Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 8, No. 3, September 2022, pp. 367-375
ISSN: 2338-3070, DOI: 10.26555/jiteki.v8i3.23684

Hybrid Approach-RSMOTE for Handling Class Imbalance with Label Noise

Hartono [1,2], Erianto Ongko [3]
[1] Department of Computer Science, Universitas Potensi Utama, Medan, 20241, Indonesia
[2] Department of Computer Science, Universitas IBBI, Medan, 20114, Indonesia
[3] Department of Informatics, Akademi Teknologi Industri Immanuel, 20114, Medan, Indonesia

ARTICLE INFO

Article history:

Received March 27, 2022
Revised July 08, 2022
Accepted September 14, 2022

Keywords:

Class Imbalance;
Hybrid Approach;
Relative Density;
SMOTE;
RSMOTE

ABSTRACT

The class imbalance problem is the main problem in classification. This issue arises because real-world datasets frequently exhibit an imbalance as a result of a class with more instances than other classes. In handling class imbalance, a Hybrid Approach that blends data-level and algorithm-level approaches produce good results. However, apart from the class imbalance, which reduces classification accuracy, the complexity of the data also has an effect. The complexity of this data causes a minority noise sample which lies between the minority and the majority. In order to determine how close minority samples are to their homogeneous and heterogeneous nearest neighbors, it is necessary to calculate the relative density. The greater the proximity to the homogeneous nearest neighbors, the greater the relative density, which causes the minority samples to be in a safe state but otherwise be categorized as noisy samples. This research will combine the application of the Hybrid Approach with A self-adaptive Robust SMOTE (RSMOTE), which is an adaptive method from SMOTE that applies the concept of relative density in the over-sampling process on minority samples. The research contribution is to implement the Hybrid Approach-RSMOTE in handling class imbalance with noise by using relative density in over-sampling and also to improve classification performance. The results showed that the Hybrid Approach-RSMOTE and Hybrid Approach-SMOTE had given good results in handling class imbalance. However, the Hybrid Approach-RSMOTE gave better results in the Precision, Recall, F1-Measure, and G-Mean and showed significant differences. Based on the results of the study, it can be stated that the performance of the Hybrid Approach in handling class imbalance is influenced by the selection of the over-sampling method. The results show that RSMOTE can be considered an over-sampling method in the Hybrid Approach.

Corresponding Author:

Hartono, Department of Computer Science, Universitas Potensi Utama, Medan, 20241, Indonesia
Email: hartonoibbi@gmail.com

1. INTRODUCTION

Real-world datasets frequently have issues with class imbalance brought about by an unequal distribution of data, which causes one class to have more instances than other classes [1]. This trend is unavoidable, as is the case in medical datasets where the number of positive samples is definitely less than the number of negative samples [2]. An imbalanced dataset is characterized by a large dataset containing uncertain and inconsistent elements [3]. This class imbalance problem has become an interesting challenge, especially in data mining and machine learning [4]. The existence of a class imbalance can reduce the accuracy of the classification, and also the most worrying thing is that often the minority class is a class with interesting instances. A number of algorithms and approaches to classification always assume that the sample is evenly distributed in each class so that if there is a class imbalance problem, the classification process will give better results for the majority

class and ignore the minority class [5]. The existence of a class imbalance can cause an interesting pattern that exists in the minority class cannot be obtained [6]. This causes it to be more important to correct misclassification in the minority class than misclassification in the majority class [7]. In handling class imbalance, it is important to pay attention to the accuracy of prediction in the minority class without neglecting the accuracy in the majority class [8].

The approach used in dealing with class imbalance can be divided into 2 (two), namely: data-level and algorithm-level [9]. By oversampling members of the minority class and undersampling members of the majority class, the data-level approach changed the distribution of the data [10]. Data-Level is classifier independent in the sense that the used classifier only classifies imbalanced datasets that have undergone resampling [11]. The algorithm-level approach is carried out by modifying the training and testing procedures on the classification algorithm, which aims to reduce the negative impact on minority class performance [12]. The workings of the algorithm level are shown in the form of using a classifier to improve the classification results for the minority class and increase the sensitivity of the classification results [13]. The hybrid approach combines the advantages offered by the data level in terms of modifying the number of instances in both majority and minority with the advantages of algorithm level in terms of generating a number of classifiers [14]. Included in the Hybrid Approach is the ensemble method using bagging or boosting the classifier [15]. Included in this Hybrid Approach are strategies for handling other perspectives, such as noise reduction [16].

Among the sampling methods, over-sampling, especially SMOTE, is the most popular method to use because the balancing results on training data are better when compared to under-sampling, as shown in the better ROC Curve (AUC) measurement results [17]. In general, SMOTE is more likely to work on feature space than data space [18]. To produce a new synthetic minority sample, the next step is to interpolate between the selected minority samples [19]. The main problem with SMOTE is the oversampling process on uninformative samples and the presence of noise [20]. Although the results of data balancing provided by SMOTE are quite good, there are other problems that arise related to dataset complexity which causes a minority noise sample that lies between the minority and the majority [21]. SMOTE's weakness in noise is caused by SMOTE not yet distinguishing the minority sample types into Safe, Border, and noisy [22]. Noise can be defined as an error in the identification of the label (noise in the class) or an error in the value of the attribute (noise in the attribute) [23]. The emergence of this noise is also caused by the tendency of lack of attention to the existence of overlapping classes and the absence of a determination of safe areas and borderline areas [24]. In addition, naturally, real-world datasets also have a tendency to have noise that can affect the performance of the classification [25]. The minority class will be more affected by noise's effects on classification outcomes than the majority class [26].

A number of researchers have attempted to modify SMOTE so that it can overcome the noise problem. One of them is Borderline-SMOTE which attempts to identify samples into 3 (three) areas, namely: safe, borderline, and noise [27]. This method has been successful in overcoming the noise problem, but another problem that arises is that when there are minority samples in the boundary area where there are no majority samples nearby, these samples will be difficult to obtain [28]. A number of other methods, such as Safe-Level-SMOTE [29], Local Neighborhood SMOTE (LN-SMOTE) [30], Over-sampling Using Propensity Scores (OUPS) [31], and Safe Level OUPS [32], are also constrained in trying to determine the ideal distance so that a sample is said to be in a safe area. Research on SMOTUNED, which is a SMOTE with hyperparameter optimization [33], gives good results but is very dependent on the distance-determining variable, so the results given can vary greatly if the distance determination is not done carefully [34].

Banerjee et al. [35] Propose the Fused Oversampling Framework by addressing the Outliers (FOFO) method, where outliers are detected to minimize them first by using IQR, and then the balancing process with SMOTE is carried out. However, this study did not consider the number of samples in the underlying dataset, so it is necessary to question the results if applied to datasets with different imbalance ratios. Other researchers use the Noise-Adaptive Synthetic Oversampling (NASO) method. This research is based on the calculation of the noise ratio for each sample in the minority class. Based on this noise ratio, positive samples will then be synthesized. The limitation of this method is that there is a limited number of samples due to the ability of the classifier to use SVM, so it is necessary to consider when using another classifier [36].

A self-adaptive Robust SMOTE (RSMOTE) uses the concept of relative density to measure the proximity of the minority samples to the heterogenous nearest neighbors and homogeneous nearest neighbors. The RSMOTE method starts by determining absolute density for homogeneous and heterogeneous nearest neighbors. Then the relative density will be determined by dividing the absolute density of the homogeneous neighbor by the absolute density value of the heterogeneous neighbor. The greater the relative density value, the greater the confidence to enter the sample into the safe area, and vice versa. The smaller the relative density value, the greater the confidence to enter the sample into the noise [28].

This research will propose the Hybrid Approach-RSMOTE method for handling class imbalance with noise. A number of assessment metrics will be used in this research, namely: Precision, Recall, F1-Measure, and G-Mean. This study will compare the results obtained with the Hybrid Approach-SMOTE. The research contribution is to implement the Hybrid Approach-RSMOTE in handling class imbalance with noise by using relative density in over-sampling and also to improve classification performance.

2. METHOD

The stages of the research can be seen in Fig. 1 which the process starts with determining the dataset that will be used in the study. After that, the research stage continues with the preprocessing process on the dataset, which is carried out using the Random Balance Ensemble Method. The purpose of this preprocessing stage is to prepare the dataset to undergo the processing stage. In this preprocessing stage, a number of classifiers are generated for both majority and minority classes. Then after that, the processing stage will be carried out using RSMOTE. The application of RSMOTE is expected to minimize noise by determining relative density so that it can determine the safe area and noise. The classification results will then be compared with the Hybrid Approach-SMOTE.

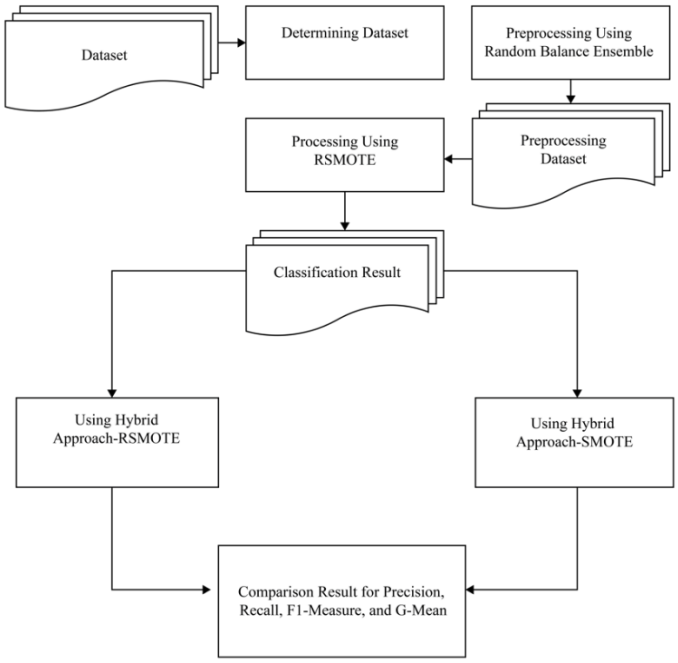

Fig. 1. Research Stages

2.1. Hybrid Approach

The pseudocode of the Hybrid Approach is as follows [37].

$Input: D_T = \{x_1, x_2, \dots, x_n\} // Training\ Dataset$
$N = Number\ of\ Classifier$
$Output: Classification\ Prediction\ P$
$Method:$
$Step\ 1\ Preprocessing\ using\ Preprocessing\ Method$
$Step\ 2\ For\ i = 1\ to\ N\ do$
$\quad i. Apply\ Machine\ Learning\ Classification\ Algorithm\ on\ The\ Attributes\ of\ D_T$
$\quad ii. Obtain\ Classification\ Prediction\ P_i\ from\ machine\ learning\ classification\ algorithm$
$End\ For$
$Step\ 3\ For\ i = 1\ to\ n$
$\quad Apply\ processing\ using\ bagging, boosting\ or\ sampling$
$End\ For$

2.2. Absolute Density

Absolute Density can be determined using Eq. (1) [28]. In the equation, p represents the dataset, k represents the nearest neighbors, $d(p,q)$ represents the distance between points p and q, and $N_k(p,D)$ represents k nearest neighbors from p in D. Absolute density represents the smallest distance between p and its nearest neighbors in D.

$$AbsoluteDensity\ (p,D) = \frac{k}{\sum d(p,q)}, for\ q \in N_k(p,D) \quad (1)$$

2.3. Relative Density

Relative Density can be determined using Eq. (2) [28]. In the equation, it can be seen that the relative density is obtained from the calculation of the distance between the minority sample p with the value of the k-nearest homogeneous neighbor q_i, $i = 1,2, \dots, k$ (absolute density from a homogeneous neighbor) divided by the calculation of the distance between the minority sample p and the value of k-nearest heterogeneous neighbors (absolute density from a heterogeneous neighbor).

$$RelativeDensity\ (p) = \frac{AbsoluteDensity(p,D+)}{AbsoluteDensity(p,D-)} \quad (2)$$

2.4. RSMOTE

The pseudocode of RSMOTE is as follows [28]. Based on the pseudocode, it can be seen that RSMOTE begins with determining the number of minority samples that need to be generated based on the result of dividing the number of majority samples with the imbalance ratio. Then do the iteration for each instance on the minority sample to calculate the relative density value of each existing instance. After the iteration process is complete, divide the existing clusters into two clusters based on the relative density value obtained. P_A cluster for instances with a higher relative density value and P_B cluster for instances with a smaller relative density. Then for each cluster, do the process of determining the number of minority samples that need to be raised. Combine the synthesis results from the two clusters to become a result dataset.

$Input: Dataset\ D, Minority\ Samples\ P, Majority\ Samples\ Q,$
$Imbalance\ Ratio\ IR, Number\ of\ Nearest\ Neighbors\ K$
$Output: Synthetic\ Samples\ Syn$
$Process:$
$N = \frac{Q}{IR}$
$for\ i = 1\ to\ number\ of\ P\ do$
  $P' \leftarrow P' \cup \{x_i\}$
  $Calculate\ RD(P')\ using\ Equation\ (2)$
$end\ for$
$Divide\ P'\ into\ two\ cluster\ C_A\ and\ C_B\ according\ to\ RD(P')\ where\ C_A \geq C_B$
$P'_A \leftarrow Minority\ Samples\ Correspond\ to\ C_A$
$P'_B \leftarrow Minority\ Samples\ Correspond\ to\ C_B$
$For\ each\ cluster\ P'_j, j \in \{A, B\}\ do$
  $N_j \leftarrow \frac{|P'_j|}{|P'|}.N$
  $For\ Each\ Sample\ x_i\ in\ P'_j\ do$
    $w_i \leftarrow \frac{k-m}{k+1}$
    $\bar{W}_i \leftarrow w_i / \sum_{i=1}^{|P'_j|} w_i$
    $N_i \leftarrow \bar{W}_i.N_j$
    $Populate\ (N_i, i, nnarray)$
  $End\ For$
  $Syn_j \leftarrow Syn \cup Generated\ Samples\ by\ every\ minority\ samples\ in\ P'_j$
$End\ For$

2.5. Dataset

The KEEL Repository contributed to the dataset that was used in this investigation [38]. Table 1 shows the dataset that was tested in this investigation which shows 6 (six) datasets that will be tested in this study. The dataset used in this study has diversity in terms of dimensions, samples, and also imbalance ratio. By using this varied dataset, it is hoped that the results of the study will describe the handling of class imbalance better.

Table 1. Dataset

Dataset	Dimension	Sample	Imbalance Ratio
Iris	4	150	2
Haberman	3	306	2.78
Vowel	13	988	10.10
Ecoli4	7	336	13.84
Shuttle0vs4	9	1829	13.87
Yeast6	8	1484	39.15

2.6. Confusion Matrix

Table 2 contains the confusion matrix [39]. The Confusion Matrix is typically used to assess how well the classification results are performed. The confusion matrix in Table 2 is a matrix with 4 distinct combinations of expected values and actual values. In the confusion matrix, the classification process' outcomes are denoted by four terms: True Positive, True Negative, False Positive, and False Negative.

Table 2. Confusion Matrix

	Predictive Positive Class	Predictive Negative Class
Actual Positive Class	True Positive (TP)	False Negative (FN)
Actual Negative Class	False Positive (FP)	True Negative (TN)

2.7. Assessment Metric

The assessment metrics used in this study are Precision, Recall, F1-Measure, and G-Mean. The determination of the assessment metric can be seen in Eq. (3-6) [40].

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - Measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{5}$$

$$G - Mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \tag{6}$$

3. RESULTS AND DISCUSSION

3.1. Experimental Setup

On the datasets listed in the previous section, performance testing of the suggested approach is carried out. Traditional performance indicators like Precision, Recall, F1-Measure, and G-Mean are used to evaluate the system. A stratified k-fold ($k = 10$) was used to conduct the analysis. This study compares the results obtained by the Hybrid Approach-RSMOTE with the results obtained by the Hybrid Approach-SMOTE as research conducted by Polat [41].

3.2. Testing Results for Precision and Recall

Table 3 shows the results of the Precision and Recall tests, which shows that in all datasets studied. The Hybrid Approach-RSMOTE outperforms the Hybrid Approach-SMOTE in terms of Precision and Recall. With an increase in the Imbalance Ratio, the Precision and Recall values achieved have a tendency to decline. Both the Hybrid Approach-RSMOTE and the Hybrid Approach-SMOTE have this problem.

3.3. Testing Results for F1-Measure and G-Mean

Table 4 shows the results of the F1-Measure and G-Mean tests. In Table 4, it can be seen that the Hybrid Approach-RSMOTE still gives better results when compared to the Hybrid Approach-SMOTE, and this occurs in all tested datasets.

3.4. Statistical Tests

Statistical Tests were carried out using the Wilcoxon Signed-Rank Test. The results of the statistical tests can be seen in Table 5. Based on Table 5, it can be seen that the statistical test results with the Wilcoxon Signed-Rank Test indicate that the results obtained by the Hybrid Approach-RSMOTE provide significant differences from the Hybrid Approach-SMOTE. This shows that the Hybrid Approach-RSMOTE provides better results in handling class imbalance and noise compared to the Hybrid Approach-SMOTE.

Table 3. Precision and Recall Test Results

Dataset	Hybrid Approach-RSMOTE		Hybrid Approach-SMOTE	
	Precision	Recall	Precision	Recall
Iris	0.967	0.937	0.957	0.915
Haberman	0.957	0.948	0.942	0.937
Vowel	0.927	0.922	0.911	0.916
Ecoli4	0.917	0.921	0.897	0.916
Shuttle0vs4	0.907	0.901	0.875	0.872
Yeast6	0.895	0.887	0.867	0.843

Table 4. F1-Measure and G-Mean Test Results

Dataset	Hybrid Approach-RSMOTE		Hybrid Approach-SMOTE	
	F1-Measure	G-Mean	F1-Measure	G-Mean
Iris	0.952	0.947	0.936	0.932
Haberman	0.952	0.948	0.939	0.941
Vowel	0.924	0.917	0.913	0.901
Ecoli4	0.919	0.921	0.906	0.903
Shuttle0vs4	0.904	0.911	0.873	0.881
Yeast6	0.891	0.889	0.855	0.851

Table 5. Statistical Test Using Wilcoxon Signed-Rank Test

	P-Value	Significant
Precision	0.0312500	Both methods show a significant difference because the P-Value>0.05
Recall	0.0312500	
F1-Measure	0.0355223	
G-Mean	0.0312500	

3.5. Discussion

Based on the test results with Precision, Recall, F1-Measure, and G-Mean, it can be seen that the Hybrid Approach-RSMOTE gives better results when compared to the Hybrid Approach-SMOTE. The tendency of these two methods is that the results obtained tend to decrease as the imbalance ratio increases. The difference given by these two methods is quite significant on the basis of the Wilcoxon Signed-Rank Test result

4. CONCLUSION

Based on the results of the study, it was found that the selection of the over-sampling method had an effect on the results obtained in handling class imbalance with the Hybrid Approach. RSMOTE, which uses the concept of relative density, tends to give better results when compared to SMOTE. This can be attributed to RSMOTE's ability to handle the complexity of data associated with noise labels by calculating the relative density to measure the proximity of the minority samples to the heterogenous nearest neighbors and homogeneous nearest neighbors. The results obtained by the Hybrid Approach-RSMOTE are better and significantly different when compared to the Hybrid Approach-SMOTE. Future research is expected to improve the ability of the Hybrid Approach, especially if there is an increase in the Imbalance Ratio so that the results obtained do not decrease.

Acknowledgments

The authors acknowledge the support of the Indonesian Ministry of Education, Culture, Research, and Technology's Directorate of Research and Development for this study.

REFERENCES

[1] S. Fan, X. Zhang, and Z. Song, "Reinforced knowledge distillation: Multi-class imbalanced classifier based on policy gradient reinforcement learning," *Neurocomputing*, vol. 463, pp. 422–436, Nov. 2021, https://doi.org/10.1016/j.neucom.2021.08.040.

[2] L.-C. Hung, Y.-H. Hu, C.-F. Tsai, and M.-W. Huang, "A dynamic time warping approach for handling class imbalanced medical datasets with missing values: A case study of protein localization site prediction," *Expert Systems with Applications*, vol. 192, p. 116437, Apr. 2022, https://doi.org/10.1016/j.eswa.2021.116437.

[3] R. Malhotra and M. Khanna, "An empirical study for software change prediction using imbalanced data," *Empir Software Eng*, vol. 22, no. 6, pp. 2806–2851, Dec. 2017, https://doi.org/10.1007/s10664-016-9488-7.

[4] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, "SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering," *Information Sciences*, vol. 291, pp. 184–203, Jan. 2015, https://doi.org/10.1016/j.ins.2014.08.051.

[5] 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, Jan. 2021, https://doi.org/10.1016/j.ins.2020.07.014.

[6] I. D. Mienye and Y. Sun, "Performance analysis of cost-sensitive learning methods with application to imbalanced medical data," *Informatics in Medicine Unlocked*, vol. 25, p. 100690, Jan. 2021, https://doi.org/10.1016/j.imu.2021.100690.

[7] J. A. Sanz, D. Bernardo, F. Herrera, H. Bustince, and H. Hagras, "A Compact Evolutionary Interval-Valued Fuzzy Rule-Based Classification System for the Modeling and Prediction of Real-World Financial Applications With Imbalanced Data," *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 4, pp. 973–990, Aug. 2015, https://doi.org/10.1109/TFUZZ.2014.2336263.

[8] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *Journal of artificial intelligence research*, vol. 61, pp. 863–905, Apr. 2018, https://doi.org/10.1613/jair.1.11192.

[9] K. De Angeli *et al.*, "Class imbalance in out-of-distribution datasets: Improving the robustness of the TextCNN for the classification of rare cancer types," *Journal of Biomedical Informatics*, vol. 125, p. 103957, Jan. 2022, https://doi.org/10.1016/j.jbi.2021.103957.

[10] M. Koziarski, "Potential Anchoring for imbalanced data classification," *Pattern Recognition*, vol. 120, p. 108114, Dec. 2021, https://doi.org/10.1016/j.patcog.2021.108114.

[11] I. Czarnowski, "Weighted Ensemble with one-class Classification and Over-sampling and Instance selection (WECOI): An approach for learning from imbalanced data streams," *Journal of Computational Science*, vol. 61, p. 101614, May 2022, https://doi.org/10.1016/j.jocs.2022.101614.

[12] F. Li, X. Zhang, X. Zhang, C. Du, Y. Xu, and Y.-C. Tian, "Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets," *Information Sciences*, vol. 422, pp. 242–256, Jan. 2018, https://doi.org/10.1016/j.ins.2017.09.013.

[13] J. Liu, "Importance-SMOTE: a synthetic minority oversampling method for noisy imbalanced data," *Soft Comput*, vol. 26, no. 3, pp. 1141–1163, Feb. 2022, https://doi.org/10.1007/s00500-021-06532-4.

[14] A. Mahmoud, A. El-Kilany, F. Ali, and S. Mazen, "TGT: A Novel Adversarial Guided Oversampling Technique for Handling Imbalanced Datasets," *Egyptian Informatics Journal*, vol. 22, no. 4, pp. 433–438, Dec. 2021, https://doi.org/10.1016/j.eij.2021.01.002.

[15] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, "Boosting methods for multi-class imbalanced data classification: an experimental review," *Journal of Big Data*, vol. 7, no. 1, p. 70, Sep. 2020, https://doi.org/10.1186/s40537-020-00349-y.

[16] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J Big Data*, vol. 6, no. 1, p. 27, Mar. 2019, https://doi.org/10.1186/s40537-019-0192-5.

[17] S. García, J. Luengo, and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," *Knowledge-Based Systems*, vol. 98, pp. 1–29, Apr. 2016, https://doi.org/10.1016/j.knosys.2015.12.006.

[18] Asniar, N. U. Maulidevi, and K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," *Journal of King Saud University - Computer and Information Sciences*, Feb. 2021, https://doi.org/10.1016/j.jksuci.2021.01.014.

[19] 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 and Information Sciences*, Jun. 2022, https://doi.org/10.1016/j.jksuci.2022.06.005.

[20] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE–Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 2, pp. 405–425, Feb. 2014, https://doi.org/10.1109/TKDE.2012.232.

[21] 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, https://doi.org/10.1109/TKDE.2008.239.

[22] F. Sağlam and M. A. Cengiz, "A novel SMOTE-based resampling technique trough noise detection and the boosting procedure," *Expert Systems with Applications*, vol. 200, p. 117023, Aug. 2022, https://doi.org/10.1016/j.eswa.2022.117023.

[23] C. M. Salgado, C. Azevedo, H. Proença, and S. M. Vieira, "Noise Versus Outliers," in *Secondary Analysis of Electronic Health Records*, pp. 163-183, 2016, https://doi.org/10.1007/978-3-319-43742-2_14.
[24] W. A. Rivera, "Noise Reduction A Priori Synthetic Over-Sampling for class imbalanced data sets," *Information Sciences*, vol. 408, pp. 146–161, Oct. 2017, https://doi.org/10.1016/j.ins.2017.04.046.
[25] S. Gupta and A. Gupta, "Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review," *Procedia Computer Science*, vol. 161, pp. 466–474, Jan. 2019, https://doi.org/10.1016/j.procs.2019.11.146.
[26] 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, May 2022, https://doi.org/10.1016/j.ins.2022.02.038.
[27] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," in *Advances in Intelligent Computing*, pp. 878–887, 2005, https://doi.org/10.1007/11538059_91.
[28] 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, Apr. 2021, https://doi.org/10.1016/j.ins.2020.10.013.
[29] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem," in *Advances in Knowledge Discovery and Data Mining*, pp. 475–482, 2009, https://doi.org/10.1007/978-3-642-01307-2_43.
[30] T. Maciejewski and J. Stefanowski, "Local neighbourhood extension of SMOTE for mining imbalanced data," in *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, Apr. 2011, pp. 104–111, Apr. 2011, https://doi.org/10.1109/CIDM.2011.5949434.
[31] W. A. Rivera, A. Goel, and J. P. Kincaid, "OUPS: A Combined Approach Using SMOTE and Propensity Score Matching," in *2014 13th International Conference on Machine Learning and Applications*, pp. 424–427, Dec. 2014, https://doi.org/10.1109/ICMLA.2014.106.
[32] W. A. Rivera and O. Asparouhov, "Safe level OUPS for improving target concept learning in imbalanced data sets," in *SoutheastCon 2015*, pp. 1–8, Apr. 2015, https://doi.org/10.1109/SECON.2015.7132940.
[33] A. Agrawal and T. Menzies, "Is "Better Data" Better than "Better Data Miners"?," *International Conference of Software Engineering (ICSE)*, pp. 1050-1061, 2018, https://doi.org/10.1145/3180155.3180197.
[34] 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, p. 106742, Feb. 2022, https://doi.org/10.1016/j.infsof.2021.106742.
[35] A. Banerjee, K. Ghosh, S. Chatterjee, and D. Sen, "FOFO: Fused Oversampling Framework by addressing Outliers," in *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pp. 238–242, Mar. 2021, https://doi.org/10.1109/ESCI50559.2021.9397056.
[36] M. T. Vo, T. Nguyen, H. A. Vo, and T. Le, "Noise-adaptive synthetic oversampling technique," *Appl Intell*, vol. 51, no. 11, pp. 7827–7836, Nov. 2021, https://doi.org/10.1007/s10489-021-02341-2.
[37] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, Jul. 2012, https://doi.org/10.1109/TSMCC.2011.2161285.
[38] J. Alcalá-Fdez, L. Sánchez, S. García, M. J. del Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit, V. M. Rivas, J. C. Fernández, and F. Herrera, "KEEL: a software tool to assess evolutionary algorithms for data mining problems," *Soft Comput*, vol. 13, no. 3, pp. 307–318, Feb. 2009, https://doi.org/10.1007/s00500-008-0323-y.
[39] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Assessing feature selection method performance with class imbalance data," *Machine Learning with Applications*, vol. 6, p. 100170, Dec. 2021, https://doi.org/10.1016/j.mlwa.2021.100170.
[40] M. Mohamad, A. Selamat, I. M. Subroto, and O. Krejcar, "Improving the classification performance on imbalanced data sets via new hybrid parameterisation model," *Journal of King Saud University - Computer and Information Sciences*, Apr. 2019, https://doi.org/10.1016/j.jksuci.2019.04.009.
[41] K. Polat, "A Hybrid Approach to Parkinson Disease Classification Using Speech Signal: The Combination of SMOTE and Random Forests," in *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pp. 1–3, Apr. 2019, https://doi.org/10.1109/EBBT.2019.8741725.

BIOGRAPHY OF AUTHORS

Hartono earned his doctorate in computer science from Universitas Sumatera Utara in 2018, after earning his master's degree in computer science from Universitas Putra Indonesia YPTK Padang in 2010 and his bachelor's degree in computer science from STMIK IBBI Medan in 2008. He now works as a lecturer at Universitas IBBI and Universitas Potensi Utama. Machine Learning, Artificial Intelligence, Data Mining, and Operational Research are some of his current research interests.

Erianto Ongko earned his master's degree in computer science from Universitas Sumatera Utara in Medan, Indonesia, in 2015 and his bachelor's degree in computer science from STMIK IBBI in 2012. He works at Akademi Teknologi Industri Immanuel as a lecturer. Machine Learning, Artificial Intelligence, and Operational Research are his current research interests.