Data Augmentation for Rainfall Classification in Bogor and Palu using SMOTE

by Arbi Haza Nasution

Submission date: 05-May-2025 11:04AM (UTC+0700)
Submission ID: 2666482262
File name: on_for_Rainfall_Classification_in_Bogor_and_Palu_Using_SMOTE.pdf (359.41K)
Word count: 5488
Character count: 28789

2024 International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics (ICoABCD)

Data Augmentation for Rainfall Classification in Bogor and Palu using SMOTE

Arbi Haza Nasution
Department of Informatics Engineering Universitas Islam Riau
Pekanbaru, Indonesia
arbi@eng.uir.ac.id

Winda Monika
Department of Library Science Universitas Lancang Kuning
Pekanbaru, Riau
windamonika@unilak.ac.id

2024 International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics (ICoABCD) | 979-8-3503-6845-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICOABCD63526.2024.10704551

***Abstract*—Weather predictions provide detailed and accurate information about weather conditions, including cloud cover, rainfall, temperature, humidity, wind conditions, and sun exposure. Indonesia, a tropical country with only rainy and dry seasons, experiences relative variations in these seasons. Due to global warming, rainfall in Indonesia has become increasingly unpredictable. This study addresses the class imbalance in rainfall data by using data from Bogor, which has the highest rainfall, and Palu, which has the lowest rainfall in Indonesia. The aim is to develop a robust classification method for imbalanced data using the Synthetic Minority Over-sampling Technique (SMOTE). We combined datasets from Bogor and Palu to create a balanced baseline dataset. We then evaluated four classification methods—Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), and Kernel Support Vector Machine (KSVM)—using precision, recall, and F1-score as evaluation metrics. Our results indicate that KSVM outperformed the other methods in terms of robustness to class imbalance with F1-score of 48.6%, 67.1%, 67.4%, 71.3% for the dataset of Bogor, Bogor with SMOTE, Palu, and Palu with SMOTE respectively. The application of SMOTE improved the F1-score by 38% for the Bogor dataset and 6% for the Palu dataset.**

Keywords*— *SMOTE, imbalanced classification, data augmentation

I. Introduction

Global warming inflicts climate change that drastically alters precipitation patterns around the world, increasing rainfall variability and the occurrence of droughts or floods caused by extreme rainfall [1]. Meanwhile, precipitation is one of the important water cycles for survival on Earth. Excessive or insufficient rainfall in an area can affect human life. Therefore, knowing the quantity of rainfall is also very substantial. This data is essential across various domains, including water resource administration, like issuing flood alerts, actively managing urban drainage networks, gauging flood hazards, and identifying potential climate conditions that could lead to rainfall [2].

The considerable influence of hydrometeorological catastrophes on tropical countries like Indonesia underscores the importance of undertaking measures to prevent such disasters. Indonesia mostly has two seasons: the rainy season and the dry season. While these are the general patterns, the intensity and duration of these seasons can vary across different regions of Indonesia, particularly between the western and eastern parts. Also, some variations and unpredictability in these patterns might occur due to climate change. Drought is a possibility in areas with low rainfall and limited water reserves. As a result, droughts are closely related to rainfall distribution throughout the year in rainfed areas, and water stress is likely to be seen in years over time [3]. One approach involves creating predictive models for gauging the intensity of rainfall. To comprehensively grasp these instances of heavy rainfall, it will be essential to conduct analyses that encompass meteorological occurrences on various levels [4]. Accurate prediction of rainfall data can model future river flow changes and can also help disaster management during floods [5]. It may be difficult to predict where and when a rainfall-induced landslide will fail, and it may be prohibitively expensive to take action to prevent erosion or protect populations from its harmful effects [6].

Machine learning models are systems that can learn automatically from pre-existing data without programming extensively through a set of commands [7]. There are as many as 16 different Machine Learning algorithms found in 24 different learning where this algorithm is divided into three types: 7 supervised learning (SL), 4 unsupervised learning (USL), and 5 semi-supervised learning (SSL) [8]. A classification is an object that is grouped into a category [9]. Machine learning techniques that can be used to classify are supervised learning techniques. The classification process uses supervised learning techniques with several classification methods. Classification using machine learning has also been carried out using several different methods. The existence of different methods can help in getting the best prediction results by comparing several classification methods with several metrics. Machine learning applications have been widely used to get good results in predicting rainfall. Numerous studies have used machine learning classifiers to combine different machine learning classification algorithms based on ensembles to predict rainfall. The strategy offers the chance to take advantage of every opportunity to boost prediction performance [10].

In this research, two cities in Indonesia are chosen as samples to be used as learning material to obtain the best classification for predicting rainfall. The selected cities are Bogor City and Palu City, which are two cities that have completely opposite rainfall, whereas Bogor City is a city with high rainfall, and Palu City has low rainfall. Thus, the data shows an imbalance that is challenging to manage in machine learning. An imbalanced data causes some issues, such as model b due to the uneven distribution of classes in your dataset; models can become biased towards the majority class, leading to poor predictive performance on the minority class.

To overcome these issues, random over-sampling (ROS) and synthetic minority oversampling techniques (SMOTE) were performed. The most common form of oversampling method is ROS, where minority class samples are randomly selected and duplicated without any special selection process [11]. With SMOTE, minority classes are repeatedly sampled

by taking each minority class sample and adding a number of replicates along the line segments connected to any/all nearest neighbors of the k minority classes [12]. Unlike ROS, SMOTE algorithm generates artificial samples based on the feature space rather than the data space, as well as the similarity between the various minority samples [13]. This study will highlight the differences in several classification methods: Section II presents the methodology used in predicting rainfall, Section III presents the use of the machine learning classification method, and Section IV presents the results and discussion of the classification method.

II. Research Methods

Machine learning allows itself to continue learning based on existing data. This allows the accuracy of the prediction results to continue to increase as more data is studied. Machine learning generally studies the research and construction of algorithms that can learn and make predictions about data. The methodology used in this research is divided into four stages. The first stage is the dataset section, which in this section analyses some details of the existing data. Next is the pre-processing stage. This section prepares the data for further processing at the cleaning and normalization stages to take care of missing data. The last step is to compare each classification method with some of the metrics displayed.

A. Dataset

In this study, the data collected is from the cities with the highest rainfall and the cities with the lowest rainfall in Indonesia, Bogor and Palu, respectively. Data collection is done by exporting the data from the official Indonesian Meteorology Climatology and Geophysics Council site: https://dataonline.bmkg.go.id/.

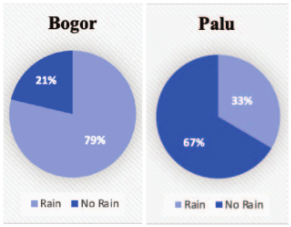

Fig. 1. Rain and no rain data in Bogor and Palu

TABLE I. Measurement Feature Details in Bogor and Palu

Attribute Name	Attribute Type	Attribute Meter
Min Temp	Continuous	°C
Max Temp	Continuous	°C
Avg Humidity	Continuous	Percentage of relative humidity (%)
Rainfall	Continuous	mm
Sunshine Duration	Continuous	Hour
Max Wind Speed	Continuous	m/s
Wind Direction at Max Speed	Continuous	°
Avg Wind Speed	Continuous	m/s
Class	Nominal	Rainfall – yes, Rain off – no

Each city has 1,451 data and is classified into 'rain' and 'no rain.' In the city of Bogor, there are 1,142 data for 'rain,' and for the city of Palu, there are only 486 data for 'rain.' A comparison of rain and no rain data in Bogor and Palu can be seen in Fig. 1. The comparison of 'rain' and 'no rain' data in these two cities is significantly imbalanced, impacting the findings for each classifier measure separately. The measurement feature details in Bogor and Palu dataset is shown in Table 1.

B. Pre-processing

This section will prepare and process data for the next stage. Noise data and incomplete data will adversely affect the results to be obtained. The application of standard pre-processing methods, which can be in the form of lowercase, stop word removal, or stemming, can cause the loss of specific features of the message the author wants to convey and also the message for the author himself [14]. At this stage, we use two pre-processing stages, namely cleaning and normalization.

C. Cleaning

In this study, there is some data on the rainfall variable with the value of 8888, which means the data is not measured, so cleaning is done by changing the data to 0, which means it is included in the 'no rain' group. The same goes for the data with a null value. In addition, the numerical data that uses a comma (,) for the decimal place is converted into a period (.) to change the data type to a number.

D. Normalization

Data normalization is an important pre-processing step involving transforming features so that a new range is obtained from an existing one. Its main purpose is to ensure that a record in the dataset remains consistent. The importance of data normalization steps in improving data quality and performance of machine learning algorithms has been demonstrated in many studies. Normalization is very useful in statistical learning methods because all the features in the data have the same contribution to the learning process [15]. At this normalization stage, Nan data or missing data will be processed by the sci-kit-learn class, namely SimpleImputer, with a strategy in the form of a mean where missing values are replaced using the average value of all data in the column. StandardScaler() is also used in this study for the next stage in normalization, where the standard scaler can change the dataset so that the result of the average value of the distribution is zero and the standard deviation is one [16].

E. Evaluation Metrics

Cross-validation is a technique used to assess the performance of a machine learning model and to ensure that the model generalizes well to unseen data. One common method of cross-validation is k-fold cross-validation. When using 4-fold cross-validation, the data is split into four parts (or "folds"). In each iteration, one of the 4 subsets is used as the test set, and the remaining 3 subsets are combined to form the training set. After each iteration, the model will be evaluated on the test set. Several metrics are used to find the best classification method for predicting rainfall. This study's three metrics include precision, recall, and f1-score. A model's precision is its capacity to recognize important items. [17]. The number of correct predictions divided by the total

number of tokens is denoted as recall [18]. F1 score can be described as the harmonic mean between precision and recall, which can indicate the overall quality of the approach [19]. Precision measures the extent to which the positive predictions from a classification model are correct. Precision calculates the percentage of true positive predictions compared to the total positive predictions made by the model.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

Recall measures how well a classification model can find all the existing positive examples. It calculates the percentage of true positive predictions compared to the total number of positive examples that exist,

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

The F1 score is a combined measure of precision and recall. It provides a harmonic average between these two metrics. The F1 score is useful when we want to find a balance between precision and recall.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

F. Machine learning model

Natural language processing (NLP), autonomous cars, robotics, image processing control, and computer vision are all examples of machine learning applications [20]. The supervised learning method will be used to find the best classification method in this study. Among them are Naïve Bayes, Logistic Regression, SVM, Kernel SVM. We use several classification models with various parameters to find out which model and parameters will perform best. The experiments were carried out to obtain the optimal value of each classifier. The performance of the four classification models will then be evaluated and compared.

1) Naïve Bayes

The Naïve Bayes method is a subset of machine learning that computes probabilities for categorizing data based on known class assignments of sample data. It determines the likelihood of an item being in a specific group based on certain attributes [21]. The Naïve Bayes classifier is a type of probabilistic classifier that predicts a probability distribution across various classes for input rather than just identifying the most probable class for the input [22]. Moreover, it assumes that the features are independent of each other.

2) Logistic Regression

Logistic Regression is a classification technique used primarily for binary classification. It employs probabilities to statistically assess the impact of each independent variable [23]. Metrics such as precision, recall, and the f1-score are utilized to determine the optimal classification approach. By invoking pre-existing functions, we can visually represent the outcomes of these metrics when using logistic regression for classification [24].

3) Support Vector Machine

In the Support Vector Machine metrics classification model, the Python programming language is used. Support Vector Machines are one of the most widely used supervised learning calculations, which are used for classification as well as regression problems [25]. The main goal of SVM algorithms is to find a hyperplane in N-dimensional space that explicitly rejects data points while minimizing the distance between two data sets [26].

4) Kernel Support Vector Machine

Kernel Support Vector Machines (Kernel SVMs) are an extension of the basic Support Vector Machine (SVM) that uses the kernel trick to transform input data into a higher-dimensional space, making it possible to handle non-linear data. Essentially, they allow SVMs to classify non-linearly separable data by implicitly mapping input features into a higher-dimensional space without having to compute the coordinates in that space. Support Vector Machine (SVM) makes it difficult to classify non-linear data. The solution that can be done is to use the kernel. Several kernels that can be used on support vector machines are linear kernels, polynomial kernels, and RBF kernels. The most popular kernel among all the kernels in support vector machines is the RBF kernel, and in previous studies, the SVM-RBF kernel has increased accuracy when compared to other algorithms [27].

5) SMOTE for Solving Imbalanced Data Problem

Rainfall data of the two cities has an imbalanced class between "rain" and "no rain" where the Bogor city has more "rain" class than the "no rain" class while the Palu city has less "rain" class than the "no rain" class. The existence of a data imbalance can affect the classification results where the classifier results typically have good accuracy on the majority class but less accuracy on the minority class because the higher majority class influences conventional training criteria in the fitness function [28]. We use three approaches to solve this problem. Our first approach is ROS, which duplicates the existing data until the amount of data becomes balanced. The second approach is to use the SMOTE technique to oversample the existing data. By combining clustered minority samples, the SMOTE method increases the sample size [29]. The third approach is to combine data from Bogor City and Palu City because the percentage of the majority data in Bogor City, which is "rain" data, is almost the same as the percentage of the majority data in Palu City which is "no rain" data so that when combined the data will be balanced. For the SMOTE technique, we can do random oversampling where the minority data will be multiplied to be as much as the majority data.

III. Result

The result of four machine learning algorithms is as follows:

A. Result of The Naïve Bayes Algorithm

The results of the naïve Bayes algorithm show that the combined data from Bogor rainfall and Palu rainfall have the highest value for all metrics used. The f1-score results for the city of Bogor using ROS show a decline, but for the city of

Palu, there is an increase of 2%. In the city of Bogor, the SMOTE dataset showed a 2% increase in the f1-score compared to the previous dataset, and in the city of Palu, the f1-score using SMOTE increased by 2%. For the ROS in the recall metric, it shows that both Bogor and Palu experienced an increase of 3%. The Recall metric in the cities of Bogor and Palu using SMOTE increased by 3% and 2.5%, respectively. For the precision metric, Bogor, which didn't use ROS and SMOTE, scored higher, while in Palu, using ROS and SMOTE, there was an increase of 0.6% and 1.3%, respectively. To see a comparison of all metrics, refer to Fig 2.

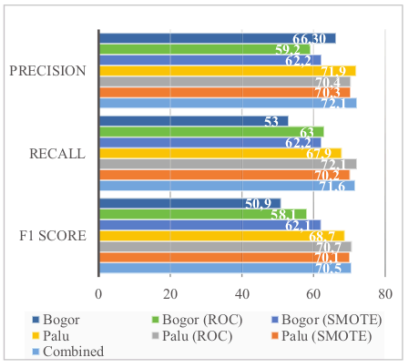

Fig. 2. Result of the Naïve Bayes Parameter Test

B. Result of The Logistic Regression Algorithm

The logistic regression algorithm indicates that the precision metric has the highest value in the combined data. The f1-score in the city of Bogor shows an increase in the dataset using ROS and SMOTE by 7.2% and 11.2%, respectively, while in the city of Palu, it increased by 2% and 1.4%. For the Recall metric, both Bogor and Palu also saw improvements in datasets using ROS and SMOTE, with increases of 10% and 9.2% for Bogor, and 4.2% and 4.2% for Palu. The Precision metric in Bogor using ROS and SMOTE showed lower values compared to datasets not using ROS and SMOTE, and the same pattern was observed for Palu. To see a comparison of all metrics, refer to Fig. 3.

C. Result of The Support Vector Machine Algorithm

The combined data from the cities of Bogor and Palu have the highest value for the F1-score metric. Rainfall data in the city of Bogor for the F1-Score metric indicates that datasets using ROS and SMOTE increased by 15.6% and 17.6%, respectively, while for the city of Palu, it increased by 2.6% and 2.4%. The Recall metric showed an increase in datasets using ROS and SMOTE for the city of Bogor by 13.6% and 11.7%, and for the city of Palu, it increased by 5% and 3.4%. The combined recall metric value for the cities of Bogor and Palu is lower by 0.1% compared to the ROS data value. For the Precision metric in Bogor using ROS and SMOTE, there was an increase of 24.3% and 22.5%, respectively, while in Palu, datasets not using SMOTE showed higher values, which are also higher than the combined data values. To see a comparison of all metrics, refer to Fig. 4.

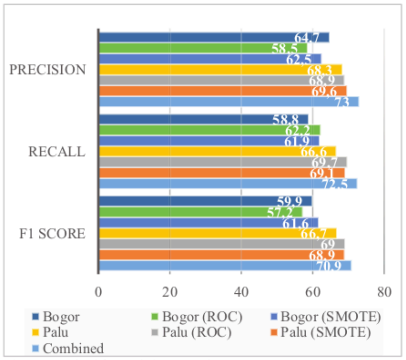

Fig. 3. Result of the Logistic Regression Parameter Test

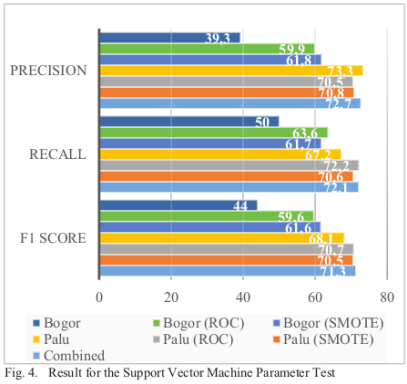

Fig. 4. Result for the Support Vector Machine Parameter Test

D. Result of The Kernel Support Vector Machine Algorithm

For this algorithm, the Precision metric shows that the highest value comes from the combined data of the cities of Bogor and Palu. For the F1-score in Bogor, data using ROS and SMOTE increased by 10.8% and 18.5%, respectively, while in Palu, it increased by 1.7% and 3.9%, with the data using ROS having the highest value compared to the others. The Recall metric in Bogor using ROS and SMOTE also showed increases of 10.1% and 15.3%, respectively, while in Palu, it increased by 4.1% and 4.7%, with the highest value being from the dataset using SMOTE in Palu. The Precision metric in Bogor using SMOTE didn't show any increase, while in Palu using SMOTE, it went up by 0.3%. To see a comparison of all metrics, refer to Fig. 5.

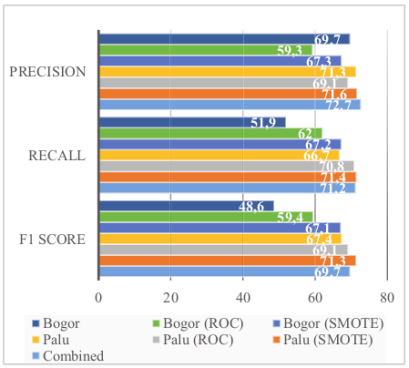

Fig. 5. Result of the Kernel Support Vector Machine Parameter Test

IV. Discussion

Rainfall prediction is essential across various sectors of human life. High rainfall in an area can lead to natural disasters, such as floods and even landslides. An appropriate method is needed to predict rainfall in a region. One approach is to use supervised learning techniques to classify the dataset. Finding a suitable classification method is essential for rainfall prediction. Therefore, comparing several classification methods is necessary. Four classification methods have been compared, revealing which is most suitable for predicting rainfall in that area. These four methods are Naïve Bayes, Logistic Regression, SVM, and kernel SVM. The data used are daily climate reports from two regions in Indonesia, Bogor, and Palu, where each area has 1,451 data points with varying rain intensities. Classification is done by dividing the data into two groups: 'rainy' and 'not rainy'. In Bogor, the 'rainy' data consists of 1,142 entries, while in Palu, there are 486 'rainy' data points.

We have compared classification methods to aid in selecting the best classification method for rainfall prediction. Three metrics are used: precision, recall, and f1-score. To address imbalanced data, the authors combine data from Bogor with data from Palu, where data from both cities shows imbalances that, when combined, balance the dataset. Additionally, ROS and SMOTE are used to increase the minority data until it is balanced. Among the three approaches, the combined data often showed higher values than the ROS and SMOTE methods.

The use of the Naive Bayes algorithm can assume the independence between features, which in some cases can help handle imbalanced data because it is not too disturbed by the proportion difference in each feature. However, in very imbalanced data, Naive Bayes tends to predict the majority class more strongly. This can result in poor performance in the minority class, especially in the recall and F1-score metrics.

In the logistic regression algorithm, incomplete data or missing values can be addressed because its method is based on the logistic function that produces probabilities. However, in imbalanced data, logistic regression tends to predict the majority class better than the minority class. This can lead to high precision but low recall for the minority class.

The support vector machine algorithm tends to provide predictions with good precision for the minority class, where this algorithm can reduce the number of false positives in the minority class. However, in very imbalanced data, that is, when the number of examples in the minority class is very small, SVM may have difficulty in building an effective model for the minority class.

The kernel support vector machine using the RBF kernel allows the algorithm to handle data that cannot be separated linearly, which is beneficial when there is data with more complex patterns. Kernel support vector machines tend to predict the majority class better than the minority class, which can lead to high precision but low recall for the minority class. However, in some metrics of certain algorithms, ROS and SMOTE methods also outperformed the combined data. As the combined data excels, it can be concluded that to achieve optimal prediction results, we can combine two datasets to balance the data first. If the available dataset is very limited, the use of ROS and SMOTE methods can still be a viable option.

V. Conclusion

To find a suitable classification model, we compared four methods: Naïve Bayes, Logistic Regression, Support Vector Machine, and Kernel Support Vector Machine using precision, recall, and f1-score for assessment. The study's findings indicate that the Kernel Support Vector Machine surpasses the other classification techniques in terms of resilience to class imbalances. Our results indicate that KSVM outperformed the other methods in terms of robustness to class imbalance with F1-score of 48.6%, 67.1%, 67.4%, 71.3% for the dataset of Bogor, Bogor with SMOTE, Palu, and Palu with SMOTE respectively. For the Bogor and Palu datasets, SMOTE improved the F1-score by 38% and 6%, respectively. In all the algorithms used, the combined data from the cities of Bogor and Palu performed better than the data using the ROS and SMOTE methods in each respective city. However, the results from the ROS and SMOTE methods in Palu have only slight differences from the results of the combined data, unlike the data using ROS and SMOTE in Bogor. The difference in results between the combined data and the ROS and SMOTE methods in Palu not only shows that the highest values are in the combined data, but in some metrics across several algorithms, it is evident that the ROS and SMOTE methods in Palu have slightly higher values.

Acknowledgment

This research was funded by Universitas Islam Riau.

References

[1] M. Personal and R. Archive, "Munich Personal RePEc Archive In-utero Exposure to Rainfall Variability and Early Childhood Health," no. 110999, 2021.

[2] J. Diez-Sierra and M. del Jesus, "Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods," *J. Hydrol.*, vol. 586, no. January, p. 124789, 2020, doi: 10.1016/j.jhydrol.2020.124789.

[3] M. F. Seleiman *et al.*, "Drought stress impacts on plants and different approaches to alleviate its adverse effects," *Plants*, vol. 10, no. 2, pp. 1–25, 2021, doi: 10.3390/plants10020259.

[4] H. Tsuguti, N. Seino, H. Kawase, Y. Imada, T. Nakaegawa, and I. Takayabu, "Meteorological overview and mesoscale characteristics of the Heavy Rain Event of July 2018 in Japan," *Landslides*, vol. 16, no. 2, pp. 363–371, 2019, doi: 10.1007/s10346-018-1098-6.

[5] Z. M. Yaseen *et al.*, "Rainfall Pattern Forecasting Using Novel Hybrid Intelligent Model Based ANFIS-FFA," *Water Resour. Manag.*, vol. 32, no. 1, pp. 105–122, 2018, doi: 10.1007/s11269-017-1797-0.

[6] E. E. Chikalamo, O. C. Mavrouli, J. Ettema, C. J. van Westen, A. S. Muntohar, and A. Mustofa, "Satellite-derived rainfall thresholds for landslide early warning in Bogowonto Catchment, Central Java, Indonesia," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 89, no. March, p. 102093, 2020, doi: 10.1016/j.jag.2020.102093.

[7] N. Khan, D. A. Sachindra, S. Shahid, K. Ahmed, M. S. Shiru, and N. Nawaz, "Prediction of droughts over Pakistan using machine learning algorithms," *Adv. Water Resour.*, vol. 139, no. January, 2020, doi: 10.1016/j.advwatres.2020.103562.

[8] M. Binkhonain and L. Zhao, "A review of machine learning algorithms for identification and classification of non-functional requirements," *Expert Syst. with Appl. X*, vol. 1, 2019, doi: 10.1016/j.eswax.2019.100001.

[9] A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 273–292, 2019, doi: 10.1007/s10462-018-09677-1.

[10] N. S. Sani, A. H. A. Rahman, A. Adam, I. Shlash, and M. Aliff, "Ensemble Learning for Rainfall Prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 11, pp. 153–162, 2020, doi: 10.14569/IJACSA.2020.0111120.

[11] T. D. Piyadasa and K. Gunawardana, "A Review on Oversampling Techniques for Solving the Data Imbalance Problem in Classification," *Int. J. Adv. ICT Emerg. Reg.*, vol. 16, no. 1, pp. 22–31, 2023, doi: 10.4038/icter.v16i1.7260.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. February 2017, pp. 321–357, 2002, doi: 10.1613/jair.953.

[13] P. Date, "UCLA UCLA Electronic Theses and Dissertations Classification of Imbalanced Data Using Synthetic Over-Sampling Techniques," 2015.

[14] V. Maslej-Krešňáková, M. Sarnovský, P. Butka, and K. Machová, "Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification," *Appl. Sci.*, vol. 10, no. 23, pp. 1–26, 2020, doi: 10.3390/app10238631.

[15] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput.*, vol. 97, no. xxxx, p. 105524, 2020, doi: 10.1016/j.asoc.2019.105524.

[16] T. D. K. Thara, P. S. Prema, and F. Xiong, "Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques," *Pattern Recognit. Lett.*, vol. 128, pp. 544–550, 2019, doi: 10.1016/j.patrec.2019.10.029.

[17] R. Padilla, S. L. Netto, and E. A. B. Da Silva, "A Survey on Performance Metrics for Object-Detection Algorithms," *Int. Conf. Syst. Signals, Image Process.*, vol. 2020-July, no. July, pp. 237–242, 2020, doi: 10.1109/IWSSIP48289.2020.9145130.

[18] A. Hard *et al.*, "Federated Learning for Mobile Keyboard Prediction," 2018, [Online]. Available: http://arxiv.org/abs/1811.03604

[19] E. F. Ohata *et al.*, "Automatic detection of COVID-19 infection using chest X-ray images through transfer learning," *IEEE/CAA J. Autom. Sin.*, vol. 8, no. 1, pp. 239–248, 2021, doi: 10.1109/JAS.2020.1003393.

[20] S. Hosseini and M. Azizi, "The hybrid technique for DDoS detection with supervised learning algorithms," *Comput. Networks*, vol. 158, pp. 35–45, 2019, doi: 10.1016/j.comnet.2019.04.027.

[21] A. B. Yilmaz, Y. S. Taspinar, and M. Koklu, "Classification of Malicious Android Applications Using Naive Bayes and Support Vector Machine Algorithms," *Int. J. Intell. Syst. Appl. Eng.*, vol. 10, no. 2, pp. 269–274, 2022.

[22] V. R. Balaji, S. T. Suganthi, R. Rajadevi, V. K. Kumar, B. S. Balaji, and S. Pandiyan, "Skin disease detection and segmentation using dynamic graph cut algorithm and classification through Naive Bayes classifier," *Measurement*, vol. 163, p. 107922, 2020.

[23] S.-H. Moon, Y.-H. Kim, Y. H. Lee, and B.-R. Moon, "Application of machine learning to an early warning system for very short-term heavy rainfall," *J. Hydrol.*, vol. 568, pp. 1042–1054, 2019.

[24] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A comparative analysis of logistic regression, random forest and KNN models for the text classification," *Augment. Hum. Res.*, vol. 5, pp. 1–16, 2020.

[25] A. Kurani, P. Doshi, A. Vakharia, and M. Shah, "A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting," *Ann. Data Sci.*, vol. 10, no. 1, pp. 183–208, 2023.

[26] B. T. Pham *et al.*, "Development of advanced artificial intelligence models for daily rainfall prediction," *Atmos. Res.*, vol. 237, p. 104845, 2020.

[27] A. P. Gopi, R. N. S. Jyothi, V. L. Narayana, and K. S. Sandeep, "Classification of tweets data based on polarity using improved RBF kernel of SVM," *Int. J. Inf. Technol.*, vol. 15, no. 2, pp. 965–980, 2023.

[28] D. Devarriya, C. Gulati, V. Mansharamani, A. Sakalle, and A. Bhardwaj, "Unbalanced breast cancer data classification using novel fitness functions in genetic programming," *Expert Syst. Appl.*, vol. 140, p. 112866, 2020.

[29] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of classification methods on unbalanced data sets," *IEEE Access*, vol. 9, pp. 64606–64628, 2021

Data Augmentation for Rainfall Classification in Bogor and Palu using SMOTE

ORIGINALITY REPORT

18%	12%	14%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Ahmad Fathan Afdhali, Hilal H. Nuha, Maman Abdulrohman. "High Wave Detection Smart Buoy using Internet of Things", 2023 International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics (ICoABCD), 2023 Publication	2%
2	Submitted to Universitas Diponegoro Student Paper	2%
3	www.mdpi.com Internet Source	1%
4	Moh Abdul Latief, Alhadi Bustamam, Titin Siswantining. "Performance Evaluation XGBoost in Handling Missing Value on Classification of Hepatocellular Carcinoma Gene Expression Data", 2020 4th International Conference on Informatics and Computational Sciences (ICICoS), 2020 Publication	1%
5	Submitted to University of Teesside Student Paper	1%
6	"Proceedings of 3rd International Conference on Smart Computing and Cyber Security", Springer Science and Business Media LLC, 2024 Publication	1%
7	Anju CP, Sunitha Subhramanian, Natalia Sizochenko, Anu R Melge, Jerzy Leszczynski,	<1%

	C.Gopi Mohan. "Multiple e-Pharmacophore modeling to identify a single molecule that could target both streptomycin and paromomycin binding sites for 30S ribosomal subunit inhibition", Journal of Biomolecular Structure and Dynamics, 2018 Publication	
8	Divyaansh Devarriya, Cairo Gulati, Vidhi Mansharamani, Aditi Sakalle, Arpit Bhardwaj. "Unbalanced Breast Cancer Data Classification Using Novel Fitness Functions in Genetic Programming", Expert Systems with Applications, 2019 Publication	<1 %
9	Thompson Stephan. "Artificial Intelligence in Medicine", CRC Press, 2024 Publication	<1 %
10	dokumen.pub Internet Source	<1 %
11	coek.info Internet Source	<1 %
12	www.iccs-meeting.org Internet Source	<1 %
13	Viera Maslej-Krešňáková, Martin Sarnovský, Peter Butka, Kristína Machová. "Comparison of Deep Learning Models and Various Text Pre-Processing Techniques for the Toxic Comments Classification", Applied Sciences, 2020 Publication	<1 %
14	publikasi.dinus.ac.id Internet Source	<1 %
15	www.icter.org Internet Source	<1 %
16	www.geeksforgeeks.org	

	Internet Source	<1%
17	pythonmania.org Internet Source	<1%
18	Submitted to Arts, Sciences & Technology University In Lebanon Student Paper	<1%
19	Suhel Ahmad Khan, Rajeev Kumar, Omprakash Kaiwartya, Mohammad Faisal, Raees Ahmad Khan. "Computational Intelligent Security in Wireless Communications", CRC Press, 2023 Publication	<1%
20	Submitted to University of Bristol Student Paper	<1%
21	www.researchgate.net Internet Source	<1%
22	journal.uad.ac.id Internet Source	<1%
23	library.cbit.ac.in Internet Source	<1%
24	web.realinfo.tv Internet Source	<1%
25	www.vtg.mod.gov.rs Internet Source	<1%
26	Seung-Hyun Moon, Yong-Hyuk Kim. "An improved forecast of precipitation type using correlation-based feature selection and multinomial logistic regression", Atmospheric Research, 2020 Publication	<1%
27	Submitted to Universitas Andalas Student Paper	<1%

28	Elene Firmeza Ohata, Gabriel Maia Bezerra, Joao Victor Souza das Chagas, Aloisio Vieira Lira Neto et al. "Automatic detection of COVID-19 infection using chest X-ray images through transfer learning", IEEE/CAA Journal of Automatica Sinica, 2020 Publication	<1 %
29	Submitted to ESoft Metro Campus, Sri Lanka Student Paper	<1 %
30	insightsociety.org Internet Source	<1 %
31	"Intelligent Distributed Computing XIII", Springer Science and Business Media LLC, 2020 Publication	<1 %
32	Arup Kumar Das, Debangshu Dey, Biswendu Chatterjee, Sovan Dalai. "A Transfer Learning Approach to Sense the Degree of Surface Pollution for Metal Oxide Surge Arrester Employing Infrared Thermal Imaging", IEEE Sensors Journal, 2021 Publication	<1 %
33	Li Tang, Shuang Xiao, Xiaodong Li. "Correlation analysis of structural characteristics of table tennis players' hitting movements and hitting effects based on data analysis", Entertainment Computing, 2023 Publication	<1 %
34	Yang, Li. "Optimized and Automated Machine Learning Techniques Towards iot Data Analytics and Cybersecurity", The University of Western Ontario (Canada), 2022 Publication	<1 %
35	aran.library.nuigalway.ie Internet Source	<1 %

36	medikom.iocspublisher.org Internet Source	<1 %
37	ojs.trigunadharma.ac.id Internet Source	<1 %
38	www.jait.us Internet Source	<1 %
39	www.w3computing.com Internet Source	<1 %
40	Amit Kumar Tyagi, Shrikant Tiwari, S. V. Nagaraj. "Quantum Computing - The Future of Information Processing", CRC Press, 2025 Publication	<1 %
41	ar5iv.labs.arxiv.org Internet Source	<1 %
42	ejurnal.stmik-budidarma.ac.id Internet Source	<1 %
43	fastercapital.com Internet Source	<1 %
44	journals.ums.ac.id Internet Source	<1 %
45	unsworks.unsw.edu.au Internet Source	<1 %
46	www.philstat.org Internet Source	<1 %
47	"Proceedings of International Conference on Smart Computing and Cyber Security", Springer Science and Business Media LLC, 2021 Publication	<1 %
48	Bhowan, U., M. Johnston, and Mengjie Zhang. "Developing New Fitness Functions in Genetic Programming for Classification With Unbalanced Data", IEEE Transactions on	<1 %

Systems Man and Cybernetics Part B (Cybernetics), 2012.
Publication

49	"Proceedings of the Second International Conference on the Future of ASEAN (ICoFA) 2017 – Volume 2", Springer Science and Business Media LLC, 2018 Publication	<1%
50	Jake Y. Chen, Stefano Lonardi. "Biological Data Mining", Chapman and Hall/CRC, 2019 Publication	<1%
51	Fatiha Belmahdi, Mourad Lazri, Fethi Ouallouche, Karim Labadi, Rafik Absi, Soltane Ameur. "Application of Dempster-Shafer theory for optimization of precipitation classification and estimation results from remote sensing data using machine learning", Remote Sensing Applications: Society and Environment, 2023 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		