Addressing Imbalance Data Label Distribution in Dengue Severity Classification with Hybrid Machine Learning Approach

Moch. Farrel Arrizal Kusuma[1,*], Wiwik Anggraeni[2]

Institut Teknologi Sepuluh Nopember, Indonesia

[1]6026231021@student.its.ac.id; [2]wiwik@is.its.ac.id;

* corresponding author

ARTICLE INFO

Article History
Received: 24 July 2024
Revised: 15 October 2024
Published: 11 December 2024

Keywords
Data Imbalanced
Dengue
Hyperparameter Tuning
Multiclass Classification
Optimization

ABSTRACT

Early detection of disease is essential for ensuring proper diagnosis and treatment, which helps to reduce complications and limit the spread of cases. Dengue fever remains a major public health issue in Indonesia, with some regions reporting case fatality rates above 1%. This study aims to develop a robust classification model to predict the severity of dengue fever based on laboratory data, addressing the issue of imbalanced label distribution. To achieve this, we applied a combination of oversampling, machine learning, and optimization techniques. Specifically, we utilized the SMOTE-ENN method to enhance the representation of minority classes and applied XGBoost for multi-class classification. Additionally, Particle Swarm Optimization (PSO) was used to fine-tune the model's hyperparameters. After testing 15 experimental scenarios, the hybrid SMOTE-ENN-XGBoost-PSO approach delivered the best performance, achieving an accuracy of 95% and an F1-score of 94%. These results demonstrate that the proposed approach effectively handles data imbalance and improves predictive accuracy. This research serves as a foundation for future work in applying advanced algorithms and Federated Learning to improve multi-class classification in healthcare.

INTRODUCTION

Dengue fever is a mosquito-borne viral infection endemic to tropical and subtropical regions [1]. It is primarily transmitted through the bites of Aedes aegypti and Aedes albopictus mosquitoes [2]. According to the Ministry of Health of the Republic of Indonesia, the case fatality rate (CFR) for dengue fever remains alarming in some regions, with figures exceeding 1% [3]. As an accessible and efficient tool, Machine Learning (ML) has become essential across numerous fields, including healthcare. In clinical settings, ML aids in diagnostics and provides high accuracy, particularly in infant and child diagnoses [4]. The application of ML in healthcare is expanding rapidly, playing a critical role in supporting healthcare professionals by facilitating decision-making and offering predictive insights from various datasets [5]. Despite its benefits, ML also faces challenges, especially with imbalanced datasets and the risk of overfitting. In healthcare, imbalanced datasets are common, often with a predominance of major classes over minor ones [6]. This imbalance can hinder the model's performance, particularly when rare but important cases, such as disease onset, are overlooked [7]. Overfitting, which occurs due to limited data, further complicates model optimization [8]. To address these challenges, several techniques have been developed, categorized into algorithm-level approaches, data-level approaches, cost-sensitive learning, and ensemble-based methods [9]. One promising approach to managing imbalanced data is the Synthetic Minority Over-sampling Technique (SMOTE), often combined with the Edited Nearest Neighbor (ENN) method [10], [11]. This hybrid approach helps create synthetic samples, improving model performance by representing minority classes more accurately [12], [13].

Machine learning models like Support Vector Machine (SVM) and XGBoost are highly effective for classification tasks in healthcare. SVM has demonstrated its reliability in solving various classification and regression problems, efficiently handling large and complex datasets [14]. XGBoost, on the other hand, is well-suited for multi-class classification and has been applied successfully in healthcare to detect diseases with high accuracy [15]. To enhance model performance, optimization techniques such as Particle Swarm Optimization (PSO) and hyperparameter tuning methods like GridSearchCV are widely used. PSO iteratively refines solutions based on global and individual best-known results, while GridSearchCV systematically searches for optimal parameters, leading to improved accuracy [16], [17].

Imbalanced data remains a challenge in healthcare applications, particularly in areas like disease prediction and fraud detection [18]. Methods like SMOTE have been employed to address this by generating high-quality synthetic samples for rare events, improving classification performance [19]. For instance, Zuherman et al. (2019) demonstrated how the SMOTE-ENN approach significantly improved accuracy in cerebral infarction classification using SVM. However, this method also has limitations, including longer computation times and a higher risk of overfitting [20]. This study aims to perform multi-class classification in the context of dengue fever, focusing on the challenge of imbalanced data. The goal is to identify the most effective model architecture and optimization techniques. The research will explore various strategies, including oversampling techniques like SMOTE and SMOTE-ENN, while utilizing advanced algorithms such as SVM and XGBoost, optimized with GridSearchCV and PSO.

METHOD

This study utilizes data obtained from the Health Department of Malang District, Indonesia, which includes details on the distribution of dengue fever in various areas of Malang Regency as well as laboratory data from infected patients. This initial data is referred to as raw data as it requires validation and pre-processing to ensure its quality before model training. In Figure 1 we proposed our research methodology.

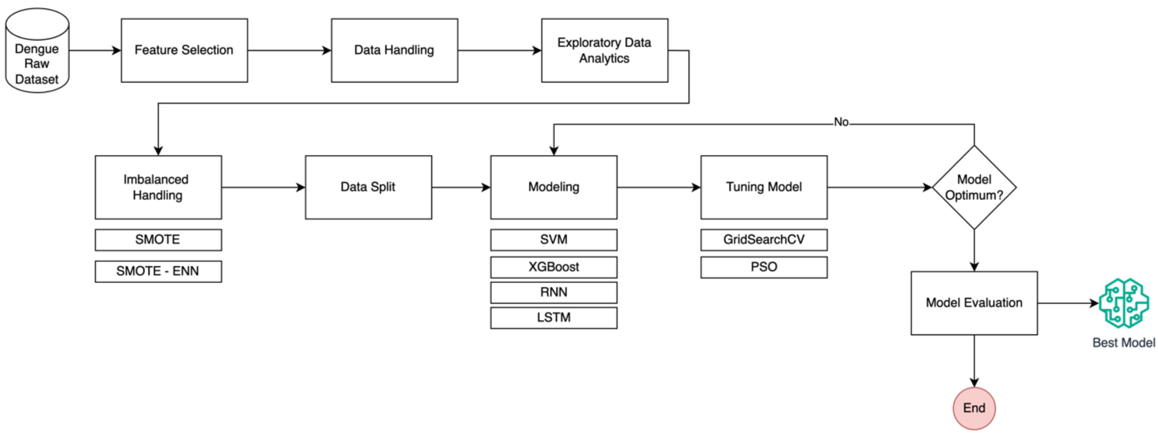

Figure 1 Research Methodology

Data Preprocessing

The raw dataset collected includes personal patient data, patient admission dates, patient status, and supporting data such as weather conditions around the patient's residence, hospital details, and patient laboratory results. For this study, which focuses on solving a multi-class classification task, we limit the dataset to personal patient data and laboratory results, such as age, gender, hematocrit (HCT), platelet count, and immunoglobulin test results (if available). Age is a particularly crucial feature, as it can reflect the body's immune response to dengue infection, including antibody production [21].

Further data filtering was applied, particularly to ensure that the platelet and HCT values were greater than zero. The dataset contains two features that describe the hematocrit level: the lowest HCT and the highest HCT for each patient. To capture the most critical point of the patient's condition, the researchers opted to use the lowest HCT value. However, if the low HCT data is missing, the HCT value from the highest HCT is used instead. Regarding the IGM, IGG, and NS1 features, these columns contain boolean values (0 or 1), indicating whether the patient's sample tested positive for these markers. Since more than half of the entries for these three features contain missing data, we did not drop the null values. Instead, we assigned a value of -1 to represent missing data. After the data cleaning process, the final dataset consists of 4,373 entries with eight key features: ID, age, gender, platelet count, HCT, IGM, IGG, NS1, and the classification label. Figure 2 shows data papeline.

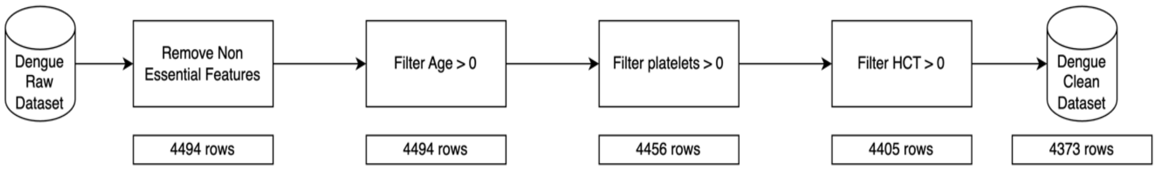

Figure 2. Data Pipeline

Figure 3 indicates that the data used has a varied age range, predominantly between 18-40 years old, and is predominantly male. There is an imbalance observed where the distribution of data on the label is not uniform. This is a common occurrence in the healthcare sector due to the varying frequencies of disease infections. Researchers can proceed with label processing decisions as a scenario design in the next stage.

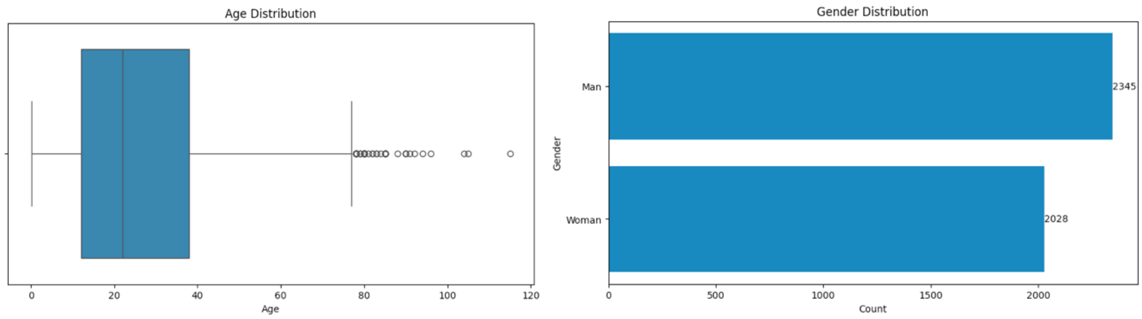

Figure 3 Age and Gender Distribution

Imbalanced Handling

Imbalanced data is a common condition frequently encountered in society. This study employs three approaches to evaluate the model's ability to understand the data. The first scenario utilizes raw data without augmentation to measure the model's performance. The second scenario applies SMOTE to increase the number of labels, particularly in the minority class. This approach has proven effective in enhancing model performance in several studies [11], [22]. In this scenario, SMOTE performs resampling with the same distribution, resulting in a total of 9588 resampled data for three labels, as shown Figure 4. The third scenario employs the SMOTE-ENN approach, which focuses on resampling the class with the least amount of data, in this case, the DSS class. After resampling, the total data for the DSS class reaches 5047, as displayed Figure 4. This approach has been shown to deliver the best performance compared to the model algorithm in previous studies [23].

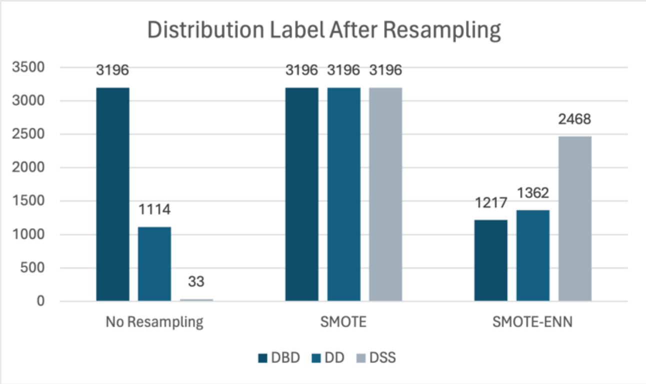

Figure 4. Label Distribution After Resampling

Splitting Dataset

Before dividing the dataset, standard scaling is applied to each feature, excluding ID and Label. Standard scaling is useful to ensure that features have a similar range of values, allowing the ML model to learn consistent patterns in the dataset more easily without being influenced by features with large value ranges, such as age and platelet. From each normalized dataset group, the data is then separated into two groups: the train set and the test set, with an 80:20 ratio.

Training

The model will be trained using an ARM-based device known for its M1 chip with an 8-core CPU, 16-core Neural Engine, and 8 GB RAM. Various studies have shown that SVM and XGBoost models can address multi-class classification problems with imbalanced class distributions [14], [24], [25], [26]. SVM is particularly effective in learning from multi-dimensional data and is better at mitigating the risk of overfitting in various studies. On the other hand, XGBoost is renowned for its reliable performance and speed, making it a popular and dependable approach for solving both classification and regression problems. This study aims to determine whether the dataset can be optimally implemented using SVM and XGBoost.

Hyperparameter Tuning

To achieve better results, the model can be further tuned in various ways. However, in this study, we attempt to optimize the model parameters using two approaches: GridSearchCV and the Particle Swarm Optimization (PSO) algorithm. GridSearchCV is used to test several common parameters to establish a baseline model for experimentation. In contrast, PSO optimizes by updating weights iteratively until the training is completed. As a result, PSO will yield the best parameter values to proceed to the performance analysis stage. Both algorithms have been proven to yield good results in optimizing SVM and XGBoost models [24], [27], [28], [29].

Model Evaluation

When dealing with imbalanced label conditions, metrics other than accuracy need to be considered to determine whether a model has good performance or not. This study also considers the F1 score, recall, and precision values.

RESULT AND DISCUSSION

Table 1 outlines the parameters defined for use in the model training scenarios. These parameters cover a range of possibilities, allowing the model to explore and optimize its values. These parameters were selected based on various scenarios from previous research, providing a broader perspective. By incorporating a diverse set of optimization scenarios, the researcher aims to thoroughly evaluate the potential of the SVM and XGBoost models.

Table 1 Parameter Settings

Model	Optimasi	Parameter Setting
SVM	PSO	K-Fold = 10 lb c = 1e-6 ub c = 1e6 lb gamma = 1e-6 ub gamma = 1e6 swarmsize = 10 scoring = f1 macro
XGBoost	GridSearchCV	max_depth: [3, 5, 7, 10] n_estimators: [50, 100, 200, 300] learning_rate: [0.001, 0.01, 0.1, 0.3, 0.5, 0.7] K-Fold = 10 scoring = f1 macro
	PSO	K-Fold = 10 lb max_depth = 1 ub max_depth = 10 lb n_estimators = 1 ub n_estimators = 100 lb learning rate = 1e-3 ub learning rate = 1 swarmsize = 10 scoring = f1 macro

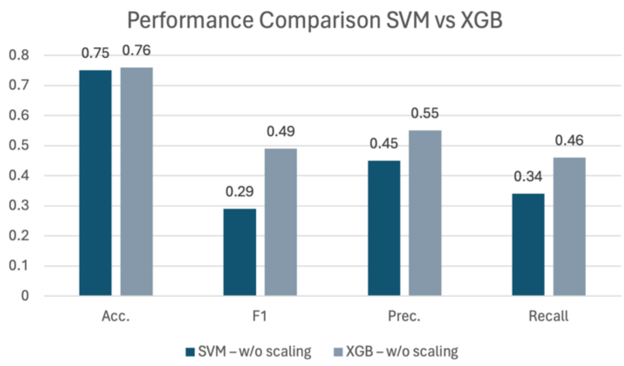

Figure 5 Performance Comparison of SVM vs. XGB

Figure 5 provides a visualization of initial experiments comparing two machine learning algorithms, SVM and XGBoost. This experiment serves as a baseline for optimization in subsequent scenarios. The data used in this experiment was not normalized using standard scaling, allowing for an assessment of the models' performance with raw data. The test results indicate that the XGBoost algorithm outperforms SVM. Due to the class imbalance in the data, accuracy is not a reliable metric for evaluating the models' performance. However, across all classification metrics, the XGBoost model consistently demonstrates superior performance compared to SVM. While the accuracy of both models does not differ significantly, a closer examination of the F1 metric reveals a 20% advantage for XGBoost, indicating its better capability to identify classification patterns compared to SVM.

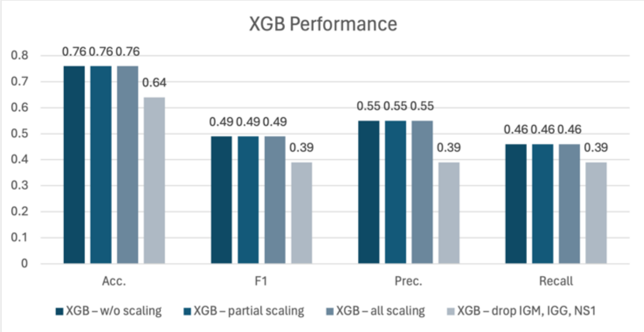

Figure 6. XGB Performance

Figure 6 explains a detailed experiment related to the XGBoost model with 4 scenarios: without normalization, partial normalization of numeric features such as age, platelet value, and hematocrit value, and finally, removing the IGM, IGG, and NS1 features to see if the model can still learn existing patterns in the dataset. The results show no significant difference regarding whether the data is normalized or not. However, consistently, the model experiences a decrease in performance when the IGM, IGG, and NS1 features are removed from the dataset.

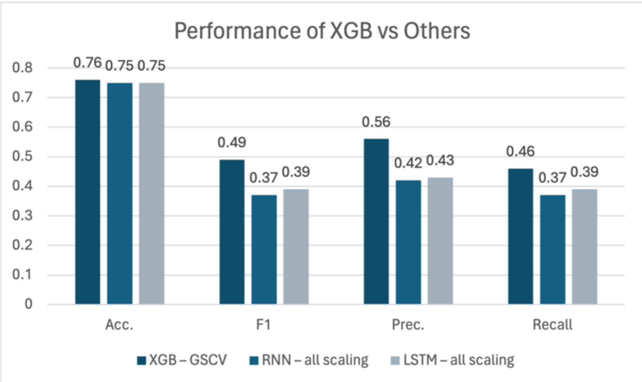

Figure 7. Performance XGB vs Deep Learning

The experiment continued by comparing the GridSearchCV-optimized XGBoost with other simple deep-learning algorithms, as shown in Figure 7. The RNN layer architecture used includes dense 128, dense 64, dropout 0.2, dense 32, dense 16, dropout 0.1, and dense out 3 with softmax activation. Meanwhile, the LSTM model architecture consists of LSTM layer 128 and LSTM layer 64 with return sequence true, dropout layer 0.2, LSTM layer 32, and dense out layer 3 with softmax activation. The results show that the XGBoost algorithm consistently outperforms other simple deep-learning architectures. Further experiments are needed to ensure performance comparisons. However, from this data, it is evident that the XGBoost algorithm remains highly reliable in identifying patterns within the dataset. Thus, the researchers decided not to proceed with experiments on deep learning-based models.

Figure 8 illustrates efforts to address the imbalanced label distribution using two approaches: SMOTE, which performs uniform resampling for all three classes, and SMOTE-

ENN, which augments the class with the least amount of data to match the largest class and applies to other classes. Figure 8 shows that XGBoost's performance consistently dominates when compared to SVM. The PSO and GridSearchCV algorithms show similar results, with an F1 score reaching 84%. The next test involves the hybrid SMOTE-ENN method to address data imbalance by making the minority class dominant, as seen in Figure 8. SVM appears to struggle to understand the data patterns compared to XGBoost. XGBoost with default parameters shows good performance. Although the GridSearchCV-optimized results appear similar, further analysis of the confusion matrix indicates that the model optimized with GridSearchCV performs better in recognizing classes.

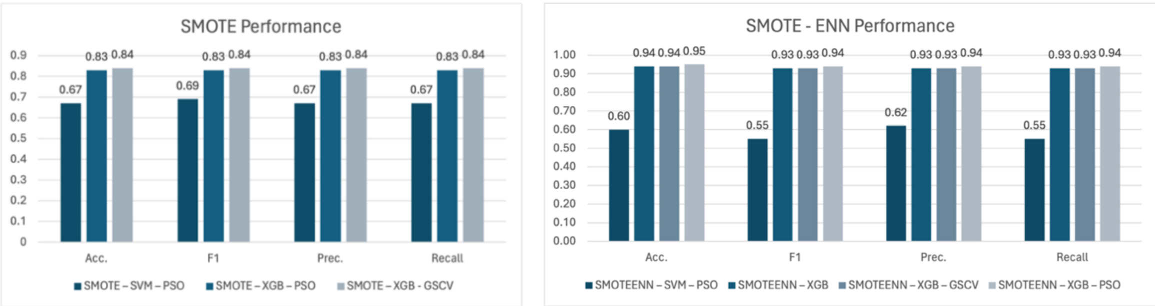

Figure 8 SMOTE and SMOTE-ENN Performance

Table 2. Experiment Results provides a summary of the results from 15 experimental scenarios to classify three classes: DBD, DD, and DSS, using laboratory data from patients infected with dengue fever as input. The features used as the baseline model include gender, age, platelet value, hematocrit value, IGM, IGG, and NS1.

Table 2. Experiment Results

Scenario Name	Acc.	F1	Prec.	Recall
SVM – w/o scaling	0.75	0.29	0.45	0.34
XGB – w/o scaling	0.76	0.49	0.55	0.46
XGB – partial scaling	0.76	0.49	0.55	0.46
XGB – all scaling	0.76	0.49	0.55	0.46
XGB – drop IGM, IGG, NS1	0.64	0.39	0.39	0.39
XGB – GSCV	0.76	0.49	0.56	0.46
RNN – all scaling	0.75	0.37	0.42	0.37
LSTM – all scaling	0.75	0.39	0.43	0.39
SMOTE – SVM – PSO	0.67	0.69	0.67	0.67
SMOTE – XGB – PSO	0.83	0.83	0.83	0.83
SMOTE – XGB - GSCV	0.84	0.84	0.84	0.84
SMOTEENN – SVM – PSO	0.60	0.55	0.62	0.55
SMOTEENN – XGB	0.94	0.93	0.93	0.93
SMOTEENN – XGB – GSCV	0.94	0.93	0.93	0.93
SMOTEENN – XGB – PSO	**0.95**	**0.94**	**0.94**	**0.94**

Table 3 shows the impact on model performance when undergoing resampling or not. From the 6 scenarios conducted, both SVM and XGB demonstrate improved performance when label resampling is applied. The SMOTE-ENN technique proves to be the most effective in enhancing accuracy and F1 score values and also directly correlates with improvements in precision and recall metrics across all scenarios.

Table 3. Effects Of Imbalance Handling

Scenario Name	Acc.	F1	Prec.	Recall
SVM	0.75	0.29	0.45	0.34
SMOTE – SVM – PSO	0.67	0.69	0.67	0.67
SMOTEENN – SVM – PSO	0.6	0.55	0.62	0.55
XGB	0.76	0.49	0.55	0.46
SMOTE – XGB – PSO	0.83	0.83	0.83	0.83
SMOTEENN – XGB – GSCV	0.94	0.93	0.93	0.93
SMOTEENN – XGB – PSO	**0.95**	**0.94**	**0.94**	**0.94**

The use of oversampling techniques, particularly SMOTE-ENN, has proven to be highly beneficial in addressing the challenge of imbalanced datasets. SMOTE-ENN not only increases the minority class samples but also removes noisy data, resulting in improved classification patterns and overall model performance. This combination of oversampling and noise removal enhances the model’s ability to recognize underrepresented patterns in the data, leading to more accurate predictions compared to other oversampling techniques, which might introduce unnecessary sample reductions or cause overfitting. Additionally, the integration of Particle Swarm Optimization (PSO) in this study has shown superior results when compared to traditional hyperparameter tuning methods such as GridSearchCV. PSO efficiently searches the solution space by balancing exploration and exploitation, which reduces the likelihood of the model being trapped in local optima. This advantage allows PSO to accelerate the training process by reducing the number of trials required to reach optimal hyperparameters, thereby speeding up the overall optimization process. By applying PSO, the training time is significantly decreased without compromising model accuracy.

Based on experiment result, the hybrid approach of SMOTE-ENN combined with XGBoost, optimized using PSO, resulted in a model that achieved an accuracy of 95% and an F1-score of 94%, with precision and recall both at 94%. These results suggest that the model is highly effective in classifying different types of dengue fever, as it is capable of learning and understanding the inherent patterns within the data. The hybrid approach of SMOTE-ENN, XGBoost, and PSO ensures that the model not only addresses the issue of data imbalance but also enhances predictive performance and generalization by optimizing the model’s hyperparameters efficiently. This hybrid method is therefore an optimal solution for dealing with complex healthcare datasets where imbalanced data and the need for efficient optimization are common challenges.

CONCLUSION

Dengue fever remains a significant public health challenge, particularly in regions with high transmission rates, where accurately predicting the severity of the disease is crucial for effective treatment and resource allocation. This study aimed to identify the most effective model architecture for predicting dengue fever severity, classifying cases into three distinct categories using laboratory data. Despite challenges posed by imbalanced data and the risk of overfitting, our findings demonstrate that the SMOTE-ENN technique effectively enhances the representation of minority classes, allowing the model to better capture important patterns and generalize across the data. Among the models tested, XGBoost consistently delivered the best performance in multi-class classification tasks. The hybrid approach combining SMOTE-ENN, XGBoost, and PSO effectively addresses the challenge of data imbalance while enhancing both predictive performance and generalization. By employing SMOTE-ENN, the model achieves better representation of minority classes, reducing the bias toward majority classes that can often hinder accurate classification. XGBoost further strengthens the model's performance with its capability to handle complex datasets and consistently deliver high accuracy in multi-class classification tasks.

The inclusion of PSO for hyperparameter optimization ensures that the model identifies optimal parameters efficiently, avoiding local optima and reducing the time required for tuning. This integrated approach results in a well-balanced and robust model, capable of making accurate predictions even in the presence of highly imbalanced data, as demonstrated by the high accuracy and F1 scores achieved in this study. Further optimization using Particle Swarm Optimization (PSO) led to the highest classification performance, achieving an accuracy of 95% and an F1-score of 94%. Additionally, this study can serve as a foundation for Federated Learning training, where models can be collaborated and aggregated with the hope of improving performance through weight aggregation from each model.

REFERENCES

[1] WHO, "Dengue and severe dengue," World Health Organization.

[2] A. Wilder-Smith, E.-E. Ooi, O. Horstick, and B. Wills, "Dengue," *The Lancet*, vol. 393, no. 10169, pp. 350–363, Jan. 2019, doi: 10.1016/S0140-6736(18)32560-1.

[3] M. R. Rondonuwu, L. Trisnantoro, Y. Pramono, I. Pambudi, and A. Utarini, "LAPORAN TAHUNAN KEMENKES 2022 - Demam Berdarah Dengue," 2022.

[4] J. W. Asare, P. Appiahene, and E. T. Donkoh, "Detection of anaemia using medical images: A comparative study of machine learning algorithms – A systematic literature review," *Inform Med Unlocked*, vol. 40, p. 101283, 2023, doi: 10.1016/j.imu.2023.101283.

[5] V. K. Verma and S. Verma, "Machine learning applications in healthcare sector: An overview," *Mater Today Proc*, vol. 57, pp. 2144–2147, 2022, doi: 10.1016/j.matpr.2021.12.101.

[6] W. Du *et al.*, "Early gastric cancer segmentation in gastroscopic images using a co-spatial attention and channel attention based triple-branch ResUnet," *Comput Methods Programs Biomed*, vol. 231, p. 107397, Apr. 2023, doi: 10.1016/j.cmpb.2023.107397.

[7] X. Chen, D. Yu, X. Fan, L. Wang, and J. Chen, "Multiclass Classification for Self-Admitted Technical Debt Based on XGBoost," *IEEE Trans Reliab*, vol. 71, no. 3, pp. 1309–1324, Sep. 2022, doi: 10.1109/TR.2021.3087864.

[8] B. Liu, B. Yan, Y. Zhou, Y. Yang, and Y. Zhang, "Experiments of Federated Learning for COVID-19 Chest X-ray Images," Jul. 2020.

[9] M. González Soto, B. Fernández Castro, R. P. Díaz Redondo, and M. Fernández Veiga, "XuILVQ: A River Implementation of the Incremental Learning Vector Quantization for IoT," in *Proceedings of the 19th ACM International Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, & Ubiquitous Networks on 19th ACM International Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, & Ubiquitous Networks*, New York, NY, USA: ACM, Oct. 2022, pp. 1–8. doi: 10.1145/3551663.3558676.

[10] M. Beckmann, N. F. F. Ebecken, and B. S. L. Pires de Lima, "A KNN Undersampling Approach for Data Balancing," *Journal of Intelligent Learning Systems and Applications*, vol. 07, no. 04, pp. 104–116, 2015, doi: 10.4236/jilsa.2015.74010.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

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

[13] T. Nguyen, K. Mengersen, D. Sous, and B. Liquet, "SMOTE-CD: SMOTE for compositional data," *PLoS One*, vol. 18, no. 6, p. e0287705, Jun. 2023, doi: 10.1371/journal.pone.0287705.

[14] B. S. Yelure, S. V. Patil, S. B. Patil, and S. B. Nemade, "Solving Multi-Class Classification Problem Using Support Vector Machine," in *2022 International Conference on Futuristic Technologies (INCOFT)*, IEEE, Nov. 2022, pp. 1–4. doi: 10.1109/INCOFT55651.2022.10094370.

[15] S. Sharma and K. Guleria, “A comprehensive review on federated learning based models for healthcare applications,” *Artif Intell Med*, vol. 146, p. 102691, Dec. 2023, doi: 10.1016/j.artmed.2023.102691.

[16] A. G. Gad, “Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review,” *Archives of Computational Methods in Engineering*, vol. 29, no. 5, pp. 2531–2561, 2022, doi: 10.1007/s11831-021-09694-4.

[17] S. Prabu, B. Thiyaneswaran, M. Sujatha, C. Nalini, and S. Rajkumar, “Grid Search for Predicting Coronary Heart Disease by Tuning Hyper-Parameters,” *Computer Systems Science and Engineering*, vol. 43, no. 2, pp. 737–749, 2022, doi: 10.32604/csse.2022.022739.

[18] N. Agrawal and S. Panigrahi, “A Comparative Analysis of Fraud Detection in Healthcare using Data Balancing & Machine Learning Techniques,” in *2023 International Conference on Communication, Circuits, and Systems (IC3S)*, IEEE, May 2023, pp. 1–4. doi: 10.1109/IC3S57698.2023.10169634.

[19] T. Kosolwattana, C. Liu, R. Hu, S. Han, H. Chen, and Y. Lin, “A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare,” *BioData Min*, vol. 16, no. 1, p. 15, Apr. 2023, doi: 10.1186/s13040-023-00330-4.

[20] Z. Rustam, D. A. Utami, R. Hidayat, J. Pandelaki, and W. A. Nugroho, “Hybrid Preprocessing Method for Support Vector Machine for Classification of Imbalanced Cerebral Infarction Datasets,” *Int J Adv Sci Eng Inf Technol*, vol. 9, no. 2, p. 685, Apr. 2019, doi: 10.18517/ijaseit.9.2.8615.

[21] Sanjeet Bagcchi, “Dengue outbreak in Peru affects adults and children,” *Lancet Infect Dis*, vol. 23, no. 9, p. e339, Sep. 2023, doi: 10.1016/S1473-3099(23)00229-3.

[22] L. Camacho, G. Douzas, and F. Bacao, “Geometric SMOTE for regression,” *Expert Syst Appl*, vol. 193, p. 116387, May 2022, doi: 10.1016/j.eswa.2021.116387.

[23] S. Parthasarathy, V. Jayaraman, and J. P. Princy R, “Predicting Heart Failure using SMOTE-ENN-XGBoost,” in *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, IEEE, Jan. 2023, pp. 661–666. doi: 10.1109/IDCIoT56793.2023.10053458.

[24] C. Sheng and H. Yu, “An optimized prediction algorithm based on XGBoost,” in *2022 International Conference on Networking and Network Applications (NaNA)*, IEEE, Dec. 2022, pp. 1–6. doi: 10.1109/NaNA56854.2022.00082.

[25] H. Limam, A. Zouhair, and W. Oueslati, “A New Hybrid Multiclass Approach Based on KNN and SVM,” *Journal of Information & Knowledge Management*, vol. 21, no. 04, Dec. 2022, doi: 10.1142/S0219649222500617.

[26] P. Zhang, Y. Jia, and Y. Shang, “Research and application of XGBoost in imbalanced data,” *Int J Distrib Sens Netw*, vol. 18, no. 6, p. 155013292211069, Jun. 2022, doi: 10.1177/15501329221106935.

[27] G. N. Ahmad, H. Fatima, S. Ullah, A. Salah Saidi, and Imdadullah, “Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV,” *IEEE Access*, vol. 10, pp. 80151–80173, 2022, doi: 10.1109/ACCESS.2022.3165792.

[28] Z. M. Alhakeem, Y. M. Jebur, S. N. Henedy, H. Imran, L. F. A. Bernardo, and H. M. Hussein, “Prediction of Ecofriendly Concrete Compressive Strength Using Gradient Boosting Regression Tree Combined with GridSearchCV Hyperparameter-Optimization Techniques,” *Materials*, vol. 15, no. 21, p. 7432, Oct. 2022, doi: 10.3390/ma15217432.

[29] J. Wang, X. Wang, X. Li, and J. Yi, “A Hybrid Particle Swarm Optimization Algorithm with Dynamic Adjustment of Inertia Weight Based on a New Feature Selection Method to Optimize SVM Parameters,” *Entropy*, vol. 25, no. 3, p. 531, Mar. 2023, doi: 10.3390/e25030531.