JOIV : Int. J. Inform. Visualization, 8(3) - September 2024 1310-1318

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Addressing Class Imbalance of Health Data: a Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies

Hairani Hairani [a,b], Triyanna Widiyaningtyas [a,*], Didik Dwi Prasetya [a]

[a] *Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia*
[b] *Department of Computer Science, Universitas Bumigora, Mataram, Indonesia*
Corresponding author: [*]triyannaw.ft@um.ac.id

***Abstract*— The Synthetic Minority Oversampling Technique (SMOTE) method is the baseline for solving unbalanced data problems. The working concept of the SMOTE method is to generate new synthetic data patterns by performing linear interpolation between minority class samples based on k-nearest neighbors. However, the SMOTE method has weaknesses, namely the problem of overgeneralization due to excessive sampling of sample noise and increased overlapping between classes in the decision boundary area, which has the potential for noise data. Based on the weaknesses of the Smote method, the purpose of this research is to conduct a systematic literature review on the Smote method modification approach in solving unbalanced data. This systematic literature review method comprises keyword identification, article search process, determination of selection criteria, and selection results based on criteria. The results of this study showed that the SMOTE modification approach was based on filtering, clustering, and distance modification to reduce the resulting noise data. The filtering approach removed the noise data before SMOTE, positively impacting resolving unbalanced data. Meanwhile, the use of a clustering approach in SMOTE can minimize the overlapping artificial minority data that has noise potential. The most used datasets are Pima 60% and Haberman 50%. The most used performance evaluation on unbalanced data is f1-measure 57%, accuracy 55%, recall 43%, and AUC 27%. The implication of the results of this literature review is to provide opportunities for further research in modifying SMOTE in addressing health data imbalances, especially handling noise and overlapping data. The thoroughness of our literature review should instill confidence in the research community.**

***Keywords*— Class imbalance; modified smote; health data; systematic literature review.**

Manuscript received 26 Oct. 2023; revised 18 Feb. 2024; accepted 2 Mar. 2024. Date of publication 30 Sep. 2024.
International Journal on Informatics Visualization is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.

I. INTRODUCTION

Machine learning methods rely heavily on datasets. Machine learning methods can obtain optimal performance when processing quality datasets without noise, outliers, missing values, and unbalanced data. Most real datasets have imbalanced data problems, such as health [1], [2], [3], education [4], and software defects [5]. Unbalanced data is when the number one class exceeds the other. The data imbalance problem occurs when one class is underrepresented (minority class) while the other class is already represented in the data (majority class) [6]. The data collection process can cause data imbalance. The problem of unbalanced data makes it difficult for classification methods to predict minority classes compared to majority classes due to the lack of representation of minority classes compared to majority classes, making it difficult to generalize. Unbalanced data can degrade the performance of classification methods and cause overfitting. In other words, predictive methods can produce high accuracy on majority data and low accuracy on minority data [7]. Whether a dataset is balanced or not can be determined by measuring the imbalance ratio. If the ratio is higher, it is more likely that the data is unbalanced. Unbalanced data also needs to be looked at in terms of its complexity, such as noise level [8], small disjunct [9], boundary area, and outlier area. Noise is a single instance of the minority class in the majority class area. Small disjunct is a collection of minority class groups in the majority class area, while outliers are foreign data far from the dataset.

Handling unbalanced data problems can use two approaches, namely the data level approach and the

algorithm level approach [10]. This research focuses on solving the unbalanced data problem using the data-level approach. The data level approach modifies the data by rebalancing the minority and majority classes [11]. Some data-level approaches that can be used are oversampling, undersampling, and hybrid (combination of both) [12], [13]. Oversampling works by balancing the minority class by duplicating the same minority class, resulting in overfitting. Under-sampling balances the minority class by removing the majority class until the distribution is balanced. The weakness of under-sampling is that it loses a lot of helpful data. In contrast, the hybrid sampling method works by adding minority classes and deleting majority class data until the class distribution is balanced. Solving the problem of unbalanced data usually uses the SMOTE method. The SMOTE method generates new synthetic data patterns by performing linear interpolation between minority class samples based on their k-nearest neighbors [6]. Some previous studies have successfully used the SMOTE method to solve unbalanced data [14]. Using the average SMOTE oversampling method can improve performance in predictive methods [15]–[17]. Also, using the number of k values to find the nearest neighbor in synthetic new data affects the level of performance of different predictive models [18], [19].

However, the SMOTE method has disadvantages, namely 1) an overgeneralization problem due to excessive sampling of sample noise [20] and 2) increased overlapping between classes around the decision boundary [21]. Not only that, but SMOTE also generates synthetic data that can cause sample noise, thus causing overfitting [22], [23], ignoring the small disjunct problem [24]. Therefore, several previous studies focus on overcoming the weaknesses of the SMOTE method using several approaches such as filtering, clustering, and distance-based. Previous research uses clustering approaches to improve the weakness of the SMOTE method, such as using the K-Means method [25], the DBSCAN method [26], research that uses the distance approach [27]–[29], and the filtering approaches in improving the SMOTE method, which is ASN-SMOTE [30] and SMOTE-LOF [31].

This research conducts a literature review on the approaches used to overcome the weaknesses of SMOTE in solving unbalanced data. The literature review results can be used as a foundation for finding gaps or gaps in problems that previous researchers have not resolved. The gaps at earlier research can be used as a reference in future research. Therefore, it is necessary to focus this research question on the literature review conducted. The questions that this research asks are namely: (1) What approach is used in the modification of the SMOTE method? (2) What is the right evaluation metric for unbalanced data? and (3) What is the most used dataset in experiments on unbalanced data?

II. Materials and Method

This study conducted a systematic literature review on the approach used to modify the SMOTE method to solve the problem of unbalanced data. This research uses three stages in conducting a systematic review: Planning, Conduction, and Reporting. The steps of this research are explained in Fig. 1.

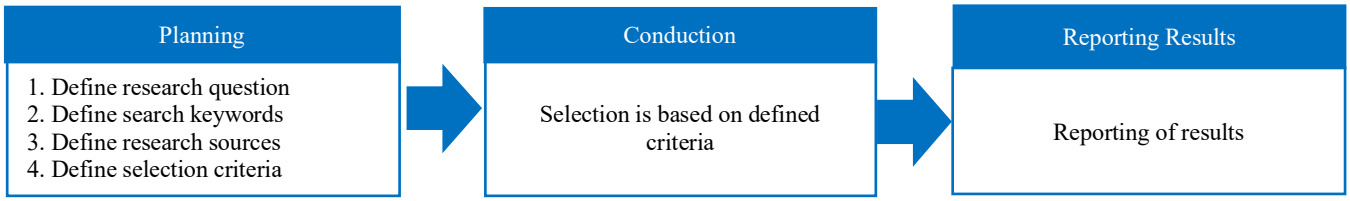

Fig. 1 Research Flowchart [32][33]

Based on Fig. 1, this process requires three stages to complete. The planning phase is used to plan research questions, define keywords used, select journal sources, and define selection criteria. Then, please proceed to the conduction phase, where the article criteria used in the systematic literature review and its publisher are selected. The final phase is reporting results based on specified criteria.

A. Planning

At this stage, the researcher defines the research question so that it focuses on the systematic literature review being carried out. After that, try to identify several works of literature related to SMOTE modification, where the search is carried out based on keywords and several previous research articles are explored. Defining research questions needs to be done to focus on the literature review conducted. The questions that this research asks are namely:

- Q1: What approach is used to modify the SMOTE method?
- Q2: What is the proper evaluation metric for unbalanced data?
- Q3: What is the most used dataset in experiments on unbalanced data?

Once the research questions have been defined, appropriate keywords were identified in the search for articles related to SMOTE modification. This research utilized the keyword "**SMOTE Modification**" in the article search. Search for articles based on keywords that had been defined previously. The article search used several reputable journal databases, such as IEEE, ScienceDirect, Scopus, and Springer

Based on the article search keywords that had been determined previously, it was necessary to define selection criteria for selecting articles relevant to the research topic from 4 reputable journal databases: ScienceDirect, Springer, IEEE, and Scopus. Some criteria that need to be defined in the selection of articles are as follows:

- Articles are published in the range of 2019 to 2023.
- The article contains a modified approach to the SMOTE method.

- The selected article is a journal article, not a proceeding.

B. Conduction

In this section, the researcher tries to collect journals based on the selection criteria determined in the planning section. During this process, articles were selected based on previously defined selection criteria. Table 1 shows the overall search results before filtering.

TABLE I
TOTAL ARTICLE SEARCH RESULTS BEFORE FILTERING

Journal Database	Unfiltered
IEEE	10
ScienceDirect	1391
Scopus	43
Springer	452
Total	**1896**

Fig. 2 summarizes the strategy used to conduct a citation literature review on modifying the SMOTE method in solving unbalanced data. This study used four databases for article selection. After that, the researcher obtained as many as 1896 articles that had not been filtered.

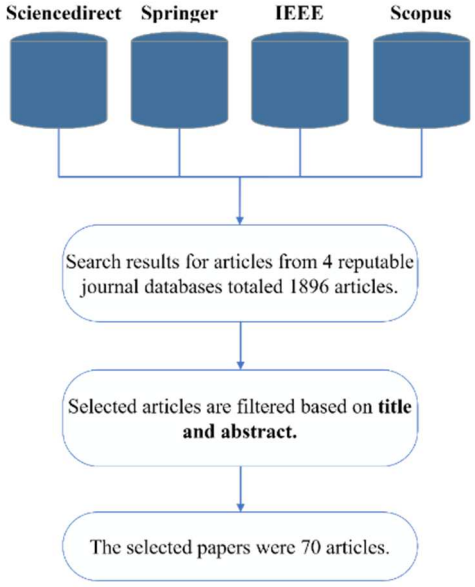

Fig. 2 Methodology Used for This Systematic Literature Review on Modification of SMOTE Method

Then, the criteria that had been made in the selection of articles were applied. The result was 70 selected articles in the last five years (2019-2023). Based on Fig. 2, the number of articles selected based on the article selection criteria was 70. Of these 70 articles, 15 are from Scopus, 21 are from Springer, 28 are from ScienceDirect, and 6 are from IEEE, which can be seen in Fig. 3. Based on the quartile distribution, 45 articles are Q1, 13 are Q2, 11 are Q3, and 1 is Q4, as shown in Fig. 4.

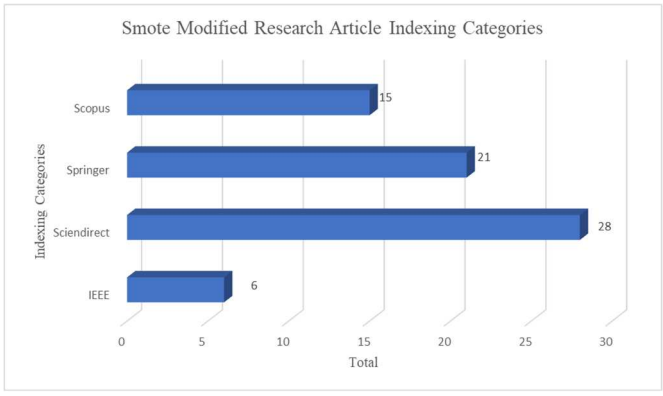

Fig. 3 Article Distribution of Each Journal Database

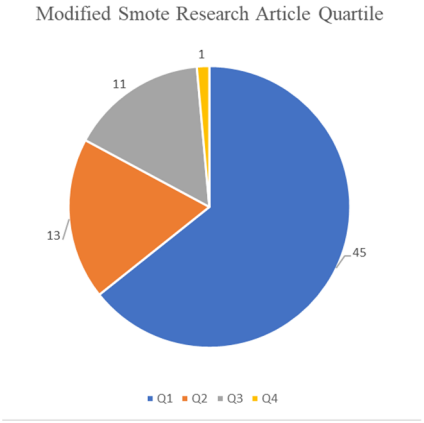

Fig. 4 Article Distribution by Quartile

C. Reporting Result

This section reports the results obtained. The process analyzes articles that have been filtered to answer research questions previously determined at the planning stage. The results of the analysis are reported in the Results and Discussion section.

III. RESULT AND DISCUSSION

There have been many previous studies on solving unbalanced data using SMOTE. SMOTE is the most commonly used oversampling method to solve data imbalance problems in machine learning modeling. The SMOTE method generates new synthetic data patterns by performing linear interpolation between minority class samples based on k-nearest neighbors [34]. Synthetic new data in the minority class using Equation (1).

$$Y' = Y^i + (Y^j - Y^i) * \gamma \quad (1)$$

Y' represents the addition of the minority class. Y^i represents the minority class, Y^j is a randomly selected value from the k-nearest neighbors of the minority class Y^i, and γ is a randomly selected vector value with a range of 0 to 1. Fig. 5 illustrates how SMOTE works.

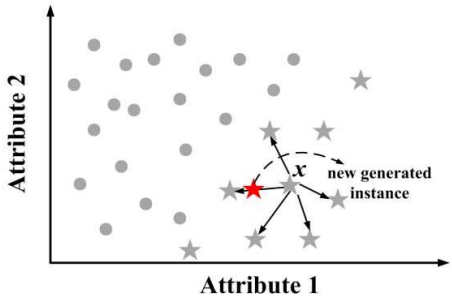

Fig. 5 How SMOTE Works

A. Q1. What Approach Is Used in the Modification of The SMOTE Method?

The SMOTE method has several weaknesses in solving unbalanced data problems, namely, producing noise data on synthetic artificial classes in the minority. Therefore, some previous studies improved the SMOTE method in solving unbalanced data, such as research [31] proposing the SMOTE-LOF approach, which combines SMOTE with the local outlier factor (LOF) algorithm to detect noise. The LOF technique is usually used to identify outliers, but LOF can detect noise because most outliers are considered noise. SMOTE-LOF focuses on detecting noise among synthetic minority samples. SMOTE-LOF applies the LOF algorithm to calculate the distance between minority instances and measure their LOF values. The minority class samples identified as outliers are removed, and the original minority class samples are retained. This method effectively prevents excessive noise removal.

Research [35] proposed the LR-SMOTE algorithm that combined the SMOTE method with the radius concept to overcome the oversampling noise problem. In contrast, Borderline SMOTE focused on changing the oversampling region. LR-SMOTE aimed to limit the artificial sampling space of oversampled minority classes to minimize noise. Research [36] proposed a cluster-based SMOTE and undersampling approach to handle cases of unbalanced data. SMOTE is used to create artificial data on the minority class. After being balanced, clustering was carried out into several majority and minority class clusters, and then the majority class for each cluster was deleted.

Research [37] proposed the Radius-SMOTE method to solve the unbalanced data problem. The Radius-SMOTE method focused on an initial selection approach and created synthetic data based on a safe radius distance, thus avoiding overlapping with other classes. Research [21] proposed a range-controlled oversampling technique (RCSMOTE), which could prevent the synthesis of minority examples in most domains. Research [38] modified the Borderline-SMOTE method with a noise reduction approach to unbalanced data. Research [39] proposed oversampling K-means clustering with SMOTE (KM-SMOTE) to solve the problem of unbalanced data on rockburst data. Research [27]–[29] used distance metrics on SMOTE in synthetic minority class data, and the results of their research can conclude that the use of multiple distance metrics on SMOTE positively impacts the performance of predictive methods. Research [40] proposed the Outlier-SMOTE oversampling technique. This method aimed to improve COVID-19 detection by overcoming the problem of a highly imbalanced dataset where only 9% of people tested positive for the virus. Outlier-SMOTE used an algorithm to create new synthetic data based on existing data, emphasizing outlier or remote samples more. The goal was to reduce negative errors in the classification of COVID-19 patients. The farthest samples had the highest priority for oversampling. This method was tested and compared with other oversampling techniques using various testing methods. The test results showed that Outlier-SMOTE performed better in most cases than SMOTE and ADASYN.

Research [41] proposed an extension of SMOTE called KSMOTE that applies the Kalman filter as a data reduction method to improve classification effectiveness. The Kalman filter in KSMOTE identified noisy data samples and removed them from the dataset. Thus, KSMOTE served to cleanse the dataset of distracting data samples, thereby reducing the impact of overfitting on the classification model and improving its performance. KSMOTE (Kalman filter-based SMOTE) removed irrelevant data samples from the minority region to remove noise. Research [42]–[45] uses a hybrid sampling approach to deal with imbalances in health data. Where the hybrid sampling method works by adding artificial data to minority data using SMOTE, then deleting the majority class that is adjacent to the minority class until it produces balanced data.

Research [46] proposed SMOTE-reverse k-nearest neighbors (SMOTE-RkNN) to solve the class imbalance problem in data. RkNN was used to estimate the probability density of each instance globally and can then be used to locate the noise accurately. The SMOTE-RkNN process in class balanced the data in a way that the data was balanced using SMOTE first, and then noise removal was carried out in the majority class using RkNN. Research [24] proposed the K-Nearest Neighbor Oversampling (KNNOR) approach to solving class imbalance. KNNOR utilized a kNN-based approach for noise filtering, a safe boundary and equal probability selection method for sample selection, and linear interpolation for sample generation. Research [4] proposed the Adaptive Synthetic (Adasyn) method with modification of the distance metric part using Manhattan distance to solve unbalanced data on student graduation data. The results showed that the modified use of Manhattan distance in the Adasyn method in creating artificial minority classes obtained better accuracy, precision, recall, and F1-measure than SMOTE and Adasyin without modification.

Research [47] the self-adaptive robust SMOTE (RSMOTE) method for unbalanced data classification involving noise data was proposed. The RSMOTE method considers three aspects of data balancing: (1) reducing the bias caused by class imbalance; (2) adaptively distinguishing between noise minority samples, dividing boundaries, and safe levels; (3) creating synthetic data on non-noise samples, generating more samples near safe level samples. Research [48] proposed the SMOTE-based Minority-prediction-probability oversampling method (MPP-SMOTE). The MPP-SMOTE method performs data balancing with four stages, namely: (1) filtering noise samples from the minority class, (2) estimating probabilities on minority class samples that do not contain noise, (3) calculating probabilities on hard-to-learn and easy-to-learn samples, and (4) generating new samples for the minority class.

Research [49] proposed natural neighbors SMOTE (NaNSMOTE) to overcome three problems, namely the selection of the k parameter, determining the number of neighbors for the artificial sample of the minority class, and being able to remove outlier data. Research [50] proposed the Importance-SMOTE approach that worked on the dividing line and edge line in the over-sampled minority class. Research [51] proposed the RD-SMOTE method to solve the problem of unbalanced data. The RD-SMOTE method performed noise filtering based on relative density to remove noise, then calculated the weights of sparsity and boundary based on relative and absolute density, and finally used the weights to generate more artificial minority samples in the class boundary region and sparse area. The RD-SMOTE method can effectively avoid sample noise generation, generate more synthetic samples in relevant regions, and do not require additional parameters. Research [52] proposed the GSMOTE-NFM algorithm that integrates the noise filter and individual sampling procedures.

Research [53] The workings of this method were (1). removing outlier data by applying the LOF method, (2). grouping the data into several clusters with the Mean-Shift algorithm, and (3). performing SMOTE on each cluster. Research [54] proposed the LoRAS method in generating new instances through an affine linear combination of several random minority class instances. Research [55] proposed the instance-weighted SMOTE (IW-SMOTE) method to overcome the limitations of SMOTE by performing noise filters to remove instances, applying soft weighting strategies to select instances in synthetic data generation, and finally using ensemble undersampling to reduce computational complexity.

Research [56] proposed the Region-Impurity Synthetic Minority Oversampling Technique (RIOT), an extension of SMOTE to overcome the class imbalance problem. RIOT used the region radius concept to identify minority instances and generate synthetic instances proportionally within the radius area. To make it easier to understand the proposed method in improving the previous SMOTE, the researcher made a comparison table, which can be seen in Table 2. Research [57] combining CNN and SMOTE in balancing image data. Research [58] proposed a firefly based on SMOTE method to solve binary class imbalance.

TABLE II
COMPARISON OF PROPOSED METHODS IN SMOTE METHOD IMPROVEMENT BASED ON FILTERING, CLUSTERING, AND DISTANCE METRIC

References	Method	Dataset	Evaluation Metrics	Weaknesses
[31]	SMOTE-LOF	Pima, Haberman, Glass	accuracy, precision, recall, f1-measure, AUC	Ignores the information in the outlier data
[30]	ASN-SMOTE	Pima, Haberman, Ecoli. New thyroid, Car, Iris, Seeds, Wine, Dermatology, Pageblock, Glass 0 – 6, Yeast 0 – 8	G-mean, f1-measure	Data noise is not removed first
[35]	LR-SMOTE	Pima, Haberman, Blood, Abalon	f1-measure, recall, G-mean	1. Determination of the optimal number of clusters 2. Ignores the minority class ratio of each cluster
[59]	MeanRadius-SMOTE	PHM	accuracy, f1-measure	3. Involves data noise
[60]	Geometric - SMOTE	Pima, Haberman, Breast, Heart, Liver, Wine, Iris, Glass, Yeast, Vehicle, New Thyroid, Ecoli, Eucalyptus, Segmentation, Page Block, Vowel Libras	accuracy. AUC, f1-measure, G-mean	G-SMOTE may produce excessive artificial minority samples in the input space.
[38]	Borderline SMOTE with Noise Reduction	Ecoli, Yeast, Abalon, Glass, Vowel, Page, Block, Statlog	accuracy	Focuses on decision boundary areas in sampling, resulting in oversampling
[61]	Spectral Clustering and SMOTE	p2p lending	accuracy, f1-measure, G-mean	1. Data noise is not addressed 2. Does not determine the optimal number of clusters and ignores the ratio of minority to majority classes in each cluster.
[62]	Minority Clustering SMOTE	Pima, Breast, Glass, Iris, Wine, Phoneme, Segment, Yeast, Vowel, Abalon, Ecoli	accuracy, f1-mesure, G-mean, recall, precision	1. Data noise is not addressed 2. Does not determine the optimal number of clusters
[63]	SMOTE and K-means	Botswana, Pavia, Kennedy, Salinas	accuracy, f1-measure, G-mean	1. Data noise is not addressed 2. Does not determine the optimal number of clusters
[36]	CUSS (Cluster-based Under-sampling and SMOTE)	Pima, Haberman, Glass, Ecoli, Vehicle, Yeast	AUC	1. Data noise is not addressed 2. Does not determine the optimal number of clusters and ignores the ratio of minority to majority classes in each cluster.
[39]	KM-SMOTE	Rockburst data	accuracy	1. Data noise is not addressed 2. Does not determine the optimal number of clusters and ignores the ratio of minority to majority classes

References	Method	Dataset	Evaluation Metrics	Weaknesses
				in each cluster.
[27]–[29]	Modification of Distance Metrics of SMOTE	Pima, Haberman, Ecoli, New Thyroid, Wisconsin, Vehicle	accuracy, f1-measure, kappa, recall specificity	-
[40]	Outlier-SMOTE	Ecoli, Abalon, Yeast, Wine, Mammography	recall, precision, f1-measure	Data noise adjacent to the majority class is not addressed.
[24]	KNNOR	Pima, Haberman, New Thyroid, Glass, Ecoli	f1-measure, G-mean	Many parameters are required in balancing the data
[64]	Improved and Random SMOTE (IR-SMOTE).	Haberman, Breast, Ecoli, Wbdc, Yeast, Abalone, Pageblock, German	accuracy, recall, specificity, f1-measure, G-mean	1. Data noise is not addressed 2. Does not determine the optimal number of clusters and ignores the ratio of minority to majority classes in each cluster.
[4]	Adasyn with Manhattan Distance	Student Graduates	accuracy, recall, precision, f1-measure	Data noise is not addressed.
[50]	Importance – SMOTE	Pima, Haberman, Glass, Ecoli, Yeas, Vehicle	f-measure, G-mean	Data noise is not addressed.
[53]	Mean-Shift SMOTE	Pima, Haberman, Blood	AUC, accuracy	1. Outlier data removed 2. Ignores the ratio of minority to majority classes in each cluster used in SMOTE.

Based on Table 2, previous research has improved the SMOTE method in solving unbalanced data problems. The most crucial weakness of SMOTE is that it produces noise data on the minority class artificial data. Previous research used the Filtering and Clustering approach to reduce the noise data generated by the SMOTE method. The filtering approach by removing data considered noise before SMOTE has a positive impact in solving unbalanced data [65], [66]. Using the Clustering method approach in SMOTE can minimize overlapping artificial minority data with the potential for noise [67]. In Table 3, gaps in previous research have not been resolved so that they can be used as opportunities for further study.

TABLE III
MODIFICATION OF THE SMOTE METHOD WITH FILTERING, CLUSTERING, AND DISTANCE METRIC APPROACHES

References	Filtering	Clustering	Modified Distance Metric
[27]–[29]	No	No	Yes
[4]	No	No	Yes
[38]	Yes	No	No
[37]	Yes	No	No
[36]	Yes	Yes	No
[30]	Yes	No	No
[31]	Yes	No	No
[40]	Yes	No	No
[39]	No	Yes	No
[61]	No	Yes	No
[62]	No	Yes	No
[63]	No	Yes	No
[64]	No	Yes	No
[65]	Yes	No	No
[26]	Yes	Yes	No
[68]	No	Yes	No
[69]	No	Yes	No
[70]	No	Yes	No
[71]	Yes	Yes	No
[72]	No	Yes	No
Proposed Method	**Yes**	**Yes**	**Yes**

B. Q2: What Is the Correct Evaluation Metric for Unbalanced Data?

Fig. 6 shows the evaluation of the metrics used to measure the success rate in solving unbalanced data. In Fig. 6, most researchers use the assessment of accuracy, f1-measure, and recall metrics on unbalanced data problems.

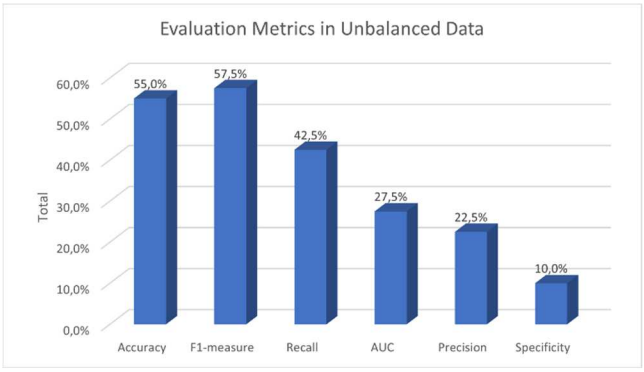

Fig. 6 Types of Metric Evaluation on Unbalanced Data

C. Q3: What Is the Most Used Dataset in Experiments on Unbalanced Data?

The dataset used as an experiment on unbalanced data is shown in Fig. 7.

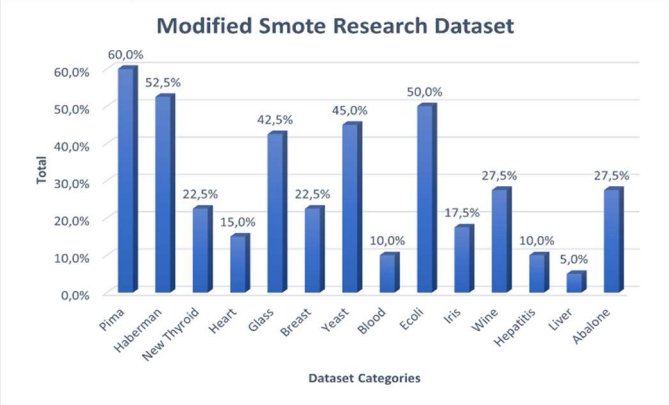

Fig. 7 Dataset Distribution as an Experiment on Unbalanced Data

IV. Conclusion

Research resolving imbalanced data is fascinating, as it involves improving methods, using experimental data, and choosing appropriate evaluation metrics. Research development regarding improving the weaknesses of the SMOTE method in resolving imbalanced data continues to develop over time. This research focuses on the approach used in SMOTE improvements to solve data noise and overlapping problems. This research found three methods most widely used in SMOTE modification in resolving unbalanced data, especially noise and overlapping data problems: filtering, clustering, and distance modification. Meanwhile, the experimental data used in the research data are Pima and Haberman. Then, regarding the most appropriate performance evaluation, this research found that the recommended evaluation metrics for imbalanced data are f1-measure, accuracy, recall, and AUC. Further research can reveal an approach to developing the SMOTE method for resolving small disjunction data in imbalanced data.

References

[1] M. Khushi *et al.*, "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.

[2] M. Naseriparsa, A. Al-Shammari, M. Sheng, Y. Zhang, and R. Zhou, "RSMOTE: improving classification performance over imbalanced medical datasets," *Health Information Science and Systems.*, vol. 8, no. 1, pp. 1–13, 2020, doi: 10.1007/s13755-020-00112-w.

[3] C. Yang, E. A. Fridgeirsson, J. A. Kors, J. M. Reps, and P. R. Rijnbeek, "Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data," *Journal of Big Data*, vol. 11, no. 1, p. 7, Jan. 2024, doi: 10.1186/s40537-023-00857-7.

[4] H. A. Gameng, B. D. Gerardo, and R. P. Medina, "A modified adaptive synthetic smote approach in graduation success rate classification," *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE).*, vol. 8, no. 6, pp. 3053–3057, 2019, doi: 10.30534/ijatcse/2019/63862019.

[5] R. Malhotra and K. Lata, "An empirical study on predictability of software maintainability using imbalanced data," *Software Quality Journal.*, vol. 28, no. 4, pp. 1581–1614, 2020, doi: 10.1007/s11219-020-09525-y.

[6] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Machine Learning.*, no. January, pp. 1–21, 2023, doi: 10.1007/s10994-022-06296-4.

[7] A. Gosain and S. Sardana, "Farthest SMOTE: A Modified SMOTE Approach," in *Advances in Intelligent Systems and Computing*, vol. 711, 2019, pp. 309–320. doi: 10.1007/978-981-10-8055-5_28.

[8] A. Puri and M. Kumar Gupta, "Knowledge discovery from noisy imbalanced and incomplete binary class data," *Expert Systems with Applications.*, vol. 181, no. March 2020, pp. 1–14, 2021, doi: 10.1016/j.eswa.2021.115179.

[9] N. A. Azhar, M. S. Mohd Pozi, A. Mohamed Din, and A. Jatowt, "An Investigation of SMOTE based Methods for Imbalanced Datasets with Data Complexity Analysis," *IEEE Transactions on Knowledge and Data Engineering.*, pp. 1–1, 2022, doi:10.1109/TKDE.2022.3179381.

[10] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of Classification Methods on Unbalanced Data Sets," *IEEE Access*, vol. 9, pp. 64606–64628, 2021, doi: 10.1109/access.2021.3074243.

[11] A. V. Vitianingsih, Z. Othman, S. S. K. Baharin, A. Suraji, and A. L. Maukar, "Application of the Synthetic Over-Sampling Method to Increase the Sensitivity of Algorithm Classification for Class Imbalance in Small Spatial Datasets," *International Journal of Intelligent Engineering and Systems.*, vol. 15, no. 5, pp. 676–690, 2022, doi: 10.22266/ijies2022.1031.58.

[12] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Information Sciences.*, vol. 505, pp. 32–64, 2019, doi:10.1016/j.ins.2019.07.070.

[13] T. Hasanin, T. M. Khoshgoftaar, J. L. Leevy, and R. A. Bauder, "Severely imbalanced Big Data challenges: investigating data sampling approaches," *Journal of Big Data*, vol. 6, no. 1, pp. 1–25, 2019, doi: 10.1186/s40537-019-0274-4.

[14] A. Anggrawan, H. Hairani, and C. Satria, "Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE," *International Journal of Information and Education Technology (IJIET).*, vol. 13, no. 2, pp. 289–295, 2023.

[15] R. Malhotra and K. Lata, "Handling class imbalance problem in software maintainability prediction: an empirical investigation," *Frontiers of Computer Science.*, vol. 16, no. 4, pp. 1–14, Aug. 2022, doi: 10.1007/s11704-021-0127-0.

[16] K. S. Babu and Y. N. Rao, "A Study on Imbalanced Data Classification for Various Applications," Revue d'Intelligence Artificielle, vol. 37, no. 2, pp. 517–524, Apr. 2023, doi:10.18280/ria.370229.

[17] P. Mooijman, C. Catal, B. Tekinerdogan, A. Lommen, and M. Blokland, "The effects of data balancing approaches: A case study," *Applied Soft Computing.*, vol. 132, p. 109853, 2023, doi:10.1016/j.asoc.2022.109853.

[18] S. Feng, J. Keung, X. Yu, Y. Xiao, and M. Zhang, "Investigation on the stability of SMOTE-based oversampling techniques in software defect prediction," *Information and Software Technology.*, vol. 139, no. June, p. 106662, 2021, doi: 10.1016/j.infsof.2021.106662.

[19] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Information Sciences.*, vol. 505, pp. 32–64, Dec. 2019, doi: 10.1016/j.ins.2019.07.070.

[20] X. Yuan, S. Chen, H. Zhou, C. Sun, and L. Yuwen, "CHSMOTE: Convex hull-based synthetic minority oversampling technique for alleviating the class imbalance problem," *Information Sciences.*, vol. 623, pp. 324–341, 2023, doi: 10.1016/j.ins.2022.12.056.

[21] P. Soltanzadeh and M. Hashemzadeh, "RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem," *Information Sciences.*, vol. 542, pp. 92–111, 2021, doi: 10.1016/j.ins.2020.07.014.

[22] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques," *Applied Soft Computing.*, vol. 143, pp. 1–23, 2023, doi: 10.1016/j.asoc.2023.110415.

[23] V. Werner de Vargas, J. A. Schneider Aranda, R. dos Santos Costa, P. R. da Silva Pereira, and J. L. Victória Barbosa, "Imbalanced data preprocessing techniques for machine learning: a systematic mapping study," *Knowledge and Information Systems.*, vol. 65, no. 1, pp. 31–57, 2023, doi: 10.1007/s10115-022-01772-8.

[24] A. Islam, S. B. Belhaouari, A. U. Rehman, and H. Bensmail, "KNNOR: An oversampling technique for imbalanced datasets," *Applied Soft Computing.*, vol. 115, pp. 1–18, 2022, doi:10.1016/j.asoc.2021.108288.

[25] Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, and X. Han, "An oversampling algorithm combining SMOTE and k-means for imbalanced medical data," *Information Sciences.*, vol. 572, no. 5, pp. 574–589, Sep. 2021, doi: 10.1016/j.ins.2021.02.056.

[26] A. Arafa, N. El-Fishawy, M. Badawy, and M. Radad, "RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification," *Journal of King Saud University - Computer science.*, vol. 34, no. 8, pp. 5059–5074, 2022, doi:10.1016/j.jksuci.2022.06.005.

[27] Q. Dai, J. wei Liu, and J. L. Zhao, "Distance-based arranging oversampling technique for imbalanced data," *Neural Computing and Applications.*, vol. 35, no. 2, pp. 1323–1342, 2023, doi:10.1007/s00521-022-07828-8.

[28] S. Feng, J. Keung, P. Zhang, Y. Xiao, and M. Zhang, "The impact of the distance metric and measure on SMOTE-based techniques in software defect prediction," *Information and Software Technology.*, vol. 142, no. January, pp. 1–14, 2022, doi:10.1016/j.infsof.2021.106742.

[29] A. Balakrishnan, J. Medikonda, P. K. Namboothiri, and M. Natarajan, "Mahalanobis Metric-based Oversampling Technique for Parkinson's Disease Severity Assessment using Spatiotemporal Gait Parameters," *Biomedical Signal Processing and Control*, vol. 86, no. September, pp. 1–14, 2023, doi: 10.1016/j.bspc.2023.105057.

[30] X. Yi, Y. Xu, Q. Hu, S. Krishnamoorthy, W. Li, and Z. Tang, "ASN-SMOTE : a synthetic minority oversampling method with adaptive qualified synthesizer selection," *Complex & Intelligent Systems.*, vol. 8, no. 3, pp. 2247–2272, 2022, doi: 10.1007/s40747-021-00638-w.

[31] A. Asniar, N. U. Maulidevi, and K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," *Journal of King Saud University - Computer science.*, vol. 34, no. 6, pp. 3413–3423, Jun. 2022, doi: 10.1016/j.jksuci.2021.01.014.

[32] R. Parente Da Costa, E. Di. Canedo, R. T. De Sousa, R. De Oliveira Albuquerque, and L. J. Garcia Villalba, "Set of Usability Heuristics for Quality Assessment of Mobile Applications on Smartphones," *IEEE Access*, vol. 7, no. April, pp. 116145–116161, 2019, doi:10.1109/access.2019.2910778.

[33] N. Hussain, H. T. Mirza, G. Rasool, I. Hussain, and M. Kaleem, "Spam review detection techniques: A systematic literature review," *Applied Sciences.*, vol. 9, no. 5, pp. 1–26, 2019, doi:10.3390/app9050987.

[34] J. Park, S. Kwon, and S. P. Jeong, "A study on improving turnover intention forecasting by solving imbalanced data problems: focusing on SMOTE and generative adversarial networks," *Journal of Big Data*, vol. 10, no. 1, pp. 1–16, 2023, doi: 10.1186/s40537-023-00715-6.

[35] X. W. Liang, A. P. Jiang, T. Li, Y. Y. Xue, and G. T. Wang, "LR-SMOTE — An improved unbalanced data set oversampling based on K-means and SVM," *Knowledge-Based Systems.*, vol. 196, no. May, pp. 1–10, May 2020, doi: 10.1016/j.knosys.2020.105845.

[36] S. Feng, C. Zhao, and P. Fu, "A cluster-based hybrid sampling approach for imbalanced data classification," *Review of Scientific Instruments.*, vol. 91, no. 5, pp. 1–9, 2020, doi: 10.1063/5.0008935.

[37] 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, doi:10.1109/access.2021.3080316.

[38] M. Revathi and D. Ramyachitra, "A Modified Borderline Smote with Noise Reduction in Imbalanced Datasets," *Wireless Personal Communications.*, vol. 121, no. 3, pp. 1659–1680, 2021, doi:10.1007/s11277-021-08690-y.

[39] Q. Liu *et al.*, "Application of KM-SMOTE for rockburst intelligent prediction," *Tunnelling and Underground Space Technology.*, vol. 138, no. October, pp. 1–10, 2023, doi: 10.1016/j.tust.2023.105180.

[40] V. P. K. Turlapati and M. R. Prusty, "Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19," *Intelligence-Based Medicine.*, vol. 3–4, no. November, pp. 1–10, 2020, doi: 10.1016/j.ibmed.2020.100023.

[41] T. G.S., Y. Hariprasad, S. S. Iyengar, N. R. Sunitha, P. Badrinath, and S. Chennupati, "An extension of Synthetic Minority Oversampling Technique based on Kalman filter for imbalanced datasets," *Machine Learning with Applications.*, vol. 8, no. January, pp. 1–12, 2022, doi: 10.1016/j.mlwa.2022.100267.

[42] H. Hairani and D. Priyanto, "A New Approach of Hybrid Sampling SMOTE and ENN to the Accuracy of Machine Learning Methods on Unbalanced Diabetes Disease Data," International Journal of Advanced Computer Science and Applications, vol. 14, no. 8, 2023, doi: 10.14569/ijacsa.2023.0140864

[43] H. Hairani, A. Anggrawan, and D. Priyanto, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link," *JOIV : International Journal on Informatics Visualization.*, vol. 7, no. 1, pp. 258–264, 2023.

[44] L. G. R. Putra, K. Marzuki, and H. Hairani, "Correlation-based feature selection and Smote-Tomek Link to improve the performance of machine learning methods on cancer disease prediction," *Engineering and Applied Science Research (EASR).*, vol. 50, no. 6, pp. 577–583, 2023, doi: 10.14456/easr.2023.59.

[45] K. Wang *et al.*, "Improving Risk Identification of Adverse Outcomes in Chronic Heart Failure Using SMOTE+ENN and Machine Learning," *Risk Management and Healthcare Policy*, vol. 14, no. June, pp. 2453–2463, Jun. 2021, doi: 10.2147/RMHP.S310295.

[46] A. Zhang, H. Yu, Z. Huan, X. Yang, S. Zheng, and S. Gao, "SMOTE-RkNN: A hybrid re-sampling method based on SMOTE and reverse k-nearest neighbors," *Information Sciences.*, vol. 595, pp. 70–88, 2022, doi: 10.1016/j.ins.2022.02.038.

[47] B. Chen, S. Xia, Z. Chen, B. Wang, and G. Wang, "RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise," *Information Sciences.*, vol. 553, pp. 397–428, 2021, doi:10.1016/j.ins.2020.10.013.

[48] Z. Wei, L. Zhang, and L. Zhao, "Minority-prediction-probability-based oversampling technique for imbalanced learning," *Information Sciences.*, vol. 622, pp. 1273–1295, 2023, doi:10.1016/j.ins.2022.11.148.

[49] J. Li, Q. Zhu, Q. Wu, and Z. Fan, "A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors," *Information Sciences.*, vol. 565, pp. 438–455, 2021, doi:10.1016/j.ins.2021.03.041.

[50] J. Liu, "Importance-SMOTE: a synthetic minority oversampling method for noisy imbalanced data," *Soft Computing.*, vol. 26, no. 2, pp. 1141–1163, 2022.

[51] R. Liu, "A novel synthetic minority oversampling technique based on relative and absolute densities for imbalanced classification," *Applied Intelligence.*, vol. 53, no. 1, pp. 786–803, 2023, doi: 10.1007/s10489-022-03512-5.

[52] K. Cheng, C. Zhang, H. Yu, X. Yang, H. Zou, and S. Gao, "Grouped SMOTE with Noise Filtering Mechanism for Classifying Imbalanced Data," *IEEE Access*, vol. 7, pp. 170668–170681, 2019, doi:10.1109/access.2019.2955086.

[53] A. S. Ghorab, W. M. Ashour, and S. I. Abudalfa, "An Adaptive Oversampling Method for Imbalanced Datasets Based on Mean-Shift and SMOTE," in *CBT 2022: Explore Business, Technology Opportunities and Challenges After the Covid-19 Pandemic*, 2023, pp. 13–23. doi: 10.1007/978-3-031-08954-1_2.

[54] S. Bej, N. Davtyan, M. Wolfien, M. Nassar, and O. Wolkenhauer, "LoRAS: an oversampling approach for imbalanced datasets," *Machine Learning.*, vol. 110, no. 2, pp. 279–301, 2021, doi:10.1007/s10994-020-05913-4.

[55] A. Zhang, H. Yu, S. Zhou, Z. Huan, and X. Yang, "Instance weighted SMOTE by indirectly exploring the data distribution," *Knowledge-Based Systems.*, vol. 249, no. August, pp. 1–24, 2022, doi:10.1016/j.knosys.2022.108919.

[56] D.-C. Li, S.-Y. Wang, K.-C. Huang, and T.-I. Tsai, "Learning class-imbalanced data with region-impurity synthetic minority oversampling technique," *Information Sciences.*, vol. 607, pp. 1391–1407, 2022, doi: https://doi.org/10.1016/j.ins.2022.06.067.

[57] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, "Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks," *Applied Sciences.*, vol. 13, no. 6, pp. 1–34, Mar. 2023, doi: 10.3390/app13064006.

[58] P. Kaur and A. Gosain, "FF-SMOTE: A Metaheuristic Approach to Combat Class Imbalance in Binary Classification," *Applied Artificial Intelligence.*, vol. 33, no. 5, pp. 420–439, 2019, doi:10.1080/08839514.2019.1577017.

[59] F. Duan, S. Zhang, Y. Yan, and Z. Cai, "An Oversampling Method of Unbalanced Data for Mechanical Fault Diagnosis Based on MeanRadius-SMOTE," *Sensors*, vol. 22, no. 14, pp. 1–15, Jul. 2022, doi: 10.3390/s22145166.

[60] G. Douzas and F. Bacao, "Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE," *Information Sciences.*, vol. 501, pp. 118–135, 2019, doi: 10.1016/j.ins.2019.06.007.

[61] P. K. Jadwal, S. Jain, S. Pathak, and B. Agarwal, "Improved resampling algorithm through a modified oversampling approach based on spectral clustering and SMOTE," *Microsystem Technologies.*, vol. 28, no. 12, pp. 2669–2677, 2022, doi:10.1007/s00542-022-05287-8.

[62] H. Yi, Q. Jiang, X. Yan, and B. Wang, "Imbalanced Classification Based on Minority Clustering Synthetic Minority Oversampling Technique with Wind Turbine Fault Detection Application," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 9, pp. 5867–5875, 2021, doi: 10.1109/TII.2020.3046566.

[63] J. Fonseca, G. Douzas, and F. Bacao, "Improving imbalanced land cover classification with k-means smote: Detecting and oversampling distinctive minority spectral signatures," *Information*, vol. 12, no. 7, pp. 1–20, 2021, doi: 10.3390/info12070266.

[64] G. Wei, W. Mu, Y. Song, and J. Dou, "An improved and random synthetic minority oversampling technique for imbalanced data," *Knowledge-Based Systems.*, vol. 248, no. July, pp. 1–13, 2022, doi:10.1016/j.knosys.2022.108839.

[65] H. Guan, Y. Zhang, M. Xian, H. D. Cheng, and X. Tang, "SMOTE-WENN: Solving class imbalance and small sample problems by oversampling and distance scaling," *Applied Intelligence.*, vol. 51, no. 3, pp. 1394–1409, Mar. 2021, doi: 10.1007/s10489-020-01852-8.

[66] Q. Chen, Z. L. Zhang, W. P. Huang, J. Wu, and X. G. Luo, "PF-SMOTE: A novel parameter-free SMOTE for imbalanced datasets," *Neurocomputing*, vol. 498, pp. 75–88, 2022, doi:10.1016/j.neucom.2022.05.017.

[67] 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, no. 7, pp. 2839–2851, 2021,

doi:10.1007/s00521-020-05130-z.
[68] W. Li, "Imbalanced data optimization combining K-means and SMOTE," *International Journal of Performability Engineering.*, vol. 15, no. 8, pp. 2173–2181, 2019, doi:10.23940/ijpe.19.08.p17.21732181.
[69] J. Arora *et al.*, "MCBC-SMOTE: A Majority Clustering Model for Classification of Imbalanced Data," *Computers, Materials and Continua.*, vol. 73, no. 3, pp. 4801–4817, 2022, doi:10.32604/cmc.2022.025960.
[70] Y. Yang, H. Akbarzadeh Khorshidi, and U. Aickelin, "A Diversity-Based Synthetic Oversampling Using Clustering for Handling Extreme Imbalance," *SN Computer Science.*, vol. 4, no. 6, pp. 1–16, 2023, doi: 10.1007/s42979-023-02249-3.
[71] K. Li *et al.*, "A hybrid cluster-borderline SMOTE method for imbalanced data of rock groutability classification," *Bulletin of Engineering Geology and the Environment.*, vol. 81, no. 1, pp. 1–15, 2022, doi: 10.1007/s10064-021-02523-9.
[72] S. Hooda and S. Mann, "Distributed synthetic minority oversampling technique," *International Journal of Computational Intelligence Systems.*, vol. 12, no. 2, pp. 929–936, 2019, doi:10.2991/ijcis.d.190719.001.