International Journal of Global Research Innovations & Technology (IJGRIT)
An International Multidisciplinary Quarterly Double Blind Open Access Peer Reviewed Refereed Journal
e-ISSN : 2583-8717, Impact Factor: (Cosmos:6.972& I2OR: 5.002), Volume 04, No. 02, April-June, 2026, pp 80-84

Study the Occurrence of Lightning and Thunderstorm during February to May, with All Environmental Influencing Factors by Neural Network and Machine Learning Technique

Sumana Chatterjee*

Ph.D. Scholar (Computer Science), Nirwan University, Jaipur, Rajasthan.

*Corresponding Author: sumana.spssaha.chatterjee@gmail.com

Citation: Chatterjee, S. (2026). Study the Occurrence of Lightning and Thunderstorm during February to May, with All Environmental Influencing Factors by Neural Network and Machine Learning Technique. International Journal of Global Research Innovations & Technology, 04(01), 80–84. https://doi.org/10.62823/IJGRIT/04.02.8921

ABSTRACT

This paper is based on the study of the occurrence of lightning and or thunderstorm, the weather phenomena, during February to May, analysis on google collaborator platform. The environmental parameters used were 'convective available potential energy(CAPE)', 'convective inhibition (CIN)', 'convective precipitation (CP)', '2 meter dew point temperature, (d2m), 'mean sea level pressure (MSL)', 'relative humidity at 500 HPA pressure level ', 'relative humidity at 850 HPA pressure level', 'earth's surface skin layer temperature (skin)', 'earth surface pressure (sp)', 'sea surface temperature (sst)', 'air temperature (t at 500 HPA)', 'air temperature (t at 850 HPA)', 'air temperature measured at 2 meters above the surface (t2m)', ' total cloud cover (tcc)', ' total precipitation (tp)', 'u, horizontal wind at 500 HPA pressure level', 'u, horizontal wind at 800 HPA pressure level', 'v, vertical wind at 500 HPA pressure level', 'v, vertical wind at 800 HPA pressure level', 'w, speed of air movement at 500 HPA pressure level', 'w, speed of air movement at 800 HPA pressure level', data source was 'Copernicus Climate Data Store (ERA5 dataset)' with sub region of data adequate for Alipore. Date wise 'Copernicus Hub' daily data merged with date wise daily surface data of Alipore to have insights of occurrence of lightning or thunderstorm during February to May as well as trend for near future. The multivariate analysis, with historical large data from 1969 to 2026, with associated variables, factors influencing lightning and thunderstorm, i.e. predictors from ERA5 data set, along with target data, i,e. surface data set of Alipore with actual data of lightning or thunderstorm occurrence, was subjected to parallel study with LSTM neural network and machine learning to study the occurrence of such phenomena, during the scheduled months.

***Keywords**: Lightning, Thunderstorm, Era5, Cape, CIN, MSL, SST, TCC, CP, LSTM, ML.*

Introduction

The study of analysis was related to the occurrence of the phenomena like lightning or thunderstorm during the months February to May. The analysis purpose was to study the frequency of occurrence of lightning or this type of special phenomena during the months February to May subjected to analysis, the period of analysis being 1969 to 2026 as per data availability. The pattern of analysis was multivariate data analysis. The ERA5 data set from Copernicus climate data store, from ECMWF data hub, was one source for collecting data of various variables influencing lightning or thunderstorm. All such influencing factors which act as predictors for development of lightning activity in atmosphere were downloaded from ERA5 data set of Copernicus, CDS (Climate Data Store). Based on availability and also as per requirement of variable for analysis, data were downloaded accordingly from that particular data hub in CDS. Some data was downloaded from ERA5 time series single level data hub, data from historical period, some were downloaded from ERA5 pressure level data hub to meet up the purpose of role of that particular variable according as the physical equations related with possibility of development of lightning or thunderstorm .The variables for which data were downloaded from CDS, were

CAPE,CIN,CP, d2m, MSL, R at 500 and 850 HPA pressure level, SKIN temperature, SP,SST,T,T2m,TCC,TP, U,V,W at 500 and 850 HPA pressure levels as mentioned full form with details in the abstract portion. Variables U (u component of horizontal velocity of air)and V (v component of vertical velocity of air)both at 10 meter height were downloaded also for the analysis as well as study for the paper. On the other hand, the data of meteorological surface weather parameters as obtained from the daily summary data of the station Alipore, station code 42807 were downloaded with data of real occurrence of lightning or thunderstorm which was applied as target parameter for analysis. The meteorological surface data and all the downloaded data files from CDS were merged together to form a composite data file to have insights of occurrence of lightning and thunderstorm phenomena.

Literature Review

Several papers on lightning and thunderstorm prediction had been studied to have idea on the same subject of this research study. Those are also dealt with machine leaning analysis or neural network, some of those were with using the ERA5 data set. In this study analysis was done by both of machine learning and neural network model and almost same accuracy for prediction from machine learning analysis was obtained as obtained by LSTM neural network model. For machine learning model technique, several machine learning models were used sequentially with updating of score card of accuracy after completion by each model. The machine learning models as used were Logistic regression, Decision Tree, Random forest Classifier, XGBoost, Naïve Bayes, Support vector machine, Hyper parameter tuning Random forest classifier by Grid search cv. The highest accuracy obtained by machine learning model was 84%. Analysis by LSTM neural network model was also performed to compare result and accuracy of models by machine learning model and neural network model. LSTM model was used to study the trend of occurrence of lightning, while for each model of machine learning, the model was used to predict the possibility of lightning or thunderstorm for next day by giving input into test data i.e. input for real data of variables under analysis for prediction of next day weather .All the basic variables related with possibility of development of lightning were applied for analysis and also the derived variables like Shear = $\sqrt{[(U500 - U850)^2 + (V500 - V850)^2]}$, R_mean = (R850 + R500) / 2, W_mean = (W850 + W500) / 2, these also were taken under consideration for analysis. For both of machine learning and neural network model, before model fitting, selection of data features removing multi-collinearity by VIF process was once performed also, to check if any significant change happened for the cause of selection of several multiple variables as features but no such significant change was noticed therefore for such reason. For neural network also the maximum accuracy as obtained was 84%, for execution of analysis by classification model. For both ML and neural network model the target variable column 'T' was converted into two types of (binary) output, either 1 or 0, 1 for occurrence of any significant weather phenomena associated with possibility of development of lightning or thunderstorm and 0 for no occurrence of any such significant phenomena. In this regard this is to mention that, the data set as obtained from meteorological parameters for surface observational daily summary data, consists of parameters indicating occurrence of weather phenomena. Daily observation is divided into four quarters with the indicator of weather phenomena as T. In this study these T had been converted into T1,T2,T3,T4 for four quarters which again converted into dummy variable T representing all these indicators related to weather occurrence. Conversion was made by python programming, lambda function. However, ultimately after pre processing all necessary variables obtained from CDS data hub and IMD Pune data set for Alipore, both of the machine learning model as well as neural network model was executed applying the independent variables as well as the downloaded derived variables. For IMD data set some selected columns, Maximum temperature, Minimum temperature, Average wind, Rainfall, Sunshine hours for daily day summary surface data were included in feature data set combined with CDS data where T was treated as target column for the whole data set.

Research Gap

The data analysis was done involving both of independent fundamental variables obtained from CDS data set as well as with the derived variables in relationship with the independent variables. The analysis was done in data diverse environment as well as in model diverse environment to verify the result accuracy from all corners. The data collection was from different source and from different hub of CDS. ERA5 data was downloaded from different data hub with variability in data format as well as temporal format. Then processing and finding mean value of every variable was performed, tackling with the problem of date parsing for the case of all individual variables, obtained from CDS as well as surface data for selected columns of meteorological parameters. All these jobs including merging of all individual data set with daily mean value of each type of variables were completed to execute the models. VIF method was also applied to check whether any accuracy reducing factor was there for the reason of

multi-collinearity among variables in feature set. More or less same accuracy was obtained for each case with variability of feature selection and also for type of model selection.

Research Questions / Hypothesis

For this research study the main objective was to study the trend of significant weather like lightning, thunderstorm for the months of February to May from historical past to near futures, analysis being performed with not only the meteorological parameters for the station Alipore, but also involving other environmental parameters obtained from Copernicus climate data set, having influence on significant weather phenomena like lightning, thunderstorm etc. So with these large set of data, in data diverse environment the study was done to observe the pattern of occurrence of these type of significant weather phenomena. Moreover involving variables from other external data set rather than only application of meteorological data and thus study the effect of these variables as obtained from CDS on the occurrence of weather phenomena, observed for meteorological station was also the objective of the study.

Methods

The analysis was done with environmental parameters and weather parameters. The environmental parameters which have influence on development of such type of weather phenomena like lightning, thunderstorm etc., were obtained from the online platform of Copernicus Climate Data store. These parameters were 'convective available potential energy(CAPE)', 'convective inhibition (CIN)', 'convective precipitation (CP)', '2 meter dew point temperature, (d2m), 'mean sea level pressure (MSL)', 'relative humidity at 500 HPA pressure level ', 'relative humidity at 850 HPA pressure level ', 'earth's surface skin layer temperature (skin)', 'earth surface pressure (sp)', 'sea surface temperature (sst)', 'air temperature (t at 500 HPA)', 'air temperature (t at 850 HPA)', 'air temperature measured at 2 meters above the surface (t2m)', ' total cloud cover (tcc)', ' total precipitation (tp)', 'u, horizontal wind at 500 HPA pressure level', 'u, horizontal wind at 800 HPA pressure level', 'v, vertical wind at 500 HPA pressure level', 'v, vertical wind at 800 HPA pressure level', ' w, speed of air movement at 500 HPA pressure level', 'w, speed of air movement at 800 HPA pressure level', u, the horizontal component of wind at 10 meter height from earth surface, v, the vertical component of wind, at 10 meter height from earth surface. All these parameters with hourly observational data, hours selected as 00, 06,12,17,23 hours, together as representative of single whole day, were downloaded for historical period since 1940. Then, thus obtained each data file for each hour were combined together to get daily date wise mean value of that particular variable. Python programme was used to get the file having such daily date wise mean value of environmental variable as mentioned earlier. The data handling, issue with date parsing all were tackled in efficient manner to perform these major tasks. Ultimately, in this way all the individual file consisting of date wise daily mean value of each variable was formed. The target file consisting of weather parameters with surface observational daily summary data for Alipore was pre-processed to select some relevant parameters such as daily maximum temperature, daily minimum temperature, daily rainfall, daily average wind and daily sunshine. The data set consists of columns associated with weather phenomena in four quarters by representing parameter as T. By python programming, with the help of lambda function, these T converted to some dummy variable to represent all daily weather phenomena in such a way that for T in between 9,6,0,5, T was considered as 1 and otherwise 0. 1 for occurrence of any significant weather phenomena and 0 for no occurrence. 9,6,0,5 codes represent for thunderstorm, rain, lightning and drizzle respectively. However thus converted daily weather data of weather station Alipore, of India Meteorological Department ultimately was merged with the previously stated environmental parameters to form a combined final data set with all variables consisting of daily date wise data. The Alipore data was available from the year 1969 and the environmental parameters obtained from CDS data set was from 1940. So the final data set was filtered for 1969 to 2026. The derived variable Shear, Delta_T, mean relative humidity, mean w, derived variables, equation as mentioned in literature review part, were included as feature instead of the basic independent variables. Then machine learning models were built one by one with model as logistic regression, Decision tree classifier, Random forest classifier, XGBoost, Naïve Bayes, Support vector machine, Hyper parameter tuning Random forest classifier by Grid search cv. The whole data set was split into train, test ratio 80%-20% . The target column kept as the T column and all other columns as feature columns for analysis. The analysis was performed with each model one by one with updating score card at the end of each completion. Then by giving input in test data, predicted possible weather pattern, either 1 or 0 for occurrence of any significant weather or no such significant weather for next day. The highest accuracy as obtained by the models Random forest classifier and XGBoost was 84%. The LSTM neural network model had been trained with same data set with same features with the variables as mentioned, the target variable remaining same as T, for T>=0.5,

T was considered as T=1, model fitting with train, test split ratio was 80-20 % ratio. After normalization and pre processing and all necessary feature engineering like dealing with missing values etc., neural network process was performed to get trend of T=1. Ultimately output with accuracy 84% was obtained. For both of machine learning models performed one by one with different models.

Significance of the Study

Each year, nowadays, unlike previous days, lightning occurrence and death case on account of lightning has been increased. So to study the trend of weather phenomena like lightning, thunderstorm, along with determination of possibility of such type of weather phenomena is necessary. Due to the effect of climate change nowadays, heat wave, severity of storm, excessive rainfall, flood, death case in lightning all are being noticed in increasing trend. So to monitor the pattern of trend of significant weather occurrence like lightning or thunderstorm etc. during the months February to May was felt as necessary.

Timeline

Obtaining data set from the site of Copernicus to download each basic parameters from that particular link as necessary in the analysis was a time consuming job. Downloading each variable with different data format, csv or Netcdf format, converting Netcdf format data into csv format for the purpose of compatibility with python programming, alignment of different data, tackling with date parsing issue for each variable to find daily average value or mean value group by date index, merging of all data to form a combined one for analysis, all these jobs took lots of time along with other procedure for data pre-processing. Ultimately one complete file with daily mean value of each variable group by daily date was processed and then the actual job of analysis started. Besides this pre-processing, other necessary feature engineering, preparing suitable environment towards compatibility for machine learning as well as neural network to get the result as intended for, along with observation of the output with counter verification followed by data diversity and algorithm diversity, all these took sufficient time to perform the whole analysis. The accuracy rate was obtained as same for both of machine learning model with best accuracy and neural network, which was 84%.

Conclusion and Future Work

As already mentioned, the analysis was performed involving all types of associated parameters influencing on development of lightning, thunderstorm etc, but the accuracy could not be obtained more than 84% for the reason of data gap. In future this research study may be tried in other manner also to obtain result with higher accuracy.

Some Screenshots of Analysis (Label for Each Figure Mentioned Below)

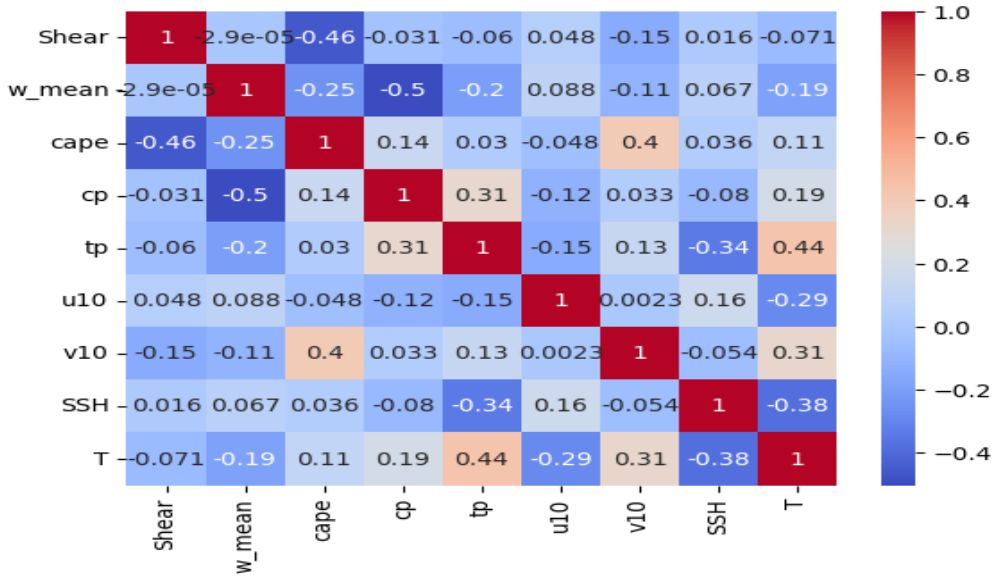

Figure 1: Correlation Matrix between Feature Variables AND T

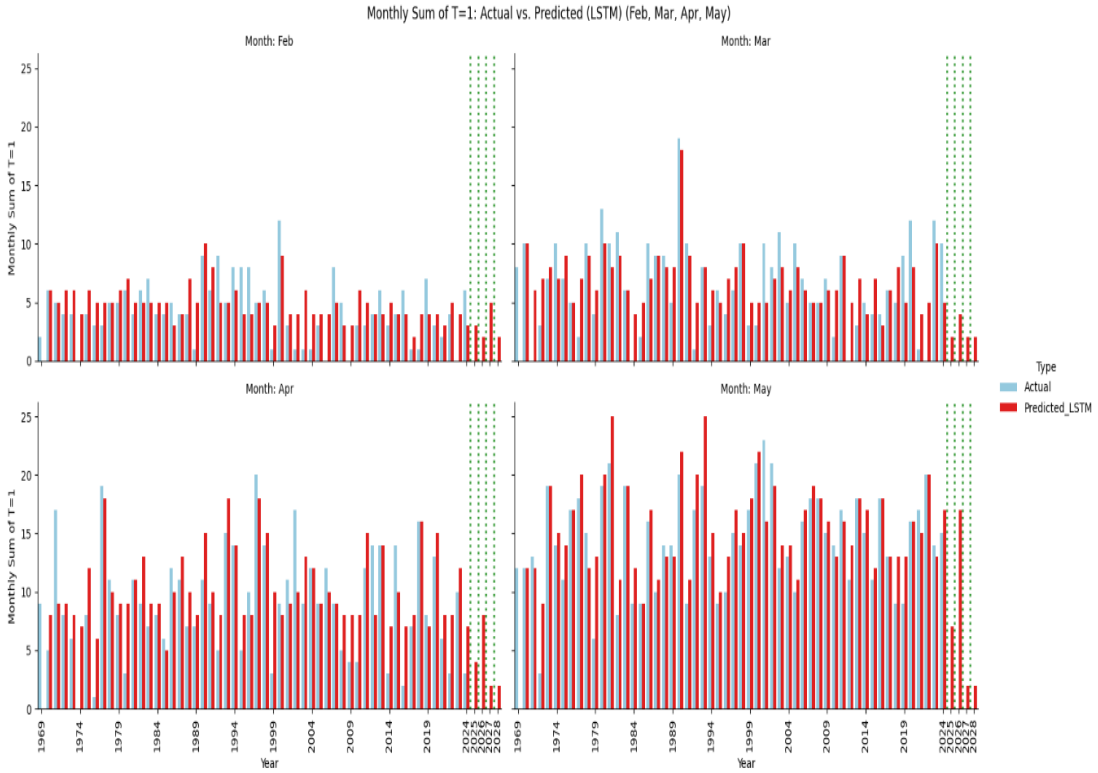

Figure 2: Monthly Trend of T=1, for Months February, March, April, May, since 1969 to 2026 with Predicted Trend till 2028.

References

1. J. M., 2024. An explainable machine learning technique to forecast lightning density over North-Eastern India. *Journal of Atmospheric and Solar Terrestrial Physics,* 259(June),p.106255.
2. Raheem, B. D., 2023. Techniques for lightning prediction: A review. *Ukrainian Journal of Educational Studies and Information Technology,* 11(4), pp. 227-241.
3. Saha, K., 2025. Prediction of lightning events over Bangladesh: A machine learning perspective. *Journal of Atmospheric and Solar Terrestrial Physics,* 268(March), p. 106448.
4. Song, G., 2023. Nowcasting lightning occurrence from commonly available. *Nature,* August, 6(126), pp. 1-10.
5. Wang, X., 2023. A Survey of Deep Learning-Based Lightning Prediction. *MDPI,* 14(1698), pp. 1-17.