML research. Imbalanced data refers to a significant difference in the number of objects between the various classes [8]. In the classification case, class imbalance occurs when the Majority Class (Maj-C) significantly outnumbers the Minority Class (Min-C). The misclassification of Min-C instances is more expensive than the misclassification of Maj-C instances [9], and this disparity results in biased model performance. As a result, various data balancing methods have been employed as crucial preprocessing steps in order to address this issue [10].

The primary motivation of this review paper is to address the persistent challenge of class imbalance in ML by analyzing and comparing existing data balancing methods with the goal of enhancing model performance, fairness, and applicability in real-world scenarios. The paper's primary objectives are:

*Corresponding author. Email: ahmed.a.n@uoanbar.edu.iq

- Review and classify the main data balancing techniques in ML, including data-level, algorithm-level, and hybrid methods.
- Analyze the effectiveness of these techniques across various datasets and ML algorithms.
- Identify and discuss limitations, including overfitting, information loss, class overlap, and computational complexity.
- Compare the results of recent studies to evaluate the practical performance and trade-offs between different approaches.
- Guide for selecting appropriate balancing techniques based on data characteristics and application requirements.
- Highlight future research directions, including integration with DL, AutoML, meta-learning, and fairness-aware approaches.

2. DATA BALANCING METHODS

Class imbalance in the real world has a significant negative impact on the generalizability of ML models. Approaches intended to address this problem (as shown in Figure 1) are primarily classified into algorithm-level methods, data-level methods, and hybrid methods.

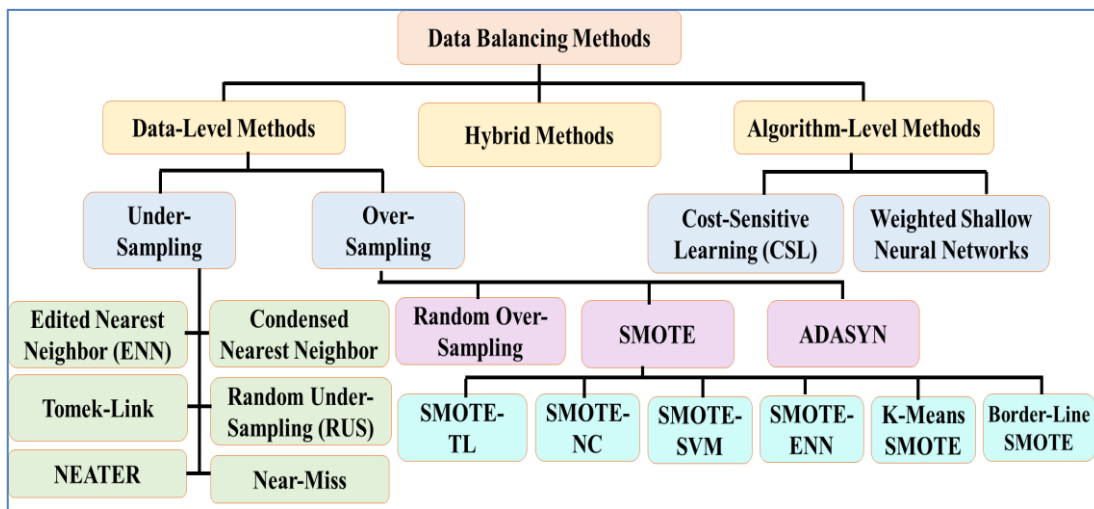


Fig. 1. Data balancing Methods

2.1 Data-Level Methods

Data-level methods address the class imbalance by modifying the training dataset through techniques such as under-sampling, over-sampling or hybrid [11]. In other words, instances are generated for the Min-C or removed from the Maj-C until the class distribution is rebalanced. These techniques target the challenge of class imbalance by rebalancing the class distribution within the training dataset. Figure 2 illustrates under-sampling and over-sampling. The main objective is to build a more balanced dataset that can effectively train a classification model [12].

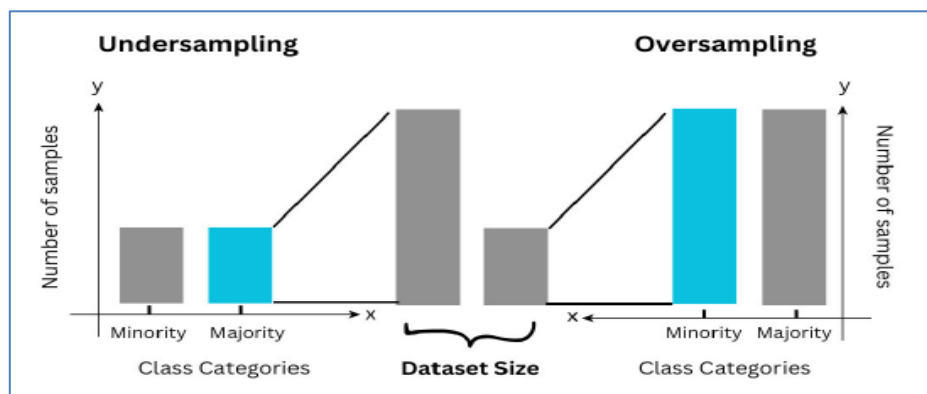


Fig. 2. Under/ Over-sampling techniques [12].

2.1.1 Under-Sampling Methods

Under-sampling balances the class distribution by removing instances from the Maj-C. To achieve a balanced dataset, redundant, noisy, or borderline Maj-C instances are discarded from the original dataset [13]. Instances are selected either at random or through informed (heuristic) criteria, such as removing instances near the class boundary. Informed selection reduces the region of the feature space dominated by the Maj-C and thereby helps the classifier recognize Min-C instances; nevertheless, any removal of data may discard useful information. The under-sampling method has several techniques, including Near-Miss, Edited Nearest Neighbors (ENN), Tomek-Link (TL), Filtering of oversampled data using non-cooperative game theory (NEATER), RUS technique, and Condensed Nearest Neighbor (CNNR) [14]. Table (1) shows the main advantages and disadvantages of under-sampling methods [15].

TABLE I. THE MAIN ADVANTAGES AND DISADVANTAGES OF UNDER-SAMPLING METHODS

Advantages	Disadvantages
Shrinks the training set, which lowers memory requirements and shortens training and tuning time, particularly for large datasets.	Discarded Maj-C instances may carry information the model needs, and random removal offers no control over what is lost.
When informed methods such as ENN or Tomek-Link are used, redundant and noisy Maj-C instances are removed, which sharpens the class boundary and reduces the classifier's preference for the Maj-C.	When the dataset is small or the imbalance ratio is extreme, the retained Maj-C subset may no longer represent its true distribution, leading to underfitting and unstable results across runs.

2.1.1.1 Near-Miss Method

The Near-Miss method is a heuristic-based approach that reduces the number of negative category instances by identifying a subset of them that are comparable to the positive category instances [16]. The primary concept behind this technique is that by maintaining only the most relevant negative category instances, the model's performance can be improved by mitigating bias towards the negative category. There are three versions of the Near-Miss technique, NearMiss-1, NearMiss-2, and NearMiss-3, each using a different criterion for selecting the negative category instances. NearMiss-1 retains the negative instances with the smallest average distance to their three nearest positive instances. NearMiss-2 retains the negative instances with the smallest average distance to the three farthest positive instances. NearMiss-3 is a two-step procedure: for each positive instance, a fixed number of its nearest negative neighbors are first shortlisted, and from these, the negative instances with the largest average distance to their nearest positive neighbors are retained [17]. Near-Miss computed the distances of K Nearest Neighbors (KNNs) of A^{negative} in A^{positive} . Then, the distances of each k-neighbor set are averaged. The instances with the smallest average distances are retained, and the rest are discarded. NearMiss may not be appropriate for datasets where categories overlap or have complex boundaries. Removing Maj-C instances exclusively based on distance to Min-C instances may not sufficiently address the challenges posed by overlapping categories [18].

2.1.1.2 Edited Nearest Neighbors (ENN)

ENN is a data-cleaning technique used to eliminate samples that include noise from a dataset [19]. It operates by reviewing each instance and corresponding it to its KNNs [20]. The Edited Nearest Neighbors (ENN) algorithm identifies instances that are noisy by comparing the class label of each sample to the k-nearest neighbor classes (using distance based on Euclidean). Instances that have labels that differ from the majority of their counterparts are removed from the dataset. In under-sampling scenarios, only noisy negative (majority) samples are removed. While ENN is effective at reducing noise in the dataset, it may also remove important samples in scenarios where the classes are overlap. The selection of the kth hyperparameter has a significant impact on this trade-off, as a result, it's necessary to assess the effect of ENN on the performance of classifiers [21].

2.1.1.3 Tomek-Link (TL)

Tomek-Link (TL) is a noise removal technique used in ML to clean noisy datasets. If two instances a and b belong to different classes and are each other's nearest neighbors, then they form a "Tomek-link" [22]. TL-based under-sampling removes such pairs, or only their Maj-C member, to reduce class overlap. This under-sampling methodology increases the distinction between classes and facilitates the learning of classifiers by removing ambiguous data points. While TL can enhance the effectiveness of classifiers by clarifying the boundary, it isn't effective all of the time. Deciding which instances to remove requires care to ensure that the dataset remains representative. Removing only the Maj-C member of each pair preserves scarce Min-C information, whereas removing both members cleans the boundary more aggressively but may discard informative Min-C instances and reduce discriminative power [23],[24].

2.1.1.4 Filtering of Oversampled Data Using Non-Cooperative Game Theory (NEATER)

NEATER is a data-cleaning filter applied after over-sampling to remove noisy or misplaced synthetic instances. It is based on the idea of a non-cooperative game, where each data point in the over-sampled dataset is assumed to be a player in the

game. The objective is for each player to determine the category label that maximizes its utility. $u_{(m+m^{syn}) \times 2}$ is a process matrix where u_j is (1,0) if positive, (0,1) if negative, and (0.5,0.5) if synthetic. Over a fixed number of iterations, each synthetic instance updates its label-probability vector according to the payoffs received from its neighbors, and synthetic instances that converge toward the negative label are removed from the dataset [25].

2.1.1.5 Random Under-Sampling (RUS) technique

RUS addresses the class imbalance issue by taking random turns in classifying instances to have a balanced distribution. While computational-efficient and simple to implement, this heuristic approach has significant flaws. The primary downside is its random selection process, which may eliminate the majority of the informative classes. This can lead to undershooting due to the lack of sufficient training data, a bias towards the minority class, and a reduced capacity to generalize on real-world data. As a result, RUS is primarily ineffective for small datasets that require the preservation of important samples [26].

2.1.1.6 Condensed Nearest Neighbor (CNNR)

The CNN rule under-samples the Maj-C by retaining only a consistent subset of the data, that is, a subset with which a 1-nearest-neighbor (1-NN) classifier still correctly classifies the training instances; redundant Maj-C instances far from the decision boundary are thereby discarded. The CNN technique follows two primary steps[27]:

- First, the condensed set is initialized with all Min-C instances and one randomly selected Maj-C instance.
- Each remaining Maj-C instance is then classified by 1-NN using the condensed set; misclassified instances are added to the condensed set, whereas correctly classified ones are discarded. The pass is repeated until no further instances are added.

2.1.2 Over-sampling Methods

The Over-sampling technique addresses the initial varying dataset by creating new synthetic instances and assigning them to the juvenile group. The key aspect of this technique is either producing existing Min-C instances or creating new ones. Duplication is performed randomly or by determining instances from the boundary between majority and minority groups. This measure drives the classifier to allocate such indefinite areas to the minority instance class. The most prevalent techniques for Over-sampling are Random Over-sampling (ROS), SMOTE, and Adaptive Synthetic Minority Sampling Technique (ADASYN) [28]. Table (2) show the main advantages and disadvantages of over-sampling methods [12].

TABLE II. THE MAIN ADVANTAGES AND DISADVANTAGES OF OVER-SAMPLING METHODS

Advantages	Disadvantages
Enhances the model's generalization capabilities by presenting a more comprehensive scope of variability within the Min-C, thereby conditioning the model for a broader range of overall scenarios.	There is a risk of overfitting, as representatives may start to learn the noise inherent in the synthetically developed instances rather than learning to generalize from the existing data.
Protection against the loss of vital information in Min-C instances is significant in datasets where each instance holds significant value.	Promotes the computational limitation, specifically when using sophisticated synthetic generation methods, which can be resource-intensive.

2.1.2.1 Random Over-Sampling (ROS) Method

ROS is the simplest over-sampling method: Min-C instances are randomly duplicated (sampled with replacement) until the desired class ratio is reached. Although fast, exact replication increases the risk of overfitting. A smoothed variant, Random Over-Sampling Examples (ROSE), generates new artificial instances through a smoothed bootstrap instead of exact copies. ROSE creates instances for both categories up to a pre-selected size. First, it determines a category to start with a likelihood of 0.5. Considering it began with a positive category, a hyper-parameter HP^{pos} specifies the possibility of determining an instance from a positive category. Then, the new positive number of instances to be created, $n^{new-pos}$, is selected based on the entire instance numbers to be created. The new negative number of instances is $n^{new-neg} = n - n^{new-pos}$, where n is the entire expected instance size. Then, it samples x^{neg} and x^{pos} using the likelihood distribution $H_{Kl}(\cdot, x_j)$, centered at x_j and dependent on the bandwidth matrix [29],[30].

2.1.2.2 Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE (Synthetic Minority Over-sampling Technique) overcomes the lack of diversity through linear interpolation between existing positive instances. For each minority sample $\hat{x}^{pos}$, the algorithm determines the k-nearest neighbors it has. A synthetic instance x^{syn} is then created along the line segment that connects x^{pos} to a random neighbor $\hat{x}^{pos}$, the computed value of this is as follows [31]:

$$X^{syn} = X_j^{pos} + (\hat{X}_j^{pos} - X_j^{pos}) \cdot \gamma \quad (1)$$

where γ is a random number with a distribution that is uniform between 0 and 1. The number of synthetic instances is decided by the desired ratio of minority to majority classes. This method effectively increases the size of the minority class region in the space of features, this will help classifiers to learn more complex boundaries for the decision. The most prominent of SMOTE Techniques are: First, Border-Line SMOTE (BL-SMOTE): This method centers on instances that are minority near the boundary of classes: these remain the most susceptible to misclassification. It produces samples that are synthetic specifically along this boundary to enhance and clarify the decision's scope (borderline Min-C instances) [32]. Second, K-means SMOTE: K-means SMOTE includes three phases: clustering, filtering, and over-sampling: (1) clustering the data using K-means, (2) selecting clusters with a high proportion of minority instances, and (3) applying SMOTE within those clusters, which distributes more samples to clusters that have a lower concentration of minority instances. This results in a balanced distribution of samples [33]. Third, SMOTR Edited Nearest Neighbors (SMOTE-ENN): A hybrid method that employs SMOTE to oversample the minority class in the initial dataset, followed by ENN to clean the augmented dataset. ENN eradicates any instance (from either class) that has a different class label than the majority of its nearest neighbors, this diminishes noise and overlap [34]. Fourth, SMOTE Support Vector Machine (SMOTE-SVM): This variant employs the support vectors of an SVM (the instances that define the class boundary) to guide synthetic data generation. It generates new minority populations that follow the lines connecting the minority class to their nearest neighbors, this increases the reinforcing the boundary [35]. Fifth, SMOTE Nominal and Continuous (SMOTE-NC): SMOTE-NC is intended for datasets that have a mixture of numerical and categorical data types (e.g. It produces synthetic numerical values via interpolation, but for categorical attributes, it chooses the most common category from the nearest neighbors. Despite its benefits for combining data, it has problems with high-dimensionality, potential noise contamination, and a malleable k-neighborhood that disregards local variations [9],[36]. Finally, SMOTE+Tomek Links: This hybrid method combines the over sampling of SMOTE with the cleaning of data via Tomek links. First, SMOTE increases the percentage of minority students. Next, Tomek links (pairs of oppositely labeled close neighbors) are recognized, and the majority portion of each pair is excised. This procedure facilitates the distinction of classes by eliminating indecipherable, boundary samples that are ambiguous, this leads to a more distinct decision boundary [37].

2.1.2.3 Adaptive Synthetic Minority Sampling Technique (ADASYN)

ADASYN is a developed Over-sampling technique created to manage class imbalance datasets, specifically for classification assignments. ADASYN aims to mitigate the effect of class imbalance by creating synthetic instances for the Min-C, generating more synthetic instances for Min-C instances that are harder to learn, i.e., those surrounded by a larger proportion of Maj-C neighbors. ADASYN assists the learning process in two ways: (1) it minimizes the bias induced by category imbalance, and (2) it adjusts the classification boundary to the complex issues. It also provides a balanced presentation of the data distribution based on the balance level set by the β coefficient. The dataset created by ADASYN enables the learning algorithm to focus on the challenging issues [38].

2.2 Algorithm-Level Methods

Algorithm-level methods focus on developing or adapting ML techniques to handle imbalanced data effectively. These techniques prioritize improving the algorithms' ability to classify samples from the Min-C accurately. By designing or adjusting algorithms to be additionally responsive to these underrepresented companies, these techniques contribute to enhancing the prevalent fairness, performance, and generalizability of ML techniques when encountered with an imbalanced dataset. Common algorithm-level techniques include CSL, class weighting in the loss function, decision-threshold moving, imbalance-aware loss functions such as focal loss, and ensemble methods that embed resampling (e.g., SMOTEBoost, RUSBoost, EasyEnsemble, and BalanceCascade); this section focuses on CSL and Weighted Shallow Neural Networks (WSNN) [10].

2.2.1 Cost-Sensitive learning (CSL)

CSL addresses issues of classifying by adding to the cost of misclassification that is asymmetric. Unlike the typical model that handles all errors equally, CSL gives higher rewards for critical errors, particularly when misclassifying positive instances as negative. This approach reduces the total cost of misclassification while also enhancing the identification of important minority classes in datasets that are imbalanced, this is particularly beneficial in domains such as medical diagnosis and spatial modeling, where the error rates are significantly different for different classes [12].

2.2.2 Weighted Shallow Neural Networks (WSNN)

WSNNs, integrated with imbalanced learning algorithms, have made comprehensive results in addressing imbalanced data issues. WSNNs have quick training ability due to the shallow layer configuration characteristics. This characteristic is particularly advantageous for applications that require fast processing of imbalanced data and real-time responses [39].

2.3 Hybrid Techniques

Hybrid methods combine various techniques, often combining algorithm-level and data-level techniques to supply a complete explanation to handle category imbalance challenges, thereby improving the reliability, effectiveness, and resilience of ML techniques. By combining multiple techniques, they objective to address issues such as information loss, overfitting, and inadequate performance in Min-Cs, thereby improving the prevalent effectiveness of ML technique in mitigating category disproportionality [11].

3. STRATEGIC CONSIDERATIONS FOR CHOOSING BETWEEN DATA BALANCING METHODS

The selection of a sampling strategy that is appropriate for the dataset's size and the criticality of the minority classes' instances is dependent on computational limitations, the dataset's size, and the importance of the classes' instances. Over-sampling is most common when it's necessary to accurately capturing rare, significant events (e.g., medical diagnoses or detection of fraud). Under-sampling is beneficial for large datasets that have sufficient redundancy, it's most effective when computational efficiency is of the essence. Hybrid methods, such as SMOTEEN, SMOTE+Tomeklinks, SMOTE-NC, or SMOTE-SVM, combine these approaches by improving the data quality while maintaining computational efficiency, making them more popular in practical applications.

4. COMPARATIVE ANALYSIS

To comprehensively evaluate the effectiveness of data balancing techniques, this section presents a comparative analysis across multiple studies, datasets, and ML algorithms. Table (3) shows some studies in balancing data and ML techniques. It includes references, publication year, dataset name, balancing data methods, type of balancing method, ML techniques, and accuracy. The comparative study emphasizes the practical performance and trade-offs among different balancing data methods. The ultimate goal is to provide insights into how different methodologies impact model performance and help determine which method is most suitable for a given dataset and desired classification task.

The practical application of data balancing techniques plays a significant role in the model's performance. Each study in the comparative analysis is examined individually. A summary of the most recent studies displays different balancing techniques applied across domains. Al- Husam A. H. Al-Najjar et al. [40] applied GANs to landslide prediction with an accuracy of 93.3% and reported instability in training. Other studies report results using SMOTE and some hybrid forms of SMOTE, such as SMOTE [41], returning an accuracy of 97.07% in a diabetes prediction application but still sensitive to overfitting and noise problems [35], [37], [42], [41], [43]. Zhang and Sun [44] proposed ESA-GCN, which extends GraphSMOTE with ENN-based cleaning and an attention mechanism for imbalanced node classification; on Cora, Citeseer, and PubMed with simulated imbalance, it achieved AUC-ROC scores of up to 0.8997 and outperformed GCN, GraphSAGE, DR-GCN, and GraphSMOTE in most settings. CSL-approaches [45] proved effective up to 93.92% accuracy with correct parameter tuning on cost. Majeed and Hwang [46] balanced poor-quality tabular data with a conditional GAN (CGAN) before applying differential privacy, showing that data balancing improves the utility of privacy-preserving data release, whereas Quantum-SMOTE [47] achieved an accuracy of 91.2% at the cost of higher computational complexity. SMOTE-based pipelines integrated with DL and ensemble models [48], [49] nearly attain perfect accuracies between 99% and 98.7%. However, generalization remains open problem. Consistently, studies emphasize that technique selection depends on data characteristics, with all methods exhibiting specific trade-offs between performance preservation, computational cost, and risk of artifacts.

TABLE III. COMPARATIVE STUDIES IN BALANCING DATA AND ML TECHNIQUES

Year+Ref.	Dataset Name	Balancing Data Technique	Type of Balancing	ML Algorithm	Accuracy Best (%)
2021 [40]	Chukha Dzongkhag Landslide Dataset	GAN	Over-sampling	ANN, RF, DT, KNN, SVM	93.3
2022 [35]	Jeeran Restaurant Reviews	SMOTE, ADASYN	Over-sampling	SVM	89.7
		SVM-SMOTE, Borderline-SMOTE	Hybrid		
2023 [42]	Trauma ICU patients, Besat Hospital, Hamadan, Iran (n = 126)	SMOTE, B-SMOTE1, B-SMOTE2, SMOTE-NC, SVM-SMOTE	Over-sampling	DT, RF, NB, ANN, SVM, XGBoost	100
2023 [36]	High School Longitudinal Study 2009	ROS	Over-sampling	RF	98.4
		RUS	Under-sampling		73.1
		SMOTE-NC + RUS	Hybrid		90.5
2023 [50]	Daraz Bengali Reviews	SMOTE	Over-sampling	SVM	92.8
2024 [37]	Kaggle Cyberbullying	Random Over-sampling	Over-sampling	SVM	91.1
		SMOTE	Over-sampling		
		SMOTE + TOMEK	Hybrid		

2024 [41]	Kaggle Diabetes Dataset	SMOTE	Over-sampling	RF, LightGBM, XGBoost, DT, MLP, SVM	97.07
2024 [43]	Assisted Reproduction Medical Data	SMOTE, ADASYN	Over-sampling	LR	84
		OSS, CNNR	Under-sampling		
2024 [44]	Cora, Citeseer, PubMed	ENN-SMOTE	Hybrid	ESA-GCN	89.97
2024 [45]	Bombay Weather Dataset (Kaggle, 2009–2020)	CSL	Algorithm-level	LR, DT, Ridge Classifier, SVM	93.92
2024 [51]	Kaggle Distribution Transformer Data	CSL	Algorithm-level	LR, SVM	80.9
2024 [9]	cleveland-0_vs_4, Haberman ecoli2, yeast-1_vs_7, glass0, abalone-21_vs_8, poker- 8_vs_6, and kr-vs-k- zero vs eight	LNE (Local Neighborhood Encodings)	Hybrid	CART, KNN, SVM, MLP	94.5
2024 [48]	KAU-BCMD	SMOTE	Over-sampling	ResNet50	99
2025 [52]	SEER Colorectal Cancer Dataset	SMOTE	Over-sampling	DT, ET, RF, GBoost, LGBost	80.69
		ENN	Under-Sampling		
		Repeat ENN (RENN), RENN + SMOTE, SMOTE + RENN	Hybrid		
2025 [46]	Adult, German Credit, Stroke Prediction	CGAN-based synthetic records	Over-sampling	RF	99.66
2025 [47]	Telco Customer Churn	Quantum-SMOTE	Hybrid	RF, LR	91.2
2025 [53]	Phishing Website Dataset	SMOTE, ROS, RUS	Over-sampling, Under-sampling	DT	95.1
2025 [54]	UCI, KEEL, Artificial (22 datasets)	MK-SMOTE	Hybrid	KNN, LR	91.0
2025 [55]	UCI Thyroid Disease Dataset	ROS	Over-sampling	Voting (RF + DT), XGBoost, RF, CatBoost, LightGBM	99.72
2025 [56]	Elderly care services survey (Guangzhou, Suzhou, Qingdao, China)	SMOTE-TomekLink	Hybrid	RF, GBDT, LightGBM	84.6
2025 [49]	Differentiated Thyroid Cancer Recurrence	SMOTE	Over-sampling	RF, LR, XGBoost, Adaboost, ANN, CatBoost, LightGBM, Bagging, ET	98.7

5. CHALLENGES AND LIMITATIONS OF DATA BALANCING TECHNIQUES

Data balancing technique plays a vital role in dealing with class imbalance. However, it also brings major problems of overfitting, loss of information as well as computational overhead. This technique improves the prediction for minor classes but tradeoffs between them should be carefully considered. Acknowledging these limitations is essential for developing more robust and adaptive solutions in future work.

5.1 Risk of Overfitting

Common Over-sampling methods can increase the risk of overfitting; since they create the synthetic cases in the MinC by replicating the minority group or perturbing it in some way, the real pattern or structure of MinC is not always captured. This highlights the extent to which balancing is a double-edged sword that can, if not properly used, harm generalization causing the problem of extracting robust, credible models even harder [57].

5.2 Information Loss

Information Loss is a problem associated more to data balancing and specifically to the under-sampling methods. A second potential limitation to the use of this class ratio balancing method is that, by discarding instances from the MajC to balance the datasets, important information and noise in the data could be removed, thus reducing the model's performance. The main issue, especially if the dropped instances include important features required to make accurate predictions [58].

5.3 Difficulty Handling Class Overlap

Difficulty dealing with class overlap is still a problem as resampling does not take into account the overlap of multi classes with similar feature space. Over-sampling or under-sampling do not deal with conflicting zones, such that fake Min-C instances can be put in ambiguous regions, or critical majority samples are removed nearby the boundaries between classes. Consequently, models can find it difficult to disambiguate classes, especially with noisy or high-dimensional datasets [59].

5.4 Computational Complexity

Although advanced balancing techniques improve data representation, their high computational overhead and lack of an inherently distributed design limit their scalability to big datasets and real-time applications. Neighbor-based methods such as SMOTE require repeated k-nearest-neighbor searches, which are costly for large, high-dimensional data, forcing a trade-off between accuracy and efficiency, most often using simpler under-sampling or selective filtering. This is a fundamental limitation that efficient, adaptive resampling algorithms with distributed operation in real-world scenarios [60].

5.5 Imbalance in Multi-Class Problems

Multi-class imbalance is more complex than binary because of interclass dependencies and multi-majority effects, where majorities suppress minorities, and with more than one minority and majority class, simple resampling becomes completely inadequate as structural relationships across classes cannot be preserved by a simplistic decomposition approach into binary. Therefore, special solutions are required to address both the lack of samples and the structural complexity in multi-class distributions [61].

6. FUTURE DIRECTIONS

Several future works trends for developing imbalanced data handling techniques are observed and summarized together with previous analyses in this paper. First, generating realistic minority-class samples by Generative Adversarial Networks (GAN) to make the final models robust and fair against minorities [40]. Second, an adaptive data balancing method can be encapsulated into Automated Machine Learning (AutoML) pipelines for automation in selecting methods and hyperparameters tuning [62]. Third, meta-learning frameworks will facilitate fast transfer of strategies between tasks on small data sets [63]. Fourth, Stream-based applications require incremental concept-drift-aware techniques due to dynamically changing class distributions typical IoT/cybersecurity environments [10]. Fifth, scalable solutions involve distributed parallel learning to address big data environments [64]. Sixth, fairness-aware balancing should integrate equity constraints directly into the resampling process so that the algorithm does not become biased against protected groups while maintaining accuracy [65]. Such actions indicate the necessity for efficient, adaptive, and mindful implementations.

7. CONCLUSION

To conclude, class imbalance is still a challenging problem for ML and has great impact on the robustness and generalization performance of predictive models. In this survey, we have investigated a variety of data balancing techniques such as under-sampling, over-sampling, algorithm-level, and hybrid methods, describing their methods, advantages, and drawbacks. While traditional approaches, such as SMOTE and RUS, offer foundational solutions, advanced hybrid, and cost-sensitive methods provide more robust handling of complex and high-dimensional imbalanced datasets. However, challenges such as overfitting, information loss, and computational overhead persist, particularly in real-time and multi-class settings. The selection of a suitable balancing method should be based on the characteristics of the dataset, the application context, and resource limitations. Future research should focus on scalable, adaptive, and fairness-aware solutions, particularly those combining DL and AutoML. Above all, this review can serve as a guide to select effective balancing strategies for constructing increasingly accurate, fair, and generalizable ML models with imbalanced data.

Acknowledgment

None.

Funding

None.

CRedit Authorship Contribution Statement

Ahmed Adil Nafea: Conceptualization, methodology, investigation, formal analysis, visualization, writing – original draft, writing – review & editing. The author has read and approved the final manuscript.

Conflicts of Interest

The author declares no conflicts of interest.

Ethical Approval

This study did not involve human or animal subjects, and ethical approval was not required.

Data Availability Statement

No new data were created or analyzed in this study. Data sharing is not applicable to this article.

References

- [1] A. G. N. Alsaedi, M. J. Varnamkhasti, H. J. Mohammed, and M. Aghajani, “Data Mining Classification Techniques to Improve Decision-Making Processes,” *International Journal of Mathematical Modelling & Computations*, vol. 14, no. 4, pp. 363–380, 2024.
- [2] K. Razzaq and M. Shah, “Machine learning and deep learning paradigms: From techniques to practical applications and research frontiers,” *Computers*, vol. 14, no. 3, p. 93, 2025.
- [3] A. A. Rabaan et al., “Unleashing the power of artificial intelligence for diagnosing and treating infectious diseases: A comprehensive review,” *J Infect Public Health*, vol. 16, no. 11, pp. 1837–1847, 2023.
- [4] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010, doi: 10.1016/j.patrec.2009.09.011.
- [5] A. K. Kareem and K. M. A. Alheeti, “Hybrid Approach for Fall Detection Based on Machine Learning,” in *International Conference on New Trends in Information and Communications Technology Applications*, Springer, 2021, pp. 111–130.
- [6] A. K. Kareem, A. M. Shaban, A. A. Nafea, M. Aljanabi, S. A. S. Aliesawi, and M. Mal-Ani, “Detecting Routing Protocol Low Power and Lossy Network Attacks Using Machine Learning Techniques,” in *2024 21st International Multi-Conference on Systems, Signals & Devices (SSD)*, IEEE, 2024, pp. 57–62.
- [7] A. R. Salehi and M. Khedmati, “A cluster-based SMOTE both-sampling (CSBBoost) ensemble algorithm for classifying imbalanced data,” *Sci Rep*, vol. 14, no. 1, p. 5152, 2024.
- [8] H. Kaur, H. S. Pannu, and A. K. Malhi, “A systematic review on imbalanced data challenges in machine learning: Applications and solutions,” *ACM computing surveys (CSUR)*, vol. 52, no. 4, pp. 1–36, 2019.
- [9] M. Kozłowski and M. Woźniak, “Local neighborhood encodings for imbalanced data classification,” *Mach Learn*, vol. 113, no. 10, pp. 7421–7449, 2024.
- [10] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, “A survey on imbalanced learning: latest research, applications and future directions,” *Artif Intell Rev*, vol. 57, no. 6, p. 137, 2024.
- [11] P. Kaur and A. Gosain, “Robust hybrid data-level sampling approach to handle imbalanced data during classification,” *Soft comput*, vol. 24, no. 20, pp. 15715–15732, 2020.
- [12] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, “Imbalanced data problem in machine learning: A review,” *IEEE Access*, vol. 13, pp. 13686–13699, 2025, doi: 10.1109/ACCESS.2025.3531662.
- [13] N. Junsomboon and T. Phienthrakul, “Combining over-sampling and under-sampling techniques for imbalance dataset,” in *Proceedings of the 9th international conference on machine learning and computing*, 2017, pp. 243–247.
- [14] G. Parmar, R. Gupta, T. Bhatt, G. J. Sahani, B. Y. Panchal, and H. Patel, “A review on data balancing techniques and machine learning methods,” in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, 2023, pp. 1004–1008.
- [15] C.-Y. Peng and Y.-J. Park, “A new hybrid under-sampling approach to imbalanced classification problems,” *Applied Artificial Intelligence*, vol. 36, no. 1, p. 1975393, 2022.
- [16] M. Peng et al., “Trainable undersampling for class-imbalance learning,” in *Proceedings of the AAAI conference on artificial intelligence*, 2019, pp. 4707–4714.
- [17] N. M. Mqadi, N. Naicker, and T. Adeliyi, “Solving misclassification of the credit card imbalance problem using near miss,” *Math Probl Eng*, vol. 2021, no. 1, p. 7194728, 2021.
- [18] F. Sağlam, “DatRel: a noise-tolerant data relocation approach for effective synthetic data generation in imbalanced classifiers,” *Mach Learn*, vol. 114, no. 5, p. 126, 2025.
- [19] R. Alejo, J. M. Sotoca, R. M. Valdovinos, and P. Toribio, “Edited nearest neighbor rule for improving neural networks classifications,” in *International Symposium on Neural Networks*, Springer, 2010, pp. 303–310.
- [20] S. Chatterjee and Y.-C. Byun, “Highly imbalanced fault classification of wind turbines using data resampling and hybrid ensemble method approach,” *Eng Appl Artif Intell*, vol. 126, p. 107104, 2023.
- [21] Q. Leng, J. Guo, J. Tao, X. Meng, and C. Wang, “OBMI: oversampling borderline minority instances by a two-stage Tomek link-finding procedure for class imbalance problem,” *Complex & Intelligent Systems*, vol. 10, no. 4, pp. 4775–4792, 2024.
- [22] G. Tuysuzoglu et al., “Joint Tomek Links (JTL): An Innovative Approach to Noise Reduction for Enhanced Classification Performance,” *IEEE Access*, 2025.
- [23] Z. Salekshahrezaee, J. L. Leevy, and T. M. Khoshgoftaar, “A reconstruction error-based framework for label noise detection,” *J Big Data*, vol. 8, no. 1, p. 57, 2021.
- [24] S. El Hajjami, J. Malki, A. Bouju, and M. Berrada, “Machine learning facing behavioral noise problem in an imbalanced data using one side behavioral noise reduction: application to a fraud detection,” *Int J Comput Info Eng*, vol. 15, no. 3, pp. 194–205, 2021.
- [25] A. Kim and I. Jung, “Optimal selection of resampling methods for imbalanced data with high complexity,” *PLoS One*, vol. 18, no. 7, p. e0288540, 2023.

- [26] T. Elhassan and M. Aljurf, "Classification of imbalance data using tometk link (t-link) combined with random under-sampling (rus) as a data reduction method," *Global J Technol Optim S*, vol. 1, no. S1, 2016.
- [27] F. Angiulli, "Fast nearest neighbor condensation for large data sets classification," *IEEE Trans Knowl Data Eng*, vol. 19, no. 11, pp. 1450–1464, 2007.
- [28] F. R. Adi Pratama and S. I. Oktora, "Synthetic Minority Over-sampling Technique (SMOTE) for handling imbalanced data in poverty classification," *Stat J IAOS*, vol. 39, no. 1, pp. 233–239, 2023.
- [29] A. D. Saputra, D. Arifianto, and R. Umilasari, "Effect Of Random Under Sampling and Random Over Sampling Method On Svm Performance," *Computing and Information System Journal*, vol. 1, no. 2, pp. 78–86, 2025.
- [30] J. Zhang and L. Chen, "Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis," *Computer Assisted Surgery*, vol. 24, no. sup2, pp. 62–72, 2019.
- [31] M. G. Karthik and M. B. M. Krishnan, "Hybrid random forest and synthetic minority over sampling technique for detecting internet of things attacks," *J Ambient Intell Humaniz Comput*, pp. 1–11, 2021.
- [32] R. M. Mathew and R. Gunasundari, "A review on handling multiclass imbalanced data classification in education domain," in *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, IEEE, 2021, pp. 752–755.
- [33] C.-C. Chang, Y.-Z. Li, H.-C. Wu, and M.-H. Tseng, "Melanoma detection using XGB classifier combined with feature extraction and K-means SMOTE techniques," *Diagnostics*, vol. 12, no. 7, p. 1747, 2022.
- [34] K. Swain et al., "Enhancing Stroke Prediction Using LightGBM With SMOTE-ENN and Fine-Tuning: A Comprehensive Analysis," *Cureus Journals*, vol. 1, no. 1, 2024.
- [35] R. Obiedat et al., "Sentiment analysis of customers' reviews using a hybrid evolutionary SVM-based approach in an imbalanced data distribution," *IEEE Access*, vol. 10, pp. 22260–22273, 2022.
- [36] T. Wongvorachan, S. He, and O. Bulut, "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining," *Information*, vol. 14, no. 1, p. 54, 2023.
- [37] M. Khairy, T. M. Mahmoud, and T. Abd-El-Hafeez, "The effect of rebalancing techniques on the classification performance in cyberbullying datasets," *Neural Comput Appl*, vol. 36, no. 3, pp. 1049–1065, 2024.
- [38] M. S. Mohammed, A. Al-Dhamari, W. Saeed, F. N. Al-Aswadi, S. A. M. Saleh, and M. N. Marsono, "Motion pattern-based scene classification using adaptive synthetic oversampling and fully connected deep neural network," *IEEE Access*, vol. 11, pp. 119659–119675, 2023.
- [39] P. Ewert, T. Orłowska-Kowalska, and K. Jankowska, "Effectiveness analysis of PMSM motor rolling bearing fault detectors based on vibration analysis and shallow neural networks," *Energies (Basel)*, vol. 14, no. 3, p. 712, 2021.
- [40] H. A. H. Al-Najjar, B. Pradhan, R. Sarkar, G. Beydoun, and A. Alamri, "A new integrated approach for landslide data balancing and spatial prediction based on generative adversarial networks (GAN)," *Remote Sens (Basel)*, vol. 13, no. 19, p. 4011, 2021.
- [41] H. Shao, X. Liu, D. Zong, and Q. Song, "Optimization of diabetes prediction methods based on combinatorial balancing algorithm," *Nutr Diabetes*, vol. 14, no. 1, p. 63, 2024.
- [42] R. Hassanzadeh, M. Farhadian, and H. Rafieemehr, "Hospital mortality prediction in traumatic injuries patients: comparing different SMOTE-based machine learning algorithms," *BMC Med Res Methodol*, vol. 23, no. 1, p. 101, 2023.
- [43] J. Zhu et al., "Processing imbalanced medical data at the data level with assisted-reproduction data as an example," *BioData Min*, vol. 17, no. 1, p. 29, 2024.
- [44] L. Zhang and H. Sun, "ESA-GCN: An Enhanced Graph-Based Node Classification Method for Class Imbalance Using ENN-SMOTE Sampling and an Attention Mechanism," *Applied Sciences*, vol. 14, no. 1, p. 111, 2024, doi: 10.3390/app14010111.
- [45] D. Subitha and G. N. Nimeshika, "Mitigating imbalance: Cost-sensitive learning for enhanced weather prediction in imbalanced datasets," in *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, 2024, pp. 1695–1698.
- [46] A. Majeed and S. O. Hwang, "Solving truthfulness-privacy trade-off in mixed data outsourcing by using data balancing and attribute correlation-aware differential privacy," *IEEE Access*, 2025.
- [47] N. Mohanty, B. K. Behera, C. Ferrie, and P. Dash, "A quantum approach to synthetic minority oversampling technique (SMOTE)," *Quantum Mach Intell*, vol. 7, no. 1, p. 38, 2025.
- [48] A. F. A. Alshamrani and F. Alshomran, "Optimizing Breast Cancer Mammogram Classification through a Dual Approach: A Deep Learning Framework Combining ResNet50, SMOTE, and Fully Connected Layers for Balanced and Imbalanced Data," *IEEE Access*, 2024.
- [49] I. Senyer Yapici and R. Uzun Arslan, "Predictive analytics for thyroid cancer recurrence: a feature selection and data balancing approach," *Eur Phys J Spec Top*, pp. 1–21, 2025.
- [50] F. U. Zaman, M. Z. Hossain, and M. K. Bhuiyan, "An Empirical Study on the Impact of SMOTE on Imbalanced Text Features for Review Rating Prediction in Bengali," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2023, pp. 1–5.
- [51] L. Shen, "Prediction of Distribution Transformer Faults by Logistic Regression Models Based on Cost-Sensitive Learning," in *2024 IEEE 7th International Conference on Information Systems and Computer Aided Education (ICISCAE)*, IEEE, 2024, pp. 510–515.
- [52] S. Soleimani, M. Bahrami, and M. Vali, "Survival prediction from imbalanced colorectal cancer dataset using hybrid sampling methods and tree-based classifiers," *Sci Rep*, vol. 15, no. 1, p. 14554, 2025.

- [53] K. Omari, C. Taoussi, and A. Oukhatar, “Comparative Analysis of Undersampling, Oversampling, and SMOTE Techniques for Addressing Class Imbalance in Phishing Website Detection,” *International Journal of Advanced Computer Science & Applications*, vol. 16, no. 2, 2025.
- [54] A. Kanwal, N. Masood, G. Mustafa, M. A. Ghafoor, and S. Ayaz, “Mk-smote and m-smote: enhanced techniques for handling class imbalance problem,” *Iran Journal of Computer Science*, pp. 1–19, 2025.
- [55] M. R. Islam, A. I. Chowdhury, S. Shama, and M. M. H. Lamyey, “Enhanced thyroid disease prediction using ensemble machine learning: a high-accuracy approach with feature selection and class balancing,” *Discover Artificial Intelligence*, vol. 5, no. 1, p. 9, 2025.
- [56] G. Yang, G. Wang, L. Wan, X. Wang, and Y. He, “Utilizing SMOTE-TomekLink and machine learning to construct a predictive model for elderly medical and daily care services demand,” *Sci Rep*, vol. 15, no. 1, p. 8446, 2025.
- [57] M. Mujahid et al., “Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering,” *J Big Data*, vol. 11, no. 1, p. 87, 2024.
- [58] M. Carvalho, A. J. Pinho, and S. Brás, “Resampling approaches to handle class imbalance: a review from a data perspective,” *J Big Data*, vol. 12, no. 1, p. 71, 2025.
- [59] N. A. Firdausanti, T. N. Fatyanosa, M. Data, I. Mendonça, and M. Aritsugi, “Two-stage sampling: A framework for imbalanced classification with overlapped classes,” in *2022 IEEE International Conference on Big Data (Big Data)*, IEEE, 2022, pp. 271–280.
- [60] W. C. Sleeman IV and B. Krawczyk, “Imbalanced Big Data Oversampling: Taxonomy, Algorithms, Software, Guidelines and Future Directions,” *arXiv preprint arXiv:2107.11508*, 2021.
- [61] B. Krawczyk, “Learning from imbalanced data: open challenges and future directions,” *Progress in artificial intelligence*, vol. 5, no. 4, pp. 221–232, 2016.
- [62] M. V. C. Aragão, M. de Freitas Carvalho, T. de Moraes Pereira, F. A. P. de Figueiredo, and S. B. Mafra, “Enhancing AutoML Performance for Imbalanced Tabular Data Classification: A Self-Balancing Pipeline,” 2024.
- [63] L. Ran, H. Sun, L. Gao, Y. Dong, and Y. Lu, “Meta-Hybrid: Integrate Meta-Learning to Enhance Class Imbalance Graph Learning,” *Electronics (Basel)*, vol. 13, no. 18, p. 3769, 2024.
- [64] A. M. Rahmani et al., “Artificial intelligence approaches and mechanisms for big data analytics: a systematic study,” *PeerJ Comput Sci*, vol. 7, p. e488, 2021.
- [65] K. Lammers, V. Vaquet, and B. Hammer, “Continuous Fair SMOTE--Fairness-Aware Stream Learning from Imbalanced Data,” *arXiv preprint arXiv:2505.13116*, 2025.