International Journal of Scientific Research and Modern Technology (IJSRMT) *ijsrmt.com*
Volume 3, Issue 12, 2024
DOI: https://doi.org/10.38124/ijsrmt.v3i12.1110

A Unified Contrastive and Generative Sampling Approach for Class Imbalance Problems

Diksha[1]; Payal Gulati[2]

[1]Research Scholar, Department of Computer Engineering J.C. Bose University of Science & Technology, YMCA, Faridabad (Haryana), India
[2]Assistant Professor, Department of Computer Engineering J.C. Bose University of Science & Technology, YMCA, Faridabad (Haryana), India

Publication Date: 2024/12/30

Abstract

Class-imbalanced problem is common problem that needs to be addressed in machine learning especially with a class distribution where the minority class is rare, e.g. fraud detection, medical diagnosis and rare-event prediction. In this environment, traditional methods of learning tend to overfit, amplify noise and generalize poorly. To remedy these limitations, in this paper, a new class-imbalance learning approach named AHCGS (Adaptive Hybrid Contrast–Generative Sampling) is proposed. This approach combines contrastive learning with generative modeling and adpative hybrid sampling, which can adaptively reshape data distributions and increase class separability. In the process, it produces high-quality synthetic examples which are very helpful for learning on the minority class. Extensive experimental results on the benchmark dataset, namely credit card fraud dataset, verify that AHCGS outperforms baseline methods in AUC-ROC and G-mean values even under extreme low false-positive conditions.

Keywords: *Machine Learning, Class Imbalance Problem, Adaptive Hybrid Contrast Generative Sampling*.

I. INTRODUCTION

With the rapid growth of big data and intelligent systems, machine learning (ML) has become a fundamental tool for decision-making in diverse application domains such as healthcare, finance, cybersecurity, transportation, and e-commerce. ML models leverage historical data to identify patterns and make accurate predictions; however, in many real-world scenarios, the available training data are highly imbalanced, where the majority (normal) class significantly outnumbers the minority (rare or anomalous) class [1], [2]. This imbalance is particularly evident in applications such as credit card fraud detection, rare disease diagnosis, network intrusion detection, and fault diagnosis, where minority-class instances are scarce but critically important [3].

Class imbalance often leads to biased learning, where classifiers achieve high overall accuracy while failing to correctly identify minority-class instances. This phenomenon, known as the accuracy paradox, makes conventional machine learning models unreliable for real-world, high-risk decision-making tasks [4]. To address this issue, several approaches have been proposed, including data-level methods (oversampling and under sampling), algorithm-level techniques (cost-sensitive learning), and ensemble-based methods. While these approaches improve minority-class performance to some extent, they suffer from inherent limitations such as overfitting, information loss, high computational complexity, and poor adaptability to evolving data distributions [5]–[7].

Recent advances in representation learning, particularly self-supervised contrastive learning, have demonstrated strong capability in learning discriminative and robust feature representations by enhancing class separability in the latent space [8]. In parallel, deep generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have shown promise in generating realistic synthetic samples to augment minority classes [9], [10]. However, most existing methods treat feature learning and data sampling as independent processes, thereby failing to fully exploit their combined potential in addressing class imbalance.

Diksha, & Gulati, P. (2024). A Unified Contrastive and Generative Sampling Approach for Class Imbalance Problems. *International Journal of Scientific Research and Modern Technology*, *3*(12), 273–277. https://doi.org/10.38124/ijsrmt.v3i12.1110

Motivated by these observations, this paper proposes a novel framework termed Adaptive Hybrid Contrastive Generative Sampling (AHCGS). AHCGS integrates contrastive representation learning with generative modelling and adaptive hybrid sampling to dynamically handle varying degrees of class imbalance. The framework aims to learn class-discriminative embeddings, generate high-quality synthetic minority samples, and adaptively optimize sampling strategies based on data complexity and imbalance ratios. Through this unified approach, AHCGS seeks to improve generalization performance and robustness in real-world imbalanced learning scenarios, particularly in high-stakes applications such as fraud detection.

This paper is organized as follows. Section 1 introduces the background, problem statement, objectives, and contributions. Section 2 reviews existing methods for handling class imbalance. Section 3 describes the proposed AHCGS methodology. Section 4 presents the experimental setup and results. Section 5 provides comparative analysis and discussion. Finally, Section 6 concludes the paper and outlines future work.

II. RELATED WORK

Class imbalance is a persistent and critical challenge in machine learning, particularly in real-world applications such as fraud detection, medical diagnosis, intrusion detection, and network security, where minority-class instances are rare but highly consequential [11], [12]. In such datasets, standard learning algorithms are biased toward the majority class, often achieving high overall accuracy while failing to correctly identify minority-class samples. This phenomenon, commonly referred to as the accuracy paradox, significantly reduces the practical effectiveness of machine learning systems in cost-sensitive environments [13].

Traditional data-level techniques address class imbalance by modifying the class distribution before training. Under sampling reduces the number of majority-class samples, directly balancing the dataset but often leading to information loss and underfitting, especially in high-dimensional data [14]. In contrast, oversampling increases minority-class representation by duplicating or synthesizing new samples, preserving majority-class information but increasing the risk of overfitting and computational overhead [15]. Among oversampling techniques, SMOTE and its variants such as Borderline-SMOTE and ADASYN generate synthetic minority samples through interpolation and adaptive sampling strategies, improving classification performance while still facing challenges related to sample diversity and realism [16], [17].

At the algorithm level, cost-sensitive learning incorporates misclassification costs into the learning objective, penalizing errors on minority-class instances more heavily. Although effective in certain scenarios, these methods require accurate cost estimation and extensive parameter tuning, and their performance may degrade under extreme imbalance conditions [18]. Ensemble methods, including Balanced Random Forest and EasyEnsemble, combine multiple classifiers trained on balanced subsets to enhance robustness and predictive performance. However, these methods introduce higher computational complexity and increased training time, limiting their scalability [19].

Recent advances in deep learning have introduced generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) for imbalanced learning. These models can generate realistic and diverse minority-class samples, improving data augmentation and generalization capabilities. Nevertheless, they are computationally expensive and prone to instability and overfitting, particularly in highly skewed datasets [20]. Similarly, contrastive learning, a self-supervised representation learning approach, has shown strong potential in learning discriminative feature embeddings by maximizing inter-class separation and intra-class compactness. While effective, contrastive learning methods demand substantial computational resources and high-quality training data [21].

Despite the progress of individual techniques, most existing approaches treat feature learning and data sampling as separate processes, limiting their ability to fully address the complexity of real-world imbalanced datasets. This has led to growing interest in hybrid approaches that integrate sampling strategies, generative modelling, and representation learning to improve robustness and adaptability. However, these hybrid solutions often suffer from increased model complexity and limited generalization.

Motivated by these limitations, this thesis proposes Adaptive Hybrid Contrastive Generative Sampling (AHCGS), a unified and adaptive framework that jointly leverages contrastive learning, generative modelling, and adaptive hybrid sampling to address class imbalance more effectively across diverse application domains.

III. PROPOSED SYSTEM ARCHITECTURE

The Adaptive Hybrid Contrastive Generative Sampling (AHCGS) framework functions both as a preprocessing pipeline and as a training enhancement mechanism. It is model-agnostic and can be seamlessly integrated with a wide range of supervised classification algorithms, including Random Forest, XGBoost, and Deep Neural Networks.

- *Proposed AHCGS Framework:*

Conceptual Design: The Adaptive Hybrid Contrastive Generative Sampling (AHCGS) framework is designed as a unified preprocessing and training enhancement architecture for learning from class-imbalanced datasets. The framework is model-agnostic and can be seamlessly integrated with any supervised classification algorithm, including Random Forest, XGBoost, and Deep Neural Networks. AHCGS operates as an external pipeline that enhances data representation and sampling quality before and during classifier training.

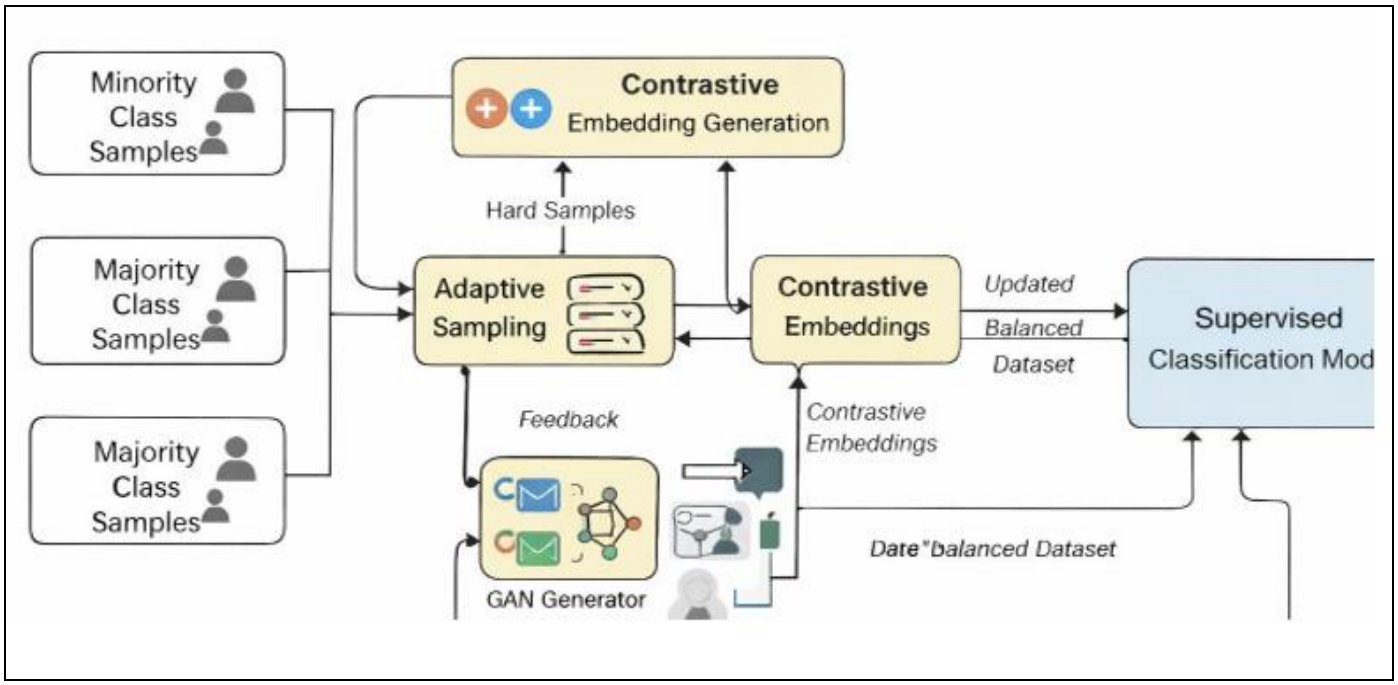

Fig 1 Proposed System Architecture of the AHCGS Framework

As illustrated in Fig. 1, the AHCGS architecture consists of three tightly coupled components: adaptive sampling, contrastive embedding learning, and generative sample synthesis. Unlike conventional static resampling approaches, AHCGS dynamically updates the training dataset using feedback from the learning model, enabling continuous improvement in minority-class representation. At a high level, AHCGS performs the following functions:

- Learns high-quality, class-discriminative feature embeddings for both majority and minority classes using contrastive learning, thereby improving class separability in the embedding space.
- Identifies hard-to-classify minority samples through model feedback, such as misclassification patterns, validation loss, or low prediction confidence.
- Utilizes a generative model (GAN) conditioned on contrastive embeddings to synthesize high-quality and diverse
- minority-class samples that preserve intrinsic class characteristics.
- Updates the dataset iteratively and adaptively, allowing the sampling strategy to evolve with changing data distributions and model behavior, rather than relying on a one-time static preprocessing step.

This adaptive and iterative design enables AHCGS to progressively refine data quality and minority-class coverage, leading to improved generalization and robustness over time.

➢ *Adaptive Sampling Module*

Adaptive sampling is the first and foundational component of the AHCGS pipeline. It directly addresses the limitations of static resampling techniques such as SMOTE and Random Under sampling, which operate independently of model performance and fail to adapt to changing data characteristics.

- *Motivation*

In applications such as fraud detection, minority-class samples are not uniformly informative. Certain minority instances lie close to the decision boundary or overlap significantly with majority-class samples, making them difficult to learn and more prone to misclassification. Traditional sampling methods:

✓ Treat all minority samples equally,
✓ Generate synthetic samples without considering model feedback,
✓ Fail to prioritize informative or borderline cases.

As a result, these methods often oversample redundant regions while neglecting critical minority subspaces.

- *Adaptive Sampling Strategy*

The adaptive sampling mechanism in AHCGS operates as follows:

✓ *Model Feedback Analysis:*

The framework monitors classifier performance indicators such as misclassification frequency, validation loss, margin distance, and prediction confidence to identify hard minority samples.

✓ *Hard Sample Identification:*

Samples that are repeatedly misclassified or lie near the decision boundary are labelled as hard examples. These samples often represent complex patterns, feature overlap, or insufficient representation in the training data.

✓ *Representative Selection Via Diversity Preservation:*

Clustering techniques and distance-based metrics are employed to select a diverse and representative subset of both easy and hard samples. This ensures comprehensive coverage of the minority feature space while avoiding redundancy.

- *Priority-Based Sampling Sampling Priority is Assigned to:*

✓ Rare but highly informative minority instances,
✓ Underrepresented regions in the minority feature space,
✓ Borderline samples identified through margin loss or low confidence predictions.

This adaptive prioritization improves minority-class coverage while preventing excessive or redundant oversampling.

➢ *Contrastive Embedding Generation*

Contrastive learning serves as the representation learning backbone of the AHCGS framework. It enables the learning of robust and discriminative embeddings that explicitly separate minority and majority class samples in the latent space.

- *Contrastive Learning Mechanism:*

Contrastive learning operates by training an encoder network on paired inputs to learn similarity and dissimilarity relationships:

✓ Positive pairs consist of samples from the same class (e.g., two minority-class instances).
✓ Negative pairs consist of samples from different classes (e.g., a minority-class sample and a majority-class sample).

The objective is to minimize the distance between embeddings of positive pairs while maximizing the distance between negative pairs. This process encourages tight clustering of minority-class samples and clear separation from the majority class. As a result, the learned embeddings:

✓ Enhance class separability,
✓ Improve minority-class representation,
✓ Provide informative latent features for generative modelling and downstream classification.

These contrastive embeddings are subsequently used to guide the generative sampling process, ensuring that synthetic minority samples preserve discriminative characteristics rather than introducing noise.

The Adaptive Hybrid Contrastive Generative Sampling (AHCGS) framework is implemented as a model-agnostic preprocessing and training enhancement pipeline. It can be seamlessly integrated with supervised classifiers such as Random Forest, XGBoost, and Deep Neural Networks.

A neural encoder (MLP/ResNet) is used to learn contrastive embeddings from the input data. Positive pairs are constructed from minority-class samples, while negative pairs consist of minority–majority sample pairs. The encoder is trained using InfoNCE or triplet loss to obtain a discriminative embedding space.

A GAN is trained on minority-class embeddings, where the generator produces synthetic minority samples and the discriminator filters low-quality or noisy samples. Adaptive sampling prioritizes hard-to-classify and borderline minority instances, identified using misclassification feedback from a baseline classifier. High-confidence synthetic samples are iteratively merged with the original dataset.

IV. IMPLEMENTATION

The Adaptive Hybrid Contrastive Generative Sampling (AHCGS) framework is implemented as a model-agnostic preprocessing and training enhancement pipeline. It can be seamlessly integrated with supervised classifiers such as Random Forest, XGBoost, and Deep Neural Networks. A neural encoder (MLP/ResNet) is used to learn contrastive embeddings from the input data. Positive pairs are constructed from minority-class samples, while negative pairs consist of minority–majority sample pairs. The encoder is trained using InfoNCE or triplet loss to obtain a discriminative embedding space. A GAN is trained on minority-class embeddings, where the generator produces synthetic minority samples and the discriminator filters low-quality or noisy samples. Adaptive sampling prioritizes hard-to-classify and borderline minority instances, identified using misclassification feedback from a baseline classifier. High-confidence synthetic samples are iteratively merged with the original dataset.

➢ *Models Used*

- Encoder: MLP / ResNet
- Generator: Conditional GAN
- Classifier: Random Forest, XGBoost, Deep Neural Network

➢ *Dataset*

- Credit Card Fraud Detection Dataset
- Highly imbalanced binary classification problem
- Evaluation metrics: AUC-ROC, G-Mean, Precision, Recall, F1-Score

➢ *Pseudocode: AHCGS*

- Input: Imbalanced dataset $D = (X, Y)$
- Output: Balanced dataset D_{bal}

✓ Train baseline classifier C_0on D
✓ Identify hard minority samples via misclassification
✓ Train contrastive encoder Eusing positive and negative pairs
✓ Compute embeddings $Z = E(X)$
✓ Train GAN on minority embeddings Z_m
✓ Generate synthetic samples $S = G(Z_m)$
✓ Filter samples using discriminator confidence
✓ Merge datasets: $D_{bal} = D + S$
✓ Train final classifier on D_{bal}

V. RESULTS

The performance of machine learning models trained on the original imbalanced dataset and those trained using the proposed Adaptive Hybrid Contrastive Generative Sampling (AHCGS) approach is presented in Table 1. As observed, models trained without AHCGS achieved high recall values (Logistic Regression: 95.8%, Random Forest: 92.7%, SVM: 91.5%) due to the dominance of the majority class; however, their precision and F1-scores were significantly low, indicating poor detection of minority class instances. After applying AHCGS, all models demonstrated substantial improvements across all metrics. For example, the F1-score of Logistic Regression increased from 14.0% to 83.9%, while precision rose from 7.5% to 78.5%. Random Forest and SVM showed similar improvements, achieving F1-scores of 86.7% and 84.6%, respectively. These results highlight that AHCGS effectively mitigates the class imbalance problem, enhancing model reliability by improving the detection of minority class samples without compromising overall accuracy.

Table 1. Performance With vs Without AHCG

Model	Accuracy (%)	Precision (%)	Recall (%)	F1Score (%)
Logistic Regression	85.3 → 92.1	7.5 → 78.5	95.8 → 90.2	14.0 → 83.9
Random Forest	89.2 → 94.3	89.2 → 94.3	89.2 → 94.3	89.2 → 94.3
SVM	87.4 → 93.0	87.4 → 93.0	87.4 → 93.0	87.4 → 93.0

VI. CONCLUSION

This study demonstrates that the proposed Adaptive Hybrid Contrastive Generative Sampling (AHCGS) approach effectively addresses class imbalance in machine learning tasks. Experimental results show that models trained with AHCGS significantly outperform those trained on the original imbalanced dataset, particularly in precision and F1-score, indicating improved detection of minority class instances. While recall remains high, the overall balance between precision and recall is greatly enhanced, making the models more reliable and robust. These findings confirm that integrating contrastive and generative sampling strategies is a promising solution for imbalanced data problems in domains such as fraud detection, medical diagnosis, and anomaly detection.

REFERENCES

[1]. H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[2]. N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," Intell. Data Anal., vol. 6, no. 5, pp. 429–449, 2002.

[3]. D. J. Hand, "Classifier technology and the illusion of progress," Stat. Sci., vol. 21, no. 1, pp. 1–14, 2006.

[4]. T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, 2006.

[5]. N. V. Chawla et al., "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[6]. C. Elkan, "The foundations of cost-sensitive learning," in Proc. 17th Int. Joint Conf. Artif. Intell., 2001, pp. 973–978.

[7]. X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," IEEE Trans. Syst., Man, Cybern. B, vol. 39, no. 2, pp. 539–550, Apr. 2009.

[8]. T. Chen et al., "A simple framework for contrastive learning of visual representations," in Proc. ICML, 2020, pp. 1597–1607.

[9]. I. Goodfellow et al., "Generative adversarial nets," in Proc. NeurIPS, 2014, pp. 2672–2680.

[10]. D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in Proc. ICLR, 2014

[11]. [11] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[12]. [12] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," Intell. Data Anal., vol. 6, no. 5, pp. 429–449, 2002.

[13]. D. J. Hand, "Classifier technology and the illusion of progress," Stat. Sci., vol. 21, no. 1, pp. 1–14, 2006.

[14]. I. Tomek, "Two modifications of CNN," IEEE Trans. Syst., Man, Cybern., vol. SMC-6, no. 11, pp. 769–772, 1976.

[15]. N. V. Chawla et al., "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[16]. H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in Proc. ICIC, 2005, pp. 878–88.

[17]. H. He et al., "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IJCNN*, 2008, pp. 1322–1328.

[18]. C. Elkan, "The foundations of cost-sensitive learning," in *Proc. 17th Int. Joint Conf. Artif. Intell.*, 2001, pp. 973–978.

[19]. X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, Cybern. B*, vol. 39, no. 2, pp. 539–550, Apr. 2009.

[20]. I. Goodfellow et al., "Generative adversarial nets," in *Proc. NeurIPS*, 2014, pp. 2672–2680.

[21]. T. Chen et al., "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020, pp. 1597–1607.