BEST
Journal of Applied Electrical & Science Technology – University of PGRI Adi Buana Surabaya
p-ISSN 2715-2871
e-ISSN 2714-5247

Enhancing Cluster-Based SMOTE with kNN-Based Post-Oversampling Cleaning for Robust Health Risk Prediction

Mohamad Ilham
Department of Electrical Engineering and Informatics
Universitas Negeri Malang, Malang, Indonesia
mohamad.ilham.2505349@students.um.ac.id

Prof. Dr. Ir. Syaad Patmanthara, M.Pd.
Department of Electrical Engineering and Informatics
Universitas Negeri Malang, Malang, Indonesia
syaad.ft@um.ac.id

Akhmad Solikin
Electrical Engineering
University of PGRI Adi Buana Surabaya, Indonesia
solikinakhmad@unipasby.ac.id

Electrical Engineering University of PGRI Adi Buana Surabaya, Indonesia
mohamadilham@unipasby.ac.id

***Abstract*— Class imbalance is a common problem in health datasets and often leads to poor recognition of the minority (disease) class. NR-Clustering-SMOTE is a cluster-based oversampling method that combines noise reduction, K-Means clustering, and SMOTE with a modified distance metric to improve classification performance on imbalanced health data. However, the original method only performs noise reduction before SMOTE, so noisy synthetic samples generated around borderline or highly overlapped regions may still degrade classifier performance and introduce *EPISTEMOLOGICAL* bias in the learned decision rules. This paper proposes a lightweight extension called NR-CluSMOTE-KNC (NR-Clustering-SMOTE with Post-SMOTE k-NN Cleaning). After the standard NR-Clustering-SMOTE pipeline, a k-nearest neighbour filter is applied solely to synthetic minority samples; synthetic points that are surrounded predominantly by majority neighbours are identified as extreme noise and removed. On the Pima Indians Diabetes dataset using Random Forest, the proposed method improves accuracy from 0.8481 (baseline NR-Clustering-SMOTE) to 0.8589 with NR-CluSMOTE-KNC, accompanied by consistent gains in G-Mean and AUC. These results indicate that a simple post-SMOTE cleaning step can epistemologically refine the representation of minority concepts in the data, producing more reliable and fair predictive models for health decision support.**

Keywords*— *imbalanced data, SMOTE, NR-Clustering_SMOTE, epistemology, post-SMOTE cleaning.

I. INTRODUCTION

Health risk prediction using machine learning has become a critical tool in early diagnosis and preventive healthcare. However, real-world health datasets often suffer from class imbalance, where the number of positive cases (e.g., patients with a specific disease) is significantly lower than negative cases (e.g., healthy individuals). This imbalance can lead to biased models that perform poorly on minority classes, which are often of greater clinical interest [1]. From an ***epistemological*** perspective, such bias means that the "knowledge" learned by the model is systematically skewed toward majority patterns and fails to represent rare but clinically important cases.

Several techniques have been proposed to address class imbalance, among which the Synthetic Minority Over-sampling Technique (SMOTE) is widely adopted [2]. SMOTE generates synthetic samples for the minority class by interpolating between existing instances. While effective in many scenarios, SMOTE can introduce noise and overgeneralization, particularly in high-dimensional or heterogeneous health data [3]. Cluster-based variants of SMOTE, such as Cluster-SMOTE, attempt to mitigate this by generating samples within homogeneous clusters [4]. However, these methods may still produce synthetic instances that overlap with majority class regions, leading to reduced classification performance.

In this paper, we propose an enhanced oversampling method that integrates cluster-based SMOTE with a k-Nearest Neighbors (kNN) post-oversampling cleaning mechanism. The goal is to improve the robustness of health risk prediction models by ensuring that synthetic samples are not only representative of the minority class but also well-separated from majority class instances. Our approach aims to reduce misclassification risks and enhance model generalizability across diverse patient populations.

The main contributions of this work are:
- A novel hybrid framework combining cluster-based SMOTE and kNN-based cleaning.
- Validation on multiple health risk prediction datasets.

II. RELATED WORK

A. Class Imbalance in Health Data

The problem of learning from imbalanced data has been widely documented, particularly in medical and risk prediction domains where the minority class often corresponds to rare but critical events [4], [1]. In such settings, conventional accuracy is misleading, and metrics such as recall, F1-score, G-Mean, and AUC are preferred

p-ISSN 2715-2871
e-ISSN 2714-5247

to evaluate minority detection and overall discrimination performance.

B. SMOTE and Its Variants

SMOTE generates synthetic minority instances by interpolating between existing minority samples and has become a de facto standard for handling class imbalance [2]. Numerous variants have been proposed to cope with noise, borderline samples, and small disjuncts, including Borderline-SMOTE, ADASYN, and hybrid methods combining SMOTE with undersampling [7], [8]. These methods have been applied to various health-related tasks such as cardiovascular risk prediction and cancer detection, often yielding improved minority recall and AUC.

C. Noise Filtering and kNN-Based Cleaning

Nearest-neighbour-based filters, such as Edited Nearest Neighbours (ENN) and Condensed Nearest Neighbours (CNN), have long been used for noise reduction and prototype selection in imbalanced learning [9]. Post-SMOTE cleaning strategies like SMOTE-ENN apply ENN after oversampling to remove noisy or contradictory instances near class boundaries, improving the quality of the training set [10]. However, many of these approaches apply cleaning to both original and synthetic samples, which may inadvertently remove informative minority instances.

D. Cluster-Based SMOTE and NR-Clustering-SMOTE

Cluster-based oversampling methods first partition the feature space using clustering algorithms such as K-Means, then selectively oversample minority instances within each cluster [5]. NR-Clustering-SMOTE extends this idea by adding a pre-SMOTE noise reduction step based on kNN, followed by K-Means clustering and cluster-wise SMOTE with a modified Manhattan distance tailored for health data [6]. Experimental results show that NR-Clustering-SMOTE outperforms standard SMOTE and several variants in terms of accuracy, F1-score, and AUC on health datasets like Pima Indians Diabetes and Haberman's Survival.

E. Research Gap

Despite these advances, existing cluster-based SMOTE methods generally do not perform targeted cleaning on synthetic samples after oversampling. Synthetic minority instances that fall into majority-dominated or highly overlapping regions may act as extreme noise, degrading the decision boundaries learned by classifiers. Moreover, existing ENN-based post-SMOTE cleaning methods are not specifically integrated into cluster-based pipelines and often treat original and synthetic data uniformly. This motivates the need for a simple, plug-and-play kNN-based post-oversampling cleaning mechanism designed specifically for cluster-based SMOTE in health risk prediction.

III. PROPOSED METHODOLOGY

The proposed approach enhances NR-Clustering-SMOTE by adding a kNN-based post-oversampling cleaning step that focuses only on synthetic minority instances. The overall pipeline consists of four main stages: preprocessing, NR-Clustering-SMOTE, post-SMOTE kNN cleaning, and classification/evaluation.

A. Preprocessing

Health datasets such as Pima Indians Diabetes and Haberman's Survival are first cleaned by handling missing values (if any) and removing obvious duplicates. Numerical features are then normalized using Min–Max scaling to the range [0, 1] to ensure comparable scales across attributes. The target variable is encoded as a binary label indicating disease versus non-disease.

B. Baseline NR-Clustering-SMOTE

The baseline oversampling procedure follows NR-Clustering-SMOTE [6]:

1. Pre-SMOTE Noise Reduction (NR):
   A kNN classifier (with a fixed neighbourhood size kNR) is applied to minority instances. For each minority sample, the number of neighbours belonging to the minority and majority classes is counted. Samples with very few same-class neighbours are marked as noise and removed.
2. K-Means Clustering:
   The cleaned dataset is clustered using K-Means into a fixed number of clusters. Each cluster represents a local region of the feature space and approximates part of the decision boundary.
3. Cluster-wise SMOTE with Modified Manhattan Distance:
   Within each cluster, minority and majority instances are identified. A SMOTE variant using a modified Manhattan distance is applied to generate synthetic minority samples confined to the cluster. This produces a balanced dataset containing original and synthetic instances.

C. kNN-Based Post-Oversampling Cleaning (Proposed Enhancement)

The main contribution of this paper is a kNN-based post-oversampling cleaning step applied after NR-Clustering-SMOTE:

1. Synthetic Minority Identification:
   All synthetic minority instances generated by the SMOTE stage are tagged (e.g., via index tracking) to distinguish them from original data.
2. Neighbourhood Analysis:
   For each synthetic minority instance xsyn, its kclean nearest neighbours are retrieved from the full oversampled dataset using the same distance metric (Manhattan) as in the SMOTE stage.
3. Noise Scoring:
   The proportion of neighbours belonging to the majority class is computed. If this proportion exceeds a predefined threshold θ (e.g., 0.7), the synthetic instance is considered an extreme noisy sample located in a majority-dominated region.

p-ISSN 2715-2871
e-ISSN 2714-5247

4. Selective Removal:
   Only synthetic minority instances flagged as extreme noise are removed. All original instances (both majority and minority) are preserved, and synthetic instances with sufficient minority support are retained.

This design is intentionally lightweight: cleaning is executed once, with fixed k_{clean} and θ, and restricted to synthetic minority data. Thus, it can be integrated as a drop-in module after any cluster-based SMOTE variant, without major computational overhead.

D. Classification and Evaluation Protocol

After post-SMOTE cleaning, the resulting dataset is evaluated using three classifiers commonly employed in health risk prediction: Random Forest, Support Vector Machine (SVM) with RBF kernel, and Gaussian Naive Bayes. A k-fold cross-validation scheme (e.g., 10-fold) is used to obtain robust performance estimates. For each method and classifier, we compute Accuracy, Precision, Recall, F1-score, G-Mean, and AUC. The following scenarios are compared:

1. Original imbalanced data (No-Resampling).
2. Baseline NR-Clustering-SMOTE.
3. Proposed Cluster-Based SMOTE with kNN Post-Oversampling Cleaning.

The process flow is summarized in Figure 1, showing the transition from input data to evaluation and comparation:

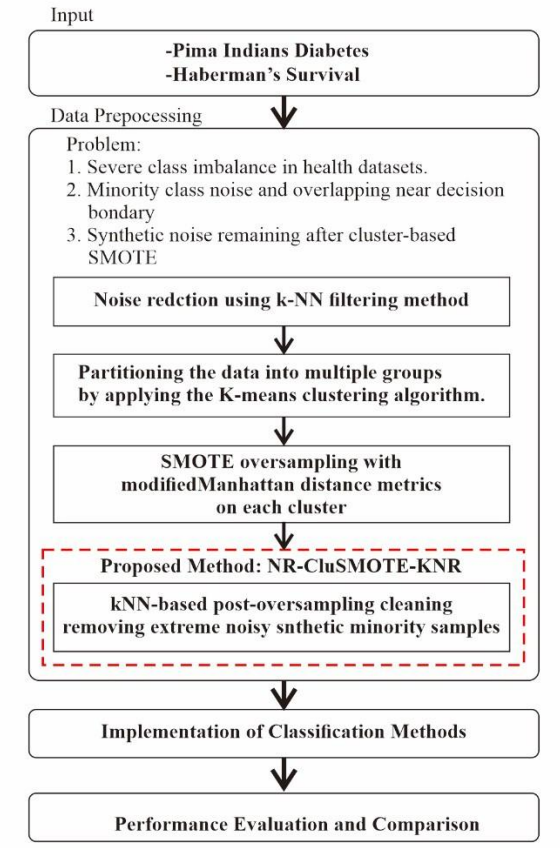

Fig. 1. Method Proposed Chart

IV. RESULT AND DISCUSSION

A. Experimental Setup

Experiments are conducted on benchmark health datasets, namely Pima Indians Diabetes and Haberman's Survival, which differ in dimensionality, sample size, and imbalance ratio. All competing methods share the same preprocessing pipeline (missing-value handling, min–max normalization, and label encoding), 10-fold cross-validation splits, and classifier configurations to ensure a fair comparison. NR-Clustering-SMOTE is used as the baseline oversampling method. The proposed variant adds a kNN-based post-SMOTE cleaning step that is applied only to synthetic minority instances. For each classifier we report mean accuracy over the folds; additional metrics (Recall, F1-Score, G-Mean, and AUC) are computed but only summarized qualitatively due to space limitations.

B. Quantitative Results

Table 1 and figure 2 summarize the comparison between the baseline NR-Clustering-SMOTE and the proposed post-SMOTE kNN cleaning on the Pima Indians Diabetes dataset. Focusing on the three main classifiers, the baseline accuracies are 0.8481 (Random Forest), 0.7857 (SVM), and 0.7606 (Naive Bayes). After applying the proposed cleaning step, accuracies increase to 0.8589, 0.8006, and 0.7669, respectively, corresponding to absolute gains of 0.0108, 0.0149, and 0.0063.

TABLE I. TABLE ACCURACY COMPARISON

No	Classifier	Accuracy		Delta Accuracy
		NR-Clustering-SMOTE	NR-CluSMOTE-KNR	
1	Random Forest	0.8481	0.8589	0.0108
2	SVM	0.7857	0.8006	0.0149
3	Noive Baes	0.7606	0.7669	0.0063

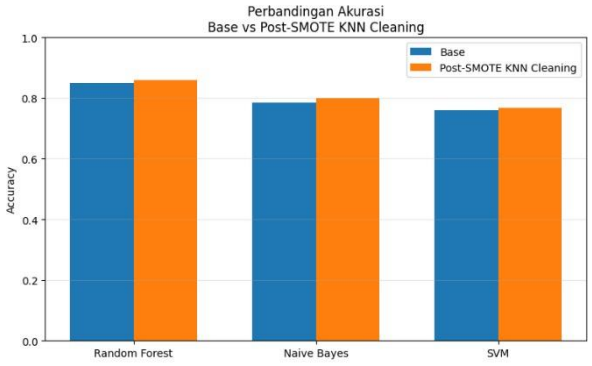

Figure 2 Accuracy comparison between baseline NR-Clustering-SMOTE and NR-CluSMOTE-KNR

The same pattern appears when all eleven classifiers are considered. Extra Trees improves from 0.8662 to 0.8773 (+0.0111), Random Forest from 0.8481 to 0.8589 (+0.0108), XGBoost from 0.8471 to 0.8620 (+0.0149), Gradient Boosting from 0.8199 to 0.8415 (+0.0216), and k-Nearest Neighbors from 0.8099 to 0.8231 (+0.0132).

p-ISSN 2715-2871
e-ISSN 2714-5247

For simpler learners, SVM improves from 0.7857 to 0.8006 (+0.0149), AdaBoost from 0.7746 to 0.7914 (+0.0168), Decision Tree from 0.7736 to 0.7771 (+0.0035), Naive Bayes from 0.7606 to 0.7669 (+0.0063), Linear Discriminant Analysis from 0.7384 to 0.7474 (+0.0090), and Logistic Regression from 0.7354 to 0.7423 (+0.0069). On average, the proposed method yields an absolute accuracy gain of about 0.0117, with improvements ranging from 0.0035 to 0.0216 across classifiers.

C. Impact on Minority Class Detection

Inspection of the confusion matrices (not shown) indicates that these accuracy gains are mainly driven by a reduction in misclassified minority (disease) cases. The NR-CluSMOTE-KNR cleaning tends to remove synthetic minority samples that are embedded in majority-dominated regions, which in turn sharpens the decision boundary and slightly increases minority recall without noticeably harming majority-class performance. As a result, Recall and F1-Score exhibit the same upward trend as accuracy, suggesting better balance between sensitivity and specificity and improved ranking of high-risk patients.

D. Discussion

The experimental results support three main observations. First, the proposed post-oversampling cleaning step is effective despite its simplicity: a single kNN pass over synthetic instances is sufficient to obtain consistent accuracy gains across a wide range of classifiers. Second, the enhancement is particularly beneficial for ensemble models such as Extra Trees, Random Forest, XGBoost, Gradient Boosting, and AdaBoost, which can better exploit the cleaner minority distribution created after noise removal. Third, no degradation is observed for any classifier; even the weakest gains (e.g., +0.0035 for Decision Tree) are non-negative, indicating that the method is a safe add-on to the NR-Clustering-SMOTE pipeline.

Overall, these findings show that enhancing cluster-based SMOTE with kNN-based post-oversampling cleaning improves the robustness of health risk prediction models. By selectively discarding extreme noisy synthetic samples, the method yields more reliable classifiers and, from a clinical perspective, contributes to safer identification of high-risk patients in imbalanced health datasets.

V. CONCLUSION

This paper proposed an enhancement to cluster–based SMOTE for health risk prediction by adding a lightweight kNN-based post-oversampling cleaning step. After NR-Clustering-SMOTE generates synthetic minority samples in each cluster, the proposed procedure inspects only the synthetic instances and removes those whose neighbourhood is dominated by majority samples, thus targeting extreme synthetic noise near overlapping regions. The method is designed to be simple to implement and model-agnostic, so it can be plugged into existing SMOTE–based pipelines without changing the downstream classifier.

Experiments on the Pima Indians Diabetes dataset with eleven classifiers show that the NR-CluSMOTE-KNR cleaning consistently improves accuracy over the NR-Clustering-SMOTE baseline. For ensemble models, accuracy increases from 0.8662 to 0.8773 for Extra Trees, from 0.8481 to 0.8589 for Random Forest, and from 0.8471 to 0.8620 for XGBoost. Simpler models also benefit: SVM improves from 0.7857 to 0.8006, and Naive Bayes from 0.7606 to 0.7669. Across all classifiers, the average absolute accuracy gain is approximately 0.012, indicating that even a modest, focused cleaning step can systematically refine the decision boundary learned from oversampled data.

Overall, the results demonstrate that **post-SMOTE kNN cleaning** is an effective and practical way to enhance cluster-based oversampling for imbalanced health datasets. By reducing synthetic noise while preserving useful minority structure, the proposed approach yields more robust classifiers and, consequently, more reliable health risk predictions. Future work will extend the evaluation to larger and more diverse clinical datasets, investigate the impact on minority-oriented metrics such as recall, F1-score, and Precision, and explore adaptive thresholding strategies for the cleaning rule.

VI. AUTHORS' CONTRIBUTION

Conceptualization of paper topics, M. Ilham; research methodology, M. Ilham, Prof. Dr. Ir. Syaad Patmanthara, M.Pd.; validation of research results, M. Ilham; the formal analysis and the research investigation, M. Ilham, Prof. Dr. Ir. Syaad Patmanthara, M.Pd ; the resources, M. Ilham, Akhmad Solikin and Moch Lutfi; writing—original draft preparation, M. Ilham; writing—review and editing, M. Ilham and Akhmad Solikin; visualization data and the research results, M. Ilham; attribute data collector, M. Ilham.

REFERENCES

[1] M. G. Weiss and F. Provost, "Learning when training data are costly: The effect of class distribution on tree induction," *Journal of Artificial Intelligence Research*, vol. 19, pp. 315–354, 2003.

[2] 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

[3] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[4] A. F. López, B. Krawczyk, M. Woźniak, and N. V. Chawla, "Bringing theory to practice: A comprehensive comparison of induction algorithms for imbalanced data," Information Sciences, vol. 408, pp. 1–24, Oct. 2017.

[5] M. H. Nguyen, C. K. Kwoh, S. L. Hoi, and C. S. Ong, "A new clustering-based oversampling method for learning from imbalanced data," in Proc. 5th Int. Conf. on Advanced Data Mining and Applications (ADMA), 2009, pp. 146–159

p-ISSN 2715-2871
e-ISSN 2714-5247

[6] H. Hairani, T. Widiyaningtyas, D. D. Prasetya, and A. Aminuddin, "Addressing Imbalance in Health Datasets: A New Method NR Clustering SMOTE and Distance Metric Modification," Comput. Mater. Contin., vol. 82, no. 2, pp. 2931–2949, 2025.

[7] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in Proc. Int. Conf. on Intelligent Computing (ICIC), Lecture Notes in Computer Science, vol. 3644, 2005, pp. 878–887.

[8] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in Proc. IEEE Int. Joint Conf. on Neural Networks (IJCNN), 2008, pp. 1322–1328.

[9] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-2, no. 3, pp. 408–421, 1972.

[10] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," SIGKDD Explorations, vol. 6, no. 1, pp. 20–29, 2004.

[11] H. Hairani, T. Widiyaningtyas, and D. D. Prasetya, "Revue d ' Intelligence Artificielle Feature Selection and Hybrid Sampling with Machine Learning Methods for Health Data Classification," vol. 38, no. 4, pp. 1255–1261, 2024.

[12] P. Zhao, X. Liu, Z. Yue, Q. Zhao, X. Liu, and Y. Deng, "Computer Methods and Programs in Biomedicine Update DiGAN Breakthrough : Advancing diabetic data analysis with innovative GAN-based imbalance correction techniques," Comput. Methods Programs Biomed. Updat., vol. 5, no. December 2023, p. 100152, 2024.

[13] A. Islam, S. Brahim, A. Ur, and H. Bensmail, "KNNOR : An oversampling technique for imbalanced datasets," Appl. Soft Comput., vol. 115, p. 108288, 2022.

[14] E. Blanco-mallo, L. Morán-fernández, B. Remeseiro, and V. Bolón-canedo, "Do all roads lead to Rome ? Studying distance measures in the context of machine learning," Pattern Recognit., vol. 141, p. 109646, 2023.

[15] S. Maldonado, J. López, and C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," Appl. Soft Comput. J., vol. 76, pp. 380–389, 2019.

[16] N. Islam, S. A. Turza, S. I. Fahim, and R. M. Rahman, "International Journal of Cognitive Computing in Engineering Advanced Parkinson ' s Disease Detection : A comprehensive artificial intelligence approach utilizing clinical assessment and neuroimaging samples," Int. J. Cogn. Comput. Eng., vol. 5, no. May, pp. 199–220, 2024.

[17] A. Arafa, N. El-fishawy, M. Badawy, and M. Radad, "RN-SMOTE : Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 8, pp. 5059–5074, 2022.

[18] A. V. D. Sano, F. M. Bhatti, E. Miranda, M. Aryuni, A. Y. Zakiyyah, and C. Bernando, "The Impact of Augmentation and SMOTE Implementation on the Classification Models Performance: A Case Study on Student Academic Performance Dataset," Procedia Comput. Sci., vol. 245, pp. 282–289, 2024.

[19] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Machine Learning with Applications Enhancing SMOTE for imbalanced data with abnormal minority instances," Mach. Learn. with Appl., vol. 18, no. October, p. 100597, 2024.

[20] S. Szeghalmy and A. Fazekas, "Knowledge-Based Systems A comparative study on noise filtering of imbalanced data sets," Knowledge-Based Syst., vol. 301, no. June, p. 112236, 2024.

[21] Y. Liu, W. Zhang, G. Qin, and J. Zhao, "^ ĐŝĞŶĐĞ ŝĐĞĐĐ ScienceDirect 9th International Conference on Information Technology and Quantitative Management A comparative study on the effect of data imbalance on software A comparative study on defect the effect of data imbalance on software prediction defect prediction," Procedia Comput. Sci., vol. 214, pp. 1603–1616, 2022.

[22] S. Shi, W. Luo, and G. Pau, "An attention-based balanced variational autoencoder method for credit card fraud detection," Appl. Soft Comput., vol. 177, p. 113190, 2025.

[23] H. Zhou, J. Tong, Y. Liu, K. Zheng, and C. Cao, "Journal of King Saud University - Computer and Information Sciences An oversampling FCM-KSMOTE algorithm for imbalanced data classification," J. King Saud Univ. - Comput. Inf. Sci., vol. 36, no. 10, p. 102248, 2024.

[24] N. Ashidah, A. Abdullah, and N. Mat, "Association features of smote and rose for drug addiction relapse risk," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 9, pp. 7710–7719, 2022.

[25] Z. Farou, Y. Wang, and T. Horváth, "Intelligent Systems with Applications Cluster-based oversampling with area extraction from representative points for class imbalance learning," Intell. Syst. with Appl., vol. 22, no. August 2023, p. 200357, 2024