Riemann: *Research of Mathematics and Mathematics Education*
Volume 8, Issue 1, 288-300, April 2026
e-ISSN: 2721-883X|p-ISSN: 2721-8848
https://journal.sanagustin.ac.id/index.php/reimann

Improving Classification Performance of Imbalanced Data Using SMOTE: empirical studies

Ulfasari Rafflesia*[1], Dedi Rosadi[2]
[1]Department of Mathematics, Universitas Bengkulu, Indonesia
[2]Department of Mathematics, Xiamen University Malaysia, Malaysia
Email Address: ulfasari@unib.ac.id

*Corresponding author

Received: February 2026 | Revised: March 2026 | Accepted: April 2026 | Published: April 2026

***Abstract*:**
Data balancing methods in multi-class settings continue to evolve as the importance of balanced data conditions for classification analysis grows. However, limited studies have provided comprehensive empirical comparisons across both binary and multi-class imbalanced datasets. Data imbalance can affect model predictions, particularly by leading to inaccurate identification of minority classes. Therefore, this study aims to evaluate the effectiveness of the Synthetic Minority Over-sampling Technique (SMOTE) in improving classification performance. Three benchmark datasets from the UCI Machine Learning Repository—Breast Cancer, Ecoli, and Glass—were selected to represent imbalanced classification problems in both binary and multi-class settings. The proposed framework addresses class imbalance during data preprocessing using SMOTE. Each dataset is first divided into training and testing subsets. SMOTE is applied only to the training data to address class imbalance, while the test data is kept unchanged for evaluation. Then, the classification process is applied to the original (imbalanced) data and to the balanced data generated by SMOTE. The classifiers used in this study are SVM, a decision tree, and AdaBoost. The classification results are evaluated based on accuracy, sensitivity, and F1-score. The results show that the decision tree and AdaBoost improve classification performance under imbalanced data conditions. In particular, AdaBoost achieves the best overall performance in terms of prediction accuracy and class balance, demonstrating the effectiveness of combining SMOTE with ensemble methods for handling imbalanced datasets.

Keywords: Imbalanced data, SMOTE, SVM, Decision Tree, AdaBoost.

Introduction

The Synthetic Minority Over-sampling Technique (SMOTE) has demonstrated efficacy in enhancing model performance in numerous instances; nevertheless, most prior research has concentrated on binary classification issues or employed a singular classification technique (Chen et al., 2022). Classification is a primary machine learning task, extensively applied across diverse domains, including medical diagnosis, pattern identification, and scientific data analysis (Kumar et al., 2020; Pratap Chandra et al., 2020; Rosadi et al., 2021). The data distribution utilized during model training significantly affects the caliber of classification outcomes. A prevalent issue is class

imbalance, characterized by a disproportionate number of samples across classes (Fernandez et al., 2018; Ha et al., 2011). This disparity may lead the model to favor the majority class, diminishing its efficacy in identifying the minority class (Chawla et al., 2008; Rezvani & Wang, 2023). This issue arises not only in binary classification but also in multi-class scenarios with more intricate data distributions.

Various approaches have been proposed to address class imbalance, including class weighting, algorithm modification, and data preprocessing techniques such as oversampling and undersampling. SMOTE is a widely utilized method that creates synthetic samples for the minority class by using data from nearest neighbors (Blagus & Lusa, 2013; Last et al., 2017). Despite its effectiveness, existing studies largely focus on binary classification or evaluate individual classifiers separately (S. Wang et al., 2021). Comprehensive studies comparing the effectiveness of SMOTE across multiple classifiers and both binary and multi-class datasets remain limited, particularly when evaluation emphasizes the identification of minority classes.

Therefore, this study aims to evaluate the effect of implementing SMOTE on the performance of classification algorithms under imbalanced data conditions. This study evaluates the effect of implementing SMOTE on the performance of three classification algorithms—Support Vector Machine (SVM), decision tree, and Adaptive Boosting (AdaBoost)—on imbalanced data. While previous studies have examined SMOTE or individual classifiers separately, this study provides a systematic comparison across multiple classifiers and datasets to determine the most effective approach for improving classification performance under imbalanced conditions. The analysis focuses on improving minority class sensitivity while maintaining overall model performance. Three benchmark datasets from the UCI Machine Learning Repository—Breast Cancer, Ecoli, and Glass—were chosen to represent imbalanced classification problems in both binary and multi-class settings.

This study contributes by: (1) integrating SMOTE with three classification algorithms and evaluating their performance across different data characteristics, (2) providing a comparative analysis between imbalanced and SMOTE-balanced data, and (3) emphasizing sensitivity and F1-score as key metrics for assessing minority class performance. The findings of this study are expected to provide practical insights for real-world applications, such as medical diagnosis and anomaly detection, where accurate identification of minority classes is critical. The remainder of this paper is organized as follows: Section 2 describes the research methods; Section 3 presents the results and discussion; and Section 4 concludes the conclusion with recommendations for future research.

Research Methods

The overall experimental procedure in this study is as follows. The datasets used in this study are obtained from the UCI Machine Learning Repository. Each dataset is divided into training and testing sets. SMOTE is applied only to the training data to

address class imbalance, while the test data remains unchanged. Subsequently, classification models (SVM, CTree, and AdaBoost) are trained using the processed training data with predefined parameter settings. The performance of each model is then evaluated on the testing data using a confusion matrix, from which accuracy, precision, recall, and F-measure are computed.

Synthetic Minority Oversampling Technique (SMOTE)

Data imbalance transpires when the quantity of observations in one data class exceeds that of another. The class with the most observations is called the majority class, and the others are designated as minority classes. This method is an advancement of oversampling, categorized within the data-level approach introduced by (Chawla et al., 2002). Conventional classification techniques may exhibit bias towards the majority class, as these methods are inclined to predict instances from the majority class, often disregarding minority classes as noise, which leads to improper classification of observations from minority classes (Galar et al., 2012). This phenomenon is frequently referred to as the accuracy paradox. Consequently, before modeling, imbalanced data management was executed using resampling (Boonamnuay et al., 2018).

A method to address data imbalance is the Synthetic Minority Oversampling Technique (SMOTE). SMOTE is an oversampling technique that generates synthetic data for the minority class using the k-nearest neighbors algorithm, thereby achieving a more equitable distribution of classes (Chawla et al., 2002). In this study, SMOTE is applied to the training data after the train–test split, using k=5 nearest neighbors with proportional oversampling and undersampling (perc.over = 10 and perc.under = 10). The distance calculation between samples depends on the attribute type; SMOTE uses the Euclidean distance for numerical variables.

$$D(x,y) = \sqrt{(x-y)'(x-y)} \tag{1}$$

For categorical attributes, SMOTE employs the Value Distance Metric (VDM):

$$D(x,y) = \sum_{i=1}^{N} \delta(v_{1i}, v_{2i}) \tag{2}$$

where the attribute-level distance is defined as:

$$\delta(v_1, v_2) = \sum_{i=1}^{S} \left| \frac{C_{1i}}{C_i} - \frac{C_{2i}}{C_2} \right| \tag{3}$$

Through these distance computations, novel minority-class samples are generated between an instance and its closest neighbors, resulting in a more equitable dataset and enhancing classifier efficacy.

Support Vector Machine (SVM)

The SVM method is a supervised machine learning technique grounded in statistical learning theory (Yang Wendong et al., 2017). It extracts a collection of distinctive subsets from the training samples, ensuring that the classification of these subsets corresponds to the segmentation of the complete dataset. The SVM has been effectively employed to address various classification challenges across numerous applications (Maulud & Abdulazeez, 2020; Tao et al., 2018).

In principle, SVM constructs an optimal separating hyperplane by maximizing the margin between two classes. However, in many real-world classification problems, including medical datasets such as breast cancer data, the classes are not linearly separable in the original feature space. To overcome this limitation, SVM applies the kernel trick to implicitly map the data into a higher-dimensional feature space, enabling the formation of a nonlinear decision boundary.

In this study, the Radial Basis Function (RBF) kernel is employed due to its capability to handle nonlinear patterns in numerical data. The SVM model is implemented using C-classification with a one-against-one multiclass strategy, where the parameters are set to C=1 (default) and γ =0.3 without additional tuning. The RBF kernel measures the similarity between two observations based on their Euclidean distance, where one observation may represent a support vector and the other a data point to be classified (Cortes & Vapnik, 1995; Park & Sandberg, 1991). The RBF kernel function is defined as:

$$K(x_i, x_j) = exp(-\gamma \left\| x_i - x_j \right\|^2), \gamma > 0 \quad (4)$$

where x_i and x_j are fature vectors, $\|.\|$ denotes the Euclidean norm, and γ is a kernel parameter controlling the area of influence of each support vector.

A larger value of γ results in a narrower area of influence, leading to a more complex decision boundary, while a smaller γ produces a smoother decision surface. The optimal value of γ is determined during the training process to achieve the best classification performance. The trained SVM model with RBF kernel is then applied to the testing dataset to evaluate its generalization ability.

Decision Tree

A decision tree is a classification model that organizes instance attribute values into class labels via a hierarchical framework. The model comprises test nodes that assess attributes and leaf nodes that denote anticipated classes (Myles et al., 2004; Quinlan, 1996).

Trees are typically constructed using a divide-and-conquer methodology: if all instances belong to a single class, a leaf node is created; if not, an attribute test partitions the data, and the procedure is recursively reiterated. For categorical features, divisions are determined by values, whereas for numerical attributes, threshold-based splits are employed.

The effectiveness of a tree largely depends on how the splitting variable is selected at each node. Various approaches exist in the literature for determining optimal splits, and simpler trees are generally preferred for their interpretability and better generalization ability.

In this study, the classification model is implemented using the Conditional Inference Tree (CTree) algorithm. Unlike traditional decision tree methods such as CART or ID3, which rely on impurity measures (e.g., the Gini index or information gain), CTree employs a statistical testing framework to select optimal splits (Hothorn et al., 2015). The model is built using default control parameters, with variable selection based on permutation tests at a significance level of 0.05 (mincriterion = 0.95), without additional tuning.

The tree construction follows a recursive partitioning procedure based on hypothesis testing. At each node, the algorithm performs the following steps:

1. Test the global null hypothesis of independence between any predictor variable and the response variable.
2. If the hypothesis cannot be rejected, the recursion stops and the node becomes a leaf.
3. Otherwise, select the predictor variable with the strongest association to the response variable based on permutation test p-values.
4. Determine the optimal binary split for the selected variable.
5. Partition the data according to the split and repeat the procedure recursively for each subset.

This statistical stopping criterion prevents overfitting and eliminates the bias toward variables with many possible splits, which is common in traditional tree algorithms. As a result, CTree produces an interpretable classification structure with improved generalisation performance.

AdaBoost

AdaBoost is the preeminent ensemble boosting technique (Cao et al., 2013). Boosting techniques enhance accuracy in classification and prediction by producing model combinations. AdaBoost is extensively utilized across various domains due to its robust theoretical framework, precise predictions, and ease of implementation (Bian et al., 2024; Gamil et al., 2024; R. Wang, 2012). The chosen classification or prediction outcome is the model characterized by attributes comprising weight values (Domingo & Watanabe, 2000; Jiang et al., 2022). The weighting in that procedure signifies the significance of an object for accurate classification. The chosen classification outcome is the model with the highest weight value.

The fundamental classifier in the AdaBoost algorithm is referred to as a weak learner, specifically a decision stump. A stump, or small tree, is a classification tree with solely a root node and two leaves. Each stump is fabricated in succession. The

prediction error produced by a stump influences the development of succeeding stumps. This indicates that misclassified observations will receive increased emphasis in the subsequent phase. The ensemble method is executed by evaluating the voting weight of each stump.

The procedure for building an AdaBoost classification model is as follows (Haties et al., 2009).

1. Initialize equal weights for each observation in the training dataset (x_i, y_i), for $i = 1, 2, \ldots, N$:

$$w_i = \frac{1}{N}. \tag{5}$$

2. For each iteration $i = 1, 2, \ldots, H$, perform the following steps:
   a. Fit a classifier $G_h(x)$ to the training data using the observation weights $w_i^{(h)}$.
   b. Compute the classification error using:

$$err_h = \frac{\sum_{i=1}^{N} w_i^{(h)} I\,(G_h(x_i) \neq y_i)}{\sum_{i=1}^{N} w_i^{(h)}}, \tag{6}$$

   where $I\,(G_h(x_i) \neq y_i)$ is an indicator function that equals 1 if the classification is incorrect and 0 otherwise.

   c. Compute the coefficient α_h using:

$$\alpha_h = log\,\left(\frac{1-err_h}{err_h}\right). \tag{7}$$

   d. Update the weights for the misclassified observations:

$$w_i^{(h+1)} = w_i^{(h)} exp(\alpha_h\, I\,(G_h(x_i) \neq y_i)),\ i = 1, 2, \ldots, N \tag{8}$$

3. The final prediction is obtained as the weighted sum of the predictions from all iterations:

$$G(x) = \text{Sign}\,(\textstyle\sum_{h=1}^{H} \alpha_h\, G_h(x)) \tag{9}$$

   The classfification rule assigns class 1 if

$$G(x) \geq \frac{1}{2}\textstyle\sum_{h=1}^{H} \alpha_h$$

and class 0 otherwise.

Results and Discussions

Data description and Software

This research utilizes three UCI Machine Learning Repository datasets: Breast Cancer, Ecoli, and Glass Identification. These three datasets have different characteristics in terms of the number of samples, the number of features, and class distribution. One of the main problems with all three datasets is class imbalance, where some classes have far fewer samples than the majority class. This condition can cause classification models to be biased toward the majority class and fail to recognize minority classes. To address this, this study applies the Synthetic Minority Over-sampling Technique (SMOTE) before performing classification. The characteristics of all datasets used in this study are summarized in Table 1. All computations in this study were performed using open source software R.

Table 1. Characteristics of the Datasets Used in the Experiment

Datasets	Instances	Features	Class	Minority Class Ratio (%)
Breast Cancer	699	9	2	34.5
Ecoli	336	7	8	0.6
Glass	214	9	6	4.2

Minority Class Ratio (%) is defined as the proportion of instances in the smallest class relative to the total number of instances in the dataset. For multiclass datasets, the smallest class is considered the minority class.

Comparative Distribution Analysis Before and After SMOTE

The Breast Cancer dataset exhibited a significant class imbalance, with 458 instances in the benign class (2) and 241 instances in the malignant class (4), resulting in a skewed distribution favoring the majority class. After applying SMOTE to the training data, the class distribution became more balanced, with approximately 57.14% benign and 42.86% malignant instances, yielding a ratio of about 1.33:1. This adjustment improved the representation of the minority class, thereby enhancing the model's sensitivity to malignant cases.

The imbalance in the Ecoli dataset was more pronounced, with several minority classes having very limited representation compared to the majority class. After applying SMOTE to the training data, the class distribution became more balanced across all classes while preserving the dataset's multiclass structure. This resampling process improved the representation of minority classes without drastically altering the overall distribution, enabling the model to capture patterns from underrepresented classes better. However, classes with very small initial sample sizes remain challenging and may still affect predictive performance.

The Glass dataset exhibited moderate class imbalance, with some classes underrepresented compared to the majority. After applying SMOTE to the training data, the class distribution became slightly more balanced while preserving the dataset's original multiclass structure. The resampling process improved the representation of minority classes without significantly altering the overall distribution, enabling the model to learn patterns from less frequent classes more effectively.

The application of SMOTE across the three datasets effectively addressed class imbalance by enhancing the representation of minority classes. Previously underrepresented classes acquired adequate synthetic samples, enabling the model to learn their patterns and enhancing predictive sensitivity. However, the potential for overfitting on synthetic data persists, especially for classes with very few original samples. Despite this limitation, SMOTE facilitated more equitable classification among all classes, mitigating bias towards the majority and enhancing predictive performance for the minority.

Experimental Results and Analysis

Three classification algorithms—Decision Tree, Support Vector Machine (SVM), and AdaBoost—were evaluated on three benchmark datasets. The Synthetic Minority Over-sampling Technique (SMOTE) was applied to address class imbalance prior to classification. Model performance was assessed using accuracy, minority-class sensitivity, and F1-score, which collectively measure overall correctness, minority-class detection, and the balance between precision and recall. Table 2 presents a comprehensive comparison of results for both scenarios, with and without SMOTE.

Table 2. Comprehensive Comparison of Results

Datasets	Methods	Method Type	Accuracy (%)	Sensitivity Minority (%)	F1 (%)
Breast Cancer	Decision Tree	Without SMOTE	95.2	95.8	93.2
		With SMOTE	97.0	98.0	97.0
	SVM	Without SMOTE	95.2	95.2	95.2
		With SMOTE	97.8	98.9	97.8
	AdaBoost	Without SMOTE	95.2	95.2	95.2
		With SMOTE	99.8	99.7	99.8
Ecoli	Decision Tree	Without SMOTE	80.4	78.5	74.3
		With SMOTE	81.0	71.4	72.7
	SVM	Without SMOTE	87.6	84.8	85.8
		With SMOTE	86.6	100.0	33.3
	AdaBoost	Without SMOTE	88.6	83.5	85.4
		With SMOTE	89.7	60.0	70.6
Glass	Decision Tree	Without SMOTE	49.1	42.3	44.8
		With SMOTE	82.0	50.0	81.0
	SVM	Without SMOTE	60.6	52.3	68.9
		With SMOTE	87.8	100.0	85.7

Datasets	Methods	Method Type	Accuracy (%)	Sensitivity Minority (%)	F1 (%)
	AdaBoost	Without SMOTE	65.5	73.4	69.3
		With SMOTE	97.0	98.0	97.0

Accuracy Analysis.

The accuracy results indicate that all classifiers achieve strong performance after applying SMOTE, particularly on the Breast Cancer dataset, where accuracy ranges from 97.0% (Decision Tree) to 99.8% (AdaBoost). In contrast, lower accuracy is observed for the Ecoli dataset due to its more complex multiclass structure. AdaBoost achieves the highest accuracy (89.7%), followed by SVM (86.6%) and Decision Tree (81.0%). A similar trend is observed in the Glass dataset, where AdaBoost again outperforms the other models, achieving 97.0% accuracy, compared to 87.8% for SVM and 82.0% for Decision Tree. Overall, AdaBoost consistently achieves the highest accuracy across all datasets, although the differences are relatively moderate.

Sensitivity Analysis.

For multiclass datasets (Ecoli and Glass), sensitivity and F1-score are computed using macro-averaging across all classes. This means that each metric is calculated per class and then averaged, ensuring equal contribution regardless of class frequency. This approach is essential in imbalanced settings to avoid dominance by majority classes.

Minority-class sensitivity reveals more pronounced differences among classifiers. SMOTE significantly improves the detection of minority instances, as demonstrated by the Breast Cancer dataset, where sensitivity exceeds 98% across all models. However, greater variability is observed in the Ecoli and Glass datasets. SVM achieves 100% sensitivity for both datasets. Since sensitivity is macro-averaged, this indicates strong recall across classes, but does not necessarily imply balanced classification performance.

In contrast, the Decision Tree shows lower sensitivity (71.4% for Ecoli and 50.0% for Glass), indicating difficulty in identifying minority classes. AdaBoost demonstrates moderate sensitivity in Ecoli (60.0%) and high sensitivity in Breast Cancer (99.7%) and Glass (98.0%). These findings suggest that while SVM maximizes recall, other methods may offer a more balanced detection capability, depending on the dataset's complexity.

F1-score Analysis.

The F1-score provides a more comprehensive evaluation by balancing precision and recall. AdaBoost achieves the highest F1-score (99.8%) on the Breast Cancer dataset, indicating excellent overall performance. However, a notable discrepancy is observed in the Ecoli dataset, where SVM achieves perfect sensitivity but a low F1-

score (33.3%). This indicates that, although the model successfully detects instances across classes (high recall), it generates a substantial number of false positives, resulting in low precision. This highlights a critical trade-off between sensitivity and precision in multiclass imbalanced settings.

In comparison, AdaBoost and Decision Tree produce more stable F1-scores (70.6% and 72.7%), suggesting a better balance between detection and misclassification. On the Glass dataset, AdaBoost again achieves the highest F1-score (97.0%), followed by SVM (85.7%) and Decision Tree (81.0%). This finding supports prior evidence that boosting-based methods are effective in improving classification balance in imbalanced datasets.

Overall Discussion and Implications.

A comprehensive evaluation across all metrics indicates that AdaBoost provides the most balanced and reliable performance among the three classifiers. While SVM consistently achieves high sensitivity, particularly in multiclass settings, this advantage is sometimes offset by lower precision, as reflected in lower F1-scores. Decision Trees show relatively stable performance but tend to decline in more complex and highly imbalanced datasets.

These results suggest that combining SMOTE with a boosting-based classifier offers a robust and effective approach for handling imbalanced data, especially when both minority detection and prediction stability are critical. This finding is consistent with previous studies that have highlighted the effectiveness of ensemble methods in improving classification performance under imbalanced conditions.

Nevertheless, limitations remain in datasets with extremely small classes, such as Glass Identification, where synthetic oversampling cannot fully address class overlap and intrinsic data complexity. Therefore, while SMOTE significantly enhances performance, its effectiveness is still influenced by the underlying data structure.

Conclusions

This study assessed the efficacy of the Synthetic Minority Over-sampling Technique (SMOTE) in enhancing classification performance on imbalanced datasets. Three benchmark datasets from the UCI Machine Learning Repository—Breast Cancer, Ecoli, and Glass Identification—were utilized to evaluate the efficacy of Decision Tree, SVM, and AdaBoost classifiers.

The findings indicate that applying SMOTE generally improves class distribution by increasing minority-class representation, which is reflected in improvements in sensitivity and F1-score for most classifiers. Table 2 presents both pre-SMOTE and post-SMOTE results, allowing direct comparison of classifier performance. While all classifiers benefit to varying degrees from SMOTE, AdaBoost demonstrates the most

stable and consistent performance across datasets, achieving higher overall accuracy while maintaining high sensitivity, particularly in imbalanced settings.

Based on these findings, it is recommended that practitioners working with imbalanced datasets consider ensemble classifiers such as AdaBoost in combination with SMOTE to achieve more reliable performance. Future research could investigate hybrid resampling techniques or alternative ensemble approaches to enhance minority-class prediction further, while continuing to evaluate models using multiple metrics to ensure comprehensive assessment.

Acknowledgements

The author would like to thank the University of Bengkulu for providing facilities and an academic environment that supported the completion of this research. The author also appreciates the constructive feedback from colleagues and seminar participants, which contributed to the development of this manuscript.

References

Bian, J., Wang, J., & Yece, Q. (2024). A novel study on power consumption of an HVAC system using CatBoost and AdaBoost algorithms combined with the metaheuristic algorithms. *Energy*, *302*, 131841. https://doi.org/10.1016/j.energy.2024.131841

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*(1), 1–16. https://link.springer.com/article/10.1186/1471-2105-14-106

Boonamnuay, S., Kerdprasop, N., & Kerdprasop, K. (2018). Classification and regression tree with resampling for classifying imbalanced data. *International Journal of Machine Learning and Computing*, *8*(4), 336–340. https://doi.org/10.18178/ijmlc.2018.8.4.708

Cao, Y., Miao, Q., Liu, J., & Gao, L. (2013). Advance and Prospects of AdaBoost Algorithm. *Acta Automatica Sinica*, *39*(6), 745–758. https://doi.org/10.1016/S1874-1029(13)60052-X

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chawla, N. V., Cieslak, D. A., Hall, L. O., & Joshi, A. (2008). Automatically countering imbalance and its empirical relationship to cost. *Data Mining and Knowledge Discovery*, *17*(2), 225–252. https://doi.org/10.1007/s10618-008-0087-0

Chen, W., Yang, K., Yu, Z., & Zhang, W. (2022). Double-kernel based class-specific broad learning system for multiclass imbalance learning. *Knowledge-Based Systems*, *253*, 109535. https://doi.org/10.1016/j.knosys.2022.109535

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Domingo, C., & Watanabe, O. (2000). MadaBoost: A modification of AdaBoost. *In Colt*, *1*, 1–26.

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905. https://doi.org/10.1613/jair.1.11192

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). 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*, *42*(4), 463–484. https://doi.org/10.1109/TSMCC.2011.2161285

Gamil, S., Zeng, F., Alrifaey, M., Asim, M., & Ahmad, N. (2024). An Efficient AdaBoost Algorithm for Enhancing Skin Cancer Detection and Classification. *Algorithms*, *17*(8), 353. https://doi.org/10.3390/a17080353

Ha, J., Kambe, M., & Pe, J. (2011). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5

Haties, T., Tibshirani, R., & Friedman, J. (2009). Springer Series in Statistics. In *The Elements of Statistical Learning* (Vol. 27, Issue 2). http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf

Hothorn, T., Hornik, K., Wien, W., & Zeileis, A. (2015). ctree: Conditional Inference Trees. *The Comprehensive R Archive Network;*, *Quinlan 1993*, 1–34. https://mirrors.nics.utk.edu/cran/web/packages/partykit/vignettes/ctree.pdf

Jiang, X., Xu, Y., Ke, W., Zhang, Y., Zhu, Q.-X., & He, Y.-L. (2022). An Imbalanced Multifault Diagnosis Method Based on Bias Weights AdaBoost. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–8. https://doi.org/10.1109/TIM.2022.3149097

Johnson, R. A. and Wichern, D. A. (2002). Applied multivariate statistical analysis. In *Prentice Hall*.

Kumar, Y., Kaur, K., & Singh, G. (2020). Machine Learning Aspects and its Applications Towards Different Research Areas. *2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, 150–156. https://doi.org/10.1109/ICCAKM46823.2020.9051502

Last, F., Douzas, G., & Bacao, F. (2017). Oversampling for Imbalanced Learning Based on K-Means and SMOTE. *Information Sciences*, *465*, 1–20. https://doi.org/10.1016/j.ins.2018.06.056

Maulud, D., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, *1*(2), 140–147. https://doi.org/10.38094/jastt1457

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics*, *18*(6), 275–285. https://doi.org/10.1002/cem.873

Park, J., & Sandberg, I. W. (1991). Universal Approximation Using Radial-Basis-Function Networks. *Neural Computation*, *3*(2), 246–257. https://doi.org/10.1162/neco.1991.3.2.246

Pratap Chandra, S., Hajra, M., & Ghosh., M. (2020). Supervised classification algorithms in machine learning: A survey and review. In *Advances in Intelligent Systems and Computing* (Vol. 937). https://doi.org/10.1007/978-981-13-7403-6_7

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, *28*(1), 71–72. https://doi.org/10.1145/234313.234346

Rezvani, S., & Wang, X. (2023). A broad review on class imbalance learning techniques. *Applied Soft Computing*, *143*, 110415. https://doi.org/10.1016/j.asoc.2023.110415

Rosadi, D., Arisanty, D., Andriyani, W., Peiris, S., Agustina, D., Dowe, D., & Fang, Z. (2021). Improving Machine Learning Prediction of Peatlands Fire Occurrence for Unbalanced Data Using SMOTE Approach. *2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*, 160–163. https://doi.org/10.1109/DATABIA53375.2021.9650084

Tao, P., Sun, Z., & Sun, Z. (2018). An Improved Intrusion Detection Algorithm Based on GA and SVM. *IEEE Access*, *6*, 13624–13631. https://doi.org/10.1109/ACCESS.2018.2810198

Wang, R. (2012). AdaBoost for Feature Selection, Classification and Its Relation with SVM, A Review. *Physics Procedia*, *25*, 800–807. https://doi.org/10.1016/j.phpro.2012.03.160

Wang, S., Dai, Y., Shen, J., & Xuan, J. (2021). Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports*, *0123456789*, 1–11. https://doi.org/10.1038/s41598-021-03430-5

Yang Wendong, Lou Zhengzheng, & Ji Bo. (2017). A multi-factor analysis model of quantitative investment based on GA and SVM. *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, 1152–1155. https://doi.org/10.1109/ICIVC.2017.7984734