A Review of Hybrid Sampling and Ensemble Learning Techniques for Imbalanced Data Classification

Puneetha B H [1], Rubini P [2]

[1] Postdoctoral Research Fellow, Lincoln University College, Malaysia.

[2] Professor ,CSE , School of Engineering and Technology, CMR University , Bengaluru.

[a]Puneethgowdar.bh@gmail.com, [b] prubini.cs@gmail.com

Abstract: Class imbalance is a significant challenge in machine learning, where the unequal distribution of classes often leads to biased prediction models and poor minority class recognition. This issue is commonly observed in critical domains such as healthcare, cybersecurity, fraud detection, and financial analysis, where accurate identification of rare events is essential. Over the years, several techniques have been developed to address this problem, including data-level methods, algorithm-level approaches, ensemble learning, and hybrid strategies. However, many existing methods still suffer from limitations related to scalability, overfitting, computational complexity, and limited adaptability to dynamic datasets.This paper presents a comprehensive review of recent hybrid learning techniques used for imbalanced data classification. The study analyzes commonly adopted sampling approaches such as SMOTE, ADASYN, Tomek Links, and Borderline-SMOTE along with ensemble techniques including Random Forest, AdaBoost, Gradient Boosting, and deep learning-based hybrid models. Furthermore, the paper discusses the strengths and limitations of existing approaches by comparing their performance using evaluation metrics such as Recall, F1-score, G-mean, and AUC. Based on the analysis, several research gaps are identified, particularly in the areas of adaptive learning, scalability, and explainability.Finally, this review highlights the need for intelligent and adaptive hybrid frameworks capable of handling evolving real-world imbalanced datasets efficiently. The findings of this study are expected to support researchers in developing robust and reliable machine learning models for future imbalance learning applications.

Keywords: Imbalanced Data, Hybrid Learning, SMOTE, Ensemble Learning, Adaptive Sampling, Machine Learning.

1. Introduction

Machine learning models have achieved remarkable success in solving complex real-world problems across domains such as healthcare, cybersecurity, finance, transportation, and industrial automation. Most machine learning algorithms are designed under the assumption that the number of samples belonging to each class is approximately Equal[1]. However, real-world datasets often contain highly uneven class distributions, commonly referred to as the class imbalance problem. In such datasets, one class, known as the majority class, contains significantly larger samples[2] compared to the minority class. This imbalance causes machine learning models to become biased toward the majority class, leading to poor detection of minority instances[3]. The class imbalance problem has become a critical issue because minority class instances are usually more important than majority class samples[4]. For example, in medical diagnosis, rare disease cases are often limited compared to healthy patient

records[5]. Similarly, in fraud detection and cybersecurity applications, fraudulent transactions or network intrusions occur less frequently than normal activities. Traditional machine learning models may achieve high overall accuracy in these situations while still failing to correctly identify minority class events[6]. Therefore, relying only on accuracy is insufficient when dealing with imbalanced datasets.

To address this issue, researchers have proposed several balancing techniques that can generally be categorized into data-level methods, algorithm-level methods[7], and hybrid approaches. Data-level techniques attempt to balance the dataset before training by applying oversampling or undersampling methods. Popular oversampling approaches such as SMOTE and ADASYN generate synthetic minority class samples, while undersampling[8] techniques reduce majority class instances to achieve balance. Although these methods improve class representation, they may introduce problems such as overfitting, information loss, or increased computational Complexity[9]. Algorithm-level techniques focus on modifying machine learning algorithms to improve minority class learning without altering the dataset directly[10]. Cost-sensitive learning, ensemble learning, boosting methods, and deep learning adjustments are commonly used algorithmic approaches. Ensemble techniques such as Random Forest, AdaBoost, and Gradient Boosting have demonstrated improved classification performance on imbalanced datasets by combining multiple weak learners into a stronger prediction model. However, many algorithm-level approaches still face challenges related to scalability, parameter tuning, and model interpretability[11]. Recently, hybrid approaches that combine data-level and algorithm-level techniques have gained significant attention. These methods attempt to utilize the advantages of both sampling strategies and intelligent learning models to improve prediction performance[12]. Hybrid frameworks integrating SMOTE with ensemble learning or deep learning techniques have shown promising results in healthcare, fraud detection, and anomaly detection applications. Despite these improvements, existing approaches still suffer from limitations such as static balancing strategies, computational overhead, limited explainability, and poor adaptability to dynamic real-world datasets[13]. Motivated by these challenges, this paper presents a comprehensive review of hybrid learning techniques for imbalanced data classification[14]. The study analyzes recent developments in sampling methods, ensemble learning, and hybrid frameworks while highlighting their strengths, limitations, and research gaps. Furthermore, this review discusses important evaluation metrics used for imbalance learning and identifies future directions toward adaptive, scalable, and explainable imbalance learning frameworks[15]. The remainder of this paper is organized as follows. Section II presents the review methodology and discusses the research questions considered in this study. Section III explains the fundamental concepts of oversampling and undersampling techniques used to address class imbalance. Section IV reviews various balancing strategies, including data-level, algorithm-level, ensemble-based, and hybrid approaches for imbalanced data classification. Section V discusses the major limitations and challenges associated with existing techniques. Section VI highlights important evaluation metrics such as Recall, F1-score, G-mean, and AUC used for assessing model performance on imbalanced datasets. Finally, Section VII concludes the paper by summarizing the findings, identifying research gaps, and presenting possible future research directions in adaptive and explainable imbalance learning frameworks.

2. Review Methodology

This review investigates recent advancements in machine learning techniques developed to address the class imbalance problem in real-world datasets. The study mainly focuses on data-level, algorithm-level, ensemble-based, and hybrid learning approaches used for imbalance classification. Research articles were collected from reputed scientific databases including IEEE Xplore, Springer, ScienceDirect, Elsevier, MDPI, and Google Scholar. The review primarily considers studies published between 2020 and 2025 to analyze recent developments and emerging trends in imbalance learning. Keywords such as "imbalanced data," "SMOTE," "ensemble learning," "hybrid approaches," "adaptive sampling," and "cost-sensitive learning" were used during the search process. To ensure quality and relevance, only peer-reviewed journal articles and conference papers related to imbalance learning and machine learning classification were selected. This review aims to identify the strengths, limitations, evaluation methods, and research gaps associated with existing imbalance handling techniques.

This review aims to address the following review questions.

RQ1: How effective are oversampling and undersampling techniques in handling imbalanced datasets?

RQ2: What are the strengths and limitations of data-level, algorithm-level, and hybrid learning approaches?

RQ3: Which evaluation metrics are most suitable for measuring performance in imbalanced classification problems?

RQ4: What research gaps and challenges still exist in current imbalance learning techniques?

By systematically answering these questions, this paper provides a comprehensive synthesis of the current state of the art, offering researchers and practitioners a structured understanding of existing imbalance learning techniques, their limitations, and potential future research directions for developing adaptive and efficient hybrid classification models.

3. Fundamental Balancing Techniques

Class imbalance significantly affects the performance of machine learning models by causing bias toward majority class instances[7]. To overcome this issue, several balancing techniques have been developed to improve the representation of minority classes in the dataset. These techniques are generally categorized into oversampling and undersampling methods. Oversampling techniques increase the number of minority class samples, whereas undersampling methods[13] reduce the number of majority class samples to achieve a balanced class distribution.Fig. 1 illustrates the basic balancing approaches used to handle class imbalance through oversampling and undersampling techniques.

3.1 Oversampling Techniques

Oversampling techniques attempt to improve minority class representation by generating additional samples for the minority class[9]. One of the most widely used oversampling methods is SMOTE (Synthetic Minority Over-sampling Technique), which creates synthetic minority samples by interpolating between neighboring minority data points[11]. Unlike random oversampling, SMOTE reduces the chances of overfitting by generating new artificial samples instead of simply duplicating existing instances. Another commonly used technique is ADASYN (Adaptive Synthetic Sampling), which generates

synthetic data adaptively based[4] on the learning difficulty of minority samples. More synthetic samples are created for difficult minority instances located near decision boundaries. Borderline-SMOTE further improves oversampling by focusing specifically on minority samples located near the classification boundary[12], where misclassification is more likely to occur. Although oversampling techniques improve minority class learning and classification performance, they may increase computational complexity and sometimes introduce noisy or overlapping samples.

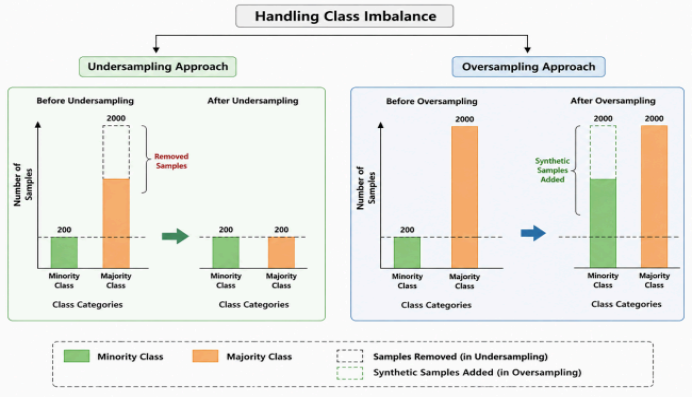

Figure. 1. Handling Class Imbalance using Oversampling and Undersampling Approaches

3.2 Undersampling Techniques

Undersampling techniques reduce the number of majority class samples to balance the dataset. Random undersampling is one of the simplest methods[3], where majority class instances are removed randomly until a balanced distribution is achieved. This method reduces training time and computational cost but may lead to information loss if important majority samples are removed. More advanced undersampling methods include Tomek Links and NearMiss. Tomek Links identify overlapping samples between majority and minority classes and remove noisy majority instances near decision boundaries[7]. NearMiss selects majority samples based on their distance from minority samples to improve class separability. Undersampling methods are computationally efficient and suitable for large datasets; however, excessive removal of majority samples may negatively affect overall classification performance[9].

3.3 Comparative Discussion

Both oversampling and undersampling techniques play an important role in handling imbalanced datasets. Oversampling methods improve minority representation without losing information but may increase training complexity[5]. In contrast, undersampling techniques reduce dataset size and computational requirements but may discard useful information. Due to these limitations, recent studies increasingly focus on hybrid approaches that combine both techniques with ensemble learning models to achieve improved classification performance and better generalization capability[3].

4. Balancing Strategies For Imbalanced Data

Several balancing strategies have been proposed in recent years to improve machine learning performance on imbalanced datasets[14]. These approaches are generally categorized into data-level techniques, algorithm-level techniques, and hybrid approaches. Data-level methods focus on modifying the training dataset through resampling, whereas algorithm-level techniques adjust the learning process of classifiers to improve minority class prediction[6]. Hybrid approaches combine both strategies to achieve better classification performance, robustness, and generalization capability. Similar categorization is also discussed in recent imbalance learning surveys[15].

4.1 Data-Level Techniques

Data-level techniques aim to balance class distribution before model training by applying oversampling or undersampling methods[2]. Oversampling methods increase minority class representation by generating synthetic samples, while undersampling methods reduce majority class instances to achieve balanced learning[4]. Popular oversampling methods include SMOTE, ADASYN, Borderline-SMOTE, and GAN-based augmentation techniques. These methods improve minority class learning and help reduce prediction bias toward majority classes[7]. Undersampling techniques such as Random Undersampling, Tomek Links, and NearMiss reduce dataset size and computational complexity. However, excessive removal of majority samples may lead to information loss[11]. To overcome the limitations of standalone resampling methods, several hybrid-sampling approaches have been developed to improve classification accuracy and data quality. Table I summarizes the effectiveness and limitations of commonly used hybrid-sampling techniques for handling imbalanced datasets[13].

TABLE I: Comparative Analysis of Data-Level Balancing Techniques

Technique	Major Contribution	Limitation
SMOTE-ENN[2]	Removes noisy samples while improving minority representation	Performance decreases with highly overlapping classes
Borderline-SMOTE[3]	Generates samples near decision boundaries for better learning	May create class overlap and noisy regions
ADASYN[4]	Produces adaptive synthetic samples for difficult instances	Increased computational complexity
D-SMOTE[5]	Controls noisy sample generation using distance measures	High-dimensional data increases processing cost
GAN-based Augmentation[8]	Generates realistic synthetic minority samples	Requires large training time and computational resources
SASMOTE[11]	Improves sample quality through adaptive neighbor selection	Requires careful parameter tuning
OOSI[14]	Enhances minority learning while reducing overfitting	Runtime efficiency decreases for complex datasets

4.2 Algorithm-Level Techniques

Algorithm-level techniques improve imbalance handling by modifying the learning behavior of classifiers without changing the original dataset distribution[6]. These approaches include cost-sensitive learning, ensemble learning, boosting algorithms, and algorithm-specific modifications. Cost-sensitive

learning assigns higher penalties to minority class misclassification, thereby forcing the classifier to focus more on minority instances[8]. Ensemble learning techniques combine multiple classifiers to improve predictive performance and reduce classification bias. Methods such as AdaBoost, Random Forest, Bagging, and Gradient Boosting have demonstrated improved performance on imbalance classification tasks[3]. Deep learning-based ensemble models further enhance prediction capability by learning complex feature representations from imbalanced datasets[2]. Table II presents the effectiveness and limitations of commonly used ensemble and algorithm-level techniques for handling imbalanced datasets.

TABLE II: Comparative Study of Algorithm-Level Learning Techniques

Technique	**Major Contribution**	**Limitation**
AdaBoost[2]	Improves learning of difficult minority instances	Sensitive to noisy data and outliers
Bagging[5]	Enhances prediction stability and reduces variance	Requires higher computational resources
Random Forest[7]	Provides robust classification performance	May still favor majority class in severe imbalance
Gradient Boosting[9]	Achieves high predictive accuracy	Risk of overfitting with excessive trees
SVM with MKL[11]	Improves flexibility and classification capability	Complex parameter optimization
Stacked Deep Learning[15]	Learns complex patterns from imbalanced data	Reduced interpretability and higher complexity

4.3 Hybrid Approaches

Hybrid approaches integrate data-level and algorithm-level techniques to utilize the strengths of both balancing strategies[14]. These methods attempt to improve minority class prediction while reducing overfitting, information loss, and model bias. Hybrid frameworks commonly combine oversampling techniques such as SMOTE with ensemble learning models including Random Forest, CNN, RNN, and boosting algorithms[15]. Recent hybrid approaches also incorporate deep learning and GAN-based models for generating realistic minority samples and improving classification robustness[11]. Such methods have shown promising performance in healthcare diagnosis, fraud detection, cybersecurity, and anomaly detection applications. However, despite improved classification performance, hybrid approaches still face challenges related to scalability, computational complexity, and explainability.Table III highlights the effectiveness and limitations of recent hybrid approaches developed for handling imbalanced data classification problems[15].

TABLE III: Comparative Analysis of Hybrid Imbalance Learning Approaches

Technique	**Major Contribution**	**Limitation**
SMOTE + Random	Improves minority class classification accuracy	Increased training complexity

Forest[3]		
SMOTEBoost[7]	Enhances minority learning using boosting	Sensitive to noisy synthetic samples
RUSBoost[9]	Combines undersampling with boosting for balanced learning	Possible information loss from majority class
DCGAN + CNN[11]	Generates realistic synthetic data for deep learning	High computational requirements
D-SMOTE + Stacked CNN/RNN[13]	Achieves strong predictive accuracy in healthcare datasets	Risk of overfitting and reduced explainability
Tomek Links + Borderline-SMOTE[14]	Improves class separation and reduces noise	Requires careful parameter tuning
Hybrid Ensemble Framework[15]	Enhances robustness and generalization capability	Complex architecture and scalability issues

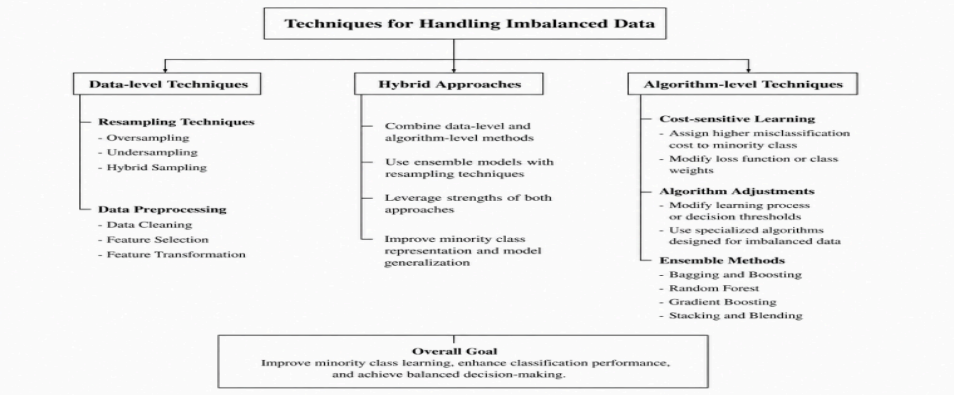

Figure 2: Overview of techniques for handling imbalanced data

5. Limitations And Research Challenges

Despite the significant progress achieved in imbalance learning techniques, several limitations and research challenges still exist in current approaches[3]. Data-level methods such as oversampling and undersampling improve class distribution; however, they may introduce new challenges related to overfitting, information loss, and computational overhead[5]. Oversampling techniques, particularly SMOTE-based methods, can generate noisy or overlapping synthetic samples when minority and majority classes are not clearly separated[7]. Similarly, undersampling approaches may remove important majority class information, negatively affecting classification performance and model generalization.Algorithm-level techniques such as cost-sensitive learning, boosting, and ensemble learning improve minority class prediction by modifying classifier behavior[9]. However, these methods often require extensive parameter tuning and high computational resources, especially for large-scale datasets. Ensemble models such as Random Forest and Gradient Boosting may still exhibit bias toward majority classes in severely imbalanced datasets[11]. Deep learning-based approaches further increase model complexity, making training expensive and reducing interpretability. Hybrid approaches combining sampling techniques with ensemble or deep learning models have shown promising results in healthcare, cybersecurity, and fraud detection applications[13]. Nevertheless, these approaches still face limitations associated with scalability, training time, and explainability. Complex hybrid frameworks

integrating CNNs, RNNs, GANs, and boosting algorithms may achieve high classification accuracy but often operate as black-box systems, making decision interpretation difficult in sensitive applications such as medical diagnosis[14].Another major challenge involves handling dynamic and evolving datasets where class distributions change over time. Most existing balancing techniques operate under static assumptions and may fail to adapt effectively to real-time imbalance variations. Furthermore, many current methods are evaluated on benchmark datasets with limited diversity, reducing their generalization capability in real-world environments.Therefore, future research should focus on developing adaptive, scalable, and explainable hybrid frameworks capable of handling dynamic imbalance scenarios efficiently while maintaining computational efficiency and reliable minority class prediction performance[15].

6. Evaluation Metrics For Imbalanced Data

Evaluating machine learning models on imbalanced datasets requires specialized performance metrics because traditional accuracy alone may provide misleading results[7]. In highly imbalanced datasets, a classifier can achieve high accuracy by correctly predicting majority class instances while failing to identify minority class samples. Therefore, evaluation metrics such as Recall, Precision, F1-score, G-mean, and Area Under the Curve (AUC) are commonly used to measure the effectiveness of imbalance handling techniques[9].Recall measures the ability of a classifier to correctly identify minority class instances. It is considered one of the most important metrics in imbalance learning because minority samples are usually more critical in applications such as disease diagnosis, fraud detection, and intrusion detection[11]. Precision evaluates how many predicted positive instances are actually correct, whereas the F1-score provides a balanced measure by combining both Recall and Precision[14].

TABLE IV : Common Evaluation Metrics for Imbalanced Classification

Metric	Purpose
Accuracy	Measures overall prediction correctness
Recall	Measures minority class detection capability
Precision	Measures correctness of positive predictions
F1-score	Balances Precision and Recall
G-mean	Evaluates balanced classification performance
AUC-ROC	Measures class separation capability

The G-mean metric evaluates the balance between classification performance on majority and minority classes[3]. It ensures that improvement in one class does not significantly reduce performance in the other class. Similarly, AUC measures the ability of a classifier to distinguish between classes across different threshold settings and is widely used for evaluating imbalance classification models. Table IV presents the commonly used evaluation metrics for imbalance learning and their significance in measuring classification performance.Among these metrics, Recall, F1-score, G-mean, and AUC are considered more reliable for imbalanced datasets because they focus on minority class prediction performance rather than overall. classification accuracy[13]. Therefore, selecting. suitable evaluation metrics is essential for developing. robust and effective imbalance learning models.

7. CONCLUSION AND FUTURE DIRECTIONS

Class imbalance remains one of the major challenges in machine learning, particularly in critical applications such as healthcare, fraud detection, cybersecurity, and anomaly detection. This paper reviewed various imbalance handling strategies, including data-level, algorithm-level, ensemble-based, and hybrid approaches. The study analyzed commonly used techniques such as SMOTE, ADASYN, Random Forest, AdaBoost, GAN-based augmentation, and hybrid ensemble models, along with their strengths and limitations. The review indicates that hybrid approaches generally provide better classification performance by combining the advantages of sampling methods and intelligent learning models. However, existing techniques still suffer from limitations related to overfitting, computational complexity, parameter sensitivity, scalability, and limited explainability. In addition, most current methods are designed for static datasets and may not adapt effectively to dynamic real-world imbalance scenarios. Based on the analysis, several research gaps were identified in adaptive learning, scalable model development, and explainable imbalance classification. Future research should focus on developing intelligent hybrid frameworks capable of dynamically adjusting to changing class distributions while maintaining reliable minority class prediction performance. Furthermore, integrating explainable artificial intelligence techniques with imbalance learning models can improve transparency and trustworthiness in sensitive application domains. Overall, adaptive and explainable hybrid learning frameworks represent a promising direction for future imbalance learning research.

References

1. M. Alamri and M. Ykhlef, "Hybrid Undersampling and Oversampling for Handling Imbalanced Credit Card Data," IEEE Access, 2024.
2. Q. Su, H. N. A. Hamed, M. A. Isa, X. Hao, and X. Dai, "A GAN-Based Data Augmentation Method for Imbalanced Multi-Class Skin Lesion Classification," IEEE Access, 2024.
3. 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.
4. T. Kosolwattana, C. Liu, R. Hu, S. Han, H. Chen, and Y. Lin, "A Self-Inspected Adaptive SMOTE Algorithm (SASMOTE) for Highly Imbalanced Data Classification in Healthcare," BioData Mining, vol. 16, no. 1, p. 15, 2023.
5. Y. Deng and M. Li, "An Adaptive and Robust Method for Oriented Oversampling with Spatial Information for Imbalanced Noisy Datasets," IEEE Access, 2023.
6. T. Suresh, Z. Brijet, and T. Subha, "Imbalanced Medical Disease Dataset Classification Using Enhanced Generative Adversarial Network," Computer Methods in Biomechanics and Biomedical Engineering, vol. 26, no. 14, pp. 1702–1718, 2023.
7. A. M. Sowjanya and O. Mrudula, "Effective Treatment of Imbalanced Datasets in Health Care Using Modified SMOTE Coupled with Stacked Deep Learning Algorithms," Applied Nanoscience, vol. 13, no. 3, pp. 1829–1840, 2023.
8. V. Werner de Vargas et al., "Imbalanced Data Preprocessing Techniques for Machine Learning: A Systematic Mapping Study," Knowledge and Information Systems, vol. 65, no. 1, pp. 31–57, 2023.
9. J. Zheng, X. Wang, D. Wei, B. Chen, and Y. Shao, "A Novel Imbalanced Ensemble Learning in Software Defect Prediction," IEEE Access, vol. 9, pp. 86855–86868, 2021.

10. G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, “Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning from Imbalanced Data,” IEEE Access, vol. 9, pp. 74763–74777, 2021
11. E. Elyan, C. F. Moreno-Garcia, and C. Jayne, “CDSMOTE: Class Decomposition and Synthetic Minority Class Oversampling Technique for Imbalanced-Data Classification,” Neural Computing and Applications, vol. 33, pp. 2839–2851, 2021.
12. N. Jiang and N. Li, “A Wind Turbine Frequent Principal Fault Detection and Localization Approach with Imbalanced Data Using an Improved Synthetic Oversampling Technique,” International Journal of Electrical Power & Energy Systems, vol. 126, p. 106595, 2021.
13. H. Zhang, H. Zhang, S. Pirbhulal, W. Wu, and V. H. C. de Albuquerque, “Active Balancing Mechanism for Imbalanced Medical Data in Deep Learning-Based Classification Models,” ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 16, no. 1s, pp. 1–15, 2020.
14. M. Khushi et al., “A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data,” IEEE Access, vol. 9, pp. 109960–109975, 2021.
15. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.