A Comparative Analysis Using Machine Learning Approach for Thunderstorm Prediction in Southern Region of Peninsular Malaysia

Shirley Anak Rufus
Faculty of Electrical Engineering
University Teknologi Malaysia (UTM)
Johor Bahru, Malaysia
Department of Electrical and Electronic Engineering
Faculty of Engineering
University Malaysia Sarawak (UNIMAS)
Kota Samarahan, Sarawak, Malaysia
rshirley@unimas.my

N. Azlinda Ahmad
Faculty of Electrical Engineering
University Teknologi Malaysia (UTM)
Johor Bahru, Malaysia
noorazlinda@utm.my

Zulkurnain Abdul-Malek
Faculty of Electrical Engineering
University Teknologi Malaysia (UTM)
Johor Bahru, Malaysia
zulkurnain@utm.my

Noradlina Abdullah
Lightning & Earthing UnitTNB
Research Sdn Bhd Selangor, Malaysia
noradlina.abdullah@tnb.com.my

2023 International Symposium on Lightning Protection (XVII SIPDA) | 979-8-3503-5779-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/SIPDA59763.2023.10349193

***Abstract*— Thunderstorms are one of the most destructive natural phenomena on the planet, as they are predominantly associated with lightning and heavy rainfall that result in human deaths, urban flooding, and agricultural damage. Thus, accurate thunderstorm prediction is essential for planning and managing agriculture, flood control, and air traffic control. This study utilized historical lightning and meteorological data from 2011 to 2018 in the southern regions of Peninsular Malaysia to predict thunderstorm occurrences. The lightning dataset is classified into three class ranges, where the high range of lightning rarely occurs in this region compared to the low and medium ranges of lightning because of the non-linear and complex characteristics of the thunderstorm and lightning itself, leading to an imbalanced dataset. The k-fold and stratified cross-validation (CV) methods and a resampling technique called SMOTE are introduced to overcome the imbalance in the training dataset. Then the dataset is trained and tested using five Machine Learning (ML) algorithms, including Decision Trees (DT), Adaptive Boosting (AdaBoost), Random Forest (RF), Extra Trees (ET), and Gradient Boosting (GB). The results have shown that the GB ML model using stratified k-fold CV and SMOTE is the best algorithm for thunderstorm prediction for this region, with accuracy ranging from 74% to 95%, recall ranging from 72% to 93%, precision ranging from 76% to 97%, and F1-Score ranging from 74% to 95%. Future thunderstorm predictions based on lightning patterns and meteorological datasets are expected to establish an early strategy to address the presence of thunderstorms by notifying the relevant authorities, to prevent any damage that may be caused by the thunderstorms.**

Keywords*— *Thunderstorm, Lightning, Machine Learning, Cross-Validation, SMOTE, Thunderstorm Prediction Model, Meteorological, Evaluation Metrics

I. Introduction

Thunderstorms are fascinating and unique natural phenomena that frequently occur around the world at any time. A thunderstorm is caused by a cumulonimbus, also known as a thundercloud, that forms from the combination of moisture, unstable air, and a lifting mechanism. Thunderstorms are characterized by the presence of strong winds, heavy precipitation, lightning, tornadoes, or hail. Thunderstorms pose significant potential to endanger humans and cause property damage, which causes significant financial losses to the country both directly and indirectly. Globally, an estimated 24 thousand fatalities and 240 thousand injuries annually are attributable to lightning [1]. A significant number of lightning strikes are reported to impact power utilities, communication networks, and infrastructure. Thunderstorms and lightning strikes in Malaysia are spectacular and very common phenomena due to their tropical climate and proximity to the Equator [2]. Malaysia has a high number of lightning days per year, with a flash density of 20 flashes/km^2/year and lightning peak currents ranging from 3 kA to 200 kA, with the average being 31 kA [3]. Malaysia has a relatively high lightning fatality rate, TD = 167 with a total of 132 deaths over a decade from 2008 until August 2019 [4]. Each year, an estimated RM 250 million in infrastructure damages and business disruptions are due to power outages caused by lightning [5]. Lightning occurrences in Malaysia are recorded by the Lightning Detection Networks System (LDNS) that is operated by TNB Research Sdn Bhd (TNBR), and the meteorological data is obtained from weather stations owned by the Department of Meteorological Malaysia (known as MetMalaysia), which has recorded millions of meteorological data for more than 20 years.

A combination of both datasets formed big data, which is useful in the process of developing a prediction model to analyze the patterns, train it, and test it to increase accuracy, and then use it to predict the new data. However, predicting a thunderstorm is a challenging task because of the dynamic, complex, nonlinear, and multi-dependencies characteristic of a thunderstorm. As such, the rise of Artificial Intelligence (AI) and ML techniques has had positive implications for monitoring and predicting thunderstorms. ML has been certified by many researchers as being able to provide a new way to solve the bottlenecks of thunderstorm prediction, whether using a pure data-driven model or improving numerical models by incorporating ML. The potential of ML has not been completely exploited, and a large amount of multi-source data has also not been fully utilized to improve the accuracy of thunderstorm prediction. The challenge is that the predictable period and stability of thunderstorm prediction

979-8-3503-5779-0/23/$31.00 ©2023 IEEE

can be difficult to guarantee because thunderstorms are different from normal weather phenomena and oceanographic processes, have complex dynamic mechanisms, and are easily influenced by many factors.

Recently, there has been a lot of interest in predicting thunderstorm-related research around the world. The availability of lightning and meteorological data either online or offline has attracted researchers to predict the upcoming event based on the historical pattern using data mining approaches and combining them with ANN, Machine Learning (ML), and Deep Learning (DL) techniques. The characterization of thunderstorms under different meteorological conditions can contribute to a better understanding of thunderstorm formation, lightning strikes, and related processes. However, the lack of available data and the fact that only a few measurements were done may hinder research progress, especially in equatorial regions. Yet, the study on predicting thunderstorms in tropical regions, especially in Malaysia, is very limited. Hence, this paper aims to predict the occurrence and non-occurrence of thunderstorms in the southern part of Peninsular Malaysia based on the reliable sources of eight years of historical lightning and a meteorological dataset by utilizing five different ML approaches. The performance of each model is measured using evaluation metrics such as accuracy, precision, recall, and F1-score. The paper is structured as follows: Section II presents the recent studies of the ML approach in predicting thunderstorms; Section III is the methodology, which includes the study region, data and software, the ML algorithms used, and the evaluation metrics; Section IV shows the results and discussions; and Section V provides the conclusion of the paper.

II. Literature Survey

In recent years, many data-driven models based on ML approaches have been developed in the domain of thunderstorm prediction. Several papers explore the use of machine learning techniques, such as SVMs, neural networks, and ensemble models, for thunderstorm prediction. [6] employed a machine-learning approach to identify weather regimes and determine skillful predictor combinations for enhancing short-term storm forecasting. The methodology involved utilizing a self-organizing map (SOM) algorithm to cluster meteorological data into distinct weather regimes and applying a random forest algorithm to identify the most informative predictor variables for predicting the storm. The ML technique identified meteorological regimes and selected important predictor combinations, improving short-term storm forecasting. [7] compared the performance of different predictive models for lightning occurrence using Artificial Neural Networks (ANNs) and Synthetic Minority Over-sampling Technique (SMOTE). The methodology involved training ANNs on SMOTE-balanced datasets and evaluated them using evaluation metrics. The combination of ANNs and SMOTE yielded better results compared to other models, demonstrating the effectiveness of this approach in predicting lightning occurrences. [8] employed ML techniques for nowcasting lightning occurrence from commonly available meteorological parameters by collecting meteorological data from weather stations, training a random forest classifier to predict the occurrence of lightning events based on the input parameters of four common surface weather variables (air pressure at station level (QFE), air temperature, relative humidity, and wind speed), and validating the predictive model using the data from lightning location systems. They used boosting based on DT and the TPOT Python Automated ML tool to fine-tune the hyperparameters. Between 2006 and 2017, their strategy was implemented at twelve Swiss meteorological facilities. The evaluation results revealed that the predictive model can provide lead times up to 30 minutes in advance within a 30 km radius. In Malaysia, a study done by [9] proposed a two-layer back-propagation neural network to predict the occurrence of lightning at least four hours before its arrival. They applied the algorithm to Malaysia, which has high lightning and thunderstorm occurrences throughout the year. By using real-time and historical lightning data, [10] applied the spatial clustering method to initiatively predict the prospective lighting area. Tracking thunderstorm groups can forecast lightning positions with 75% accuracy in central China. A prediction model to forecast the lightning in the Korean Peninsula by using Support Vector Machines (SVMs) and under-sampling techniques was developed [11]. The methodology involved training SVMs on the under-sampled data to classify lightning occurrences based on atmospheric variables from ECMWF data. The findings revealed that the proposed approach improves the accuracy of lightning prediction compared to traditional methods. The performance of six statistical and ML techniques for distinguishing between non-lightning and lightning days across Australia using lightning-flash counts and atmospheric variables from the ERA-Interim dataset was explained in [12]. The LR prediction model was found to have superior prediction skills, with atmospheric instability, lifting potential, and water content as the key factors in the final models. The lightning prediction models for the province of Alberta, Canada, based on CG lightning data from the Canadian LDNS, were developed and validated [13]. The models paired geographic and temporal covariates with meteorological observations and achieved high accuracy with hit rates of 85% in the Rocky Mountain and Foothills Natural Regions, with the RF approach identified as a viable modeling method.

Another study done by [14] used ML to generate binary predictions of lightning occurrence within a specific location and time interval based on weather variables from the European Centre for Medium-Range Weather Forecasts and compared the results with lightning reports from a region including the Korean Peninsula and found equitable threat scores of 0.0885 and 0.0828 for support vector machines and random forests, respectively. A short-term lightning prediction model was developed for the Amazon region using ground-based weather station data and ML techniques [15].

Weather station data was used to train a random forest classifier to predict lightning within a short timeframe. The findings showed the feasibility of using ground-based weather station data and ML techniques for short-term lightning prediction in the Amazon region, offering valuable insights for improving lightning forecasting and mitigation strategies in the area. A nowcasting model for predicting the occurrence of cloud-to-ground lightning strikes was developed using a RF classifier and the application of available meteorological parameters. The ML model is a useful tool for short-term lightning prediction and enhancing measures for lightning protection [16].

Different geographic regions were investigated, including the Korean Peninsula, Beijing area, Amazon region, Australia, China, Switzerland, and South Africa. Among the papers is the utilization of ML techniques, such as SVMs, neural

networks, and ensemble models, to improve lightning prediction accuracy and nowcasting capabilities. Additionally, meteorological data and atmospheric variables are widely used as inputs for training prediction models. Moreover, the methodologies vary, incorporating approaches such as postprocessing ECMWF data, lightning tracking algorithms, and the use of ground-based weather station data. Due to the random nature and highly imbalanced data of thunderstorms, the k-fold CV, stratified k-fold CV, and SMOTE are applied to the training dataset in this study. By using K-Fold CV, the whole dataset is partitioned into K parts of equal size called fold, where 1-fold is used as a validation set and the remaining K-1 folds are used as the training set. Stratified K-Fold is an enhanced version of K-Fold CV where each fold will have the same ratio of instances of the target variable as in the whole dataset [17]. The SMOTE oversampling method is used to generate synthetic minority class samples. SMOTE creates synthetic examples by identifying and connecting the nearest minority class instances and producing synthetic data, avoiding overfitting that occurs with random replication of existing minority class samples [18].

In summary, all the researchers contribute to the advancement of thunderstorm prediction techniques by employing ML, analyzing model performance, and developing region-specific methodologies. However, the studies vary in terms of geographical focus, methodology, and specific findings, providing a diverse perspective on thunderstorm prediction research.

III. Methodology

A. Study Region

The state of Johor in Malaysia is selected in this study because it is situated at the equatorial line with 1.561871 North latitude and 103.636179 East longitude, as depicted in Fig. 1. The location of the study region is unique because it is surrounded by three main seas (the Strait of Malacca at the west, the South China Sea at the east, and the Strait of Tebrau at the south) and characterized by two monsoon seasons (southwest monsoon from May to September and the northeast monsoon from November to March). This region also receives high annual rainfall (2000-4000 mm), with the monsoons being the controlling feature of climatic variability. The warm and humid climate (25°C until 32°C), combined with the mountainous and lowland terrain, is well-suited for the formation of thunderclouds, which leads to an increase in thunderstorms and lightning development [19]. Malaysia is in the top three among Southeast Asian countries for recorded high lightning activity, after Indonesia and Vietnam. For the selected study region, the state of Johor has averaged 95.92 lightning events per km^2 each year.

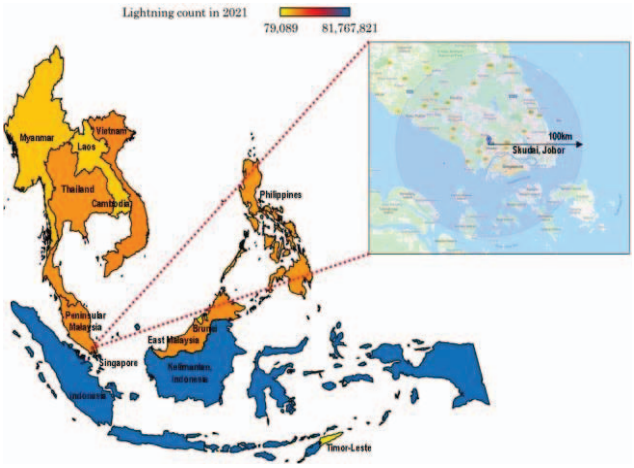

Fig. 1. Lightning counts in 2021 of Southeast Asia. (Source from Vaisala interactive lightning map) [20].

B. Data and Software

This study utilized eight years of historical lightning data from southern peninsular Malaysia obtained from the LDNS-TNBR. LDNS-TNBR located eight sensors around Peninsular Malaysia that utilized the Time of Arrival (ToA) and Magnetic Direction Finding (MDF) techniques. The lightning dataset contains the following main attributes: the geographical location (longitude and latitude) of the lightning strike; the date and time (down to millisecond resolution) of the lightning strike; and the polarity and amplitude of the lightning strike (in kA). The atmospheric conditions in the respective region were obtained from MetMalaysia, including the vertical profiles such as the average relative humidity (%), total rainfall (mm), average wind speed (km/h), and average, maximum, and minimum temperature (°C). Microsoft Excel and Jupyter Notebook were used to analyse the datasets. The Jupyter Notebook in Python is a development platform for Python-based development that provides an interpreted object-oriented programming language with many built-in functions and libraries [21], such as the sci-kit-learn library (sklearn) [22], the Pandas library [23], the Matplotlib library [24], TensorFlow [25], and Keras [26]. Tensorflow is a backend library that is used to create neural networks. Keras is an improved version of the tensor flow library, which is frequently used in deep learning due to its simplicity.

C. Workflow

Initially, it is crucial that all raw datasets undergo the critical step of data pre-processing in ML, which includes data cleaning and data replacement. Data cleaning is the process of eliminating noise and inconsistencies from incomplete data. Data replacement is the process of replacing nominal data with numerical data that accurately reflects it. Following pre-processing and feature selection, the dataset was partitioned into two datasets: The training dataset and the testing dataset. The training dataset is divided into three classes, namely Class 0 (Low range of lightning), Class 1 (Medium range of lightning), and Class 2 (High range of lightning), as depicted in Table I. However, this reveals a significant imbalance in the distribution of the three classes, as shown in Fig. 3. The SMOTE resamples the dataset to mitigate the imbalance in the training dataset. The training dataset is subjected to two CV techniques: the k-fold CV and the stratified k-fold CV. The k-fold CV procedure is randomly partitioned into five parts, with four-fold used for training and the other one-fold for

validation. The Stratified k-fold CV procedure is similar to k-fold CV, the difference is that each fold has the same ratio of instances of the target variable as in the whole dataset. CV is a commonly used method because it ensures the distribution of the samples for each class is balanced and optimal [17]. Subsequently, five ML algorithms are executed to compare the performance of the model. The optimal hyperparameters for each ML approach are determined through a CV grid search. The aforementioned steps are visually depicted in Fig. 2.

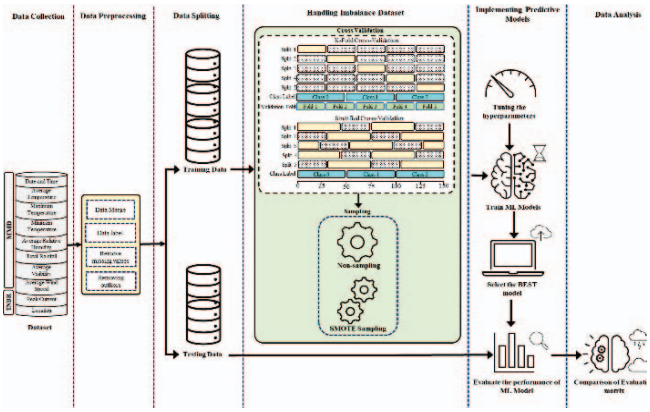

Fig. 2. Workflow for Thunderstorm Prediction Model.

D. ML Algorithms and Evaluation Metrics

The performance of various ML models within the context of thunderstorm prediction is evaluated based on an imbalanced and balanced dataset. The DT model and the GB model are both ML techniques that can be used for classification or regression. The DT model is implemented using a tree-based data structure consisting of the root (parent), internal, and leaf (child) nodes that represent the class labels. The DT model is popular due to its ease of implementation, interpretation, and robustness when dealing with non-linear data [27]. However, the DT model is highly sensitive to small changes in the data, leading to instability and a tendency to overfit. In contrast, the GB model combines multiple weak models, usually DT, to create a robust model that makes accurate predictions [28]. Like the RF model, the GB model is known for its capacity to manage high-dimensional data and avoid overfitting. The RF algorithm is an ensemble ML algorithm that can be used for regression tasks by creating multiple DT, each based on a subset of the features, and then averaging their predictions [29]. The AdaBoost model combines weak learners by iteratively reweighting training examples to prioritize misclassified ones, thereby increasing accuracy by adjusting the weights of relevant features. It is used to predict lightning and thunderstorms, but it is susceptible to chaotic data and outliers [30]. The ET model is an ML algorithm that constructs an ensemble of DT by randomly selecting subsets of features and splitting points, and then aggregating the results using a voting scheme. Like the RF model, it introduces additional randomness by splitting nodes without considering the optimal point. It is scalable and capable of handling large datasets with high-dimensional features, enabling accurate prediction of lightning and thunderstorms. Nevertheless, overfitting occurs if the quantity of trees is excessive [31].

The evaluation metrics used in this work are all based on three classes of lightning events: Low-range lightning (Class 0), Medium-range lightning (Class 1), and High-range lightning (Class 2), which can be represented in the confusion matrix. It reflects the data in connection with the true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN) [32]. The TP and TN are represented as true positive and true negative events, while the FP and FN represent false positive and negative events. The elements in the confusion matrix can be used to calculate the evaluation metrics such as Accuracy, Precision, Recall, and F-Score [33], as in Table II. Accuracy shows the overall effectiveness as the proportion of the correct predictions to the overall events. Precision is the ratio of correct positive events. Recall, also known as sensitivity, is the ratio of correctly predicted positive events. F1-Measure, or F1-Score, is defined as the harmonic mean for both precision and recall.

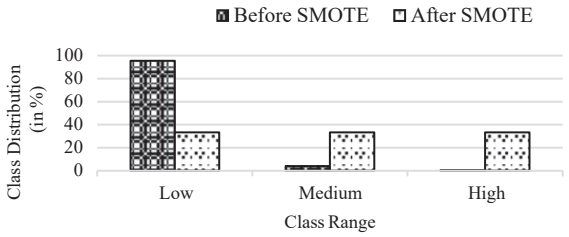

Fig. 3. Class distribution of Low, Medium, and High Range Lightning.

TABLE I. PROPOSED CATEGORY AND CLASS RANGE OF LIGHTNING FOR THUNDERSTORM PREDICTION MODEL

Class	*Category*	*Range*
Class 0	Low	[-60 kA, +60kA]
Class 1	Medium	[-120 kA, -60 kA] and [+60k kA, +120 kA]
Class 2	High	[-180 kA, -120 kA] and [+120 kA, 180 kA]

TABLE II. EVALUATION METRICS, FORMULA, AND OPTIMUM VALUE

Metric	*Formula*	*Range*	*Optimum*
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	[0,1]	1
Precision	$\frac{TP}{TP+FP}$	[0,1]	1
Recall	$\frac{TP}{TP+FN}$	[0,1]	1
F1-Measure	$2*\frac{Precision*Recall}{Precision+Recall}$	[0,1]	1

IV. RESULTS AND DISCUSSIONS

The comparison between CV and sampling methods for the thunderstorm prediction models is shown in this section. The experimental dataset includes about 1.3 million, which are recognized with the effect of the skewed class distribution, with 1244486 as the Low-class range lightning (Class 1), 53182 as the Medium-class range lightning (Class 1), and 5556 as the High-class range lightning (Class 2). SMOTE can achieve the desired ratio (1:1:1) through the training process, where k is set to 5 for both CV techniques. The capability of each ML algorithm using a combination of (1) k-fold CV with and without SMOTE and (2) Stratified k-fold CV with and without SMOTE to predict thunderstorms based on the three

class ranges of lightning is evaluated in terms of accuracy, precision, recall, and F1-score, as depicted in Table III and Table IV, respectively. The overall comparison is summarized, and the highest performance of evaluation metrics is marked in bold. All the details including the confusion matrix, can be obtained from the classification report that was generated using the sci-kit-learn library in Python. In Table III, the result of all evaluation metrics except the accuracy of the SMOTE-GB model is the highest compared to other ML models, which ranged from 93% to 97%. Followed by the SMOTE-ET model, which gave an accuracy of 0.883, a precision of 0.879, a recall of 0.839, and an F1- score of 0.894. The accuracy of the SMOTE-AdaBoost model was 0.904, the precision was 0.795, the recall was 0.794, and the F1-score was 0.882, respectively. The SMOTE-RF model achieved the highest accuracy equal to 0.964 compared to others, while the precision, recall, and F1-score are 0.842, 0.854, and 0.858, respectively. The SMOTE-DT model achieved accuracy equal to 0.931, precision was 0.752, recall was 0.736, and the F1-score was 0.806.

The combination of stratified CV and SMOTE in Table IV reveals that the SMOTE-GB model gave the best accuracy of 0.952, the precision was 0.933, the recall was 0.973, and the F1-score was 0.953. Followed by the SMOTE-RF model, which gave an accuracy of 0.883, a precision of 0.882, a recall of 0.893, and an F1-score of 0.888. The accuracy of the SMOTE-AdaBoost model was 0.813, the precision was 0.835, the recall was 0.814, and the F1-score was 0.824. The accuracy of the SMOTE-DT and SMOTE-ET models was equal to 0.771 and 0.743, respectively. These two models achieved precision, recall, and F1-score within the range of 72% to 79%.

TABLE III. COMPARATIVE OF EVALUATION METRICS USING K-FOLD CROSS VALIDATION WITHOUT AND WITH SMOTE

ML Model	SMOTE Sampling	Evaluation Metrics			
		Accuracy	*Precision*	*Recall*	*F1-Score*
DT	Without	0.925	0.919	0.930	0.728
	With	0.931	0.752	0.736	0.806
Ada Boost	Without	0.895	0.902	0.887	0.775
	With	0.903	0.795	0.794	0.882
RF	Without	0.954	0.836	0.840	0.814
	With	**0.964**	0.842	0.854	0.858
ET	Without	0.864	0.839	0.840	0.864
	With	0.883	0.879	0.839	0.894
GB	Without	0.900	0.922	0.960	0.896
	With	0.912	**0.939**	**0.971**	**0.942**

TABLE IV. COMPARATIVE OF EVALUATION METRICS USING STRATIFIED K- CROSS VALIDATION WITHOUT AND WITH SMOTE

ML Model	SMOTE Sampling	Evaluation Metrics			
		Accuracy	*Precision*	*Recall*	*F1-Score*
DT	Without	0.765	0.769	0.741	0.704
	With	0.771	0.792	0.775	0.783
Ada Boost	Without	0.794	0.811	0.786	0.786
	With	0.813	0.835	0.814	0.824
RF	Without	0.853	0.873	0.875	0.750
	With	0.883	0.882	0.893	0.888
ET	Without	0.742	0.714	0.760	0.696
	With	0.743	0.727	0.769	0.748
GB	Without	0.939	0.915	0.931	0.928
	With	**0.952**	**0.933**	**0.973**	**0.953**

V. CONCLUSIONS

This comparative study is based on the various ML methods using lightning and meteorological data in the southern part of Peninsular Malaysia. Data pre-processing was done, including data cleaning, reduction, and replacement. The most important findings revealed that baseline models tend to ignore the minority class, leading to higher error rates in imbalanced data. Therefore, to avoid the overfitting problem the training dataset was resampled using a combination of k-fold CV and stratified k-fold CV techniques and SMOTE. The dataset was divided into training (70%) and testing datasets (30%). Five different ML models, such as DT, AB, RF, ET, and GB, are used to predict thunderstorm occurrences in the southern region of Peninsular Malaysia by using lightning and meteorological data. Using the stratified k-fold CV and SMOTE technique significantly provides better handling of the class imbalance problem, with more than 70% accuracy (ranged from 74% to 95%), recall (ranged from 72% to 93%), precision (ranged from 76% to 97%), and F1-Score (ranged from 74% to 95%); moreover, they determine the best prediction across each sample. Furthermore, the SMOTE-GB model has the highest F1-score among all the others, with 95.29 %, so it is the best prediction model for thunderstorm prediction for this region regarding the evaluation metrics. The dataset used for this study represents only 8 years of information in a specific region. This fact suggests that the results of the predictive models created can only represent the thunderstorms that occurred during this period. Therefore, to generalize the results, it is necessary to train the algorithms for longer periods and take into account the seasonality of convective events in different regions of Peninsular Malaysia. Our future work includes testing other optimization methods on the prediction models and comparing the results. Besides, the prediction results based on the lightning pattern and meteorological dataset are hoped to alert the related authorities to make an effective strategy to handle the occurrence thunderstorms.

ACKNOWLEDGMENT

The authors would like to thank the Research Management Centre (RMC), Universiti Teknologi Malaysia (UTM), and the Research, Innovation and Enterprise Centre (RIEC), Universiti Malaysia Sarawak (UNIMAS), for the financial and management support under research grant number 06G50. The cooperation from lightning technical expert, Ir. Noradlina Abdullah and other technical staff of TNBR is highly appreciated. This research uses data provided by the Lightning Detection System Network (LDSN), managed by the Lightning Detection System Laboratory, TNB Research Sdn. Bhd (TNBR), and meteorological data from Met Malaysia for research and educational purposes.

REFERENCES

[1] R. L. Holle, "Annual Rates of Lightning Fatalities by Country," 20th International Lightning Detection Conference, no. January 2008, pp. 1–14, 2015, [Online]. Available: http://es.vaisala.com/VaisalaDocuments/Scientific papers/Annual_rates_of_lightning_fatalities_by_country.pdf

[2] M. Abdul Rahim, A. N. Abdul Ghani, and M. A. Che Munaaim, "Lightning Protection System in Malaysia: Materials Selection for Down Conductor," Jurnal Teknologi, vol. 78, no. 5, Apr. 2016, doi: 10.11113/jt.v78.8229.

[3] A. H. A. Bakar, D. N. A. Talib, H. Mokhlis, and H. A. Illias, "Lightning back flashover double circuit tripping pattern of 132kV lines in Malaysia," International Journal of Electrical Power & Energy

Systems, vol. 45, no. 1, pp. 235–241, Feb. 2013, doi: 10.1016/j.ijepes.2012.08.048.

[4] A. R. Syakura, M. Z. A. Ab Kadir, C. Gomes, A. B. Elistina, and M. A. Cooper, "Comparative Study on Lightning Fatality Rate in Malaysia between 2008 and 2017," in 2018 34th International Conference on Lightning Protection (ICLP), IEEE, Sep. 2018, pp. 1–6. doi: 10.1109/ICLP.2018.8503420.

[5] M. Ab-Kadir, "Lightning severity in Malaysia and some parameters of interest for engineering applications," Thermal Science, vol. 20, no. suppl. 2, pp. 437–450, Jan. 2016, doi: 10.2298/TSCI151026028A.

[6] J. K. Williams et al., "A machine learning approach to finding weather regimes and skillful predictor combinations for short-term storm forecasting," 2008. [Online]. Available: https://www.researchgate.net/publication/259869325

[7] E. Alves, A. Leal, M. Lopes, and A. Fonseca, "Performance Analysis Among Predictive Models of Lightning Occurrence Using Artificial Neural Networks and SMOTE," IEEE Latin America Transactions, vol. 19, no. 5, pp. 755–762, May 2021, doi: 10.1109/TLA.2021.9448309.

[8] A. Mostajabi, D. L. Finney, M. Rubinstein, and F. Rachidi, "Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques," NPJ Clim Atmos Sci, vol. 2, no. 1, p. 41, Nov. 2019, doi: 10.1038/s41612-019-0098-0.

[9] A. F. Ali, D. Johari, N. F. Nik Ismail, I. Musirin, and N. Hashim, "Thunderstorm forecasting by using artificial neural network," 2011 5th International Power Engineering and Optimization Conference, PEOCO 2011 - Program and Abstracts, no. June, pp. 369–374, 2011, doi: 10.1109/PEOCO.2011.5970391.

[10] G. Juntian, G. ShanQiang, and F. Wanxing, "A lightning motion prediction technology based on spatial clustering method," in 2011 7th Asia-Pacific International Conference on Lightning, IEEE, Nov. 2011, pp. 788–793. doi: 10.1109/APL.2011.6110234.

[11] S.-H. Moon and Y.-H. Kim, "Forecasting lightning around the Korean Peninsula by postprocessing ECMWF data using SVMs and undersampling," Atmos Res, vol. 243, p. 105026, Oct. 2020, doi: 10.1016/j.atmosres.2020.105026.

[12] B. C. Bates, A. J. Dowdy, and R. E. Chandler, "Lightning Prediction for Australia Using Multivariate Analyses of Large- Scale Atmospheric Variables," J Appl Meteorol Climatol, vol. 57, no. 3, pp. 525–534, Mar. 2018, doi: 10.1175/JAMC-D-17-0214.1.

[13] K. D. Blouin, M. D. Flannigan, X. Wang, and B. Kochtubajda, "Ensemble lightning prediction models for the province of Alberta, Canada," Int J Wildland Fire, vol. 25, no. 4, p. 421, 2016, doi: 10.1071/WF15111.

[14] S. H. Moon and Y. H. Kim, "Forecasting lightning around the Korean Peninsula by postprocessing ECMWF data using SVMs and undersampling," Atmos Res, vol. 243, Oct. 2020, doi: 10.1016/j.atmosres.2020.105026.

[15] A. F. R. Leal and W. L. N. Matos, "Short-term lightning prediction in the Amazon region using ground-based weather station data and machine learning techniques," in 2022 36th International Conference on Lightning Protection (ICLP), IEEE, Oct. 2022, pp. 400–405. doi: 10.1109/ICLP56858.2022.9942500.

[16] A. La Fata, F. Amato, M. Bernardi, M. D'Andrea, R. Procopio, and E. Fiori, "Horizontal grid spacing comparison among Random Forest algorithms to nowcast Cloud-to-Ground lightning occurrence," Stochastic Environmental Research and Risk Assessment, vol. 36, no. 8, pp. 2195–2206, Aug. 2022, doi: 10.1007/s00477-022-02222-1.

[17] X. Zeng and T. R. Martinez, "Distribution-balanced stratified cross-validation for accuracy estimation," Journal of Experimental & Theoretical Artificial Intelligence, vol. 12, no. 1, pp. 1–12, Jan. 2000, doi: 10.1080/095281300146272.

[18] 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, June. 2002, doi: 10.1613/jair.953.

[19] N. Yusop, M. Riduan Ahmad, T. Shea Ching, S. Ammar Shamsul Baharin, M. Riza Mohd Esa, and M. Abu Bakar Sidik, "Correlation analysis between lightning flashes and rainfall rate during a flash flood thunderstorm," Indonesian Journal of Electrical Engineering and Computer Science, vol. 28, no. 3, p. 1322, Dec. 2022, doi: 10.11591/ijeecs.v28.i3.pp1322-1329.

[20] "Welcome to Vaisala's Interactive Global Lightning Density Map!," https://interactive-lightning-map.vaisala.com/, Jan. 2022. https://interactive-lightning-map.vaisala.com/?_ga=2.251216771.448933359.1670901480-1779715020.1670901112

[21] X. Cai, H. P. Langtangen, and H. Moe, "On the Performance of the Python Programming Language for Serial and Parallel Scientific Computations," Sci Program, vol. 13, no. 1, pp. 31–56, 2005, doi: 10.1155/2005/619804.

[22] Pedregosa F. et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011, [Online]. Available: http://scikit-learn.sourceforge.net.

[23] W. McKinney, "Data Structures for Statistical Computing in Python," 2010, pp. 56–61. doi: 10.25080/Majora-92bf1922-00a.

[24] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," Comput Sci Eng, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.

[25] Martín Abadi et al., "A system for large-scale machine learning," in 12th USENIX symposium on operating systems design and implementation (OSDI 16) (, Savannah, GA, USA, 2016, pp. 265–283. [Online]. Available: https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

[26] Piotr Szyma´nski and Tomasz Kajdanowicz, "scikit-multilearn: A scikit-based Python environment for performing multi-label classification," Journal of Machine Learning Research, vol. 1, pp. 1–22, 2016.

[27] D. J. Gagne, A. McGovern, and J. Brotzge, "Classification of Convective Areas Using Decision Trees," J Atmos Ocean Technol, vol. 26, no. 7, pp. 1341–1353, Jul. 2009, doi: 10.1175/2008JTECHA1205.1.

[28] Jerome H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, [Online]. Available: http://www.jstor.org/stable/2699986

[29] K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: from early developments to recent advancements," Systems Science & Control Engineering, vol. 2, no. 1, pp. 602–609, Dec. 2014, doi: 10.1080/21642583.2014.956265.

[30] C. Schön, J. Dittrich, and R. Müller, "The Error is the Feature," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, NY, USA: ACM, Jul. 2019, pp. 2979–2988. doi: 10.1145/3292500.3330682.

[31] G. R. Herman and R. S. Schumacher, "'Dendrology' in Numerical Weather Prediction: What Random Forests and Logistic Regression Tell Us about Forecasting Extreme Precipitation," Mon Weather Rev, vol. 146, no. 6, pp. 1785–1812, Jun. 2018, doi: 10.1175/MWR-D-17-0307.1.

[32] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, p. 6, Dec. 2020, doi: 10.1186/s12864-019-6413-7.

[33] C. Goutte and E. Gaussier, "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation," 2005, pp. 345–359. doi: 10.1007/978-3-540-31865-1_25.