Thunderstorm Prediction Model Using SMOTE Sampling and Machine Learning Approach

Shirley Anak Rufus
Department of Electrical and Electronics Engineering, Universiti Malaysia Sarawak
Sarawak, Malaysia
rshirley@unimas.my
Institute of High Voltage and High Current (IVAT), Faculty of Electrical Engineering, Universiti Teknologi Malaysia
Johor, Malaysia
ashirley@graduate.utm.my

N.A. Ahmad
Institute of High Voltage and High Current (IVAT), Faculty of Electrical Engineering Universiti Teknologi Malaysia
Johor, Malaysia
noorazlinda@utm.my

Z. Abdul-Malek
Institute of High Voltage and High Current (IVAT), Faculty of Electrical Engineering Universiti Teknologi Malaysia
Johor, Malaysia
zulkurnain@utm.my

Noradlina Abdullah
Lightning and Earthing Unit TNB Research Sdn. Bhd.
Selangor, Malaysia
noradlina.abdullah@tnb.com.my

***Abstract*—Thunderstorms are one of the most destructive phenomena worldwide and are primarily associated with lightning and heavy rain that cause human fatalities, urban floods, and crop damage. Therefore, predicting thunderstorms with reasonable accuracy is one of the crucial requirements for the planning and management of many applications, including agriculture, flood control, and air traffic control. This study extensively applied the historical lightning and meteorological data from 2011 to 2018 of the southern regions of Peninsular Malaysia to predict thunderstorm occurrence. Positive CG lightning rarely occurs compared to negative CG lightning and also due to the non-linear and complex characteristics of the thunderstorm and lightning itself, leading to an imbalance in the dataset. The resampling technique called SMOTE is introduced to overcome the imbalance of the training dataset. Then the dataset is trained and tested with 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 a good prediction with accuracy (74% to 95%), recall (72% to 93%), precision (76% to 97%), and F1-Score (74% to 95%) with SMOTE. The SMOTE and GB model prediction model is the best algorithm for thunderstorm prediction for this region in terms of performance metrics. In the future, the prediction results based on the lightning pattern and weather dataset will likely alert the related authorities to make an early strategy to handle the occurrence of thunderstorms.**

Keywords—Thunderstorm, Lightning, Machine Learning, SMOTE, Thunderstorm Prediction Model, Meteorological, Performance Metrics

I. INTRODUCTION

A thunderstorm is caused by a cumulonimbus cloud that produces the electric discharge. Typically, the thunderstorm is associated with lightning and accompanied by heavy rainfall and wind. Thunderstorms adversely impact humans, industries, infrastructure, and other related sectors that directly cause human injuries, fatalities, and financial losses. An estimated 24 thousand fatalities and 240 thousand injuries annually are attributable to lightning [1]. Malaysia has an average of 204 days of thunderstorms which is equivalent to 40 strikes per kilometre per year [2]. In Malaysia, a total of 132 deaths over ten years from 2008 until August 2019 led to a very high lightning fatalities rate, TD = 167 [1]. About RM 250 million losses in infrastructure damages and business disruption due to power outages caused by lightning each year [2]. Lightning occurrences recorded by the Lightning Detection Networks System (LDNS) operated by Tenaga Nasional Berhad-Research (TNBR) and meteorological data obtained from weather stations owned by the Department of Meteorological Malaysia (known as MetMalaysia) have millions of recorded data. A combination of both datasets formed big data which is useful in the process of developing a prediction model to analyse the patterns, trained, and tested to increase the accuracy and for predicting 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 Intelligent (AI) and Machine Learning (ML) techniques have shown positive implications in monitoring and predicting thunderstorms. Typically, the dataset used by researchers consists of historical lightning and weather parameters of one region. In recent years, some ML-based approaches have been developed in the domain of thunderstorm prediction. Tervo et al. [15] monitored the trajectory of storm objects and assessed the potential impact on the power grid. Juntian et al. [16] presented a spatial clustering method to predict lightning motion that used real-time and historical lightning data to initiatively predict the prospective lighting area. The position prediction of lightning can be achieved by tracking these thunderstorm groups, and the method was applied to a certain region of central China with an average prediction accuracy of 75%. Mostajabi et al. [3] applied a set of single-site observations of meteorological parameters to develop the Boost model to perform up to 30 minutes in advance in an area of 30 km around the 12 locations in Switzerland. Alves et al. [4] used cluster information as input for a classification method aiming to generate warning alerts to three Target areas (TA). Bryson C. Bates et. al. [5], had compared the performance of six statistical and ML techniques such as the combination of principal component analysis (PCA) and LR, classification and regression trees, RF, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and logistic regression (LR) for distinguishing between non-lightning and lightning days across Australia using lightning-flash counts and atmospheric variables from the ERA-Interim dataset. 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. Blouin et al. [6] developed and validated lightning prediction models for the province of Alberta, Canada, based on CG lightning data