Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2025. Vol. 4. No. 3; e-ISSN: 2808-4519

Performance Analysis of SMOTE and SMOTEN Techniques for Daily Rainfall Classification using XGBoost

Najwa Laila Aggraini[1]*, Basuki Rahmat[2], Achmad Junaidi[3]

[1,2,3] *Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jawa Timur*
21081010191@student.upnjatim.ac.id[1], basukirahmat.if@upnjatim.ac.id[2], achmadjunaidi.if@upnjatim.ac.id[3]*

Abstract

The vital function that rainfall patterns fulfill in diverse sectors of life, including agriculture, water management, and disaster mitigation, has engendered the necessity for an accurate rainfall classification system to facilitate early warning and decision-making. However, the development of a classification system is often encumbered by various obstacles, with data imbalance being a prominent one. The objective of this study is to analyze two data resampling techniques, namely SMOTE and SMOTEN, with the aim of improving the performance of the XGBoost classification model. The dataset utilized is accessible on the BMKG website and is classified into five categories. Subsequent to the preprocessing stage, the data is divided by two schemes: 70:30 and 80:20. The determination of the sensitivity of each dataset is achieved through variations in the number of folds in cross validation and the use of learning rates. The experimental results indicate that the SMOTE configuration, with a data division proportion of 80:20 using 10 folds and a learning rate of 0.15, attains the maximum accuracy value of 92.92%. This represents a substantial enhancement from the original dataset accuracy result of 75.36% and surpasses the SMOTE experimental results with an accuracy of 90.58%. Consequently, SMOTEN was found to be superior and effective in managing the imbalance of numerical and categorical datasets, thereby enhancing the performance of the XGBoost model in daily rainfall classification.

Keywords: *Daily Rainfall, XGBoost, SMOTE, SMOTEN, Classification*

1. Introduction

Precipitation plays a pivotal role in diverse domains of life, including agriculture, water resource management, and the miti gation of hydrometeorological disasters. According to a report from the Meteorology, Climatology and Geophysics Agency (BMKG), there has been a 2,784-mm increase in the rate of change in annual rainfall over the past 30 years, with a minimum decrease of 750 mm[1]. Changes in rainfall patterns have been demonstrated to exert an influence on the intensity of natural disasters. According to a report by the Indonesian National Board for Disaster Management (BNPB) in 2023, East Java experienced a total of 41 flood events, 5 landslides, and 59 extreme weather events[2].

In light of the observed phenomenon, the development of a daily rainfall classification system is imperative. Such a system would facilitate the prediction of daily rainfall levels, thereby enabling the implementation of proactive planning and disaster management st rategies. A prevalent approach in classification systems is extreme gradient boosting, also known as XGBoost. XGBoost employs an ensemble technique that is predicated on decision trees. However, a significant challenge arises in the pursuit of high accuracy values, namely the issue of data imbalance. In order to address the issue of data imbalance, the research employed resampling techniques, specifically SMOTE and SMOTEN.

Research conducted by Haumahu et al., 2021 [3] sought to predict hoax news and valid news using the XGBOOST method. The findings of the present study demonstrate that XGBoost is capable of attaining 89% accuracy, 90% precision, and 80% recall, in addition to true positive results. The study also reveals that there are 44 true negative results, 6 false negative results, and 2 false positive results. Wiwaha's research examined the optimization of XGBoost accuracy in rainfall classification using resampling techniques. The study demonstrated that SMOTE and XGBoost achieved the highest accuracy of 99.981%, while SMOTE attained an accuracy result of 99.9435 [4]. Subsequently, Yang and Guan (2023) implemented the XGBoost algorithm to predict heart disease through SMOTE to address the data imbalance problem. The accuracy result obtained by SMOTE-XGBoost is 93.44% [5]. Sapari et al. also conducted research to classify air quality using XGBoost, achieving an accuracy of 98.61% [6].

Preliminary research indicates that employing the XGBoost model with datasets that have been balanced using SMOTE and SMOTEN resampling techniques results in enhanced model performance. SMOTE and SMOTEN are algorithms that generate synthesis samples by calculating interpolation from the nearest neighbours of the minority class. Nevertheless, SMOTEN incorporates a voting system into categorical features.

According to the explanation, the objective of this study is to make a comparison between the use of SMOTE and SMOTE data resampling techniques in improving the accuracy of daily rainfall classification using the XGBoost model. The dataset under consideration comprises a series of columns containing meteorological data, including but not limited to date, minimum temperature, average temperature, average humidity, duration of irradiation, and rainfall. The following assessment parameters were utilized in the evaluation: accuracy, precision, recall, F1-score, and AUC score.

2. Research Method

The research phase commenced with the acquisition of datasets from the BMKG's official website, accessed through https://dataonline.bmkg.go.id/data-harian. The data set under consideration enc ompasses daily rainfall measurements collected from 2009 to 2024. Subsequently, data preprocessing entails the cleansing of data and its standardization. Subsequently, resampling techniques with SMOTE and SMOTEN were implemented in a series of experiments. Subsequent to achieving a balanced proportion of class distribution, the data was divided into three segments: 80% for training, 10% for validation, and 10% for testing. The train data that has been divided is then trained using the XGBoost model for classification. Subsequently, the model is evaluated using 10-fold cross-validation. Subsequently, the trained model is subjected to analysis to ascertain the evaluation results based on the specified assessment parameters, namely accuracy, precision, recall, f1-score, and AUC score. The flow of this research methodology can be seen in Fig 1.

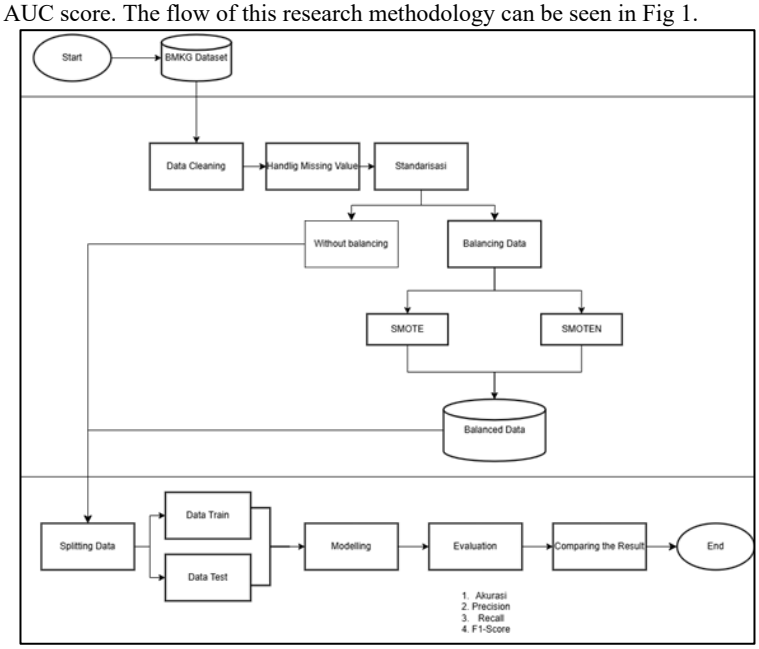

Fig. 1: System Design

2.1. Datasets

In this research, rainfall data of East Java Province from 2009 to 2024 is used. The data is provided in .csv and .pdf formats, in the form of rainfall records in millimeters and some other information such as temperature, humidity, length of irradiation, wind speed, and wind direction with numerical data types. Table 1 showsthe number of datasets before and after the resampling process.

Table 1: Data Shape

Preprocessing	Data Shape
Before Resampling	5.804 × 9
After Resampling	8.580 × 8

As illustrated in Table 1, the initial dataset contains 5,804 data points prior to the implementation of the resampling process. This data set has been meticulously compiled from a variety of relevant sources. However, due to an imbalance in the number of classes in the target variable, a resampling technique using the SMOTE and SMOTEN methods was employed to balance the class distribution. Following the implementation of the resampling process, the dataset increased to 8,580 data points. This increase is attributed to the utilization of SMOTE and SMOTEN, which generate synthetic samples from the minority class, thereby enhancing the proportion of distribution between classes and supporting the model training process to ensure that it is not biased towards the majority class.

Table 2: Rainfall Category

Rainfall Intensity	Description
0-5 mm/day	Not Rainy
5-20 mm/day	Light Rain
20-50 mm/day	Moderate Rain
50-100 mm/day	Heavy Rain
>100 mm/day	Extreme Rain

Table 2 presents the classification of rainfall intensity in this study refers to the standard grouping based on the amount of daily rainfall observed. The table indicates that rainfall is categorized into five classes, namely: Rainfall was categorized as follows: No Rain (0-5 mm/day), Light Rain (5-20 mm/day), Moderate Rain (20-50 mm/day), Heavy Rain (50-100 mm/day), and Extreme Rain (>100 mm/day). This grouping is employed to transform numerical rainfall data into a categorical form, thereby enabling the application of classification techniques with algorithms such as XGBoost. Moreover, this classification system enables the interpretation of prediction results and the implementation of practical decisions, such as the provision of early warnings in disaster situations or the adjustment of planting schedules in the agricultural sector.

2.2. SMOTE (Synthetic Minority Oversampling Technique)

The synthetic minority over-sampling technique (SMOTE) is a popular method employed to address data imbalance[7]. SMOTE is a data mining technique that is implemented through the generation of new samples using interpolation techniques from the nearest neighbors of the minority class. SMOTE is a simple concept with the potential to enhance the efficacy of the model through the optimization of data proportion[8]. SMOTE is a data mining technique that involves two primary steps: first, calculating the Euclidean distance between the original data points; and second, generating synthetic samples based on the calculated distance. The following functions are available:

$$d = \sqrt{(x_1 + y_1)^2 + (x_2 + y_2)^2 + \cdots + (x_n + y_n)^2} \quad (1)$$

Where:
d: Euclidean distance
x: coordinate values of the first point
y: coordinate values of the second point

$$x_{synthetic} = x_i + \lambda \,.\, (y_i - x_i) \quad (2)$$

Where:
X synthetic: synthesis sample
xi: original minority sample
yi: nearest neighbour
λ : a random number between 0 and 1

2.3. SMOTEN (Synthetic Minority Oversampling for Nominal)

SMOTEN (Synthetic Minority Oversampling for Nominal) is a variant of the SMOTE method specifically designed to handle classification datasets with more than two labels and categorical features. Its primary function is to generate synthetic data for the minority class, thereby helping to reduce model bias towards the majority class and improve classification performance. A multitude of studies have demonstrated the efficacy of SMOTEN in various domains, including drug and alcohol rehabilitation outcome prediction by Robert-Licklider & Trafalis, 2024[9], cardiovascular disease classification by Garcia-Vincente., et al 2023[10] and cyberattack detection by Alabrah 2023[11] . SMOTEN's integration with various machine learning and normalization techniques demonstrates its flexibility across multiple application domains. SMOTEN has been shown to be a very important tool in overcoming data imbalance and ensuring a fairer and more accurate analysis process, due to its ability to precisely generate synthetic samples on categorical data[4].

2.4 XGBOOST

XGBoost is an ensemble-based machine learning algorithm that functions by iteratively combining weak models to form an optimal model. The process entails the calculation of the residual error from each iteration, with the objective of constructing a new model that corrects the previous error. In order to circumvent the issue of overfitting, XGBoost has been developed to incorporate L1 and L2 regulation techniques. Furthermore, XGBoost is characterised by two primary categories of parameters: general parameters and hyperparameters. The latter are employed to establish the configuration of the decision tree prior to the initiation of the training process. The following components are integral to the XGBoost algorithm:

- Gradient and hessian of the loss function:

$$g_i = \hat{y}_i^{(t-1)} - y_i \quad (3)$$

$$h_i = \hat{y}_i^{(t-1)} \cdot (1 - \hat{y}_i^{(t-1)}) \quad (4)$$

- Gain:

$$Gain = \frac{g_L^2}{hL + \lambda} + \frac{g_R^2}{hR + \lambda} - \frac{(gL + gR)^2}{hL + hR + \lambda} \quad (5)$$

- Node:

$$w = \frac{\sum g_i}{\sum hi + \lambda} \quad (6)$$

- Update model:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \cdot w \quad (7)$$

3. Result and Discussion

Each dataset was implementedin 8 experiments to see the generalization and stability of the resampling technique. The following experimental results are presented in Table 3, Table 4, and Table 5.

Table 3: Imbalanced dataset Experiment

Dataset	Data Split	Fold	Learning rate	Accuracy	Precision	Recall	F1Score	AUC Score	komputation(s)
Imbalanced	80:20	5	0.15	75,28	33,15	25,18	25,59	54,43	7,8
	70:30	5	0.15	75,28	33,15	25,18	25,59	54,43	7,71
	80:20	10	0.15	75,29	29,64	24,82	24,97	54,25	11,54
	70:30	10	0.15	75,29	29,64	24,82	24,97	54,25	11,65
	80:20	5	0.3	74,81	29,98	25,05	25,37	54,44	7,31
	70:30	5	0.3	74,81	29,98	25,05	25,37	55,05	6,81
	80:20	10	0.3	75,36	34,64	25,76	26,43	54,94	8,99
	70:30	**10**	**0.3**	**75,36**	**34,64**	**25,76**	**26,43**	**54,94**	**8,46**

As illustrated in Table 3, the accuracy value demonstrates a high degree of stability across all experiments, remaining consi stently low. Among the eight experiments, the highest level of accuracy was recorded at 75.36%, while the lowest was 74.81%, indicating a marginal difference of 0.55%. This finding suggests the presence of a bias effect on the majority class, a phenomenon that can be attributed to the limitations inherent in a learning model that is exclusively focused on the majority class. Therefore, the precision of this method is notably high, despite the suboptimal prediction outcomes observed for other classes. Conversely, the precision and recall values demonstrate suboptimal performance across experiments utilizing this imbalanced dataset. The precision value ranges from 29.98% to 34.64%, while the recall value ranges from 24.82% to 25.76%. The observed phenomenon can be attributed to the model's conservative learning approach, which has led to diminished recall and an reduced F1-score. This is a result of the model's inability to effectively detect minority classes. The findings further suggest an overfitting issue in the model, which exhibits a tendency to favor the minority class. Despite the augmentation of the learning rate to 0.3, the outcomes remain incapable of enhancing the recall value, signifying that the model is unable to predict the minority class. The area under the curve (AUC) score derived from the imbalanced dataset exhibits a similar trend, with a range from 54.25 to 54.94. This value suggests that the model exhibits a marked inability to differentiate between minority and majority classes. Subsequently, the imbalance dataset demonstrates a comparatively brief mean computation duration. The experiment with the combination of split data 70:30, fold 10, and learning rate 0.3 is the best-performing model. The accuracy obtained was 75.36%, the precision was 34.76%, the recall and f1-score were 26.43%, and the AUC score was 54.94. The computation time on the model is also relatively short, at only 8.46 seconds. The result obtained in this instance exhibits notable similarity to that of Experiment 7, with a data proportion of 80:20 being observed for all assessed parameters. However, it should be noted that the computational time incurred was marginally elevated, with a recorded duration of 0.53 seconds.

Table 4: SMOTE Technique Experiment

Technique	Data Split	Fold	Learning rate	Accuracy	Precision	Recall	F1Score	AUC Score	komputation(s)
SMOTE	80:20	5	0.15	90,07	90,07	90,07	89,97	93,79	28,11
	70:30	5	0.15	90,07	90,07	90,07	89,97	93,79	27,72
	80:20	**10**	**0.15**	**90,58**	**90,57**	**90,58**	**90,49**	**94,11**	**45,63**
	70:30	10	0.15	90,58	90,57	90,58	90,49	94,11	47,14
	80:20	5	0.3	89,41	89,37	89,41	89,3	93,37	24,26
	70:30	5	0.3	89,41	89,37	89,41	89,3	93,37	21,98
	80:20	10	0.3	90,04	90	90,04	89,94	93,37	36,3
	70:30	10	0.3	90,04	90	90,04	89,94	93,37	37,92

marginally extended, ranging from 21.98 seconds to 47.14 seconds. This phenomenon can be attributed to the fact that the resampling data set also increases, necessitating additional iterations during the training process to accommodate the full data set. In all experiments utilizing the SMOTE dataset, the optimal outcomes were attained by the model employing a combination of 80:20 for the proportion of split data, 10-fold cross validation, and a learning rate of 0.15 in the third experiment. The accuracy, precision, and recall values were 90.58%, which is the highest of the eight experiments.

Table 5: SMOTEN Technique Experiment

Technique	Data Split	Fold	Learning rate	Accuracy	Precision	Recall	F1Score	AUC Score	komputation(s)
SMOTEN	80:20	5	0.15	92,7	92,72	92,7	92,69	95,43	26,55
	70:30	5	0.15	92,7	92,72	92,7	92,69	95,43	48,16
	80:20	**10**	**0.15**	**92,92**	**92,94**	**92,92**	**92,92**	**95,57**	**37,16**
	70:30	10	0.15	92,92	92,94	92,92	92,92	95,57	38,94
	80:20	5	0.3	92,5	92,51	92,5	92,49	95,31	21,3
	70:30	5	0.3	92,5	92,51	92,5	92,49	95,31	20,69
	80:20	10	0.3	92,7	92,71	92,7	92,69	95,94	33,45
	70:30	10	0.3	92,7	92,71	92,7	92,69	95,43	34,13

Table 5 shows the accuracy results from SMOTEN experiments. Out of the 8 experiments, the resulting accuracy results reached the highest value of the previous two datasets, which was in the range of 92.5% to 92.92%. The highest accuracy is achieved by most of the configurations, which shows the accuracy of the model in performing classification. The values of precision, recall and f1-score were all consistent at 92.51% to 92.94%. In fact, almost all experiments showed a balance in the model's ability to predict all classes, both majority and minority. The AUC value also increased from 95.31% to 95.57%. As far as the experiments are concerned, the use of the SMOTEN dataset outperforms the results of other datasets such as SMOTE and imbalance. The best configuration time of the model using SMOTEN is about 37.16 seconds less than the best configuration of SMOTE. With a learning rate of 0,15, a configuration of 80:20 to 70:15 split data with a 10-fold cross-validation produces best results with highest accuracy, precision, recall, f1 value of 92,92%. and AUC value of 37,16 seconds in the third experiment and 38,94 seconds in the fourth experiment. results.

3.1. Performance Comparison of Each Techniques

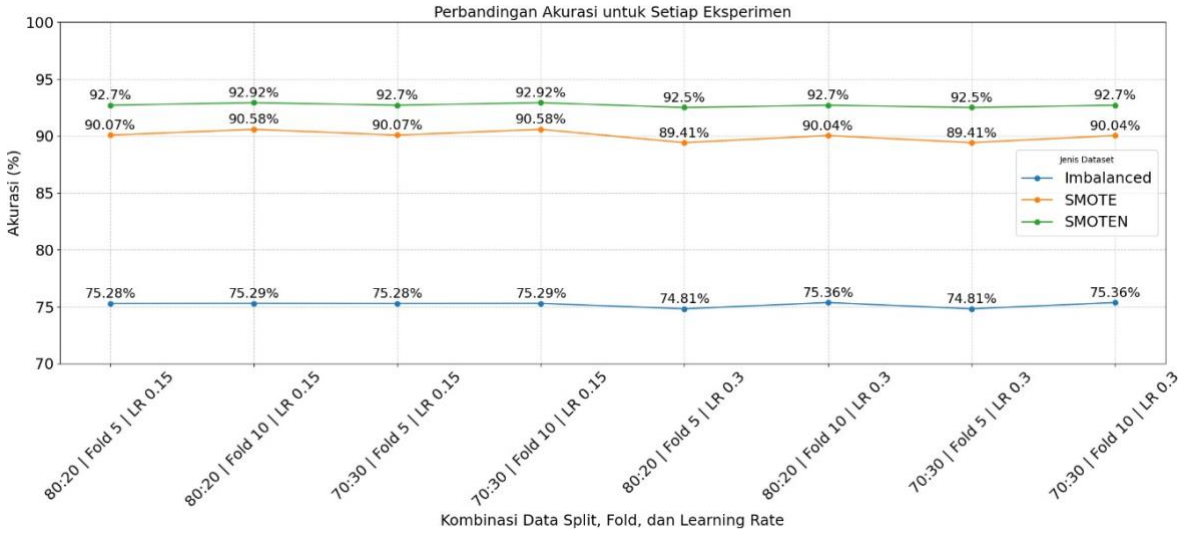

Fig. 2: Accuracy comparison chart

Figure 2 presents a comparison of model accuracy values from all experiments utilizing the three datasets. The blue line in the graph, which represents the model utilizing the imbalanced dataset, demonstrates an accuracy range of 74.81% to 75.35%. This finding indicates that variations in data division, fold, and learning rate differences are unable to counteract the model bias that is associated with the class balance problem in the dataset.

Experiments employing the SMOTE dataset, as depicted by the orange line, yielded favorable outcomes. The accuracy value exhibited a substantial increase, reaching 89.41% and achieving a maximum of 90.58%. However, the graph reveals fluctuations in the combination of using a learning rate of 0.3 and 5 fold. The application of a learning rate set at a value of 0.15 has been demonstrated to consistently yield optimal results, irrespective of the utilized configuration.

Consequently, the efficacy of SMOTEN in addressing data imbalance problems characterized by categorical features is marginally superior and more consistent in comparison to the SMOTE technique.

3.2. Comparison of Model Computational Time

In addition to the previously analysed performance metrics, other metrics such as computation time are also important considerations in determining the best model combination.

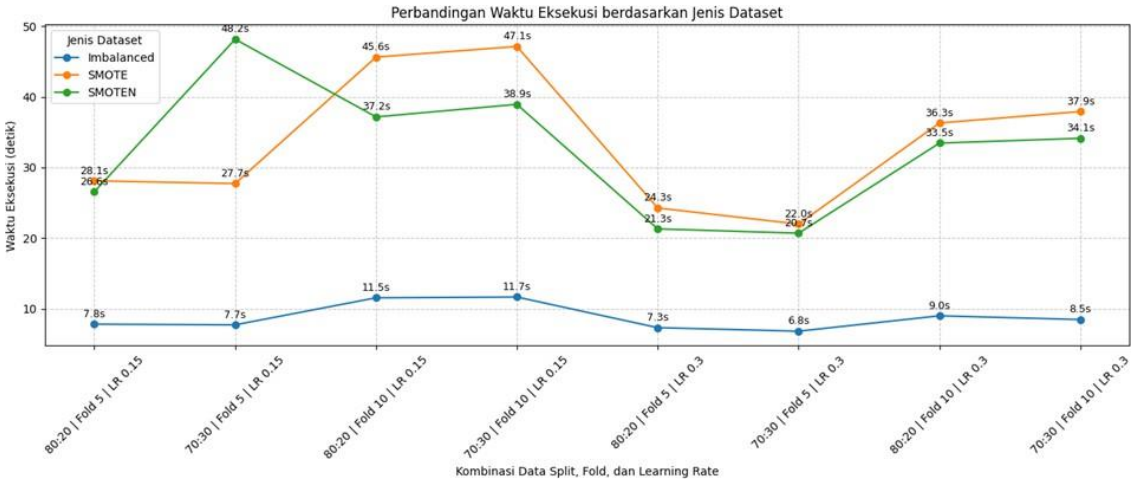

Fig. 3: Model computation timeline diagram

As illustrated in Figure 3, the computational time for each experiment is presented. The graph illustrates that the duration for the eight experiments utilizing the original or unbalanced dataset demonstrates the most efficient performance across all combinations conducted. This phenomenon can be attributed to the smaller size of the utilized dataset, which limits the capacity of the model to learn from the additional synthetic data.

However, it should be noted that once SMOTE and SMOTEN are applied, the computational time required is nearly equivalent. However, SMOTE, represented by the orange line, exhibited the longest average time of all the experiments. The SMOTE range is 21.98 for the sixth experiment with a data split of 75:15:15, 5 folds, and a learning rate of 0.3. The SMOTEN technique generates an average time that is shorter than SMOTE and longer than the original dataset. The shortest computation was generated by the fifth experiment in the SMOTEN dataset scenario, which took 21.3 seconds. Experiment 2, which required the most time, required 48.16 seconds to complete, with a 70:30 dataset split, 5 folds, and a learning rate of 0.15.

4. Conclusion

This study aims to compare the performance of SMOTE and SMOTEN resampling techniques in handling data imbalance in categorical features using the XGBoost algorithm. Eight experimental scenarios with imbalanced datasets demonstrate that while the accuracy reaches approximately 75%, the precision, recall, F1-score, and AUC score are all remarkably low, with the majority being below 40%. This finding suggests a bias toward the majority class, indicating that variations in parameters such as data split, fold, and learning rate do not yield positive outcomes without the implementation of resampling techniques. The application of SMOTE to the eight experimental scenarios demonstrated a substantial enhancement in performance, with accuracy reaching 89.41% to 90.57%. The optimal scenario attained 90.58% accuracy, 90.57% precision, 90.58% recall, 90.49% F1-score, and 94.11% AUC score. Despite the enhanced performance, the model with SMOTE remains vulnerable to parameter alterations, and it requires a greater amount of computation time compared to the other scenarios. SMOTEN demonstrated the optimal and most consistent performance among the three approaches evaluated. The optimal results were obtained with 92.92% accuracy, 92.94% precision, 92.92% recall, 92.92% F1-score, and 95.57% AUC score. SMOTEN demonstrated superior stability in the face of parameter variations, including learning rate, data split, and fold, while exhibiting more efficient computation time compared to SMOTE.

Acknowledgement

I would like to express my sincere appreciation to all those who have provided support during this research process. My special thanks go to my supervisors and mentors for their valuable guidance and input. I also appreciate the participation of various parties who contributed to the implementation of this research. I hope this report can provide benefits, and I am open to all constructive suggestions for future improvements.

References

[1] BMKG, “Analisis Laju Perubahan Curah Hujan Tahunan.” Accessed: Dec. 17, 2024. [Online]. Available: https://www.bmkg.go.id/iklim/analisis-laju-perubahan-curah-hujan

[2] “Data Bencana di Tingkat Kabupaten/Kota dapat dilihat file pdf Buku Data Bencana Indonesia 2023 pada qris berikut ini.”

[3] J. P. Haumahu, S. D. H. Permana, and Y. Yaddarabullah, "Fake news classification for Indonesian news using Extreme Gradient Boosting (XGBoost)," *IOP Conf Ser Mater Sci Eng*, vol. 1098, no. 5, p. 052081, Mar. 2021, doi: 10.1088/1757- 899x/1098/5/052081.

[4] D. D. Wiwaha, D. A. Gafyunedi, Z. M. Mahdi, I. W. Putro, B. A. Pramudita, and D. P. Setiawan, "Enhancing Rainfall Prediction Accuracy through XGBoost Model with Data Balancing Techniques," in *2024 20th IEEE International Colloquium on Signal Processing and Its Applications, CSPA 2024 - Conference Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 120–125. doi: 10.1109/CSPA60979.2024.10525558.

[5] J. Yang and J. Guan, "A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm," *Information (Switzerland)*, vol. 13, no. 10, Oct. 2022, doi: 10.3390/info13100475.

[6] A. M. Sapari, A. Id Hadiana, and F. R. Umbara, "Air Quality Classification Using Extreme Gradient Boosting (XGBOOST) Algorithm ARTICLE INFORMATION ABSTRACT," 2023. [Online]. Available: http://innovatics.unsil.ac.id

[7] M. Alamri and M. Ykhlef, "Hybrid Undersampling and Oversampling for Handling Imbalanced Credit Card Data," *IEEE Access*, vol. 12, pp. 14050–14060, 2024, doi: 10.1109/ACCESS.2024.3357091.

[8] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," *IEEE Trans Neural Netw Learn Syst*, vol. 34, no. 9, pp. 6390–6404, Sep. 2023, doi: 10.1109/TNNLS.2021.3136503.

[9] K. Roberts-Licklider and T. Trafalis, "Machine Learning Techniques with Fairness for Prediction of Completion of Drug and Alcohol Rehabilitation," Apr. 08, 2024. doi: 10.21203/rs.3.rs-4208301/v1.

[10] C. García-Vicente *et al.*, "Evaluation of Synthetic Categorical Data Generation Techniques for Predicting Cardiovascular Diseases and Post-Hoc Interpretability of the Risk Factors," *Applied Sciences (Switzerland)*, vol. 13, no. 7, Apr. 2023, doi: 10.3390/app13074119.

[11] A. Alabrah, "An Improved CCF Detector to Handle the Problem of Class Imbalance with Outlier Normalization Using IQR Method," *Sensors*, vol. 23, no. 9, May 2023, doi: 10.3390/s23094406.