Analysis of ensemble machine learning classification comparison on the skin cancer MNIST dataset

Poetri Lestari Lokapitasari Belluano, Reyna Aprilia Rahma, Herdianti Darwis, Abdul Rachman Manga
Department of Computer Science, Faculty of Computer Science, Universitas Muslim Indonesia, Makassar, Indonesia

Article Info

Article history:

Received Jan 21, 2024
Revised Jul 8, 2024
Accepted Sep 8, 2024

Keywords:

Ensemble machine learning
Imbalanced data
Performance comparison
Skin cancer

ABSTRACT

This study aims to analyze the performance of various ensemble machine learning methods, such as Adaboost, Bagging, and Stacking, in the context of skin cancer classification using the skin cancer MNIST dataset. We also evaluate the impact of handling dataset imbalance on the classification model's performance by applying imbalanced data methods such as random under sampling (RUS), random over sampling (ROS), synthetic minority over-sampling technique (SMOTE), and synthetic minority over-sampling technique with edited nearest neighbor (SMOTEENN). The research findings indicate that Adaboost is effective in addressing data imbalance, while imbalanced data methods can significantly improve accuracy. However, the selection of imbalanced data methods should be carefully tailored to the dataset characteristics and clinical objectives. In conclusion, addressing data imbalance can enhance skin cancer classification accuracy, with Adaboost being an exception that shows a decrease in accuracy after applying imbalanced data methods.

Corresponding Author:

Reyna Aprilia Rahma
Department of Computer Science, Faculty of Computer Sciences, Universitas Muslim Indonesia
St. Kakatua II Lr. 3B No. 29D, Makassar, South Sulawesi-90133, Indonesia
Email: reynarahma.iclabs@umi.ac.id

1. INTRODUCTION

Skin cancer is a type of cancer characterized by the growth of abnormal skin lesions [1], such as nodules, which can potentially threaten human health [2], [3]. Its causes vary, including exposure to toxic substances from food, the environment, and excessive sunlight [4], [5]. Identifying and classifying types of skin cancer is crucial for patient treatment and early detection. In the era of information technology, machine learning can improve accuracy [6] and efficiency in classifying types of skin cancer. Skin cancer MNIST is a valuable resource that contains various skin images to train classification models. Ensemble machine learning techniques, such as adaboost, bagging, and stacking, are an interesting approach to improving skin cancer classification accuracy.

Previous research has shown that the use of ensemble methods improves classification performance. For example, in previous research, bagging and boosting with decision trees as base learners successfully increased accuracy in classifying diabetes. The research results also show that using ensemble methods with base learners such as naïve bayes or random forest can achieve the best accuracy scores of up to 98.65% in various contexts [6]. In another study, the bagging method utilizing random forest as the base learner achieved the highest accuracy rate of 96.66%. Conversely, the decision tree algorithm yielded the lowest accuracy rate, approximately 73.33% [7].

Although previous research has explored ensemble machine learning techniques in various domains, there has been no comprehensive study comparing the performance of various ensemble methods in the context of skin cancer MNIST. Therefore, this is the aim of the research machine learning methods, such as adaboost, bagging, and stacking, using various decision tree algorithms as base learners. This study will also investigate the impact of handling dataset imbalance on the classification model's performance by applying methods such as random under sampling (RUS), random over sampling (ROS), synthetic minority over-sampling technique (SMOTE), and synthetic minority over-sampling technique with edited nearest neighbor (SMOTEENN). Through this comprehensive comparative analysis, this research will provide a deeper understanding of the strengths and weaknesses of various ensemble machine learning methods in classifying types of skin cancer. It is hoped that the research findings will serve as a valuable guide for practitioners and researchers in efforts to enhance the effectiveness of machine learning technology [8] in diagnosing and treating skin cancer.

2. METHOD

The research method consists of four stages, as shown in Figure 1. From Figure 1, it is known that the research process starts from dataset collection to performance analysis. Each stage is meticulously crafted to guarantee a thorough and methodical evaluation of the research goals.

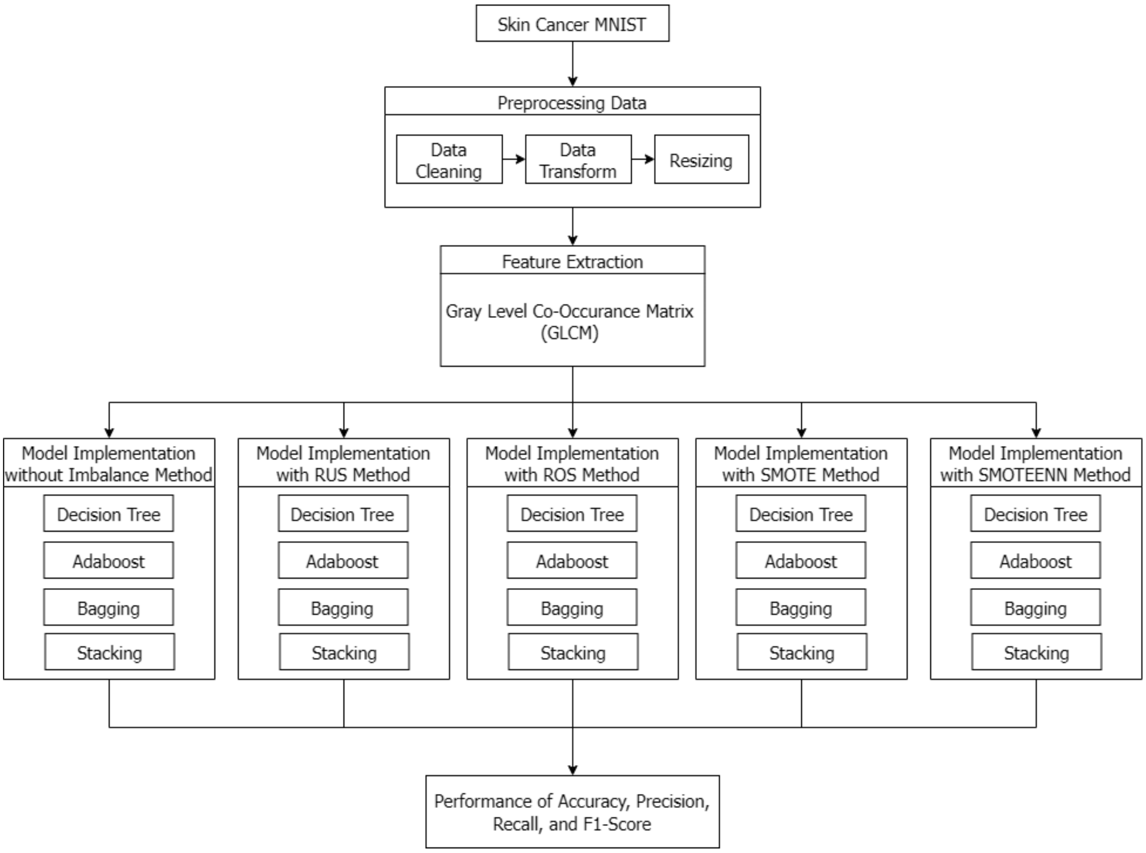

Figure 1. Research step

In this study, the skin cancer MNIST image dataset was used, which is available for download from [9]. This dataset is structured into two main components: a folder containing skin images and clinical data in CSV file format obtained [10]. Seven different types of images are identified and documented, as seen in Figure 2, namely Melanocytic Nevi (nv), Melanoma (mel), Benign Keratosis-Like Lesions (bkl), Basal Cell Carcinoma (bcc), Actinic Keratoses (akiec), Vascular Lesions (vasc), and Dermatofibroma (df). The first and second folders contain various skin images taken from different individuals with various skin cancer conditions. The CSV file contains clinical data relevant to the same patients as the skin images in the first and second folders. This clinical data consists of 10,015 entries with seven different attributes. A complete explanation of these seven attributes can be found in Table 1.

Table 1. Attribute explanation

Attribute	Value
lesion_id	Unique identifier
image_id	Unique identifier
dx	“nv”, “mel”, “bkl”, “bcc”, “akiec”, “vasc”, dan “df”
dx_type	“histo”, “consensus”, “confocal”, dan “follow_up”
age	Patient’s age
sex	“male” dan “female”
localization	“scalp”, “ear”, “face”, “back”, “trunk”, “chest”, “abdomen”, “unknown”, “genetal”, “neck”, “hand”, “foot”, “acral”, “upper extremity”, dan “lower extremity”

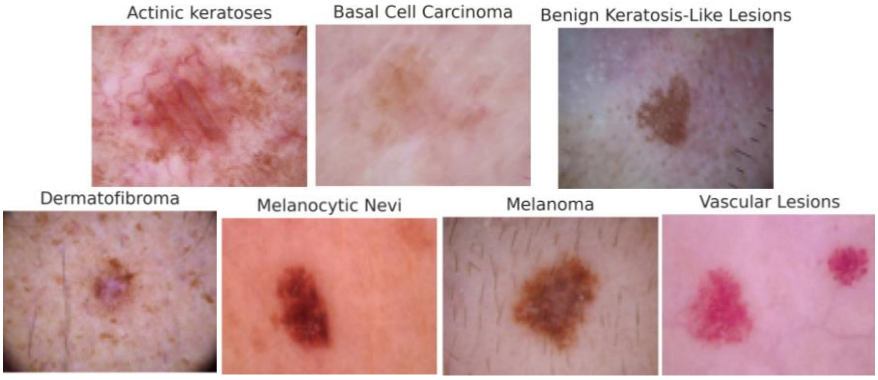

Figure 2. Types of skin cancer

Data preprocessing is the initial preparation step before performing data analysis, aimed at addressing potential issues that affect the results of data analysis. The first step is feature cleaning, where missing values in the “age” column are replaced with the mean value of that column without missing values. The next step is data transformation, where categorical data is converted into numeric format by replacing the label “Male” with “0” and the label “Female” with “1”. Subsequently, the third step involves image resizing to 100×75 pixels to reduce data complexity and prevent overfitting. The fourth step is feature extraction, which involves extracting important information from complex or high-dimensional data [11]. The method used is gray co-occurrence level matrix (GLCM), used to extract texture features [12], [13], which are used to extract texture features such as energy, correlation, dissimilarity, homogeneity, and contrast [14]. These features describe the texture characteristics in the image and aid in identifying desired regions in an image. The formulas used for the five GLCM features can be seen in the following (1) to (5):

$$Energy = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} P_{d,\theta}(i,j)^2 \quad (1)$$

$$Correlation = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} P_{d,\theta}(i,j) * \frac{(i-\mu_x)(i-\mu_y)}{\sigma_x \sigma_{\sigma y}} \quad (2)$$

$$Dissimilarity = \sum_{i,j=0}^{N-1} P_{i,j}|i-j| \quad (3)$$

$$Homogen\ (\sigma_x) = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} \left(\frac{1}{1+(i-j)^2}\right) * \left(P_{d,\theta}(i,j)\right) \quad (4)$$

$$Contrast = \sum_{n=0}^{N_g-1} |i,j|^2 * \left\{\sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} P_{d,\theta}(i,j)\right\} \quad (5)$$

The final step in data preprocessing in this study is to address data imbalance resulting from differences in the number of samples among different types of skin cancer. Some types of skin cancer are more common than others, leading to an imbalance in the data count between classes, as shown in Figure 3.

This imbalance has the potential to affect the performance of machine learning models in classifying skin cancer images. Therefore, this study employs four data imbalance handling techniques: RUS, which enhances the majority class samples [15], [16]. ROS, which augments the minority class samples [17], [18]. SMOTE, which creates new minority class samples [19]–[21], and SMOTEENN, which combines the SMOTE technique with edited nearest neighbor (ENN) to address overfitting issues [22], [23]. The implementation of these techniques aims to improve the model's performance and reliability in classifying skin cancer images. Nevertheless, it is essential to compare the results with the case of no data imbalance handling.

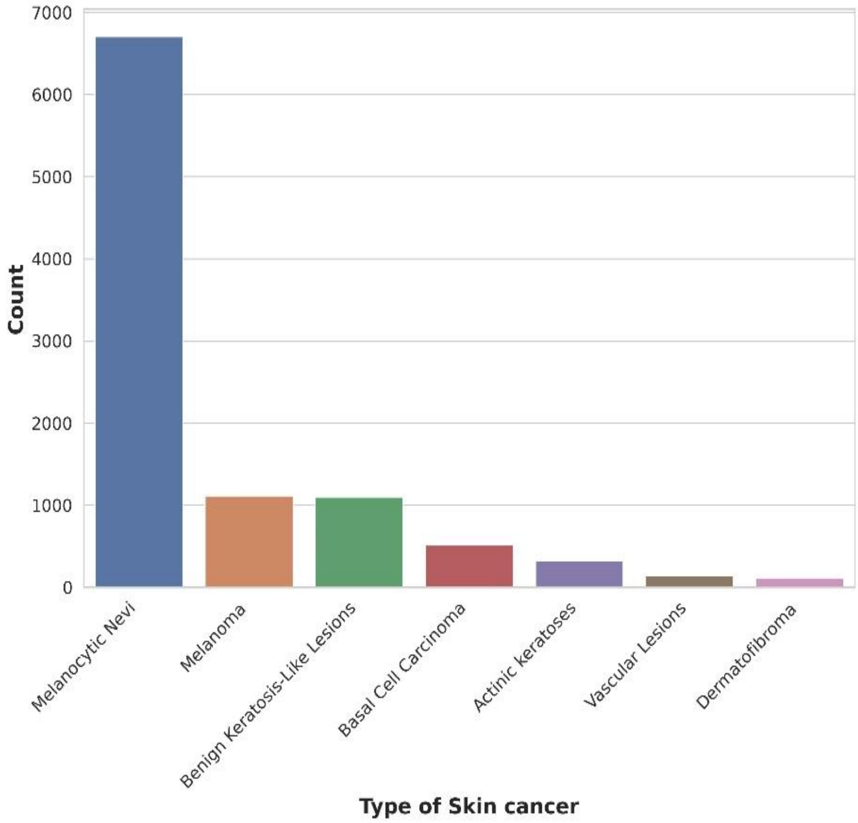

Figure 3. Skin cancer MNIST classes

Ensemble methods are machine learning techniques that combine multiple learning models into a single, stronger model, allowing the final decision to be based on the combination of classifiers used [24], [25]. One effective ensemble method is Adaboost, which assigns different weights to data [26], [27] and incrementally improves the model [28], [29]. Bagging is used to reduce model sensitivity to data variations by creating multiple models on randomly sampled training subsets [30]. Stacking uses a first-level base learner to make predictions [31], [32], and these predictions are then used as training data for a meta-learner [33]. Decision trees [33], [34] are machine learning algorithms used to classify data based on specific conditions in the decision-making process.

Performance analysis is the process of evaluating a system or process to understand the extent to which its performance meets predefined goals or standards. It is main goal is to identify potential issues, understand existing trends, and design necessary improvements. Model performance evaluation is done using metrics such as accuracy, precision, recall, and F1 score [35]. To calculate accuracy, precision, recall, and F1 Score, the following formulas are used:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (8)$$

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

$$F1\ Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (11)$$

3. RESULTS AND DISCUSSION

The performance evaluation results of each classification method in classifying skin images on the test data are presented in Tables 2 and 3. In Table 2, the classification performance results without data balancing reveal that the decision tree algorithm achieved the highest accuracy, which is 74.6%. On the other hand, among the ensemble methods, Adaboost obtained the highest accuracy, which is 73.5%. Adaboost is effective in handling datasets with class imbalance.

Table 2. Comparison of classification performance of ensemble methods without imbalanced data handling

Method	Performance			
	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	**74.6**	68.1	74.6	70.6
Adaboost	73.5	70.2	73.5	70.5
Bagging	73	69.2	73	69.2
Stacking	70.2	62.3	70.1	61.7

Based on Table 3, it can be seen that imbalanced data handling methods have a significant impact on the classification model's performance. In some cases, such as using ROS with decision tree, bagging, and stacking, accuracy reaches 100%, indicating a drastic improvement compared to the method without imbalanced data handling. However, it's important to note that such high results can also be due to overfitting for the minority class. On the other hand, there is a significant decrease in accuracy when the Adaboost method is applied to certain imbalanced data methods, especially with the use of RUS and SMOTE. This underscores the importance of selecting the appropriate imbalanced data method based on the dataset's characteristics and classification goals. In the context of medical datasets like this, where high precision and recall may be more important than overall accuracy, the choice of an imbalanced data method should be carefully considered to meet clinical needs.

Table 3. Comparison of classification performance of ensemble methods without imbalanced data handling

Imbalanced data handling method	Method	Performance			
		Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RUS	Decision tree	44	44.1	44	43.3
	Adaboost	37.8	38.7	37.8	35.8
	Bagging	50.9	50.8	50.9	50
	Stacking	47.2	47.5	47.2	44.5
ROS	Decision tree	**100**	100	100	100
	Adaboost	54	51.4	54	51.1
	Bagging	**99.9**	99.9	99.9	99.9
	Stacking	**99.8**	99.8	99.8	99.8
SMOTE	Decision tree	85.7	85.5	85.7	85.5
	Adaboost	55.2	54.2	55.2	53.1
	Bagging	92.7	92.7	92.7	92.6
	Stacking	90.5	90.6	90.6	90.2
SMOTEENN	Decision tree	90.5	90.4	90.5	90.4
	Adaboost	49	49	49	45.8
	Bagging	95.3	95.3	95.3	95.3
	Stacking	97.6	97.6	97.6	97.6

The results of the research indicate that the performance classification model in the skin cancer MNIST dataset is significantly affected by data imbalance. High accuracy may be reached using Ensemble Methods such as ROS. However, it should be emphasized that this could be related to potential overfitting of the minority class. The importance of selecting the right imbalanced data method depends greatly on the dataset's characteristics and classification goals. In a medical dataset context, high precision and recall may be more relevant than overall accuracy. Therefore, the choice of imbalanced data method should be carefully considered according to clinical needs. From the classification results in Tables 2 and 3, it can be concluded that handling data imbalance can improve accuracy [36]–[39]. However, imbalanced data methods do not seem to affect the Adaboost method, as evidenced by the decrease in accuracy after using imbalanced data methods with Adaboost.

4. CONCLUSION

The results of this study show that the decision tree algorithm achieved the highest accuracy of 74% on the imbalanced dataset, while in the ensemble methods, Adaboost reached the highest accuracy of 73.5%. Adaboost has proven to be effective in dealing with datasets with class imbalance. However, there is a significant impact of data imbalance on the classification model's performance. The application of imbalanced data methods such as ROS with decision tree, bagging, and stacking resulted in accuracy improvements of up to 100%, which may be related to potential overfitting on the minority class. However, Adaboost experienced a decrease in accuracy when imbalanced data methods were used, especially with RUS and SMOTE. This emphasizes the importance of carefully selecting imbalanced data methods based on dataset characteristics and classification goals, especially in the context of medical datasets where precision and recall may be more relevant than accuracy. In conclusion, handling data imbalance can improve accuracy, but the choice of imbalanced data methods should be carefully considered according to clinical needs, with Adaboost being an exception that shows a decrease in accuracy after applying imbalanced data methods. Further studies should explore the underlying reasons for Adaboost's decreased performance with certain imbalanced data methods and investigate alternative techniques to mitigate this issue. Additionally, examining the impact of other Ensemble Methods and more sophisticated data balancing techniques on a wider range of medical datasets could provide deeper insights into optimizing classification performance while maintaining clinical relevance.

ACKNOWLEDGEMENTS

We express our profound gratitude to the Faculty of Computer Science at Universitas Muslim Indonesia. Their guidance, expertise, and steadfast support have been pivotal in bringing this research to fruition.

REFERENCES

[1] G. I. Sayed, "A novel melanoma prediction model for imbalanced data using optimized SqueezeNet by bald eagle search optimization," *Comput. Biol. Med.*, vol. 136, 2021, doi: 10.1016/j.compbiomed.2021.104712.

[2] A. Murugan, S. A. H. Nair, and et all Preethi, A A P, "Diagnosis of skin cancer using machine learning techniques," *Microprocess. Microsyst.*, 2021, doi: https://dx.doi.org/10.1016/j.micpro.2020.103727.

[3] A. C. Salian, S. Vaze, S, Singh, P, Shaikh, G N, Chapaneri, and D. Jayaswal, "Skin lesion classification using deep learning architectures," *Int. Conf. Commun. Syst. Comput. IT Appl.*, 2020, doi: https://dx.doi.org/10.1007/978-981-15-8391-9_42.

[4] A. K. Verma and S. Pal, "Prediction of skin disease with three different feature selection techniques using stacking ensemble method," *Appl. Biochem. Biotechnol.*, 2020, doi: 10.1007/s12010-019-03222-8.

[5] P. Banasode, M. Patil, and N. Ammanagi, "A melanoma skin cancer detection using machine learning technique: support vector machine," *IOP Conf. Ser.*, 2021, doi: 10.1088/1757-899X/1065/1/012039.

[6] P. Y. Taser, "Application of bagging and boosting approaches using decision tree-based algorithms in diabetes risk prediction," *Proceedings*, 2021, doi: https://dx.doi.org/10.3390/proceedings2021074006.

[7] M. A. Al-Hagery, E. I. Al-Fairouz, and N. A. AL-Humaidah, "Improvement of Alzheimer disease diagnosis accuracy using ensemble methods," *Indones. J. Electr. Eng. Informatics*, 2020, doi: https://dx.doi.org/10.11591/ijeei.v8i1.1321.

[8] H. Darwis, Z. Ali, Y. Salim, and P. L. L. Belluano, "Max Feature Map CNN with Support Vector Guided SoftMax for Face Recognition," *Int. J. Informatics Vis.*, vol. 7, no. 3, pp. 959–966, 2023, doi: 10.30630/joiv.7.3.1751.

[9] K SCOTT MADER, "Skin Cancer MNIST: HAM100," 2018. https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000 (accessed Apr. 05, 2023).

[10] S. M. K, "Skin Cancer MNIST: HAM10000," 2018. https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000 (accessed Apr. 05, 2023).

[11] Ş. Öztürk, "Application of Feature Extraction and Classification Methods for Histopathological Image using GLCM, LBP, LBGLCM, GLRLM and SFTA," *Procedia Computer Science*, vol. 132. pp. 40–46, 2018. doi: 10.1016/j.procs.2018.05.057.

[12] P. K. Mall, "GLCM based feature extraction and medical X-RAY image classification using machine learning techniques," *2019 IEEE Conference on Information and Communication Technology, CICT 2019*. 2019. doi: 10.1109/CICT48419.2019.9066263.

[13] M. Vidya and M. V Karki, "Skin cancer detection using machine learning techniques," *2020 IEEE Int. Conf. …*, 2020, doi: https://dx.doi.org/10.1109/CENTCON56610.2022.10051495.

[14] L. S. Wei, "Skin Disease Recognition Method Based on Image Color and Texture Features," *Comput. Math. Methods Med.*, vol. 2018, 2018, doi: 10.1155/2018/8145713.

[15] N. S. Sudaryanto, "Synthesis Ensemble Oversampling and Ensemble Tree-Based Machine Learning for Class Imbalance Problem in Breast Cancer Diagnosis," *Proceeding of the International Conference on Computer Engineering, Network and Intelligent Multimedia, CENIM 2022*. pp. 110–116, 2022. doi: 10.1109/CENIM56801.2022.10037251.

[16] Y. Wu, "Stroke prediction with machine learning methods among older chinese," *Int. J. Environ. Res. Public Health*, vol. 17, no. 6, 2020, doi: 10.3390/ijerph17061828.

[17] T. Riston, “Oversampling Methods for Handling Imbalance Data in Binary Classification,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14105. pp. 3–23, 2023. doi: 10.1007/978-3-031-37108-0_1.

[18] F. Kamalov, “Ensemble Learning with Resampling for Imbalanced Data,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12837. pp. 564–578, 2021. doi: 10.1007/978-3-030-84529-2_48.

[19] Q. Wang, “A novel ensemble method for imbalanced data learning: Bagging of extrapolation-SMOTE SVM,” *Comput. Intell. Neurosci.*, vol. 2017, 2017, doi: 10.1155/2017/1827016.

[20] J. Tanha, “Boosting methods for multi-class imbalanced data classification: an experimental review,” *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00349-y.

[21] K. N. V. P. S. Rajesh, “Classification of imbalanced ECG beats using re-sampling techniques and AdaBoost ensemble classifier,” *Biomed. Signal Process. Control*, vol. 41, pp. 242–254, 2018, doi: 10.1016/j.bspc.2017.12.004.

[22] M. Khushi, “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.

[23] E. Mbunge, “Implementation of ensemble machine learning classifiers to predict diarrhoea with SMOTEENN, SMOTE, and SMOTETomek class imbalance approaches,” *2023 Conference on Information Communications Technology and Society, ICTAS 2023 - Proceedings*. 2023. doi: 10.1109/ICTAS56421.2023.10082744.

[24] A. Mosavi, “Ensemble Boosting and Bagging Based Machine Learning Models for Groundwater Potential Prediction,” *Water Resour. Manag.*, vol. 35, no. 1, pp. 23–37, 2021, doi: 10.1007/s11269-020-02704-3.

[25] A. R. Javed, “Ensemble Adaboost classifier for accurate and fast detection of botnet attacks in connected vehicles,” *Trans. Emerg. Telecommun. Technol.*, vol. 33, no. 10, 2022, doi: 10.1002/ett.4088.

[26] S. Rahman, “Performance analysis of boosting classifiers in recognizing activities of daily living,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, 2020, doi: 10.3390/ijerph17031082.

[27] C. Iwendi, “The use of ensemble models for multiple class and binary class classification for improving intrusion detection systems,” *Sensors (Switzerland)*, vol. 20, no. 9, 2020, doi: 10.3390/s20092559.

[28] P. Bobade, “Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data,” *Proceedings of the 2nd International Conference on Inventive Research in Computing Applications, ICIRCA 2020*. pp. 51–57, 2020. doi: 10.1109/ICIRCA48905.2020.9183244.

[29] Q. Huang, “On Combining Biclustering Mining and AdaBoost for Breast Tumor Classification,” *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 4, pp. 728–738, 2020, doi: 10.1109/TKDE.2019.2891622.

[30] S. K. Gunturi, “Ensemble machine learning models for the detection of energy theft,” *Electr. Power Syst. Res.*, vol. 192, 2021, doi: 10.1016/j.epsr.2020.106904.

[31] Z. Faska, “A robust and consistent stack generalized ensemble-learning framework for image segmentation,” *J. Eng. Appl. Sci.*, vol. 70, no. 1, 2023, doi: 10.1186/s44147-023-00226-4.

[32] F. He, W. Zhang, and Z. Yan, “A novel multi-stage ensemble model for credit scoring based on synthetic sampling and feature transformation,” *J. Intell. &Fuzzy Syst.*, 2022, doi: https://dx.doi.org/10.3233/JIFS-211467.

[33] A. Kumar and J. Mayank, “Ensemble learning for AI developers,” *BApress Berkeley, CA, USA*, 2020, doi: 10.1007/978-1-4842-5940-5.

[34] C. S. Lee, P. Y. S. Cheang, and M. M, “Predictive analytics in business analytics: decision tree,” *Adv. Decis. Sci.*, 2022, doi: https://dx.doi.org/10.47654/V26Y2022I1P1-30.

[35] R. E. Saragih and A. W. R. Emanuel, “Banana ripeness classification based on deep learning using convolutional neural network,” *East Indones. Conf. onComputer Inf. Technol.*, 2021, doi: https://dx.doi.org/10.11591/ijeei.v8i1.1321.

[36] Y. Dong, H. Jiang, Y. Liu, and Z. Yi, “Global wavelet-integrated residual frequency attention regularized network for hypersonic flight vehicle fault diagnosis with imbalanced data,” *Eng. Appl. Artif. Intell.*, 2024, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095219762400126X

[37] S. Xie, J. Zhang, and A. T. Lawniczak, “Handling Data Imbalance In Linear Modelling of Fatality Rate of Auto Collision,” *Proc. 2023 6th Int. Conf. Inf. Sci. Syst.*, 2023, doi: 10.1145/3625156.3625169.

[38] M. R. Indrahimawan, P. I. Santosa, and T. Adji, “Handling Data Imbalance Using Text Augmentation For Classifying Public Complaints,” *2023 Int. Conf. Comput. Control. Informatics its Appl.*, 2023, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10285813/

[39] R. Bounab, B. Guelib, and K. Zarour, “A Novel Machine Learning Approach For handling Imbalanced Data: Leveraging SMOTE-ENN and XGBoost,” *2024 6th Int. Conf. Pattern Anal. Intell. Syst.*, 2024, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10541220/

BIOGRAPHIES OF AUTHORS

Poetri Lestari Lokapitasari Belluano obtained a Bachelor’s degree in Computer Science (S.Kom) from the Universitas Muslim Indonesia in 2006 and a Master's degree in Electrical Engineering (M.T) from Universitas Hasanuddin in 2014. Currently, she is an active lecturer in the field of computer engineering, specializing in software engineering. Her current research interests include the development of microservice-based applications, startup technology, expertise application development (artificial intelligence), and information system engineering. She can be contacted at email: poetrilestari@umi.ac.id.

Reyna Aprilia Rahma graduated from Universitas Muslim Indonesia (Makasar, Indonesia) with a Bachelor of Computer Science in the Faculty of Computer Science. Her research interests involve various topics, particularly in the field of data science. She can be contacted at email: reynarahma.iclabs@umi.ac.id.

Herdianti Darwis graduated from Hasanuddin University (Makassar, Indonesia) with a B.S. in Statistics. She received an M.Eng. in Computer Science and Engineering at Kyushu Institute of Technology, Fukuoka, Japan. She is currently working as a lecturer in the Faculty of Computer Science, Universitas Muslim Indonesia, South Sulawesi, Indonesia. Her research interests encompass data science, artificial intelligence, and computer vision including statictical learning, machine learning, and deep learning approaches. She can be contacted at: herdianti.darwis@umi.ac.id.

Abdul Rachman Manga’ graduated from Universitas Muslim Indonesia (Makassar, Indonesia) with a Bachelor of Computer Sience in Faculty of Computer Science. He received a Master of Engineering at Hasanuddin University, Makassar, Indonesia. And currently pursuing his Doctore degree at University of Malang, Malang, Indonesia. And currently working as a lecturer in the Faculty of Computer Science, Universitas Muslim Indonesia, South Sulawesi, Indonesia. His research interests cover Natural Languange Processing and Artificial intelligence. He can be contacted at: abdulrachman.manga@umi.ac.id.