Thunderstorms and Lightning Over South and West Bengal Using Machine Learning

Submitted by:
Srinjayee Ganguly
Department of Data Science
B.C. ROY ENGINEERING COLLEGE

Supervised by:
Dr. Sourish Bandopadhyay
IMD, Kolkata

July 2025

Abstract

The goal of this research project is to develop a machine learning-based prediction system that can forecast thunderstorm and lightning events throughout southern West Bengal, with a focus on Kolkata and its nearby districts. Key thermodynamic and dynamic indicators such as. CAPE, CIN, Lifted Index, K-Index, Theta-e, D-CAPE, cloud base and top pressures, dew point temperature, are included in the gridded meteorological variables to efficiently devise the best system for predicting such weather phenomena.

Keywords: Thunderstorm Prediction, Lightning Forecasting, Machine Learning, XGBoost, SMOTE, South Bengal

1 Introduction

Among the most hazardous and frequent convective weather events that impact tropical and subtropical climates are thunderstorms, which are characterized by lightning, torrential rainfall, gusty winds, and occasionally hail. It encompasses not only the congested urban sprawl of Kolkata, the seventh-largest metropolis in India in terms of population, but also wetlands, riverine habitats, and productive agricultural areas. The area is particularly vulnerable to pre-monsoon convection, moisture surges, and land-sea breeze interactions that trigger thunderstorms, as it is situated at the intersection of the Bay of Bengal and the Indo-Gangetic plains. Its dependence on outdoor labor-intensive industries (such as construction and farming), high population density, and deteriorating infrastructure further make it more vulnerable to lightning-related risks. In this area, severe convective events often result in power outages, communication system damage, and even fatalities.

Conventional thunderstorm forecasting techniques rely on empirical indices and numerical weather prediction (NWP) models, which often display poor localized forecasting accuracy, large processing costs, and narrow spatial resolution. Furthermore, real-time lightning observational data is often absent or under-reported, which further complicates accurate forecasting. Data-driven techniques for weather prediction have become possible in recent years due to advancements in machine learning and the availability of reanalysis datasets. These techniques offer an additional tool to traditional forecasting systems by learning intricate, non-linear correlations between atmospheric components and observed weather outcomes.

Using an array of meteorological variables, this study intends to investigate the application of supervised machine learning techniques—especially, eXtreme Gradient Boosting (XGBoost) and Random Forest classifiers—for forecasting the daily occurrence of lightning. In the critical pre-monsoon months of March through May, the emphasis area is Southern West Bengal, which includes the Kolkata metropolitan area.

Additionally, the study contributes to the developing subject of climate informatics by establishing a replicable framework that can be extended to various regions, time frames, or weather occurrences.

2 Literature Survey

The use of machine learning in meteorological forecasting has seen remarkable growth in recent years, especially in the context of severe convective weather events such as thun-

derstorms and lightning. Traditional physics-based models, although foundational, often struggle with high-resolution short-term predictions due to their reliance on numerical approximations and computational constraints.

A number of deep learning-based approaches have demonstrated promise in overcoming such limitations. For instance, some studies [8] effectively used deep learning to predict convective initiation and storm evolution by identifying patterns in numerical weather prediction (NWP) data, while others [6] focused on extracting meaningful predictors from storm-scale NWP models, improving both real-time forecasting and interpretability. Additional efforts [12, 3, 13] have explored radar and satellite-based inputs to build robust nowcasting models using convolutional and recurrent neural networks. While these methods benefit from the scalability of deep learning, their dependence on large-scale NWP input data and significant training time can be barriers in data-scarce or real-time applications.

In contrast, classical machine learning models like decision trees, random forests, and support vector machines have been applied in more region-specific contexts. For example, in India, [2] employed decision trees and random forest classifiers for lightning prediction in eastern India using surface meteorological variables, emphasizing the role of high-resolution spatial and temporal feature selection. Similarly, support vector machines and logistic regression have also been used [11], demonstrating that relatively simple models can perform competitively when backed by carefully engineered features. While computationally efficient and interpretable, these models often lack the capacity to generalize across complex convective environments.

Some researchers have focused on the use of thermodynamic indices and satellite-derived features. Studies such as [9, 10] leveraged indices like the K-Index and SWEAT Index from satellite and radiosonde data to estimate thunderstorm potential. These domain-specific indicators offer insights into the atmospheric stability and moisture content, enhancing forecasting reliability in data-limited regions.

Ensemble-based and hybrid machine learning techniques have also gained traction. For instance, [1] proposed ensemble classifiers combining random forests and gradient boosting, trained on features derived from atmospheric profiles. Similar works [5, 7] highlighted the advantage of combining multiple learners to improve generalization across convective regimes. These models achieved improved accuracy, but often relied on high-quality satellite and radar data.

Finally, a comprehensive review by [4] summarizes the trajectory of machine learning applications in meteorology, highlighting the evolving trends, strengths, and limitations of data-driven approaches.

Overall, these studies underscore the shift in thunderstorm prediction from purely physics-driven models to hybrid and data-driven methodologies. Each approach brings its own trade-offs in terms of complexity, interpretability, input requirements, and real-time applicability.

3 Methodology

The overall objective is to design an accurate and interpretable machine learning pipeline to predict thunderstorm and lightning events during the pre-monsoon season in Southern West Bengal using meteorological and lightning data sourced from the **India Meteorological Department (IMD).** over four years, from March 2020 to May 2024.

3.1 Data Acquisition and Preprocessing

The preprocessing phase involved several crucial steps aimed at preparing the heterogeneous meteorological and lightning datasets for model training. The lightning occurrence data, sourced from the India Meteorological Department (IMD), was available in NetCDF format, while the majority of meteorological variables—such as temperature, pressure, humidity, and wind components—were originally provided as GRIB files. To enable consistent and efficient handling of both datasets, all GRIB files were first converted into NetCDF format using appropriate tools and libraries.

Following conversion, the datasets were temporally filtered to coincide with March to May, which is traditionally when Southern and Western Bengal have the most thunderstorm activity. Every day throughout this time frame was given a binary label, with 1 indicating the presence of lightning and 0 denoting its absence. To create a clean, daily label series, the lightning datasets needed different pre-processing methods depending on the year, such as column reassignments, date standardization, and duplicate entry removal.

Pressure-level variables such as temperature, relative humidity, specific humidity, and wind speed and direction components were used to derive meteorological features. Using domain-specific methods, a set of advanced meteorological indices was generated from these fundamental variables. Cloud base and top pressure levels, Lifted Index (LI), K-Index (KI), Total Totals Index (TTI), Downdraft CAPE ($DCAPE$), Lightning Potential Index (LPI), Convective Available Potential Energy ($CAPE$), Convective Inhibition (CIN), and *Theta-E* were among them. These features were calculated manually to optimize computational efficiency.

To understand the temporal and spatial dynamics of thunderstorms in the region, several feature-wise exploratory visualizations were generated. One of the key indicators of storm likelihood is dew point temperature, which reflects the atmospheric moisture content. As seen in *Figure 1*, elevated dew point temperatures clustered over coastal South Bengal, particularly around the Kolkata region, highlight the area's vulnerability to moisture-driven convective activity.

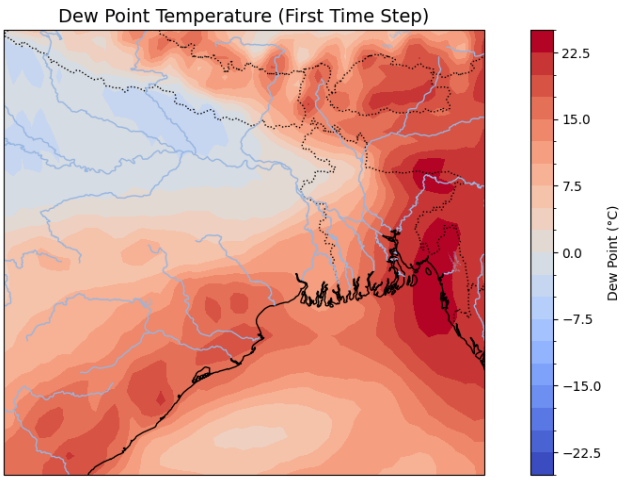

Figure 1: Dew Point Temperature over Southern Bengal

A density map of lightning strikes across the region from March to December 2024 (*Figure 2*) revealed intense activity over South Bengal and eastern Jharkhand, reinforcing the need for focused predictive modeling in this area.

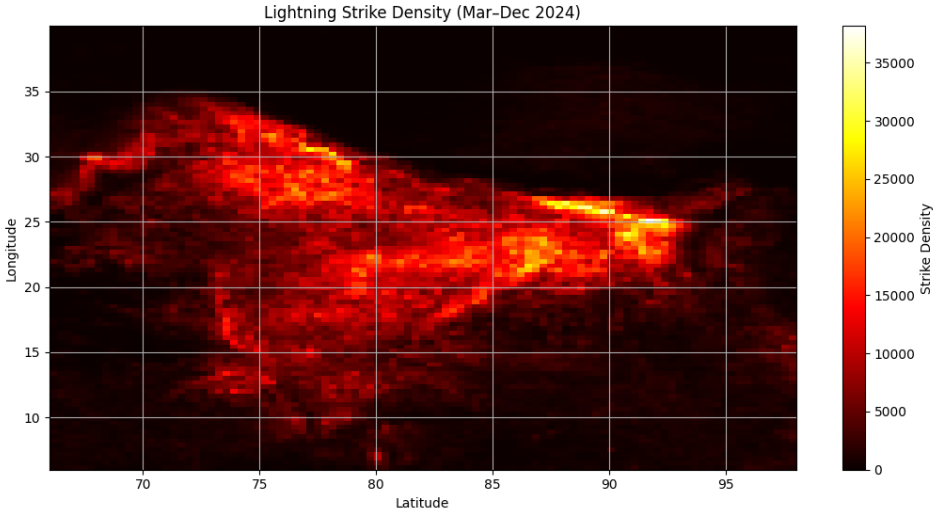

Figure 2: Lightning Strike Density

Further examination of temperature patterns at mid-level atmospheric pressure (850 hPa) revealed stark contrasts in the thermal gradient, which is often associated with vertical instability. *Figure 3* presents a temperature distribution map that delineates warmer southern regions from relatively cooler northern ones. This spatial gradient is crucial in storm formation dynamics.

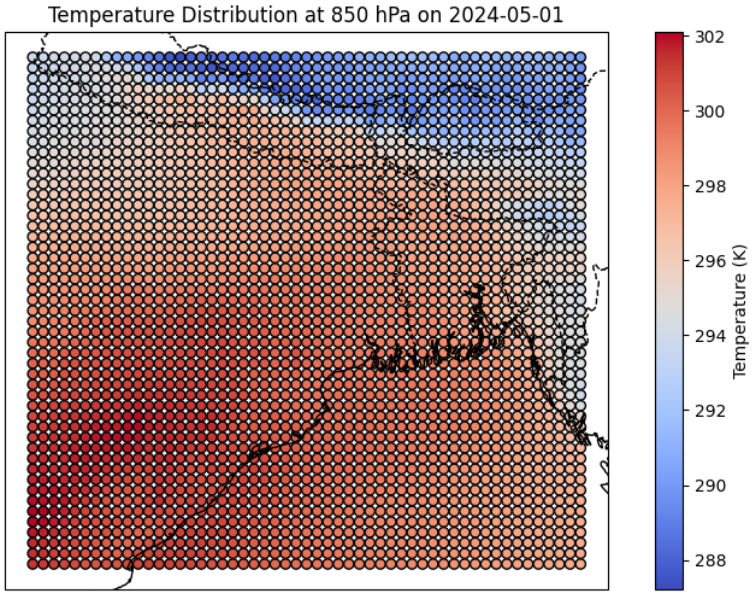

Figure 3: Temperature at 850 hPa

Wind patterns also play a crucial role in the development of convective systems. The U and V wind vectors visualized in *Figure 5* captured the flow of air masses toward South Bengal from the Bay of Bengal. This convergence of warm, moist air supports the buildup of thunderstorm conditions.

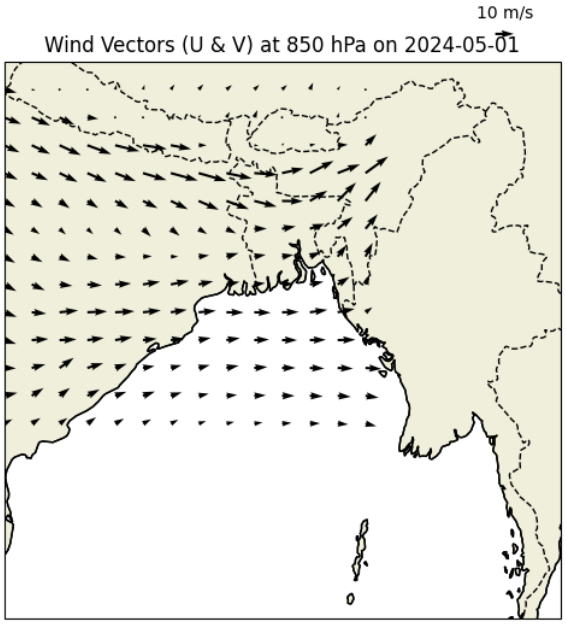

Figure 4: Wind vector patterns

3.2 Meteorological Feature Significance and Formulation

A variety of thermodynamic, and moisture-related parameters were extracted from gridded meteorological data supplied by the India Meteorological Department (IMD) in order to represent the convective properties of the atmosphere over South Bengal. These factors were picked because they have been experimentally proven to be linked to lightning activity and thunderstorm formation in tropical and subtropical regions. To enable multidimensional analysis and variable computation utilizing scientific Python modules like *xarray, NumPy*, and *MetPy* were used.

The estimation of **Convective Available Potential Energy (CAPE)**, which measures an air parcel's buoyancy and represents the atmospheric instability accessible for deep convection, was a key step in the feature engineering process. With regard to ambient temperature, CAPE was computed by integrating the positive buoyancy between the Level of Free Convection (LFC) and the Equilibrium Level (EL). **Convective Inhibition (CIN)**, which measured the negative energy a parcel must overcome to ascend from the surface to the LFC, was a complement to this.

The **Lifted Index (LI)**, a measure of the temperature differential between an elevated air parcel and the surrounding air at 500 hPa, was another important characteristic that was determined. High instability is indicated by strong negative LI values, which are generally linked to observed lightning activity. Additionally included were other thermodynamic indicators, including the **Total Totals Index (TTI)** and the **K-Index (KI)**. TTI takes into account both temperature and dew point gradients to characterize

atmospheric instability, whereas KI incorporates lapse rates and moisture content across numerous layers, which makes it effective in identifying the likelihood for widespread thunderstorms.

To better quantify the total latent energy in an air parcel, **Theta-E** was computed. This parameter combines temperature and moisture into a single measure of instability and was particularly useful in identifying regions with high convective potential. In addition, the **SWEAT**(Severe Weather Threat) Index was included as a composite indicator that accounts for moisture, lapse rate, and wind shear across the lower and mid-troposphere. Cloud base and top pressures were calculated using pressure and temperature criteria to evaluate cloud shape.

Finally, to determine which features had the most predictive influence, a feature importance analysis was conducted using the internal scoring mechanism of XGBoost, the primary model used. This analysis revealed that CAPE, LI, low-level convergence, and Theta-E consistently ranked among the most significant features across training years.

3.3 Dataset Construction and Label Integration

After all of the years from 2020 to 2024 had their features extracted, the next significant phase was to combine the obtained meteorological features with the data from lightning events. First, the format and resolution of the lightning occurrence data set were temporarily standardized by preprocessing. The lightning dataset's records were binary-labeled as either 1 (lightning present) or 0 (no lightning), indicating whether or not a lightning incident was seen at any location on the grid on that specific day.

The date field was used to temporarily combine the lightning labels with the meteorological features. This made sure that a lightning label was associated with each daily set of atmospheric conditions, which were determined by factors like cloud base pressure, wind divergence, CAPE, CIN, LI, and so on. Importantly, the feature data was limited to the pre-monsoon months (March to May), which are known to have higher thunderstorm and lightning activity in South Bengal.

The training dataset was subjected to the **Synthetic Minority Oversampling Technique** (SMOTE) method in order to remedy imbalance caused. By interpolating between pre-existing samples in the feature space, SMOTE creates artificial examples of the minority class. This approach increases the representation of lightning days while maintaining the original data's structure. To further strengthen the model's capacity to learn intricate lightning patterns without overfitting to particular years or conditions, additional custom oversampling was used.

3.4 Model Training

The thunderstorm prediction framework was built primarily using the XGBoost algorithm, chosen for its superior performance on structured tabular datasets. As a benchmark comparison, a Random Forest classifier was also trained on the same dataset. Data from the years 2020 to 2023 was used exclusively for training, while 2024 was held out as an unseen test set to evaluate real-world model performance.

Given the inherent class imbalance in lightning data (many more non-lightning days than lightning days), the training set was balanced using SMOTE. This oversampling created synthetic lightning cases based on feature-space interpolation, resulting in a near

1:1 ratio between lightning and non-lightning days. This balancing was applied only to the training data, preserving the integrity of the test set for unbiased evaluation.

Both the XGBoost and Random Forest models were trained on the SMOTE-balanced dataset. Additionally, for both models, the optimum decision threshold was determined by selecting the threshold that maximized the F1 score on validation data.

For the XGBoost model, a grid search strategy was employed using GridSearchCV from scikit-learn. A hyperparameter grid was defined, such as `n_estimators`, `max_depth`, `learning_rate`, `subsample`, and `colsample_bytree` to optimize the learning process of the model.

After training, both models were evaluated on the 2024 test set using a per-day evaluation strategy—the predicted label for each date was taken as the mode of its hourly predictions. Evaluation metrics included not only standard classification scores such as Accuracy, Precision, Recall, and F1 Score, but also meteorology-specific skill scores: Critical Success Index (CSI), False Alarm Ratio (FAR), Bias, Miss Rate, and True Skill Statistic (TSS).

4 Results

To evaluate the effectiveness of the lightning prediction framework, both the XGBoost and Random Forest classifiers were trained using data from 2020 to 2023 and tested on unseen data from 2024 (March–May). The evaluation was carried out daily, and the final predictions were compared with the ground-truth lightning occurrence data provided by the India Meteorological Department.

4.1 XGBoost Model Results

The XGBoost model, after hyperparameter tuning and SMOTE-based balancing, demonstrated superior performance across all metrics. The confusion matrix in Figure 5 illustrates the model's predictive capability.

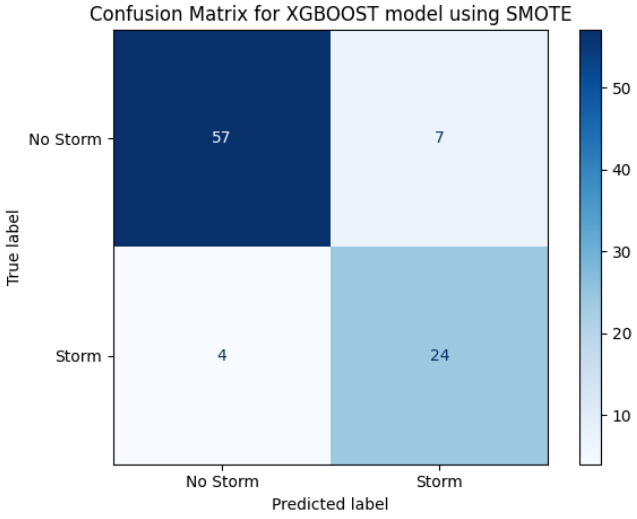

Figure 5: Confusion Matrix of the XGBoost model (Per-Day Evaluation)

The contingency table for the XGBoost model is presented below in Table 1.

Table 1: Contingency Table – XGBoost Model

	Observed Lightning	**No Lightning**
Predicted Lightning	TP = **24**	FP = **7**
Predicted No Lightning	FN = **4**	TN = **57**

Based on this contingency matrix, the following evaluation metrics were computed:

- **Accuracy:** 88.04
- **Precision:** 77.42
- **Recall (POD):** 85.71
- **F1 Score:** 81.30
- **False Alarm Ratio (FAR):** 22.58
- **Critical Success Index (CSI):** 68.57
- **Bias Score:** 1.11
- **Miss Rate:** 14.29
- **Correct Non-Occurrence:** 89.06
- **True Skill Statistic (TSS):** 0.748

4.2 Random Forest Model Results

For comparison, a Random Forest classifier was also trained on the same dataset with the same features and SMOTE. Its confusion matrix is shown in Figure 6.

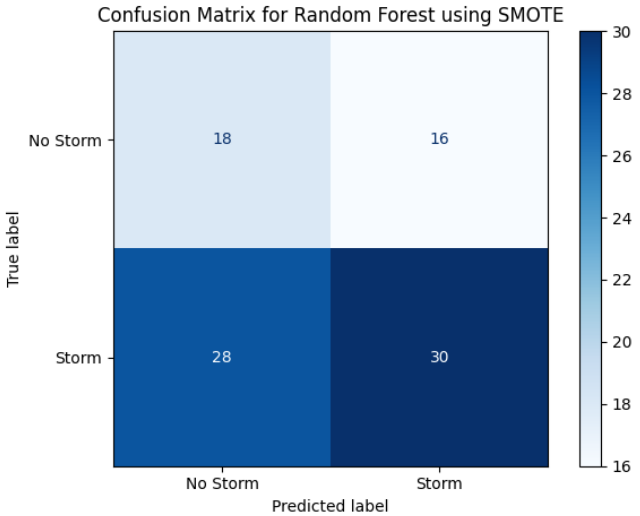

Figure 6: Confusion Matrix of the Random Forest model

The contingency matrix and associated performance scores for the Random Forest model are provided in Table 2 and the list below.

Table 2: Contingency Table – Random Forest Model

	Observed Lightning	**No Lightning**
Predicted Lightning	TP = **30**	FP = **16**
Predicted No Lightning	FN = **28**	TN = **18**

- **Accuracy:** 52.17
- **Precision:** 65.22
- **Recall (POD):** 51.72
- **F1 Score:** 57.78
- **False Alarm Ratio (FAR):** 34.78
- **Critical Success Index (CSI):** 40.54
- **Bias Score:** 0.793
- **Miss Rate:** 48.28
- **Correct Non-Occurrence:** 52.94
- **True Skill Statistic (TSS):** 0.047

4.3 Comparison and Interpretation

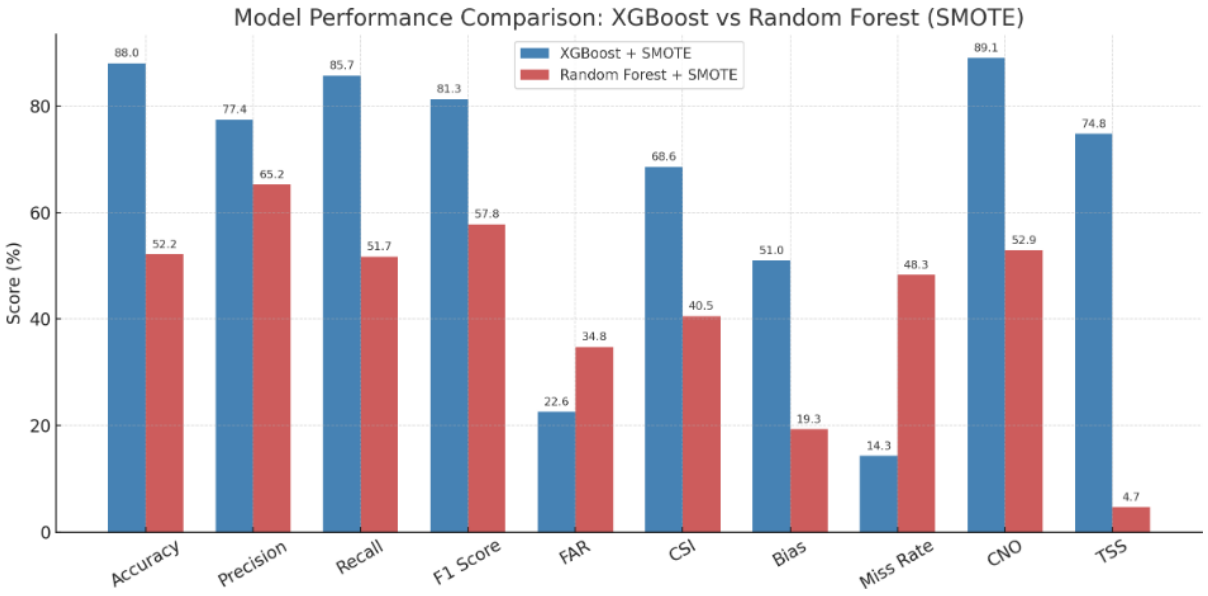

Figure 7: Comparison between the XGBoost and Random Forest models

Table 3: Comparison of Performance Metrics between XGBoost and Random Forest

Metric	**XGBoost**	**Random Forest**
Accuracy	88.04%	52.17%
Precision	77.42%	65.22%
Recall (POD)	85.71%	51.72%
F1 Score	81.36%	57.69%
False Alarm Rate (FAR)	22.58%	34.78%
Critical Success Index (CSI)	68.57%	40.54%
Bias	1.11	0.79
Miss Rate	14.29%	48.28%
Correct No Lightning (CNO)	89.06%	52.94%
True Skill Statistic (TSS)	0.748	0.047

The usage of XGBoost in conjunction with SMOTE to handle unbalanced meteorological data is highly supported by this comparison. It indicates that thunderstorm prediction performance over Southern Bengal is greatly enhanced by appropriate resampling, model selection, and feature integration.

5 Limitations

This study has a number of shortcomings despite its encouraging results. The model's capacity to represent intra-day thunderstorm dynamics is limited by the lightning labels' daily generation. The analysis's relevance to other seasons was not explored because it was restricted to the pre-monsoon season (March–May). Furthermore, the model does not include real-time satellite or radar inputs that could improve prediction accuracy; instead, it only uses ground-based meteorological data. Although SMOTE is useful for class balance in XGBoost, it generates artificial patterns that could not accurately depict atmospheric processes. Unbalanced data was used to train the Random Forest model for benchmarking in order to prevent the decrease in performance, but that didn't serve as an equal in comparison.

6 Conclusion

Despite the promising results achieved through this machine learning-based framework, several limitations should be acknowledged. Firstly, while SMOTE was employed to address class imbalance in both the XGBoost and Random Forest models, this synthetic oversampling method may introduce noise by generating minority class samples that do not fully represent the natural variability of convective environments. This can occasionally lead to overfitting, particularly in Random Forests, which are sensitive to data duplication and noise, potentially reducing generalization in unseen conditions.

Secondly, although the threshold for classification was optimized based on F1 score balancing precision and recall—this approach does not necessarily align with real-world cost sensitivities, such as prioritizing fewer false negatives in disaster management scenarios. Additionally, the model performance is currently constrained to the pre-monsoon season and trained on a geographically limited domain (South and West Bengal), which

may limit the generalizability of the approach to other regions or seasons without domain adaptation.

Furthermore, while the models integrate a variety of atmospheric indices, the lack of vertical resolution (e.g., pressure-level profiles beyond a few summary features) and real-time satellite data limits the system's ability to fully resolve storm-scale processes. The reliance on surface and mid-level reanalysis data, although pragmatic, may underrepresent mesoscale dynamics critical to short-term thunderstorm prediction.

Lastly, although XGBoost and Random Forest provide strong performance and interpretability, the study does not yet incorporate temporal sequence learning or spatial dependencies. Future work can extend this framework through recurrent or convolutional architectures, attention mechanisms, or hybrid physics-ML fusion models to better capture the evolving spatio-temporal structure of convective systems.

References

[1] Arka Banerjee, R Basu, and R Chakraborty. Thunderstorm forecasting using ensemble classifiers and feature engineering from atmospheric profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022.

[2] Sayantan Chowdhury, Swagat Mahato, Shreya Mitra, and Prasanta Bandyopadhyay. Lightning prediction over indian region using machine learning techniques. *Natural Hazards*, 109(1):163–182, 2021.

[3] David J Gagne II, Amy McGovern, and Sue Ellen Haupt. Storm-based probabilistic hazard guidance using deep learning. *Weather and Forecasting*, 35(4):1415–1432, 2020.

[4] S Kamini, Ankita Bansal, and Anju Vyas. Machine learning techniques in weather prediction: A review. *International Journal of Scientific & Technology Research*, 8(12):3400–3405, 2019.

[5] Ali Karimian, Ali Zare, and Udaysankar S Nair. Improving severe thunderstorm prediction using ensemble learning and feature selection. *Atmospheric Research*, 238:104879, 2020.

[6] Ryan Lagerquist, Amy McGovern, and Timothy Smith. A method for extracting predictors from storm-scale numerical weather prediction models for real-time severe weather prediction. *Weather and Forecasting*, 34(4):1271–1293, 2019.

[7] Hao Li, Chen Liu, Yu Zhang, and Yaowen Qiu. An ensemble machine learning approach for convective storm nowcasting using satellite and radar data. *Remote Sensing*, 13(11):2114, 2021.

[8] Amy McGovern, David J Gagne, Grant E Jergensen, Kim L Elmore, Cameron R Homeyer, and Ryan Lagerquist. Deep learning to predict convective initiation, evolution, and hazards. *Journal of Advances in Modeling Earth Systems*, 9(3):1721–1738, 2017.

[9] U C Mohanty, M Pradhan, S Pattnaik, and S K Sahu. Satellite-based estimation of thunderstorm potential using k-index and sweat index over india. *Mausam*, 71(2):375–386, 2020.

[10] P Mukhopadhyay, S Abhilash, and R P M Krishna. Forecasting of pre-monsoon thunderstorms in the gangetic west bengal region using conventional and satellite data. *Natural Hazards*, 49:181–197, 2009.

[11] G Suresh, D C Raju, V S Rao, and V Krishnamurthy. Thunderstorm prediction using support vector machine and logistic regression models. *Meteorological Applications*, 26(4):689–698, 2019.

[12] Sarah Taylor, David Gagne, Amy McGovern, Ming Xue, and Bryan Gallo. Deep-learning techniques for convective forecasting. *Bulletin of the American Meteorological Society*, 99(9):1847–1859, 2018.

[13] Feng Zhang, Xue Liu, and Yun Wang. A deep-learning-based convective storm nowcasting model using radar and satellite data. *Atmospheric Research*, 229:222–229, 2019.