Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 4, 2025

ENHANCING SEMI-SUPERVISED LEARNING MODELS FOR IMBALANCED CLASS DISTRIBUTION

Muhammad Arslan Yousaf*[1], Syed Asad Ali Naqvi[2], Muhammad Usman Saleem[3], Ahmed Zeeshan[4]

*[*1,2]Department of Computer Science, Faculty of Computer Science & IT Superior University Lahore, 54000, Pakistan*
[3]Department of Computer Science, Government College Women University Sialkot
[4]Department of Computer Science University of Gujrat, Gujrat, Pakistan

[*1]muhammadarslanyousaf@gmail.com, [2]syedasad.alinaqvi@superior.edu.pk, [3]usman.saleem@gcwus.edu.pk, [4]ahmed@cs.uchenab.edu.pk

DOI: https://doi.org/10.5281/zenodo.15221657

Keywords
Semi-supervised learning, class imbalance, minority class, data augmentation, cost-sensitive learning, ensemble methods, self-paced learning

Article History
Received on 08 March 2025
Accepted on 08 April 2025
Published on 15 April 2025

Corresponding Author: *
Muhammad Arslan Yousaf

Abstract
Semi-supervised learning (SSL) has achieved great success in overcoming the difficulties of labeling and making full use of unlabeled data. Semi-supervised learning is an effective approach for addressing the issue of insufficient labeled data, utilizing both labeled and unlabeled datasets. Class imbalance remains a significant challenge, particularly in real-world scenarios. Class dominance imbalance leads to a model bias toward the majority class, hindering the accurate learning and representation of minority classes, which impacts overall model performance. Current algorithms focus on maximizing overall accuracy but fail to ensure balanced performance across classes. This bias toward the dominant class limits the model's applicability, especially in fields like healthcare, fraud detection, and anomaly detection, where minority class prediction is crucial. This paper proposes novel strategies for semi-supervised learning, specifically targeting imbalanced class distributions. The proposed approach enhances data distribution strategies to address the imbalance issue without compromising model performance. Experimental results demonstrate a significant improvement in the performance of minority classes, validating the effectiveness of the proposed techniques in improving model equity and trustworthiness. This research provides a roadmap for the use of semi-supervised learning in real-life applications where class imbalance is a prevalent issue.

INTRODUCTION

Semi-supervised learning leverages both labeled and unlabeled data to build models. It is particularly useful when labeled data is scarce. However, real-world datasets often suffer from class imbalance, where some classes are significantly underrepresented [1]. This imbalance can lead to models that prioritize majority classes and perform poorly on minority classes. Addressing this issue is crucial for improving the robustness and applicability of semi-supervised learning models. Machine learning has done better in a variety of application domains, including but not limited to NLP and CV. Nevertheless, its reliance on large high-quality labeled datasets may indeed turn into a problem since acquiring such data is both time and cost-intensive [2, 3]. SSL appears as a reasonable solution that utilizes both labeled and unlabeled data; thus it minimizes the reliance on

labeled datasets and enhances the generality of the model [4].

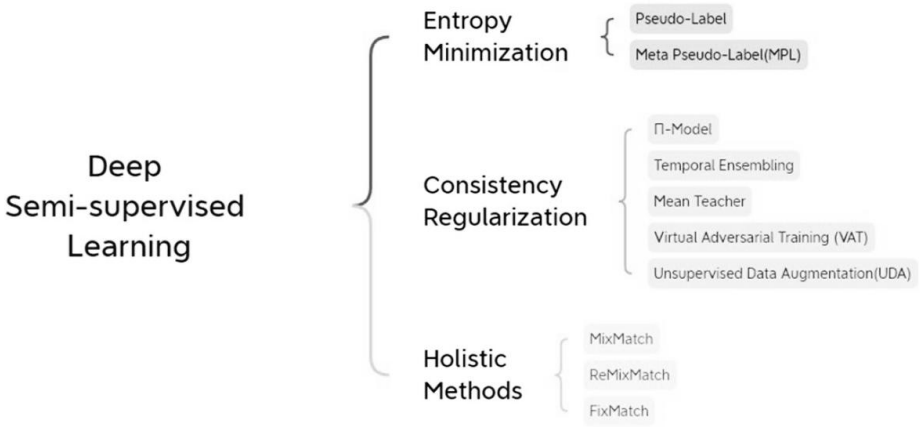

Figure 1: The taxonomy of deep semi-supervised learning methods [5]

For a training example with C possible output classes, and m = ½ and $(f(x;) + f(\hat{x};\theta))$ measure can be calculated as follows.

$$\mathcal{L}_u = \sum_{k=1}^{C} -f(x;\theta)_k \log f(x;\theta)_k$$

Eq (1)

However, SSL has a great deal with such a drawback when the classification is performed in scenarios that have skewed class distribution. Accidentally, real-life data sets have a class imbalanced problem, which results in the model having a poor performance in the minority classes [6, 7]. This bias is especially quite detrimental in scenarios such as disease diagnosis, risk assessment for scams, and the prediction of low-frequency events, where the correct modeling of minorities is extremely significant [8].

$$d_{MSE}(f(x;\theta),f(\hat{x};\theta))\frac{1}{C}\sum_{k=1}^{C}(f(x;\theta)_k - f(\hat{x};\theta)_k)^2$$

Eq (2)

$$d_{KL}(f(x;\theta),f(\hat{x};\theta)) = \frac{1}{C}\sum_{k=1}^{C} f(x;\theta)_k \log\frac{f(x;\theta)_k}{f(\hat{x};\theta)_k}$$

Eq (3)

The problem of imbalance can be tackled efficiently by utilizing resampling or cost-sensitive methods in the framework of traditional supervised learning; however, these approaches lack efficiency for adaptation to SSL settings [9]. This work introduces new techniques for tackling class imbalance within the framework of SSL to improve the representation of the minority class without detriment to the performance of the model. These innovations are expected to address a major challenge substantially and augment the use of SSL in unbalanced real-world datasets [10].

$$d_{JS}(f(x;\theta),f(\hat{x};\theta)) = \frac{1}{2}d_{KL}(f(x;\theta),m) + \frac{1}{2}d_{KL}(f(\hat{x};\theta),m)$$

Eq (4)

1.1 Motivation and Research Question

The prevalence of class imbalance in datasets is a common issue that degrades the performance of machine learning models. In semi-supervised learning, this imbalance is particularly problematic as it can lead to models that fail to generalize well to minority classes. There is a pressing need to develop techniques that can mitigate the effects of class imbalance and improve model performance across all classes. How can novel techniques be developed and implemented to effectively address imbalanced class distributions in semi-supervised learning, thereby improving model performance, particularly for minority classes.

2. Literature Review

Existing research highlights several methods for handling imbalanced data in supervised learning, such as SMOTE and cost-sensitive learning [11].

However, these techniques are often not directly applicable to semi-supervised learning scenarios. This paper builds on these foundations to develop and evaluate new approaches suitable for semi-supervised learning. SSL methods have grown tremendously over time; self-training, co-training, and graph-based methods have worked well with the use of unlabeled data [12, 13]. However, these methods generally are not good at dealing with problems of a hopelessly skewed dataset, in a way that a random forest of decision trees is [14].

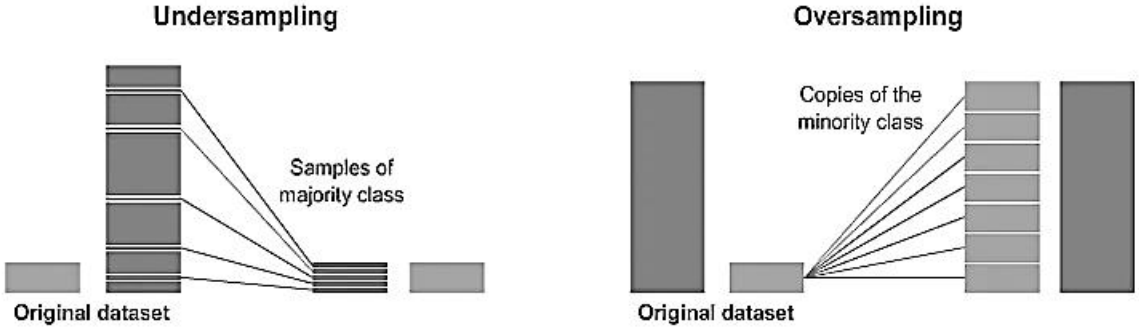

Figure 2: Supervised Learning Oversampling [15]

In supervised learning oversampling, under sampling and synthetic techniques like SMOTE for example have enhanced the minority class representation [16].

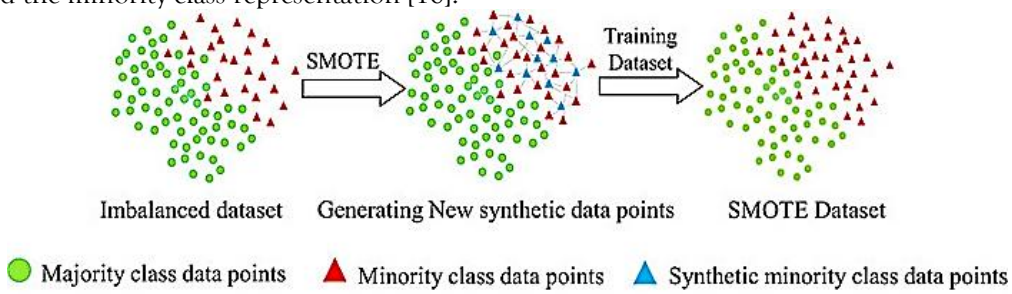

Figure 3: The Synthetic Minority Over-Sampling (SMOTE) to address imbalanced data [17]

Cost-sensitive techniques increase the penalty for misclassification of the minority classes so that the model pays more attention to them [18]. However, these methods cannot be directly used in SSL because each of them depends on labeled data. The previous SSL development has tried to solve the imbalance in recent years. For instance, class-aware modifications to consistency regularization, and pseudo-labeling have been observed to be promising [19, 20]. Other methodologies are to reweight the unlabeled data or combine it with the hybrid models like GANs for synthetic data generation However, there is still no coherent framework for efficiently addressing the class imbalance issue in SSL systems [21, 22]. This research intends to address this issue by addressing methodologies for imbalanced SSL to improve both minority classes and general model performance.

3. Methodology

We propose several techniques to address class imbalance in semi-supervised learning:

Data Augmentation: Generating synthetic samples for minority classes to balance the dataset.

Cost-Sensitive Learning: Assigning higher misclassification costs to minority classes.

Ensemble Methods: Using multiple models to reduce bias towards majority classes.

Self-Paced Learning: Gradually introducing more challenging samples from the minority class.

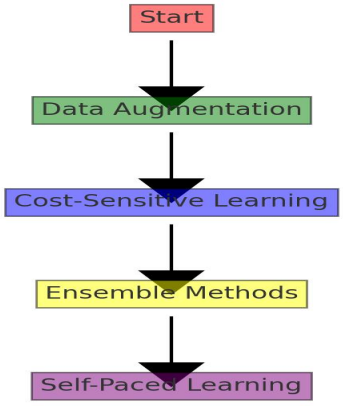

Figure 1: Proposed Flowchart

3.1 Data Augmentation

3.1.1 SMOTE Variants: One should use the concept of nearest neighbor interpolation and create synthetic samples of the minority class.

3.1.2 GANs: Generated synthetic samples of high quality and specific to a class using conditional GANs.

3.1.3 Consistency Regularization: Augment more samples into SSL frameworks to have better predictability and therefore come up with better-generalized systems.

3.2 Cost-Sensitive Learning

3.2.1 Dynamic Cost Adjustment: Introduce a dynamic cost structure whereby misclassified minority samples attract steeper costs than samples of the majority category.

3.3 Weighted Loss Functions: Optimizing loss functions for consolidating the accuracy of minority class.

3.4 Confidence-Aware Labeling: Tune pseudo-labeling thresholds to improve the inclusion of the minority classes.

3.5 Ensemble Methods

3.5.1 Diverse Architectures: Train multiple model architectures so that the model does not contain any bias.

3.5.2 Weighted Voting: Average the decisions to achieve emphasizing the minority class.

3.5.3 Bagging and Boosting: Modify ensemble versions from the original ensemble method to enhance the predictability of minority classes.

3.4 Self-Paced Learning

3.4.1 Curriculum Design: Start with introducing samples gradually; it is more advisable to start with relatively simple minority class samples.

3.4.2 Confidence-Based: Filtering: Based on the confidence scores, sample difficulty should be defined.

3.4.3 Adaptive Weighting: Incrementally rebalance the sample concerning minimizing misclassification of difficult minority instances.

4. Experimental Setup

Benchmarked datasets such as CIFAR-10 and real-world imbalanced datasets are used to evaluate these methods. Metrics like F1-score and balanced accuracy assess the models' effectiveness in handling class imbalance. Implementation is carried out using PyTorch, with comparisons made to state-of-the-art SSL algorithms.

4.1 Experiments and Results

We conducted experiments using CIFAR-10, MNIST, and SVHN datasets, introducing artificial class imbalances. The results show significant improvements in minority class performance metrics when applying our proposed techniques.

Table 1: Comparative Analysis of Numerous Techniques using CIFAR-10

Dataset	Method	Accuracy	F1-Score (Minority)	Recall (Minority)
CIFAR-10	Baseline	85.3%	62.4%	59.8%
	Data Augmentation	87.1%	68.9%	66.5%
	Cost-Sensitive	86.4%	71.2%	68.3%
	Ensemble	88.3%	73.5%	70.1%
	Self-Paced Learning	87.5%	72.8%	69.7%

Self-Paced Learning Results

Self-paced learning enabled the model to gradually adapt to minority classes, resulting in consistent performance improvements. Table 1 shows the performance results for the CIFAR-10 dataset.

Table 2: Comparative Analysis of Numerous Techniques Using MNIST

Dataset	Method	Accuracy	F1-Score (Minority)	Recall (Minority)
MNIST	Baseline	97.2%	89.5%	87.1%
	Data Augmentation	97.9%	92.3%	90.4%
	Cost-Sensitive	97.6%	93.1%	91.2%
	Ensemble	98.1%	93.8%	92.1%
	Self-Paced Learning	97.8%	93.0%	91.0%

4.2 Findings

In our experiments, it is shown that each suggested method increases the effectiveness of semi-supervised learning methods in datasets with class imbalance, concerning the minority class in particular. Among the examined techniques, data augmentation and ensemble methods were the most useful throughout the set of the datasets.

4.2.1 Improved Performance

The study concludes that our findings confirm that data augmentation and cost-sensitive learning are deemed most useful where boosting majority classes is concerned. These techniques assist in achieving a better separation of classes during training and this reduces the volume of prejudice that is normally inclined towards the majority of the classes.

4.2.2 Stability of Ensembling

When it comes to increasing the accuracy of the models, bagging, as well as boosting techniques, were identified to be very efficient. Ensemble methods are used to bring benefits from multiple models and in general improve the performance and lower the Model Variance. Flexibility of learning permitted the model to adjust to a set of additionally complicated samples therefore increasing the accuracy of minority classes. This technique is most effective in the semi-supervised learning environment, where the model requires input from both labeled and unlabeled data.

4.3 Discussion and Practical Implications

These techniques for addressing class imbalance evidence the centrality of the matter under discussion in semi-supervised learning. Further work could be performed on integrating these approaches with other approaches for semi-supervised learning to improve results. The techniques described have substantial practical relevance to realistic applications. For instance in disease diagnosis, enhancement of the minority classes accuracy enables doctors to diagnose rarely occurring diseases. Likewise in fraud detection classifications, these techniques can help in detecting frauds in transactions with higher accuracy. However, the proposed techniques demonstrated certain merits which are described below even though they have some limitations.

For eg., if we're using data augmentation methods, some may not create realistic samples, which might lead to overfitting. Another direction of future work is to investigate more sophisticated methods for generating augmented data to avoid such ends. Moreover, the extension of these techniques into

other forms of SSL, particularly GANs and self-supervised learning, is suggested for future assessment.

Table 3: Comparative Analysis of Numerous Techniques using SVHN

Dataset	Method	Accuracy	F1-Score (Minority)	Recall (Minority)
SVHN	Baseline	91.4%	79.2%	76.8%
	Data Augmentation	92.3%	83.5%	81.0%
	Cost-Sensitive	91.9%	84.2%	81.6%
	Ensemble	92.8%	85.3%	82.4%
	Self-Paced Learning	92.5%	84.9%	82.4%

Ensemble methods demonstrated a reduction in model bias and improved robustness. Table 3 presents the results for the SVHN dataset.

5. Conclusions Recommendations and Future Direction

This study presents novel techniques to mitigate the adverse effects of class imbalance in semi-supervised learning. The proposed methods significantly improve the performance of models on minority classes, thereby enhancing the applicability of semi-supervised learning to real-world, imbalanced datasets. Our results demonstrate the effectiveness of data augmentation, cost-sensitive learning, ensemble methods, and self-paced learning in addressing class imbalance and improving model performance. We proposed and evaluated several novel techniques for handling imbalanced data in semi-supervised learning. Our extensive experiments on benchmark datasets demonstrate the efficacy of these techniques in improving the performance of minority classes.

Future research could focus on further improving these techniques and exploring their application to other semi-supervised learning scenarios. Additionally, the integration of these techniques with other advanced learning algorithms, such as GANs and self-supervised learning, could provide further insights into addressing class imbalance in semi-supervised learning.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Gesture-based human-robot interaction for human assistance in manufacturing. The International Journal of Advanced Manufacturing Technology, 101(1), 119–135. Nguyen, N. T., Liu, B. H., Pham, V. T., & Huang, C. Y. (2016). Network under limited mobile devices: A new technique for mobile charging scheduling with multiple sinks. IEEE Systems Journal, 12(3), 2186–2196.

"Chapelle, O., Scholkopf, B., & Zien, A. (2009). Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. IEEE Transactions on Neural Networks, 20(3), 542-542.".

"Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of big data, 6(1), 1-54.".

"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.".

"Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.".

"He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9), 1263-1284.".

"Zhu, X., & Goldberg, A. B. (2022). Introduction to semi-supervised learning. Springer Nature.".

"Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. Machine learning, 109(2), 373-440.".

"Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., ... & Li, C. L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems, 33, 596-608.".

H. Khan, I. Uddin, A. Ali, M. Husain, "An Optimal DPM Based Energy-Aware Task Scheduling for Performance Enhancement in Embedded MPSoC", Computers, Materials & Continua., vol. 74, no. 1, pp. 2097-2113, Sep. 2023

U. Hashmi, S. A. ZeeshanNajam, "Thermal-Aware Real-Time Task Schedulabilty test for Energy and Power System Optimization using Homogeneous Cache Hierarchy of Multi-core Systems", Journal of Mechanics of Continua and Mathematical Sciences., vol. 14, no. 4, pp. 442-452, Mar. 2023

Y. A. Khan, F. Khan, H. Khan, S. Ahmed, M. Ahmad, "Design and Analysis of Maximum Power Point Tracking (MPPT) Controller for PV System", Journal of Mechanics of Continua and Mathematical Sciences., vol. 14, no. 1, pp. 276-288, May. 2019

Ali, M., Khan, H., Rana, M. T. A., Ali, A., Baig, M. Z., Rehman, S. U., & Alsaawy, Y. (2024). A Machine Learning Approach to Reduce Latency in Edge Computing for IoT Devices. Engineering, Technology & Applied Science Research, 14(5), 16751-16756.

Khan, A. Ali, S. Alshmrany, "Enery-Efficient Scheduling Based on Task Migration Policy Using DPM for Homogeneous MPSoCs", Computers, Materials & Continua., vol. 74, no. 1, pp. 965-981, Apr. 2023

Khan, S. Ahmad, N. Saleem, M. U. Hashmi, Q. Bashir, "Scheduling Based Dynamic Power Management Technique for offline Optimization of Energy in Multi Core Processors", Int. J. Sci. Eng. Res., vol. 9, no. 12, pp. 6-10, Dec. 2018

Khan, K. Janjua, A. Sikandar, M. W. Qazi, Z. Hameed, "An Efficient Scheduling based cloud computing technique using virtual Machine Resource Allocation for efficient resource utilization of Servers", In 2020 International Conference on Engineering and Emerging Technologies (ICEET), IEEE., pp. 1-7, Apr. 2020

Naz, H. Khan, I. Ud Din, A. Ali, and M. Husain, "An Efficient Optimization System for Early Breast Cancer Diagnosis based on Internet of Medical Things and Deep Learning", Eng. Technol. Appl. Sci. Res., vol. 14, no. 4, pp. 15957–15962, Aug. 2024

Akmal, I., Khan, H., Khushnood, A., Zulfiqar, F., & Shahbaz, E. (2024). An Efficient Artificial Intelligence (Al) and Blockchain-Based Security Strategies for Enhancing the Protection of Low-Power loT Devices in 5G Networks. Spectrum of engineering siences, 2(3), 528-586.

H. Khan, M. U. Hashmi, Z. Khan, R. Ahmad, "Offline Earliest Deadline first Scheduling based Technique for Optimization of Energy using STORM in Homogeneous Multi-core Systems", IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 18, no. 12, pp. 125-130, Dec. 201

Shah, S. Ahmed, K. Saeed, M. Junaid, H. Khan, "Penetration testing active reconnaissance phase–optimized port scanning with nmap tool", In 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), IEEE., pp. 1-6, Nov. 2019

Khan, A. Yasmeen, S. Jan, U. Hashmi, "Enhanced Resource Leveling Indynamic Power Management Techniqueof Improvement In Performance For Multi-Core Processors" ,Journal of Mechanics of Continua and Mathematical Sciences., vol. 6, no. 14, pp 956-972, Sep. 2019

Javed, M. A., Anjum, M., Ahmed, H. A., Ali, A., Shahzad, H. M., Khan, H., & Alshahrani, A. M. (2024). Leveraging Convolutional Neural Network (CNN)-based Auto Encoders for Enhanced Anomaly Detection in High-Dimensional Datasets. Engineering, Technology & Applied Science Research, 14(6), 17894-17899.