Vol. 16, No. 46, Spring 2026, Pages 31-48

Journal of Analytical and Numerical Methods in Mining Engineering

Journal home page: http://anm.yazd.ac.ir/

Research article

Application of resampling algorithms in the imbalanced geochemical data classification: Case study; Geochemical data of Qayen 1:100000 sheet

Hamid Geranian[1*]

1- Dept. of Mining Engineering, Birjand University of Technology, Birjand, Iran

(Received: 19 January 2025, Revise: 17 April 2025, Accepted: 27 May 2025)

DOI: 10.22034/anm.2025.22666.1661

Keywords

SMOTE algorithm
ADASYN algorithm
RUS algorithm
OSS algorithm
SMOTE-Tomek algorithm
ADASYN-CNN algorithm
Qayen Sheet

English Extended Abstract

Summary

The geochemical data exhibit an imbalance, with a high number of samples belong to the low-grade or background class, and a low number of samples in the high-grade or anomaly class. This imbalance in the dataset can lead to the development of a biased model, decreasing the likelihood of new samples belonging to classes with fewer representations and resulting in reduced accuracy and precision of the model. This paper introduces oversampling techniques such as SMOTE and ADASYN, undersampling methods like RUS and OSS, and hybrid-sampling approaches such as SMOTE-Tomek and ADASYN-CNN to address this data imbalance. The performance of these algorithms on geochemical data from the Qayen sheet is evaluated using SVM and ANN classification methods. The results demonstrate that data balancing leads to a significant increase in metrics such as accuracy, sensitivity, specificity, precision, F-score, F-value, G-mean, and AUC by 10 to 50 percent, while reducing error metrics by about 10 percent. The oversampling, hybrid-sampling, and undersampling algorithms exhibit high performance levels in improving the classification results, respectively. The geochemical anomaly maps generated with the help of these balancing algorithms show a greater number of anomaly areas in the study area, effectively aligning with mineralized rock units. Notably, oversampling techniques like SMOTE and ADASYN, followed by the hybrid-sampling method ADASYN-CNN, demonstrate superior performance in this regard. Therefore, this paper recommends oversampling and then hybrid-sampling algorithms before classifying exploration data to enhance model accuracy and better identify geochemical anomalies in the target area.

Introduction

Classification is a supervised machine learning technique that seeks to build a model from a dataset comprising multiple classes by analyzing the statistical relationships between them, determining the probability of a new sample belonging to each class. The accuracy of model estimation heavily relies on the distribution of samples within the dataset. A balanced dataset indicates an equal distribution of samples across different classes. However, in real-world data, we often encounter imbalanced datasets. Various approaches have been proposed to address the issue of classifying imbalanced datasets, including data sampling, algorithmic modification, and cost-sensitive learning methods [1,2]. This paper focuses on the algorithms associated with the first approach. To achieve this, geochemical data from the 1: 100,000 Qayen sheet in South Khorasan Province were utilized, employing support vector machine (SVM) and artificial neural network (ANN) classification methods.

*Corresponding author: E-mail: h.geranian@birjandut.ac.ir

Vol. 16, No. 46, Spring 2026, Pages 31-48

Journal of Analytical and Numerical Methods in Mining Engineering

Journal home page: http://anm.yazd.ac.ir/

Methodology and Approaches

In a two-class training dataset with n samples, if the number of samples in the first class (n1) significantly exceeds the number of samples in the second class (n2) such that n1≫n2 and n=n1+n2, we are faced with the challenge of classifying an imbalanced dataset. The majority class is represented by the first class while the second class is considered the minority class, and the ratio n1/n2 is referred to as the imbalance ratio. Three approaches have been suggested to address this issue: oversampling techniques, undersampling techniques, and hybrid methods. This study introduces oversampling methods such as SMOTE and ADASYN, undersampling techniques like RUS and OSS, as well as hybrid-sampling approaches including SMOTE-Tomek and ADASYN-CNN, in order to achieve dataset balance [3,4,5]. Furthermore, the performance of these methods on the geochemical data of stream sediments from the Qayen sheet has been examined using SVM and ANN classification methods. Accuracy, sensitivity, specificity, precision, F-score, F-value, G-mean, and AUC metrics have been employed to assess the confusion matrix of the classification techniques [6,7].

Results and Conclusions

Initially, the dataset is randomly split into two parts: training data (comprising approximately 80 percent of the data, specifically 450 samples from the majority class and 71 samples from the minority class, totaling 521 samples) and testing data (comprising around 20 percent of the data, which equates to 113 samples from the majority class and 18 samples from the minority class, totaling 131 samples). The classification outcomes of the imbalanced dataset reveal that despite the high average classification accuracy in both methods, approximately 87 percent, the classification accuracy of the minority class is notably low. Consequently, there is a necessity to balance the training data. Tables 1 and 2 exhibit the classification outcomes of the testing data using balanced models and the metrics from the confusion matrix, respectively. These tables illustrate that data balancing has succeeded in enhancing the value of all metrics and diminishing the error metric's value. Overall, the ADASYN-CNN algorithm within the SVM method and the SMOTE algorithm within the ANN method can be recommended as the top methods, as they possess the highest cumulative metric values.

Table 1. Classification results of testing data with balanced models using the SVM and ANN methods

		Predicated samples							
Type of data		SMOTE				ADASYN			
		SVM		ANN		SVM		ANN	
		Background	Anomaly	Background	Anomaly	Background	Anomaly	Background	Anomaly
True samples	Background	102	11	111	2	100	13	112	1
	Anomaly	1	17	2	16	1	17	3	16
	Accuracy	90.3	64.4	98.2	89.9	88.5	94.4	99.1	89.9
		RUS				OSS			
True samples	Background	88	25	95	18	89	24	99	14
	Anomaly	۴5	13	2	16	5	13	2	16
	Accuracy	77.9	72.2	84.1	89.9	78.8	72.2	87.6	89.9
		SMOTE-Tomek				ADASYN-CNN			
True samples	Background	93	20	107	6	101	12	110	3
	Anomaly	1	17	1	17	1	17	2	16
	Accuracy	82.3	94.4	94.7	94.4	98.3	94.4	97.3	89.9

Vol. 16, No. 46, Spring 2026, Pages 31-48

Journal of Analytical and Numerical Methods in Mining Engineering

Journal home page: http://anm.yazd.ac.ir/

Table 2. Values of the confusion matrix metrics for the testing data

Method		Type of data						
		Imbalanced	Balanced					
			SMOTE	ADASYN	RUS	OSS	SMOTE-Tomek	ADASYN-CNN
SVM	AC	0.809	0.908	0.893	0.771	0.779	0.839	0.901
	ER	0.191	0.092	0.107	0.229	0.221	0.161	0.099
	S	0.867	0.903	0.885	0.885	0.779	0.823	0.983
	SP	0.444	0.994	0.994	0.994	0.722	0.994	0.994
	P	0.907	0.990	0.990	0.946	0.947	0.989	0.990
	F-Score	0.887	0.945	0.935	0.914	0.855	0.898	0.986
	F-Value	0.889	0.971	0.967	0.933	0.908	0.951	0.989
	G-Mean	0.620	0.923	0.914	0.914	0.750	0.934	0.963
	AUC	0.656	0.923	0.915	0.750	0.755	0.884	0.919
ANN	AC	0.855	0.969	0.977	0.847	0.878	0.947	0.962
	ER	0.145	0.031	0.023	0.153	0.122	0.053	0.038
	S	0.903	0.982	0.991	0.841	0.876	0.947	0.973
	SP	0.555	0.899	0.899	0.899	0.899	0.944	0.899
	P	0.927	0.982	0.974	0.979	0.980	0.991	0.982
	F-Score	0.915	0.982	0.982	0.905	0.925	0.969	0.977
	F-Value	0.922	0.982	0.977	0.948	0.957	0.982	0.981
	G-Mean	0.708	0.939	0.943	0.870	0.887	0.945	0.935
	AUC	0.729	0.936	0.912	0.865	0.882	0.946	0.931

In the following, models generated from a balanced dataset utilized to produce a composite geochemical anomaly map of the study area. To achieve this, we employed the inverse square distance method to estimate the concentration of each element in 500×500 m cells. The Qayen sheet comprises a total of 12221 cells, resulting in a data matrix of 27×12221. Figures 1 and 2 display the geochemical anomaly map of the study area generated using SVM and ANN classification methods, utilizing imbalanced and balanced datasets, respectively. The anomaly zones in the SVM and ANN maps cover approximately 16 and 50.25 km2, respectively. Figure 1 highlights that the identified geochemical anomalies exhibit limited overlap with the igneous rock units within the study area. The balancing algorithms utilized in modeling the geochemical anomaly maps expanded the anomaly areas and effectively aligned them with mineralized rock units. Furthermore, oversampling algorithms such as SMOTE and ADASYN, followed by the hybrid ADASYN-CNN algorithm, demonstrated superior performance in achieving this outcome.

*Corresponding author: E-mail: h.geranian@birjandut.ac.ir

Vol. 16, No. 46, Spring 2026, Pages 31-48

Journal of Analytical and Numerical Methods in Mining Engineering

Journal home page: http://anm.yazd.ac.ir/

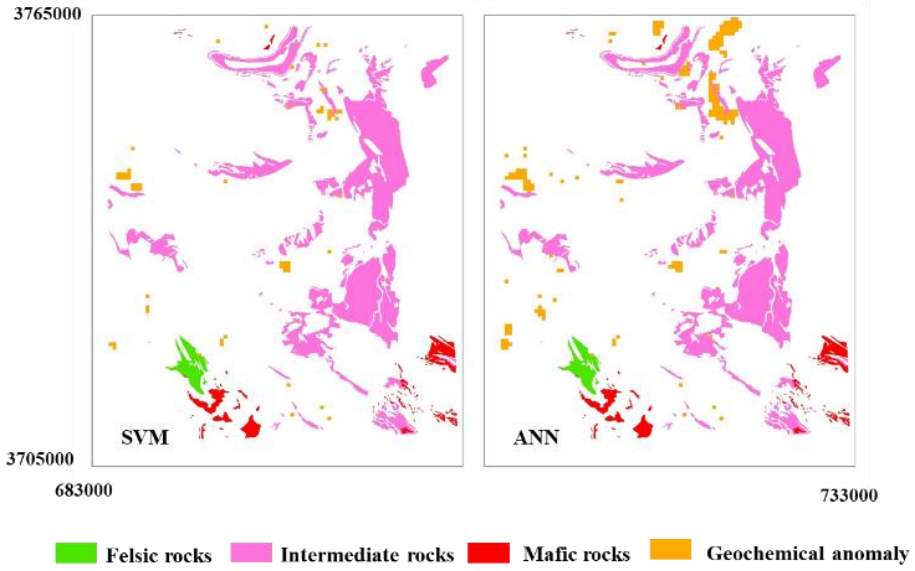
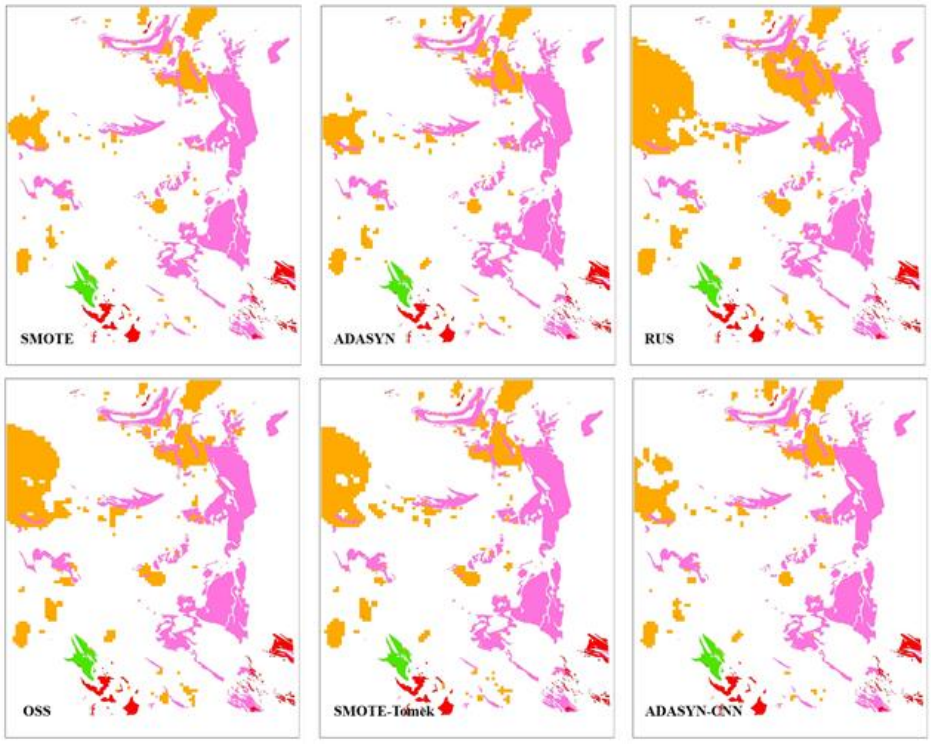


Fig. 1. Map of composite geochemical anomalies estimated with the imbalanced dataset along with the location of igneous rock units in the study area



Vol. 16, No. 46, Spring 2026, Pages 31-48

Journal of Analytical and Numerical Methods in Mining Engineering

Journal home page: http://anm.yazd.ac.ir/

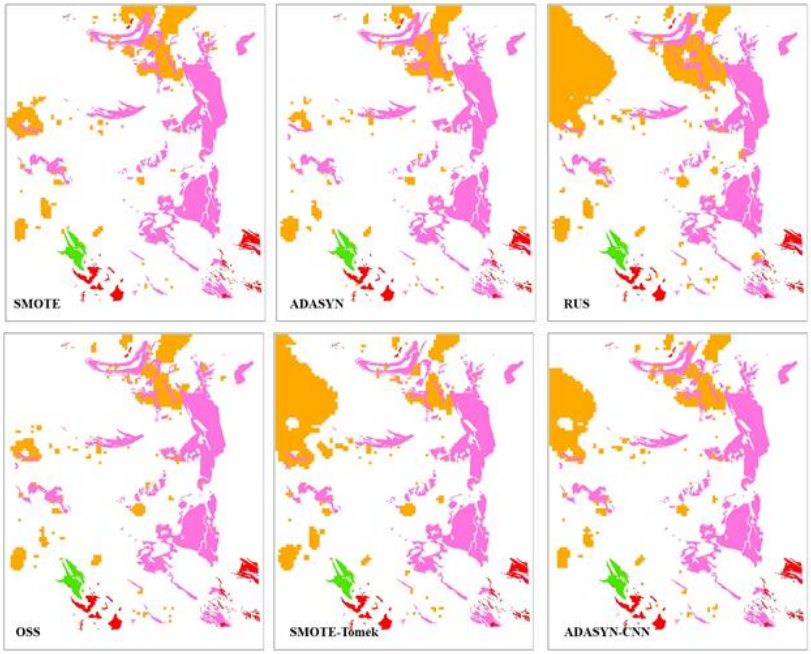


Fig. 2. Map of composite geochemical anomalies estimated with the balanced dataset (Legend is the same as in Fig. 1)

Hence, this paper recommends balancing the dataset before classification to enhance the accuracy, precision, and efficiency of the classification model. Additionally, another suggestion outlined in this paper is to utilize oversampling algorithms followed by hybrid-sampling algorithms prior to classifying the exploration data.

References

[1] Yuan, Y., Wei, J., Huang, H., Jiao, W., Wang, J. and Chen, H. (2023). Review of resampling techniques for the treatment of imbalanced industrial data classification in equipment condition monitoring. Engineering Applications of Artificial Intelligence 126: 106911.

[2] Payal Gulati, P. (2020). Hybrid Resampling Technique to Tackle the Imbalanced Classi¦cation Problem. Computer Science, Corpus ID: 241168640.

[3] Brownlee, J. (2021). Imbalanced Classification with Python, Machine Learning Mastery, 463 P.

[4] Wongvorachan, T., He, S. and Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. Information 14: 54.

[5] Brandt, J. and Lanzén, E. (2021). A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification. Department of Statistics, Uppsala University, 42 P.

[6] Han, J., Kamber, M. and Pei, J. (2022). Data mining: concepts and techniques, 4th Edition, Morgan Kaufmann, 752 P.

[7] Zoynul Abedin1, M., Guotai, C., Hajek, P. and Zhang, T. (2023). Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. Complex & Intelligent Systems 9: 3559–3579.

[8] Zaki, M.J. and Meira, W. (2020). Data Mining and Machine Learning: Fundamental Concepts and Algorithms, Cambridge University Press, New York, 777 P.

[9] Cerulli, G. (2023). Fundamentals of Supervised Machine Learning: With Applications in Python, R, and Stata, Springer Cham, 391 P.

*Corresponding author: E-mail: h.geranian@birjandut.ac.ir

Vol. 16, No. 46, Spring 2026, Pages 31-48

Journal of Analytical and Numerical Methods in Mining Engineering

Journal home page: http://anm.yazd.ac.ir/

[10] Moradzadeh, A., Zare, M., Kamkar Rouhani, A. and Doulati Aredehjan, F. (2019). Classification of environmental geochemical data using discriminant analysis and neural network in carbonate-sulfide waste dumps of lead and zinc mines. Journal of Mining Engineering 14(44): 12-25 [In Persian].

[11] Geranian, H., Tabatabaei, S.H., Asadi, H.H. and Carranza, E.J.M. (2016). Application of discriminant analysis and support vector machine in mapping gold potential areas for further drilling in the Sari-Gunay gold deposit, NW Iran. Nat. Resour. Res. 25: 145–159.

[12] Zaremotlagh, S. and Hezarkhani, A. (2017). The use of decision tree induction and artificial neural networks for recognizing the geochemical distribution patterns of LREE in the Choghart deposit, Central Iran. Journal of African Earth Sciences 128: 37-46.

[13] Degtyareva. K., Kukartseva, O., Tynchenko, V., Mariupolskiy, T. and Pereverzev, D. (2024). Analysis of geochemical characteristics of rocks using machine learning methods. E3S Web of Conferences 583, 01007.

[14] Geranian, H., Tabatabaei, S.H. and Asadi, H.H. (2013). Application of classifiers based on Bayes decision theory in gold potential mapping in Sari Gunay epithermal gold deposit. Geochemistry Journal 1(4): 347-355 [In Persian].

[15] Ziaii, M., Abedi, A. and Ziaei, M. (2009). Geochemical and mineralogical pattern recognition and modeling with a Bayesian approach to hydrothermal gold deposits. Applied Geochemistry 24(6): 1142-1146.

[16] Yin, S., Lin, X., Huang, Y., Zhang, Z. and Li, X. (2023). Application of improved support vector machine in geochemical lithology identification. Earth. Sci. Inform. 16: 205–220.

[17] Mahdiyanfar, H., Mohammadpoor, M. and Mahdavi, M. (2022). Determination of alteration genesis and quantitative relationship between alteration and geochemical anomaly using support vector machines. International Journal of Mining and Geo-Engineering 56(1): 33-391.

[18] Trott, M., Leybourne, M., Hall, L. and Layton-Matthews, D. (2022). Random forest rock type classification with integration of geochemical and photographic data. Applied Computing and Geosciences 15: 100090.

[19] Zhang, Y., Ye, X., Xie, S., Dong, J., Yaisamut, O., Zhou, X. and Zhou, X. (2023). Prediction of Au-Polymetallic Deposits Based on Spatial Multi-Layer Information Fusion by Random Forest Model in the Central Kunlun Area of Xinjiang, China. Minerals 13(10): 1302.

[20] Chen, Y. and Zhao, Q., (2021). Mineral exploration targeting by combination of recursive indicator elimination with the ℓ2-regularization logistic regression based on geochemical data. Ore Geology Reviews 135: 104213.

[21] Hanson, D.R. and Lawson, H.E. (2023). Using Machine Learning to Evaluate Coal Geochemical Data with Respect to Dynamic Failures. Minerals 13(6): 808.

[22] Puzyrev, V., Zelic, M. and Duuring, P. (2023). Applying neural networks-based modelling to the prediction of mineralization: A case-study using the Western Australian Geochemistry (WACHEM) database. Ore Geology Reviews 152: 105242.

[23] Tahmooresi, M., Babaei, B. and Dehghan, S. (2022). Geochemical exploration numerical modeling using convolutional neural network (Case study: Gonabad region). Analytical and Numerical Methods in Mining Engineering 12(31): 47-58.

[24] Chen, Y., Zhao, Q. and Lu, L. (2022). Combining the outputs of various k-nearest neighbor anomaly detectors to form a robust ensemble model for high-dimensional geochemical anomaly detection. Journal of Geochemical Exploration 231(1):106875.

[25] Chen, Y. and Lu, L. (2023). The Anomaly Detector, Semi-supervised Classifier, and Supervised Classifier Based on K-Nearest Neighbors in Geochemical Anomaly Detection: A Comparative Study. Math. Geosci. 55: 1011–1033.

Vol. 16, No. 46, Spring 2026, Pages 31-48

Journal of Analytical and Numerical Methods in Mining Engineering

Journal home page: http://anm.yazd.ac.ir/

[26] Parsa, M. (2021). A data augmentation approach to XGboost-based mineral potential mapping: An example of carbonate-hosted Zn-Pb mineral systems of Western Iran. Journal of Geochemical Exploration 228: 106811.

[27] Ibrahim, B., Majeed, F., Ewusi, A. and Ahenkorah, I. (2022). Residual geochemical gold grade prediction using extreme gradient boosting. Environmental Challenges 6: 100421.

[28] Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., and Japkowicz, N. (2024). The class imbalance problem in deep learning. Machine Learning 113: 4845–4901.

[29] Khushi, M., Shaukat, K., Mahboob Alam, T., Hameed, I.A., Uddin, S., and Luo, S. (2021). A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. IEEE Access 9: 109960-109975.

[30] Altalhan, M., Algarni, A., and Turki-Hadj Alouane, M. (2025). Imbalanced Data Problem in Machine Learning: A Review. IEEE Access 13: 13686-13699.

[31] Liu, L., Wu, X., Li, S., Tan, S., and Bai, Y. (2022). Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection. BMC Medical Informatics and Decision Making volume 22: Article number: 82.

[32] Wang, W., and Sun, D. (2021). The improved AdaBoost algorithms for imbalanced data classification. Information Sciences 563: 358-374.

[33] Salehi, A. R., and Khedmati, M. (2024). A cluster-based SMOTE both-sampling (CSBBoost) ensemble algorithm for classifying imbalanced data. Scientific Reports 14(1): 5152.

[34] Araf, I., Idri, A., and Chairi, I. (2024). Cost-sensitive learning for imbalanced medical data: a review. Artificial Intelligence Review 57(4): 80.

[35] Xiao, J., Li, S., Tian, Y., Huang, J., Jiang, X., and Wang, S. (2025). Example dependent cost sensitive learning based selective deep ensemble model for customer credit scoring. Scientific Reports 15(1): 6000.

[36] Liu, Y., Li, Z., Chen, J., Zhang, T., Pan, T., and He, S. (2025). A batch-adapted cost-sensitive contrastive feature learning network for industrial diagnosis with extremely imbalanced data. Measurement 244: 116478.

[37] Abhishek, K. and Abdelaziz, M. (2023). Machine Learning for Imbalanced Data: Tackle imbalanced datasets using machine learning and deep learning techniques, Packt Publishing, 344 p.

[38] Yang, Y., Akbarzadeh Khorshidi, H. and Aickelin, U. (2024). A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems. Front. Digit. Health 26: 1430245.

[39] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research 16: 321–357.

[40] Hu, S., Liang, Y., Ma, L. and He, Y. (2009). MSMOTE: Improving Classification Performance when Training Data is imbalanced. 2009 Second International Workshop on Computer Science and Engineering, 13-17.

[41] Tahmooresi, M., Babaei, B. and Dehghan, S. (2022). Geochemical exploration numerical modeling using convolutional neural network (Case study: Gonabad region). *Journal of Analytical and Numerical Methods in Mining Engineering* 12(31): 47-58.

[42] Kosolwattana, T., Liu, C., Hu, R. Han, S., Chen, H. and Lin, Y. (2023). A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare. BioData Mining 16: 15.

[43] Hengyu, Z. (2020). Improved SMOTE algorithm for imbalanced dataset. Chinese Automation Congress (CAC), Shanghai, China, 693-697.

*Corresponding author: E-mail: *h.geranian@birjandut.ac.ir*

Vol. 16, No. 46, Spring 2026, Pages 31-48

Journal of Analytical and Numerical Methods in Mining Engineering

Journal home page: http://anm.yazd.ac.ir/

[44] Lee, H., Kim, J. and Kim, S. (2017). Gaussian-Based SMOTE Algorithm for Solving Skewed Class Distributions. Int. J. Fuzzy Log. Intell. Syst. 17(4): 229-234.

[45] He, H., Bai, Y., Garcia, E.A. and Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning, IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 1322-1328.

[46] Kurniawati, Y.E., Permanasari, A.E. and Fauziati, S. (2018). Adaptive Synthetic-Nominal (ADASYN-N) and Adaptive Synthetic-KNN (ADASYN-KNN) for Multiclass Imbalance Learning on Laboratory Test Data, 4th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, 1-6.

[47] Qing, Z., Zeng, Q., Wang, H., Liu, Y., Xiong, T. and Zhang, S. (2022). ADASYN-LOF Algorithm for Imbalanced Tornado Samples. Atmosphere, 13(4): 544.

[48] Devi, D., Biswas, S.K. and Purkayastha, B. (2020). A Review on Solution to Class Imbalance Problem: Undersampling Approaches, International Conference on Computational Performance Evaluation (ComPE), Shillong, India, 626-631.

[49] Mazhari, S.A. and Safari, M. (2013). High-K Calc-alkaline Plutonism in Zouzan, NE of Lut Block, Eastern Iran: An Evidence for Arc Related Magmatism in Cenozoic. Journal Geological Society of India 81: 698-708.

[50] Geranian, H. and Carranza, E.J.M. (2022). Mapping of Regional-scale Multi-Element Geochemical Anomalies Using Hierarchical Clustering Algorithms. Natural Resources Research 31(4): 1841-1865.

[51] Seyedrahimi-Niaraq, M., Mahdiyanfar, H. and Mokhtari, A. R. (2023). Application of geochemical structural methods to determine lead-contaminated areas related to mining activities. *Journal of Analytical and Numerical Methods in Mining Engineering* 13(34): 41-55.

[52] Kubat, M. and Matwin, S. (1997). Addressing the course of imbalanced training sets: One-sided selection. Proceedings of the 14th international conference on machine learning, Morgan Kaufmann, pp. 179-186.

[53] Jia, C. and Zuo, Y. (2017). S-SulfPred: A sensitive predictor to capture S-sulfenylation sites based on a resampling one-sided selection undersampling-synthetic minority oversampling technique. Journal of Theoretical Biology 422: 84-89.

[54] Batista, G., Bazzan, A. and Monard, MC. (2003). Balancing Training Data for Automated Annotation of Keywords: A Case Study. II Brazilian Workshop on Bioinformatics, 10-18.

[55] Hart, P.E. (1968). The Condensed Nearest Neighbour Rule. IEEE Transactions on Information Theory 14(5): 515-516.

[56] Hassani Pak, A.A. (2016). Principles of Geochemical Exploration. Tehran University Press, Tehran [In Persian].

[57] Fakhari, S., Jafarirad, A., Afzal, P., and Lotfi, M. (2019). Delineation of hydrothermal alteration zones for porphyry systems utilizing ASTER data in Jebal-Barez area, SE Iran. Iranian Journal of Earth Sciences, 11: 80-92.

[58] Mokhtari, Z., and Seifi, A. (2021). Detection of Hydrothermal Alteration Zones Using ASTER Remote Sensing Data in Turquoise mine of Neyshabur. Journal of Analytical and Numerical Methods in Mining Engineering, 11(28): 1-22 [In Persaian].