Journal of Data, Information and Management
https://doi.org/10.1007/s42488-022-00071-9

ORIGINAL ARTICLE

Interpretation of black box for short-term predictions of pre-monsoon cumulonimbus cloud events over Kolkata

Debashree Dutta[1] · **Sankar K. Pal**[1]

Received: 30 April 2022 / Accepted: 19 May 2022

Abstract
Thunderstorms are meso-scale systems that are characterised by deep convective cumulonimbus (Cb) clouds associated with torrential rain, lightning, hail, dust storms, strong winds, downbursts, and tornadoes. As Gangetic West Bengal is prone to thunderstorm, early forecasting is imperative in order to protect life and property, and prevent the damage caused by these intense storms. The present study comprises two issues on model evaluation and interpretability by applying popular machine learning algorithms, viz., Extreme Gradient Boosting (XGBoost) and Logistic Regression (LR) with potentially predictive thermodynamic indices and parameters for short-term predictions of pre-monsoon thunderstorms over Kolkata. The thermodynamic indices and parameters employed include convective available potential energy (CAPE), convective inhibition (CIN), Bulk Richardson number (BRN), K-Index (KI), lifted index (LI), total totals Index (TT), Showalter index (SI),temperature (TEMP), relative humidity (RELH), dew point temperature (DWPT), wind direction (DRCT), wind Speed (SKNT), mixing ratio (MIXR), severe weather threat index (SWI), potential temperature (THTA), equivalent potential temperature (THTE), and virtual potential temperature (THTV). In the proposed approach, we place a greater emphasis on the concept of Explainable artificial intelligence (XAI) to apply SHapley Additive exPlanations (SHAP), a Shapley-value-based explanation method based on the coalitional game theory. The SHAP approach, as a primary interface, enables the identification and prioritization of features that determine the occurrences of pre-monsoon thunderstorms and compares the two different machine learning algorithms. SHAP can quantify the contribution of predictor variables to each data point and rank the importance of predictor variables in terms of their contributions to the model output. It also facilitates the computation of different plots on both global and local levels. Accordingly, it can help determine the validity of the model based on domain expertise by identifying the most important variables. The results indicate that both XGBoost and LR support the dominant positive influence of the convective available potential energy (CAPE), while the ranks and interpretations of the other predictor variables differ. Although, these two models perform well in predicting the pre-monsoon thunderstorms, they may favour different predictor variables due to their varying natures, thereby resulting in different explainability.

Keywords Thunderstorm · SHapley Additive exPlanations (SHAP) · XGBoost · Logistic Regression · CAPE

1 Introduction

The cumulonimbus cloud (Cb), or thunderstorm, is a convective cloud or cloud system that produces rainfall, lightning, and often develops large hail, severe wind gusts, tornadoes, and heavy rainfall. It normally lasts less than an hour and can range in size from a few kilometres to a few hundred kilometres, posing a serious threat to human life, the economy, agriculture, and infrastructure. This weather occurrence has piqued the interest of professional meteorologists for nearly a century because of its potential to wreak harm to people and property on the ground as well as aircraft. Severe thunderstorms known as "Nor'westers" and locally known as "Kalbaisakhi" hit Gangetic West Bengal (GWB) and nearby districts during the pre-monsoon months of March, April, and May. Thunderstorms are most common in the afternoons and evenings. It arises when warm, moist air rises from the Earth's surface due to strong subsurface heating, which causes the ascending air to cool due

✉ Debashree Dutta
debashree.120@gmail.com

Sankar K. Pal
sankarpal@yahoo.com

[1] Center for Soft Computing Research, Indian Statistical Institute, Kolkata 700108, India

Published online: 22 June 2022

to adiabatic expansion and condensate. As the water in the rising air mass condenses and transforms from a gas to a liquid, it transfers energy into the surrounding air, further heating it and causing further convection and the cloud mass to rise to higher heights. This results in the formation of a towering cumulonimbus cloud (Cb). The clouds can rise to the top of the troposphere (18 km in the tropics) and the movement of water droplets and ice crystals within a cumulonimbus cloud (thunderhead) creates an electrical charge, which causes lightning. The positive charge forms at the top, while the negative charge develops at the bottom, forming at the cloud's base. Severe thunderstorms and lightning forecasting are difficult for meteorologists and atmospheric scientists around the world, since such highly nonlinear and chaotic occurrences can have serious effects.

Extreme weather events have also increased because of climate change around the world. Warming, and hence higher available moisture in the atmosphere, increases latent instability, favouring convection. Pre-monsoon season (March–May) in India is the most common season in which ~ 76% of the tornadoes occur in India. Further, ~ 72% of the reported tornadoes in South Asia occur in north-east India and Bangladesh, and are connected with Nor'westers (STORM 2005). An effective early warning system is essential to reduce the disaster risk. However, the forecasting of thunderstorms is a challenging task as it belongs to the cloud-scale or meso-scale. Besides, the event is extremely complicated, distinct, and highly unstable. As a result, accurate forecasting with sufficient lead time is a primary responsibility of the meteorological community. Synoptic weather chart analysis, thermodynamic diagram studies (T– ϕ gram), satellite imagery, Doppler Radar observations, and statistical and numerical models are used to forecast thunderstorms on a daily basis. Predicting thunderstorms by numerical weather prediction (NWP) models using the atmospheric variables influenced by past conditions, means predicting thunderstorms is based on the initial condition of the atmosphere. In reality, one of the drawbacks of NWP models is the predictability of a chaotic atmosphere, and this increases intrinsic model error (Collins and Tissot 2016).

Over the past few decades, numerous machine learning (ML) algorithms have been developed to solve various problems arising in real-life applications (Pathak and Pal 1986; Sen and Pal 2010; Pal et al. 2015; Pal and Pal 2017; McGovern et al. 2017; Nie et al. 2018; Maddalena et al. 2020; Jiang et al. 2021). Since the 1990s, the use of artificial intelligence in atmospheric sciences has received much interest. As there is no local network of observatories, traditional methods of forecasting small-scale weather events have some limitations and may diverge from accurate forecasts. In addition to the traditional approaches, machine learning techniques have been widely adopted to synthesize large amounts of atmospheric data due to their capability to capture the complex relationships in data and deliver accurate forecasts without explicit prediction assumptions (Jergensen et al. 2019). ML models map a set of inputs to a certain output by optimizing the model's structure so that the disparities between the ML predictions and the output observations, or "ground truth," are as small as possible. ML models enhance the prediction accuracy of forecast, often used as a substitute for NWP with post-processing and interpretation of predictions (McGovern et al. 2017; Rasp and Lerch 2018; Schultz et al. 2021). Many researchers have used machine learning techniques to nowcast the occurrences of lightning (Mostajabi et al. 2019, 2020; Zhou et al. 2020; Shrestha et al. 2021). In the present study, Logistic Regression (LR) and Extreme Gradient Boosting (XGBoost), the two mostly used data-driven approaches, are applied to assess and predict the occurrences of thunderstorms. Logistic regression analysis is used to examine the association (categorical or continuous) of independent variable(s) with one dichotomous dependent variable. Sánchez et al. (1998) applied logistic regression to the short-term forecast of hail risk in the province of Leon in the northwestern Iberian Peninsula of Spain. Dasgupta and De (2007) considered binary logistic regression models for the prediction of convective developments from a prior knowledge of the values of certain dynamic and thermodynamic parameters. Lee et al. (2020) used logistic regression for convective detection in Korea. However Logistic Regression is a generalized linear model (McCullagh and Nelder 1989; Smita 2021). As an alternative, Gradient Boosting (GB) is a machine learning algorithm that utilises decision trees that are trained iteratively with each successive tree rectifying the errors of the previous trees. Leinonen et al. 2022 used LightGBM, implementation of the GB algorithm for nowcasting thunderstorm hazards in an area in the Northeastern United States. XGBoost is an algorithm that has recently been dominating applied machine learning. It provides a powerful and effective implementation of the gradient boosting ensemble algorithm which is one of the most effective algorithms for supervised learning (Ibrahem Ahmed Osman et al. 2021). It has emerged as a potential tool for short-term forecasting, not only because of its capacity to describe complicated nonlinear systems in a flexible manner, but also because of its ability to directly map input variables to output variables. Although more non-linear regression algorithms such as neural networks have been investigated to predict thunderstorms and convective weather systems (McCann1992; Litta et al. 2013; Collins and Tissot 2016; Zhou et al. 2019; Kamangir et al. 2020; Stankova et al. 2021), XGBoost and Logistic Regression are proved to be effective methods in data-driven weather forecasting.

These ML models are like 'Black Box', constructed directly from the data by an algorithm. That means the users, or even those who develop them have no idea how variables are combined or interact to make predictions, given a list of

input variables. For the task of model interpretation (explanation), we propose to apply SHapley Additive exPlanations (SHAP) to quantify the contributions of different predictor variables. It is based on game-theoretic approach that quantifies the contribution of each predictor variable to the prediction in terms of Shapley values, which can explain how predictions are made from any machine learning algorithm. It also correctly identifies the ranking of the importance of the predictor variables in terms of their contributions to the model output. In this paper, we focus on both the two approaches, viz., global explanations and local explanations. Local explanations try to interpret individual predictions at the single statistical unit level, whereas global explanations characterize the model as a whole in terms of which the explanatory variables mostly determine its predictions for all statistical units. The Shapley value technique, first proposed by Shapley (1953) and applied by Lundberg and Lee (2017a, b), and Strumbelj and Kononenko (2010), is gaining a lot of attention among local explanation approaches because of its significant advantages. Shapley based XAI methods may be used to assess the contribution of each explanatory variable to each point prediction of a machine learning model, regardless of the underlying model or principle. Their interpretation tools are not restricted to their specific model classes or data, allowing for greater applicability and personalization of their findings.

In this study, we emphasise on the concept of "Explainable artificial intelligence" (XAI) for short-term predictions of pre-monsoon thunderstorms over Kolkata (22°33′N and 88°20′E) during the pre-monsoon season (March, April, and May). This is unique.

Here, SHAP is used as a unifying framework to quantify the contributions and relevance of different predictor variables of an ML model; thereby providing the interpretation of the model's decision toward thunderstorm prediction. In terms of the said explainability, we have compared two machine learning models viz., XGBoost and Logistic Regression, with the thermodynamic indices and parameters as input variables. Besides, different performance metrices, namely, relative operating characteristic (ROC) curve, area under the curve (AUC), Matthews correlation coefficient (MCC), root mean square error (RMSE), mean absolute error (MAE), and R^2 (coefficient of determination) have been used to quantify the prediction capability of these models for the purpose of evaluation and comparison. Finally, using the SHAP-based most important variables, the validity of the models is tested with domain knowledge.

We have employed shap Python packages for implementing SHAP. For visualization, we have applied SHAP feature attributions, SHAP explanation force plots, and SHAP summary plots to explore Global and Local explanations.

The remainder of the paper is structured as follows: Section "Materials and methods" presents an overview of the study area, data collection, and detailed architecture and algorithms used. In Section "Results and discussion", experimental results are analyzed, and compared with the domain knowledge. Finally, Section "Conclusions" provides concluding remarks on this research, and suggestions for future investigation.

2 Materials and methods

2.1 Study area

Kolkata is the capital city of the East Indians state of West Bengal, and one of the large metropolitan areas of India. Kolkata is in eastern India in the Ganges Delta at 22°33′N and 88°20′E, along the east bank of the Hooghly River. The Chota Nagpur Plateau is in the western/SW section of the study zone, and the Bay of Bengal is in the southern part. This area belongs to tropical monsoon climate. There is a transition from winter monsoon to summer monsoon circulations during the pre-monsoon season, and the region receives highly intense insolation, resulting in the development of a heat flow. The season is characterised by high surface temperatures and severe convective activity. Over the West Bengal region, two distinct air masses coexist: land-based west to northwest winds and moist winds from the Bay of Bengal. During this time, a low-pressure system is present over the Chota Nagpur Plateau region, West Bengal, Assam, Bangladesh, and the surrounding areas, and a seasonal high-pressure system develops over the Bay of Bengal (Lohar and Pal 1995; Tyagi et al. 2011). During this time, shallow layer of wet southerlies or south-westerlies from the Bay of Bengal near the ground and dry westerlies aloft characterise the upper airflow over Gangetic West Bengal and the surrounding areas. The atmosphere becomes potentially unstable in the pre-monsoon months of March, April and May due to a weak surface pressure field, weak surface and lower atmospheric winds, and intense daytime heating. However, thunderstorms may not occur daily, even though the conditions are ideal for them to occur practically every day throughout these months. Synoptic systems or features provide the trigger for the occurrences of thunderstorms.

2.2 Data collection, data pre-processing and implementation of methodology

The present work is based on meteorological data of Radiosonde (RS) or Rawinsonde (RW) upper-air observations and records of thunderstorms that occurred in Kolkata (22°33′N and 88°20′E) (station no. 42809) during the pre-monsoon months of March, April and May between 2011 and 2020. The RS/RW sounding observations of both 00 UTC and 12 UTC at a height

of six metres are collected from the website of the University of Wyoming, Department of Atmospheric Science (http://weather.uwyo.edu/upperair/sounding.html). The stability indices like convective available potential energy (CAPE), convective inhibition (CIN), Bulk Richardson number (BRN), K-Index (KI), lifted index (LI), total totals Index (TT), Showalter index (SI),temperature(TEMP), relative humidity (RELH), dew point temperature (DWPT), wind direction (DRCT), wind Speed (SKNT), mixing ratio (MIXR), severe weather threat index (SWI), potential temperature (THTA), equivalent potential temperature (THTE), and virtual potential temperature (THTV) are considered for our study. Dates of thunderstorm activity over Kolkata are collected from the STORM Project Report, issued by Regional Meteorological Centre Kolkata. Figure 1 represents, as an example, the thermodynamic diagram of two thunderstorm days over Kolkata. In this study, we have considered both thunderstorm (TS) and non-thunderstorm(non-TS) days. The RS/RW data during TS are collected before the occurrence of thunderstorms. To convert the data on the same scale and make the data a pure dimensionless quantity or value, we have used max–min normalization as follows:

$$z(x) = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

Where min and max are the minimum and maximum values of x, given its range, and $z(x)$ is the normalized value of x.

Seventeen thermodynamic indices and parameters are used as input for the two models in the present study consisting of DRCT, SKNT, THAT, THTE, THTV, CAPE, CIN, BRN, TEMP, DWPT, RELH, MIXR, SI, KI, TT, SWI and LI. These thermodynamic indices and parameters were collected at 00 and 12 UTC before the occurrence of pre-monsoon thunderstorms during the aforementioned months over Kolkata. If a thunderstorm happened between 00 and 12 UTC on a particular day, the data at 00 UTC is used, whilst 12 UTC data is used for the thunderstorms that occur after 12 UTC. The forecast is done with a 0–12 h lead time prior to the occurrence of thunderstorms. The overall research framework is illustrated in Fig. 2. The learning process for prediction (forecasting) of the thunderstorms over Kolkata involves the following steps:

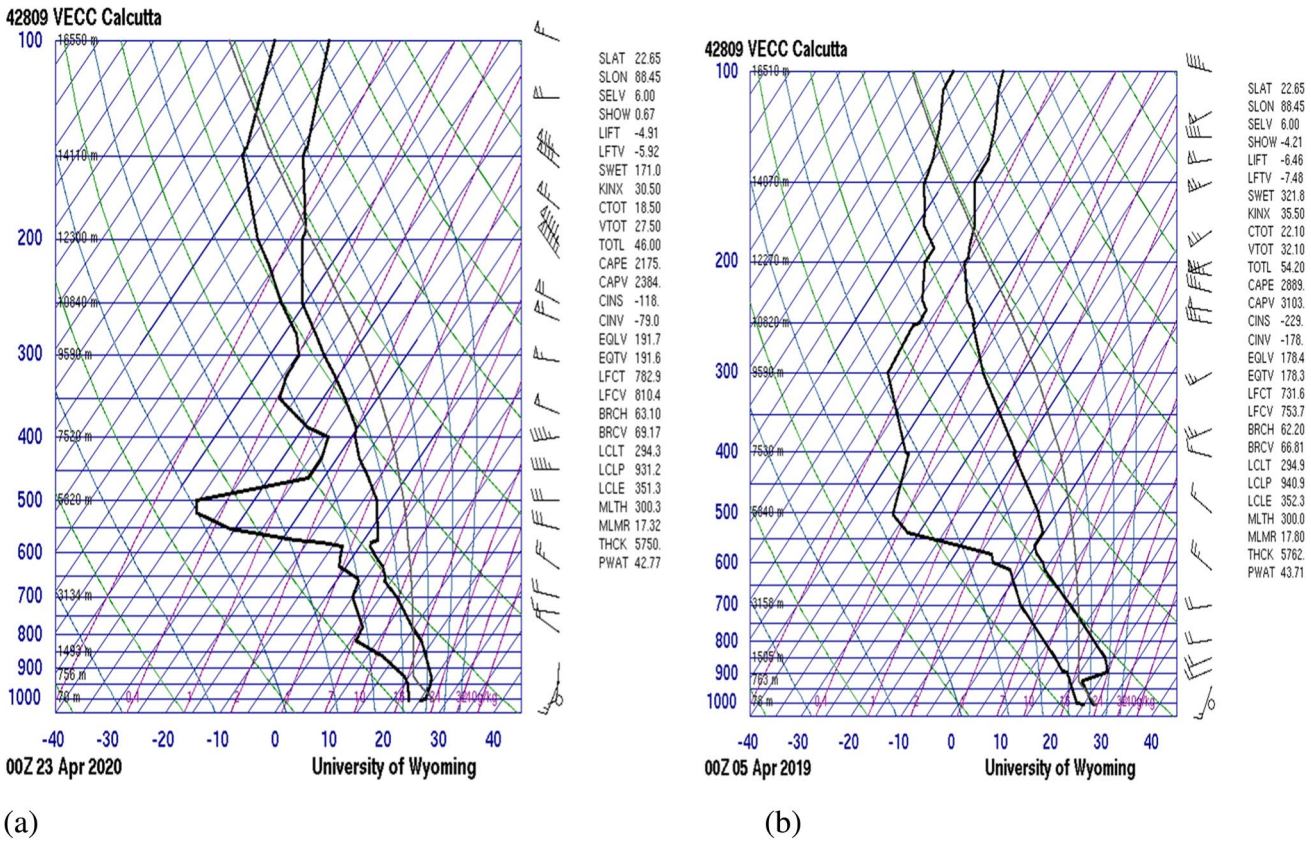

Fig. 1 Thermodynamic diagrams on (**a**) 23 April 2020 at 00:00 UTC. (**b**) 05 April 2019 at 00:00 UTC for VECC Calcutta (Kolkata) from University of Wyoming

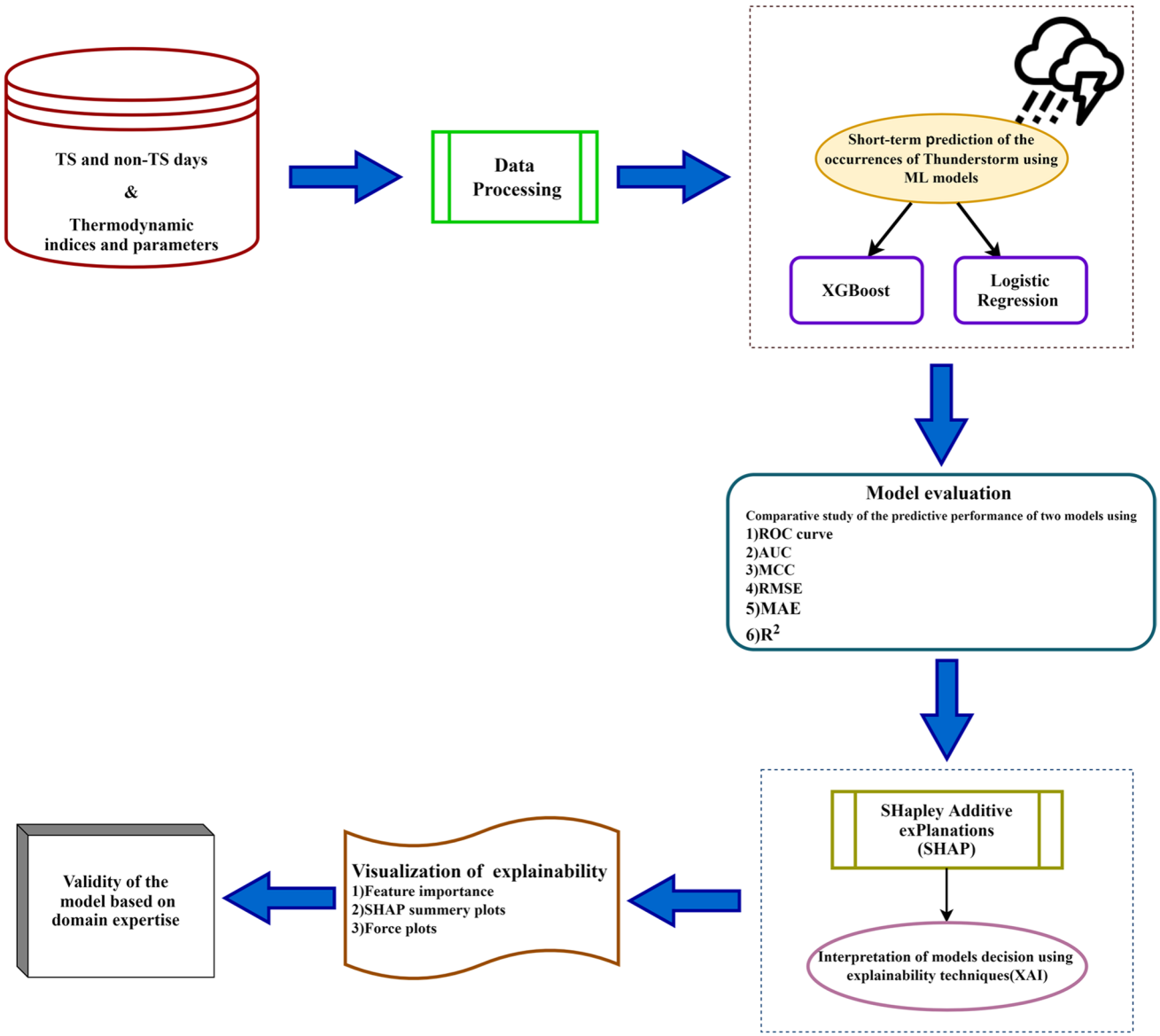

Fig. 2 Block-diagram showing proposed research framework for prediction of pre-monsoon thunderstorms over Kolkata

1) We have compared the prediction performance of two machine learning algorithms, XGBoost and Logistic Regression, for short-term predictions of pre-monsoon thunderstorms. For evaluation and comparison of the models' performance, we have incorporated some performance metrices, namely, relative operating characteristic (ROC) curve, area under the curve (AUC), Matthews correlation coefficient (MCC), root mean square error (RMSE), mean absolute error (MAE), and R^2 (coefficient of determination).
2) We have implemented a Shapley-value-based explanation technique using coalitional game theory to assess and quantify the contribution of each predictor variable to predict the occurrences of thunderstorms. For the purpose of visualization, we have used SHAP feature attributions, SHAP explanation force plots, and SHAP summary plots to explore Global and Local explanations.
3) Finally, the validity of the models is evaluated using the SHAP-based important variables and domain knowledge with explanation.

2.3 Explainable machine learning

Explainable artificial intelligence (XAI) deals with a set of processes and strategies that allows human users to understand and trust the results and output created by machine learning algorithms. It aids in the evaluation of model

accuracy, fairness, transparency, and outcomes in AI-powered decision-making.

As AI advances, humans will find it more difficult to comprehend and retrace how the algorithm arrived at a conclusion. Powerful AI/machine learning (ML) models are so complicated that understanding their internal dynamics is nearly difficult for their creators, hence these are an unexplainable and uninterpretable black box. Understanding how an AI-enabled system arrived at a particular result offers numerous advantages. There have been different approaches and implementations for explaining complex machine learning models. SHAP (SHapley Additive exPlanations) is a game-theoretic approach to explain the output of any machine learning model (Fig. 3). It uses the traditional Shapley values from game theory and aims to explain the individual prediction by computing the contribution of each predictor variable. The Shapley value of a feature indicates its contribution to the output value, weighted and summed over all possible feature combinations. Consider a model where a group N (with n features) is used to predict an output $v(N)$. In SHAP, the contribution of each feature (ϕ_j is contribution of feature j) on the model output $v(N)$ is assigned based on their marginal contribution (Shapley 1953). Using several axioms to enable equitably allocate the contribution of each feature, Shapely value for feature j is determined as follows:

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{j\} - v(S)] \tag{2}$$

Here S represents a subset of the features used in the model and $v(S)$ represents the prediction for feature values in set S.

The total contribution (or Shapley value ϕ_j) of feature j is determined as the average of its contributions across all possible permutations over the feature set S. As a result, features are individually added to the set and the change in the model output indicates their relevance. Additionally, this approach takes into account the feature ordering, which has an impact on the observed changes in a model's output when correlated features are present.

The SHAP technique is an additive feature attribution method that defines the output of a model as the sum of the real values attributed to each input feature. The explanatory model for additive feature attribution methods is specified as a linear function of binary features, as seen in the following equation:

$$g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j x'_j \tag{3}$$

Where $x' \in \{0, 1\}^M$ is the coalition vector, equals to 1 when a feature is observed, otherwise it is 0. M is the maximum coalition size and $\phi_j \in \mathbb{R}$ is the feature attribution for feature j (Lundberg and Lee 2017a, b).

The following are the advantages of SHAP: (1) global interpretability—the aggregate SHAP value can identify the positive or negative relationship between each variable and the target; and (2) local interpretability—each feature of an instance has its own SHAP values.

Furthermore, we can also measure the global importance of features by computing the absolute Shapley value for feature j as follows:

$$I_j = \sum_{i=1}^{n} \left| \phi_j^{(i)} \right| \tag{4}$$

where $\phi_j^{(i)}$ represents the SHAP value of the j-th feature for instance i.

2.4 Logistic Regression (LR)

Logistic regression is considered as a generalized linear model. It is employed when the output is categorical. In the present study, logistic regression has been used for forecasting the probability of occurrences of thunderstorm which in turn decides whether it will be occurred or not. The following function is used in logistic regression (Jafari Goldarag et al. 2016)

$$f(z_i) = \frac{1}{1 + e^{-Z_i}} \tag{5}$$

In Eq. (5), $f(z_i)$-value represents the estimate of the probability of occurrence of a certain binary outcome (occurrences of thunderstorm in the present study) and 1-$f(z_i)$ represents the probability of the opposite outcome (non-occurrences of thunderstorm in the present study). z_i is a linear function of the independent variables. For each observation i, z_i is expressed as follows:

$$z_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i \tag{6}$$

where β_i are the model parameters (known as regression coefficients) which are estimated using the maximum likelihood method.

2.5 Extreme Gradient Boosting (XGBoost)

XGBoost, as introduced by Chen and Guestrin (2016), is a powerful ensemble learning technique that achieves state-of-the-art results in various machine learning challenges. Ensemble learning is a method for combining the predictive abilities of multiple learners in a systematic way. The purpose of XGBoost is to keep adding trees and performing feature splitting in order to grow a tree. When a tree is added, it learns a new function to fit the residuals of the previous prediction. The score of a sample is predicted, when the training is completed. The sample will fall to the corresponding leaf node in each tree based on its attributes. Each leaf node corresponds to a score, and the final score for each tree is the predicted value of the

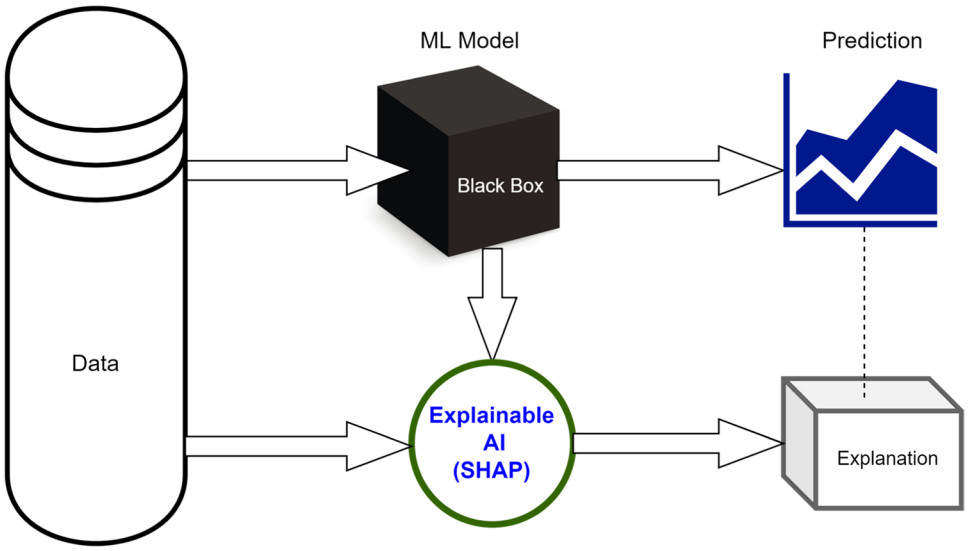

Fig. 3 Interpretation of Black Box using SHapley Additive exPlanations (SHAP)

sample. The trees are built sequentially in boosting, with each consecutive tree aiming to reduce the previous tree's errors. Each tree builds on the knowledge of its predecessor and corrects any lingering faults. As a result, the following tree in the sequence will learn from an updated set of residuals.

Suppose that the model generates t decision trees. Its prediction value for sample i can be represented as (Chang et al. 2019):

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i), f_k \in F, i \in n, \tag{7}$$

Here, $\hat{y}_i^{(t)}$ denotes the predicted value of sample i, which is based on the sum of the predicted values of t decision trees. n is the total number of samples, and the subscript i denotes the i-th sample. f_t indicates the t-th classification tree, and F is the set space of all trees.

The loss function is shown as follows:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \tag{8}$$

Where l denotes the degree of deviation between the predicted value $\hat{y}_i^{(t)}$ and the true value y_i; the second part of Eq. 8 is the sum of the complexity of each tree and $\Omega(f_k) = \gamma * T + 1/2\lambda||\omega||^2$.Here, T denotes the number of leaf nodes,γ is the weight of leaf nodes, and λ and ω are regular coefficients.

Combining Eq. 7 and Eq. 8, and using Taylor expansion of the loss function, we get Eq. 9 as follows

$$\begin{aligned} L^{(t)} &= \sum_{i=1}^{n} l[y_i, \hat{y}_i^{(t-1)} + f_i(x_i)] + \Omega(f_t) + \sum_{k=1}^{t-1} \Omega(f_k) \\ &= \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2h_i f_t^2(x_i)} \right] + \Omega(f_t) + \sum_{k=1}^{t-1} \Omega(f_k) \\ &= \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2h_i f_t^2(x_i)} \right] + \gamma T + \frac{1}{2\lambda w_j^2} + C \end{aligned} \tag{9}$$

where g_i represents the first derivative, h_i represents the second derivative and C is a constant. These are defined as follows:

$$g_i = \partial_{\hat{y}_i^{[t-1]}} l\left(y_i, \hat{y}_i^{t-1}\right), \tag{10}$$

$$h_i = \partial^2_{\hat{y}_i^{[t-1]}} l\left(y_i, \hat{y}_i^{t-1}\right), \tag{11}$$

$$C = \sum_{i}^{n} l(y_i, y_i^{[t-1]}) + \sum_{k=1}^{t-1} \Omega\left(f_k\right), \tag{12}$$

Definition $I_j = \{i|q(x_i) = j\}$ denotes the sample set of leaf node j. After removing the constant term from Eq. 9, the derivative term is 0, and we get the optional solution w_j^* as follows:

$$w_j^* = -\frac{G_j}{H_j + \lambda'} \tag{13}$$

$$G_j = \sum_{i \in I_j} g_i \tag{14}$$

$$H_j = \sum_{i \in I_j} h_i \tag{15}$$

Applying the optimal solution w_j^* into Eq. 9, we get Eq. 16 as follows

$$L^{(t)} = -1/2 \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T + C \tag{16}$$

XGBoost applies the greedy algorithm to segment the existing nodes each time. Let IL and IR be the sets of left and right nodes after segmentation, and $I = IL \cup IR$. Then the information gain after segmentation is:

$$L_{(split)} = Gain = 1/2[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} + \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}] - \gamma \tag{17}$$

$$G_L = \sum_{i \in I_L} g_i, G_R = \sum_{i \in I_R} g_i, H_L = \sum_{i \in I_L} h_i, H_R = \sum_{i \in I_R} h_i \tag{18}$$

From Eq. 17, XGBoost identifies whether a node is splitting by subtracting the unsplit node score from the left and right splitting node score. Furthermore, XGBoost takes account of the complexity of the model and brings a regular term λ to limit the growth of the tree. When the gain is less than λ, no node splitting is performed (Chang et al. 2019).

2.6 Performance indices

We have implemented different indicators, namely, relative operating characteristic (ROC) curve and area under the curve (AUC), Matthews correlation coefficient (MCC), root mean square error (RMSE), mean absolute error (MAE), and R^2 (or coefficient of determination) to analyse and interpret the experimental results. These are defined as follows:

2.6.1 ROC curve and AUC

The performance measurement for machine learning models is represented using confusion matrix (Table 1). The relative operating characteristic (ROC) curve is a highly adaptable tool for analyzing the quality of dichotomous, categorical, continuous, and probabilistic forecasts. ROC is a representation of the skill of a forecasting system used to evaluate the quality of probable forecasts. It is a plot of the true positive rate (POD or Probability of Detection) on x-axis versus false positive rate (FAR or False Alarm Rate) on y-axis.

$$TruePositiveRate/POD = \frac{TP}{TP + FN} \tag{19}$$

$$FalsePositiveRate/FAR = \frac{FP}{TN + FP} \tag{20}$$

Table 1 The standard confusion matrix

Forecast	*Observed*		
	Positive	*Negative*	*Total*
Positive	True Positive (TP)	False Positive (FP)	TP + FP
Negative	False Negative (FN)	True Negative (TN)	FN + TN
Total	TP + FN	FP + TN	TP + FP + FN + TN

Here, true positives (TP) and true negatives (TN) are the correct predictions, while false negatives (FN) and false positives (FP) are the incorrect predictions.

The area under the curve (AUC), measures the entire two-dimensional area underneath the entire ROC curve. The higher the AUC is, the better is the prediction of a model. It takes values from 0 to 1, where a value of 0 indicates a perfectly inaccurate forecast and a value of 1 reflects a perfectly accurate forecast. The model is usually said to perform well when the AUC value lies between 0.8 and 0.9, and when it is greater than 0.9, the prediction ability is considered to be excellent (Metz 1978; El Khouli et al. 2009).

2.6.2 Matthews correlation coefficient (MCC)

The Matthews correlation coefficient is used in machine learning to evaluate the accuracy of binary (two-class) classifications (Matthews 1975). It is a correlation coefficient between the predicted values and the true values.

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP + FP).(TP + FN).(TN + FP).(TN + FN)}} \tag{21}$$

The MCC can be used to determine how well a classification model or function is functioning (worst value = − 1; best value = + 1).

2.6.3 Root mean square error (RMSE)

RMSE shows the gap between the observed value and the predicted value, and it is more sensitive than other measures to the occasional large error.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \tag{22}$$

The predicted and observed values of the parameters are denoted by $\hat{y}$ and y respectively, and n is the number of cases.

2.6.4 Mean absolute error (MAE)

MAE is a natural and unambiguous measure of average error, and it indicates the robustness of models.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{23}$$

2.6.5 The coefficient of determination (R^2)

R^2 statistic measures the predictive accuracy of a statistical model. It illustrates the proportion of variance in the outcome variable that is explained by the predictions.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{24}$$

where the mean value of the parameters is denoted by $\bar{y}$.

3 Results and discussion

The results of the analysis of the relevant data for the 10-year period 2011–2020 over Kolkata are provided here. As mentioned earlier, seventeen thermodynamic indices and parameters are used as input for the two models in the present study, consisting of DRCT, SKNT, THAT, THTE, THTV, CAPE, CIN, BRN, TEMP, DWPT, RELH, MIXR, SI, KI, TT, SWI and LI. The histogram plot is used to determine the range of different input variables for the pre-monsoon months of Mach, April, and May over Kolkata.The analysis reveals that SI, KI, TT, SWI and LI range from -10.4 to 62, -61.9 to 45.3, -9.8 to 169.8, 12.92 to 639, and -14.1 to 19.54 respectively (Fig. 4). The results further show that CAPE, CIN, BRN, TEMP, DWPT, RELH, MIXR are varying from 0 to 6254, 0 to -807, 0 to 5165, 11.2 to 37.4, 8 to 28.6, 44 to 98 and 6.7 to 25.57 respectively. Accordingly, from Fig. 4 the ranges of DRCT, SKNT, THAT, THTE and THTV are observed to be within 0 to 270, 0 to16, 283.1 to 311.5, 303.7 to 386.7, and 284.4 to 316.

3.1 Prediction performance and comparative study

The dataset is divided into two parts to see the accuracy of the classification results: 75% for training data and 25% for testing data. In the present study, we have used the ROC curve (Fig. 5), which is a graphical representation of the diagnostic ability of a binary classifier system, to measure the ability of forecast to discriminate between two alternative outcomes. The AUC is a numerical index that is used to evaluate the predictive ability and effectiveness of the model, as well as its quality and accuracy. As shown in Fig. 5 and Table 2, both XGBoost and Logistic Regression models have excellent prediction skills with AUCs of 0.982 and 0.976 respectively. The other error metrics are also applied similarly for each forecasting step. The K-fold cross-validation technique demonstrates robust performance for the accuracy of any machine learning model (Varma and Simon 2006; Gupta et al. 2021). In our study, tenfold cross-validation process is applied to the dataset to test the stability of the model performance. In this case, the data is randomly divided into ten sub-samples, and ten models are trained in such a way that each time nine sub-samples are used to train a model and one subsample is used to test a model. The accuracy of both the XGBoost and Logistic Regression models in the form of a line graph for short-term predictions of thunderstorms is analysed which depicts the consistent performances for both the models (Fig. 6). However, the accuracy is close to 97% for XGBoost, whereas it is close to 95% in the case of Logistic Regression. Table 2 depicts the values of other performance indices using XGBoost and Logistic Regression models for the prediction of thunderstorms. The values of MAE, RMSE, R^2 and MCC for the prediction of pre-monsoon thunderstorm are 0.030, 0.173, 0.885, 0.954 respectively using XGBoost model (Table 2). The same values for MAE, RMSE, R^2 and MCC using logistic regression are 0.050, 0.223,0.791 and 0.899 respectively.

The above results reveal that both the XGBoost and Logistic Regression models perform well in the prediction of pre-monsoon thunderstorms over Kolkata, although the former has an edge over the latter. This may be due to the fact that Logistic Regression is based on some assumptions, such as the linear relationship between input and output and/or multicollinearity between the features. But the dataset of real-life problems such as thunderstorms, which is complex, does not meet all the assumptions of Logistic Regression.

As a comparison of our investigation using XGBoost and Logistic Regression models for thunderstorm prediction over Kolkata, one may consider two previous studies, namely, based on ANN model by Basak et al. (2012) and multivariate statistical analysis by Chatterjee et al. (2009). The overall percentage of correct prediction by ANN model was around 68–72% in the morning and around 60% in the afternoon. Using a multivariate statistical analysis of 20 different thermodynamic and dynamic parameters to predict the convective development at Kolkata, the correct prediction was around 72.7% in the morning and 56.6% in the evening. The results obtained by our models (e.g., about 97% for XGBoost and 95% for Logistic Regression) are therefore better.

3.2 Importance of predictors using SHAP

SHapley Additive exPlanations (SHAP) is a method that uses Shapley values as a measure of feature importance, based on game-theory. Possibly, this is the only locally accurate and globally consistent feature attribution method (Lundberg and Lee 2017a, b). Using SHAP, the feature contributions to each state can be evaluated and visualized. Now our aim is to measure the global importance of characteristics using Eq. 4. Based on the matrix of SHAP values, the absolute SHAP values per characteristic all through the data were calculated. The

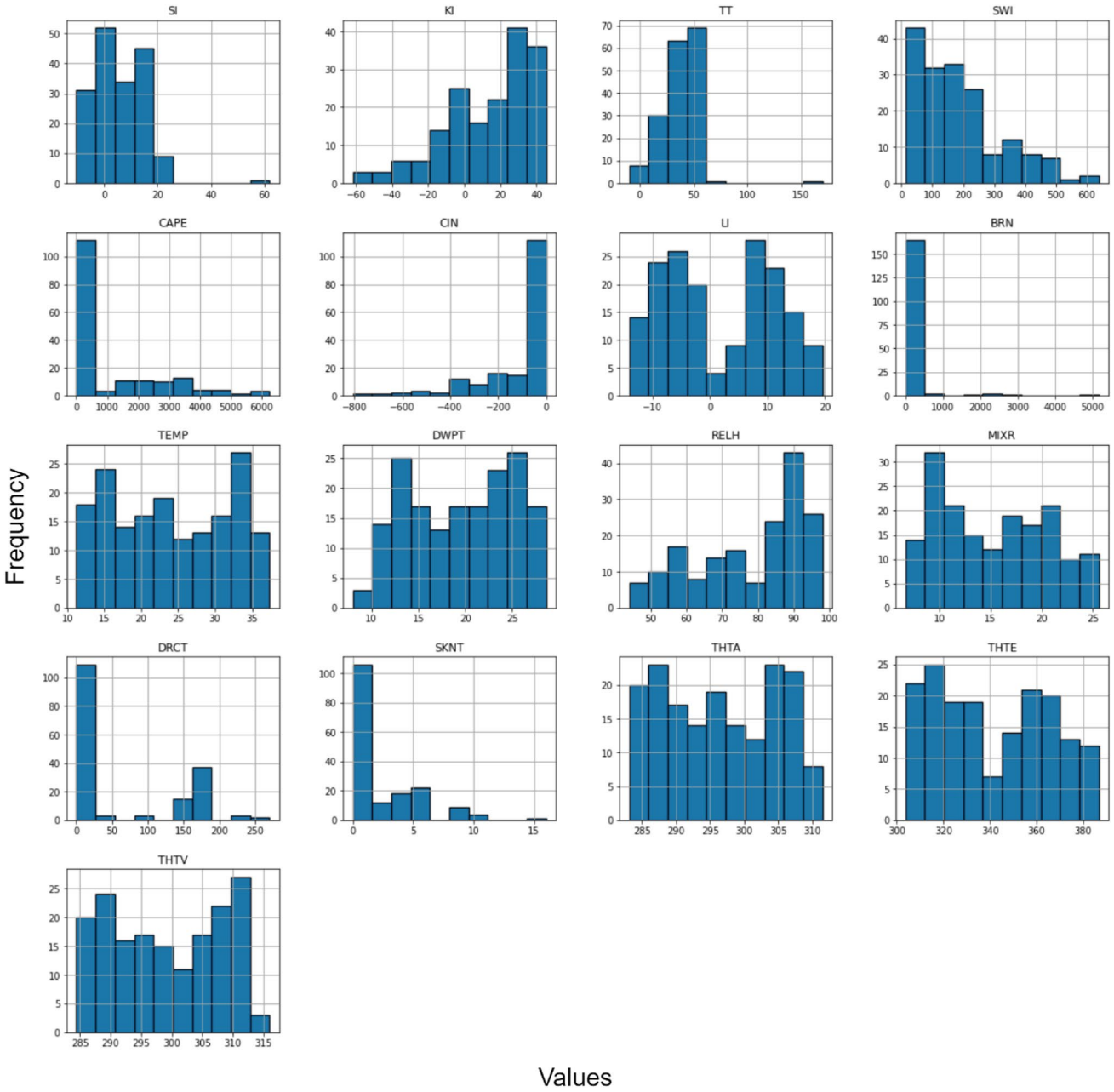

Fig. 4 The histogram plots showing the ranges of different thermodynamic indices and parameters during the study period

features are sorted in decreasing importance and plotted in Fig. 7. The top features contribute more to the model than the bottom ones and hence have a greater impact on the occurrence and non-occurrence of thunderstorms. The ranking of features for the models XGBoost and Logistic Regression is seen to be CAPE > BRN > LI and CAPE > CIN > BRN > SWI respectively, where the symbol > denotes superior ranking.

This indicates that convective available potential energy (CAPE) has the highest impact on both the models, XGBoost and Logistic Regression, whereas the ranks of the other features are different. But, the representation in Fig. 7 does not include positive or negative influence of the features.

Given that SHAP-feature importance only contains the absolute value of feature contributions, feature attributions like Shapley values can be visualized as "forces". Each feature value is a force that either increases or decreases the prediction. The prediction process begins with the baseline. SHAP-value is able to quantify the contribution (positive or negative) of each predictor variable to prediction based on the average training output. i.e., how much it pushes the model output from the base value (the average

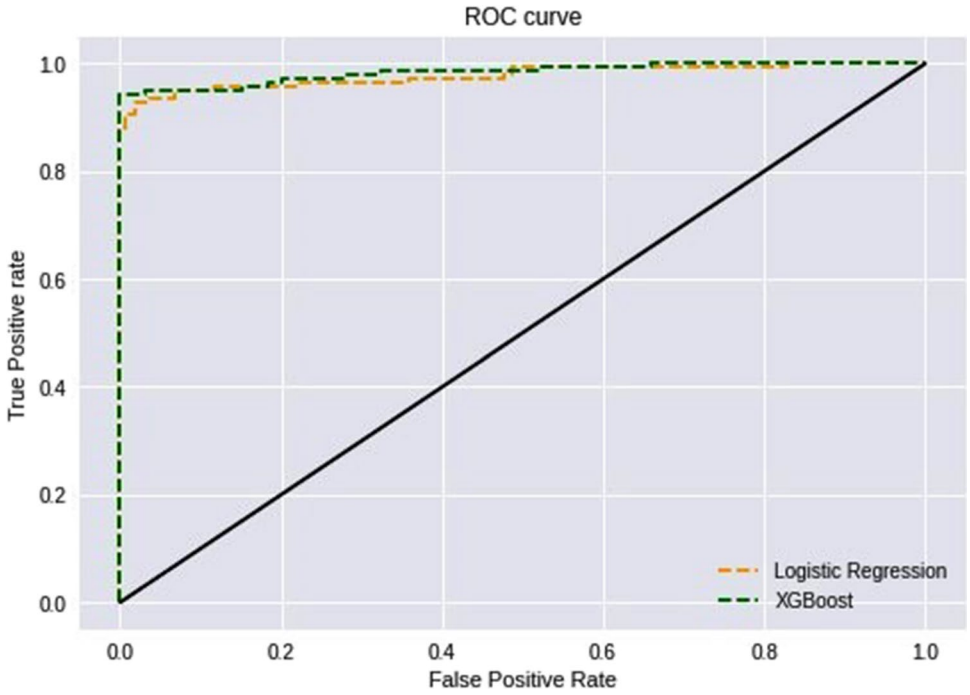

Fig. 5 The diagram showing the receiver operator characteristic (ROC) curve of XGBoost and Logistic Regression

training output). Figure 8 shows the force plot that illustrates that the features have a positive impact pushing the prediction higher (indicated in red), while those have a negative impact pushing the prediction lower (indicated in blue). The length of the arrow represents the SHAP-value of each feature. The longer the arrow, the higher the impact. The prediction of XGBoost is seen to be 1.00 (Fig. 8a), which differs from the base value (0.5117). Here, CAPE, BRN and LI are the powerful positive forces to drive the forecast accuracy up. Similarly, for Logistic Regression, the base value is 60.07 and the predicted value is 95.64 (Fig. 8b). Here, the SHAP-value for CAPE is positive which means boosting the forecast accuracy higher, whereas the SHAP-values for BRN and CIN are negative which slows down the prediction to a certain extent.

Table 2 Performance of XGBoost and Logistic Regression for the prediction of pre-monsoon thunderstorms over Kolkata

Accuracy Measure	*Model*	
	XGBoost	Logistic Regression
MAE	0.030	0.050
RMSE	0.173	0.223
R^2	0.885	0.791
AUC	0.982	0.976
MCC	0.954	0.899

The summary plot combines the feature importance with the effects of features on the prediction. Here, the features are ranked in descending order. The SHAP-value for a feature or an instance is represented by each point on the summary plot. The position on the y-axis is indicated by the feature and on the x-axis by the SHAP-value. For a particular observation, the color denotes the value of the feature, which ranges from low (represented in blue) to high (represented in red). Further, the horizontal location indicates whether the effect of that value is associated with a higher or lower prediction. More specifically, if the SHAP-value increases with the increase of the corresponding feature value, this characteristic has a positive impact on the occurrences of thunderstorm events.

When comparing Fig. 9a and b, the XGBoost and Logistic Regression models appear to be similar in terms of quantifying the importance of predictor variables, but the ranks of their contributions differ. The dominant higher positive contribution of convective available potential energy (CAPE) is confirmed by both XGBoost and Logistic Regression. In a part of the investigation, for more emphasis on our findings we have tested the experimental data with another tree-based machine learning algorithm, namely, Random Forest (RF). The results show that CAPE is the only variable that impacts the decision of the model. Furthermore, CAPE has the dominant higher positive contribution which confirms that the interpretation of most dominant variable chosen by the ML models for the prediction of pre-monsoon thunderstorms over Kolkata remains the same.

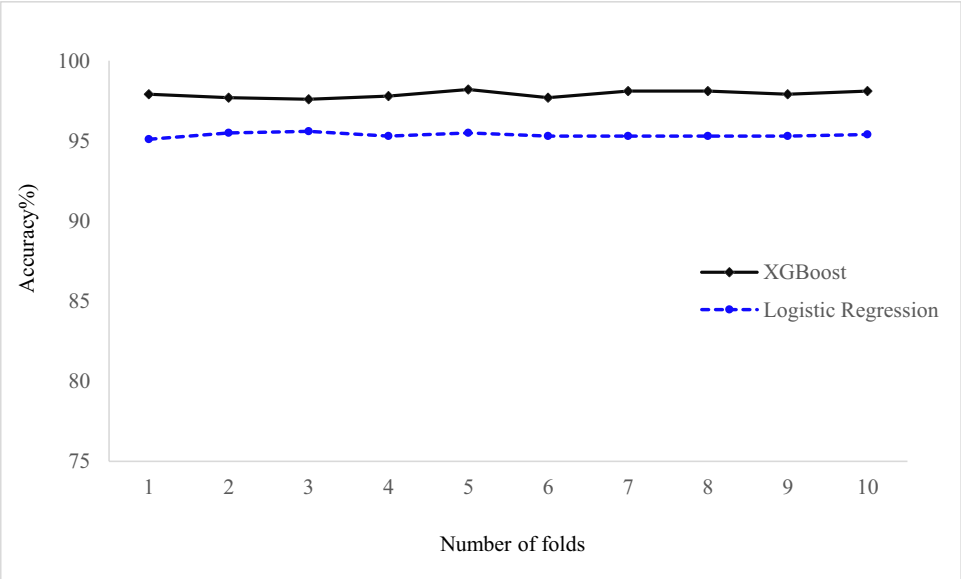

Fig. 6 Results of K-fold cross-validation

Identifying the most important variables using SHapley Additive exPlanations can aid in determining the model's validity based on domain expertise. CAPE, a measure of the amount of energy available for convection, plays an important role in meso-scale convective systems. More precisely, it describes atmospheric instability and provides an estimate of updraft strength within a thunderstorm. Mathematically it is represented as

$$CAPE = \int_{LFC}^{LNB} g\left(\frac{T_{v,parcel} - T_{v,env}}{T_{v,env}}\right)dz \tag{25}$$

Where

$T_{v,parcel}$ = virtual temperature of the air parcel

$T_{v,env}$ = virtual temperature of the environment.

g = acceleration by gravity

The integration normally starts from the level of free convection (LFC) which is the level above which the lifted parcel is warmer than the environment. The upper limit for the integration is the level of neutral buoyancy (LNB) in which the parcel and environment curves meet again. CAPE is efficiently characterizing severe weather events such as thunderstorms, tornedos, hailstorms (Blanchard 1998; Brooks and Dotzek 2007; Ukkonen et al. 2017; Lin and Kumjian 2022). A higher CAPE value indicates that the atmosphere is more unstable, resulting in a stronger updraft. Large CAPE is generally found to be associated with severe thunderstorms. Thus, CAPE is the most important parameter for thunderstorm dynamics. Raman and Raghvan (1961), Rao et al. (1971), Koteshwaram and Srinivasan (1958) and Krishna Rao (1966) have analysed the role of conditional instability (CAPE) in the outbreak of severe thunderstorms. Basu and Mondal (2002) showed a forecasting aspect for propagation speed of thunderstorm cell in pre-monsoon season at Kolkata examined with respect to CAPE. Bondyopadhyay et al. (2021) found that thunderstorm days have higher mean values for the indices CAPE compared to non-thunderstorm days over Kolkata.

The Bulk Richardson number (BRN) is a dimensionless number that relates vertical stability and vertical shear (more precisely stability divided by shear). This indicates the ratio of thermally produced turbulence and turbulence generated by vertical shear (Das 2017). For mesoscale forecasting purposes it relates buoyancy through CAPE to vertical wind shear for a 5.5 km thickness and is simply defined as (Das 2017):

$$BRN = CAPE/(0.5 * (u6km - u500m)^2) \tag{26}$$

where, $u6km$ is the wind speed at 6 km above ground level (AGL) and $u500m$ is the wind speed at 500 m AGL. The values within the range 10–45 indicate favourable environmental conditions for supercell development. Bulk Richardson Number (BRN) ranks second in XGBoost model and ranks third in Logistic Regression model for the prediction of pre-monsoon thunderstorm events over Kolkata. It shows more positive influences on thunderstorm occurrences over Kolkata (Fig. 9a and b).

Figure 9 indicates that LI (XGBoost model) and CIN (Logistic Regression model) have higher negative influences

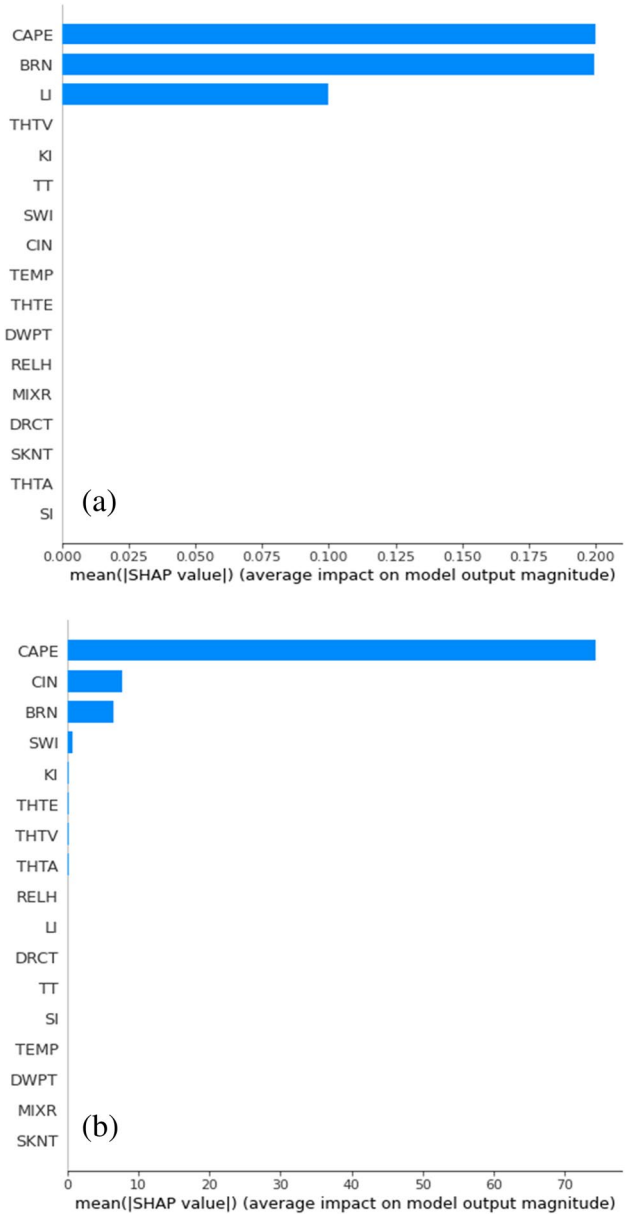

Fig. 7 Ranking of features importance (mean SHAP value) for final outcome of the model (**a**) XGBoost and (**b**) Logistic Regression

on the occurrence of thunderstorm. CIN is a measure of the energy required by the atmosphere, at a specific time and location, to prevent an air parcel to rise (because of convection). CIN behaves as a possible barrier to the generation of convection, even when the CAPE-value is high. High values of CAPE do not necessarily lead to strong convection (Riemann-Campe et al. 2009), as the simulated air parcel needs to overcome a usually stable layer between the surface free convection (SFC) and level of free convection (LFC). The following is a mathematical expression for CIN

$$CIN = -\int_{SFC}^{LFC} g\left(\frac{T_{v,parcel} - T_{v,env}}{T_{v,env}}\right)dz \tag{27}$$

It is important to note that XGBoost does not consider CIN as the negative influencing feature for the occurrences of thunderstorms over Kolkata.

The lifted index (LI) is the difference between the observed temperature at 500 hPa and the temperature of an air parcel lifted to 500 hPa from near the surface (Sahu et al. 2020; Bondyopadhyay et al. 2021). It may be expressed as

$$LI = T_{500} - T_{parcel,500} \tag{28}$$

where LI (°C) is the lifted index, T_{500} is the 500 hPa environmental temperature (°C) and $T_{parcel,500}$ is the 500 hPa temperature (°C) which a parcel will acquire if it is lifted dry adiabatically from the surface to its lifting condensation level (LCL) and then moist adiabatically to 500 hPa pressure level. When a parcel of air is warmer than the surroundings, it rises freely. When a parcel is lifted, it acquires an upward vertical velocity, which can be caused by a surface front or trough, orographic features, upper-level short waves, or convection. When the value is positive, the atmosphere (at the respective height) is stable and negative value LI indicates that the boundary layer is unstable with respect to the middle troposphere (Galway 1956). This instability represents an environment in which convection can occur. The more is the negative value, the more unstable the air is, and the stronger is the updrafts likely to be with any developing thunderstorms. LI proves its usefulness as a dichotomous predictor of a thunderstorm (Kunz 2007; Kunz et al. 2009; Sahu et al. 2020). Pradhan et al. (2012) developed nowcasting technique and convective indices like CAPE, CIN, LI and BRN had been evaluated and analyzed statistically for thunderstorm prediction in Gangetic West Bengal (India) using Doppler Weather Radar and upper air data. The validity of these convective indices has been checked with 34 occurrences of thunderstorms during 2006–2007 recorded by Doppler Weather Radar Kolkata. UmaKanth et al. (2019) studied climatological aspects of the convective systems over five major cities of West Bengal region including Kolkata with thermodynamic atmospheric stability indices where it is found that convective available potential energy (CAPE) helped us to study the convective systems both seasonally and decadal.

Figure 9b illustrates that severe weather threat index (SWI) has little positive impact on model decision (Logistic Regression) for prediction of pre-monsoon thunderstorm over Kolkata. The SWI (Miller 1972; Das 2017; Rabbani et al. 2020) determines the potential for severe weather by combining several parameters into one index. These parameters include low-level moisture (850 mb dewpoint), instability (total totals index), lower and middle-level (850

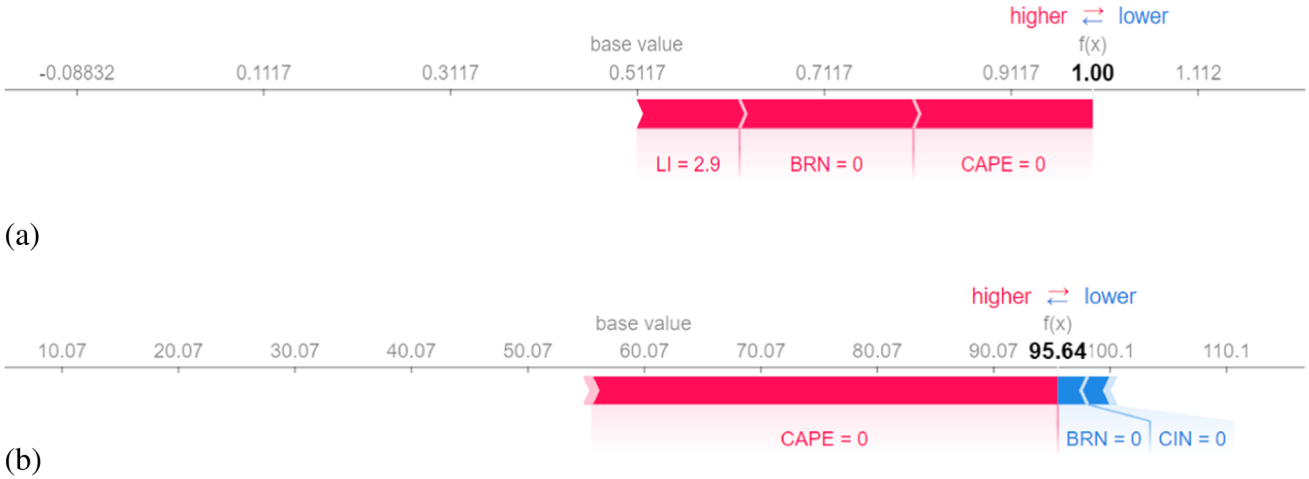

Fig. 8 SHapley Additive exPlanations (SHAP) explanation force plots for prediction of thunderstorms over Kolkata using (**a**) XGBoost and (**b**) Logistic Regression (April 17, 2018 at 18:55 IST)

and 500 mb) wind speeds, and warm air advection (veering between 850 and 500 mb). Therefore, an attempt is made to incorporate kinematic and thermodynamic information into one index. As a consequence, rather than assessing ordinary thunderstorm potential, the SWI should be used to estimate severe thunderstorm potential. The following is a mathematical expression for SWI.

$$SWI = 12[T_d(850mb)] + 20(TT - 49) + 2(f8) + f5 + 125(S + 0.2) \quad (29)$$

Where TT indicates the total totals index value, $f8$ and $f5$ represent 850mb and 500mb wind speed in kont respectively and $s = sin\,(500mb\ wind\ direction - 850mb\ wind\ direction)$, i.e., the sine of the angle between the 500mb and 850mb wind directions (the shear term). This index takes into account of other thunderstorm development factors such as low-level moisture and instability; however, it also incorporates some other severe thunderstorm parameters.

From the performance analysis it is clear that both the models perform well for the prediction of pre-monsoon thunderstorm over Kolkata, but the variables are used in different ways by XGBoost and LR. This means the interpretation of these two models are different. The experimental results are validated with the knowledge based on domain expertise.

4 Conclusions

In recent years, an increase in the number of thunderstorm initiation events have occurred as a result of urbanization than would have occurred over natural vegetation (Haberlie et al. 2015). Every year, many people die as a result of lightning strikes, damaging winds, and flash floods caused by strong rainfall associated with nor'westers. Despite technological advancements, lightning strikes continue to be the most common cause of mortality caused by natural forces in India. Precise forecasting of cumulonimbus cloud events or thunderstorms helps to mitigate the risk of disasters and economic losses. Machine learning helps create advanced analytics that are much easier to understand than traditional physical models. They use fewer resources and may be installed on almost any computer, even mobile devices.

In this paper, prediction and assessment of occurrences of pre-monsoon thunderstorms over Kolkata have been made using thermodynamic indices and parameters, and user-friendly computational models. Finally, interpretation of decisions of the models is done using explainable artificial intelligence (XAI). The investigation emphasizes a comparative study between two types of machine learning approaches, viz., XGBoost and Logistic Regression for short-term predictions of pre-monsoon cumulonimbus cloud event or thunderstorm. Both XGBoost and Logistic Regression predictive models perform well for thunderstorm prediction, though the former has an edge over the latter. We have applied SHAP, a Shapley-value-based explanation method using the coalitional game theory, to evaluate the contribution of each predictor variable to forecast short-term occurrences of pre-monsoon thunderstorms. We have quantified how the most important variables, as identified by SHAP- value, can help in explaining the model's validity based on domain knowledge. The ML-based interpretability using SHAP, as described here, is unique. The results reveal that CAPE (convective available potential energy) is the most important parameter for pre-monsoon thunderstorms over Kolkata, as assessed by both the models. The other predictive variables have different importance in ranking and interpretations. Indices CAPE, CIN, LI, BRN and SWI

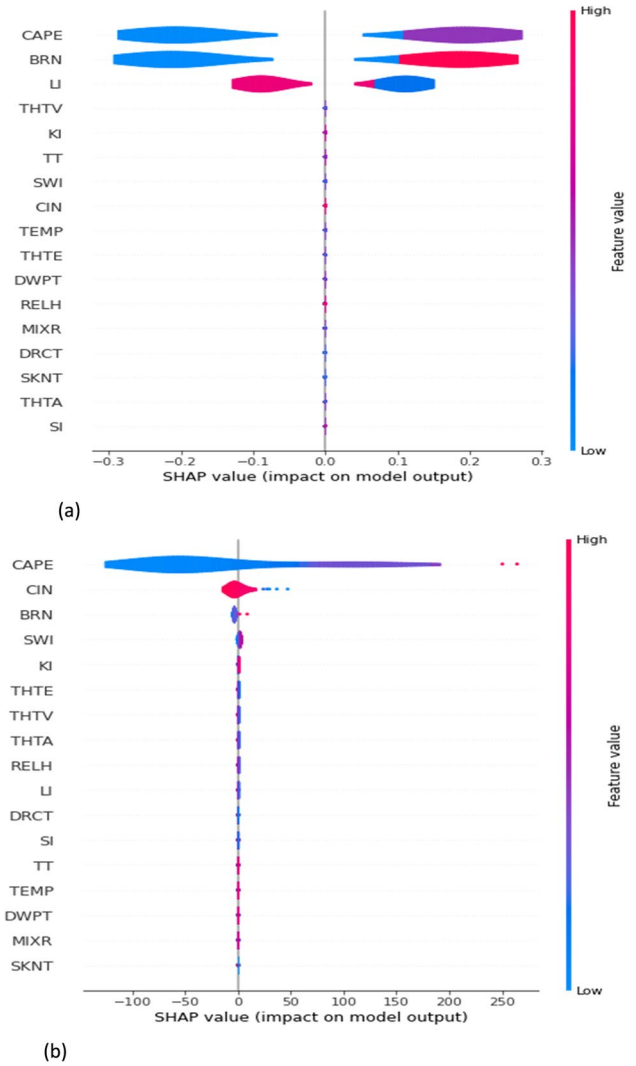

Fig. 9 SHAP summery plot showing the impact of predictor variables for the prediction of thunderstorms using (**a**) XGBoost and (**b**) Logistic Regression

truly reflect the decisions of the two models in prediction of thunderstorms.

Traditional importance measures can only represent the relative importance of predictor variables based on their predictive strength, but they cannot estimate the contribution of each predictor variable to the model output. Furthermore, different models may not all use the same approach, making the task of model comparison challenging in regards of interpretability. We have demonstrated that SHAP is capable of both detecting and measuring the contribution of predictor variables, as well as being a consistent method for the interpretation of various machine learning models. In addition, the ordering of variables according to their relevance aids in selecting in which sequence one should do further model investigation. In the present study, the two models perform equally good for the prediction of thunderstorms. However, due to the diverse natures of the models, they may prefer to have different predictor variables, resulting in different explainability. The findings of the present study confirm the validity of the SHAP technique, enables better understanding of the ML models, and improves the model interpretability. These characteristics may be used in various data driven scientific research in the future.

Acknowledgements The work was done when Prof. S.K. Pal held a National Science Chair, SERB-DST, Govt. of India.

Author contribution DD formulated the research problem, wrote the programs and made the first draft. SKP is the mentor and Principal Investigator who gave the guidance in machine learning aspects of the paper and provided corrections of the overall manuscript for better organization and understanding.

Funding SERB-National Science Chair, Govt. of India awarded to Prof. Sankar K. Pal.

Data availability The corresponding author will provide data supporting the findings of this study upon reasonable request.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Yes.

Consent to participate Not applicable.

Competing interests The authors declare that they have no competing interests.

References

Bondyopadhyay S, Mohapatra M, Sen Roy S (2021) Determination of suitable thermodynamic indices and prediction of thunderstorm events for Kolkata, India. Meteorol Atmos Phys 133:1367–1377. https://doi.org/10.1007/s00703-021-00813-1

Basak P, Sarkar D, Mukhopadhyay AK (2012) Estimation of thunderstorm days from the radio-sonde observations at Kolkata (22.53oN, 88.33oE), India during pre-monsoon season: an ANN based approach earth science India. Earth Sci 5(4):139–151

Basu GC, Mondal DK (2002) A forecasting aspect of thunder squall over Calcutta and its parameterization during pre-monsoon season. Mausam 53(3):271–280

Blanchard DO (1998) Assessing the vertical distribution of convective available potential energy. Wea Forecasting 13:870–877. https://doi.org/10.1175/1520-0434(1998)013%3c0870:atvdoc%3e2.0.co;2

Brooks HE, Dotzek N (2007) The spatial distribution of severe convective storms and an analysis of their secular changes. Clim Extremes Soc 35–53. https://doi.org/10.1017/cbo9780511535840.006

Chang W, Liu Y, Xiao Y et al (2019) A machine-learning-based prediction method for hypertension outcomes based on medical data. Diagnostics 9:178. https://doi.org/10.3390/diagnostics9040178

Chatterjee S, Ghosh S, De SK (2009) Reduction of number of parameters and forecasting convective developments at Kolkata (22.35 N, 88.33E), India during pre-monsoon season: an application of multivariate technique. Indian J Radio Space Phys 38:275–282

Chen T, Guestrin C (2016) Xgboost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939785

Collins WG, Tissot P (2016) Thunderstorm predictions using artificial neural networks. Artif Neural Netw - Model Appl. https://doi.org/10.5772/63542

Das S (2017) Severe Thunderstorm observation and modeling – AReview. VayuMandal 43:1–29

Dasgupta S, De UK (2007) Binary logistic regression models for short term prediction of premonsoon convective developments over Kolkata (India). Int J Climatol 27:831–836. https://doi.org/10.1002/joc.1449

El Khouli RH, Macura KJ, Barker PB et al (2009) Relationship of temporal resolution to diagnostic performance for dynamic contrast enhanced MRI of the breast. J Magn Reson Imaging 30:999–1004. https://doi.org/10.1002/jmri.21947

Galway JG (1956) The lifted index as a predictor of latent instability. Bull Am Meteorol Soc 37:528–529. https://doi.org/10.1175/1520-0477-37.10.528

Gupta VK, Gupta A, Kumar D, Sardana A (2021) Prediction of Covid-19 confirmed, death, and Cured cases in India using random forest model. Big Data Min Anal 4:116–123. https://doi.org/10.26599/bdma.2020.9020016

Haberlie AM, Ashley WS, Pingel TJ (2015) The effect of urbanisation on the climatology of thunderstorm initiation. Q J R Meteorol Soc 141:663–675. https://doi.org/10.1002/qj.2499

Ibrahem Ahmed Osman A, Najah Ahmed A, Chow MF, Feng Huang Y, El-Shafie A (2021) Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Eng J 12:1545–1556. https://doi.org/10.1016/j.asej.2020.11.011

Jafari Goldarag Y, Mohammadzadeh A, Ardakani AS (2016) Fire risk assessment using neural network and logistic regression. J Ind Soc Remote Sens 44:885–894. https://doi.org/10.1007/s12524-016-0557-6

Jergensen GE, McGovern A, Lagerquist R, Smith T (2019) Classifying convective storms using machine learning. Wea Forecasting 35:537–559. https://doi.org/10.1175/waf-d-19-0170.1

Jiang M, Chen W, Li X (2021) S-GCN-GRU-NN: A novel hybrid model by combining a Spatiotemporal Graph Convolutional Network and a Gated Recurrent Units Neural Network for short-term traffic speed forecasting. J Data Inf Manag 3:1–20. https://doi.org/10.1007/s42488-020-00037-9

Kamangir H, Collins W, Tissot P, King SA (2020) A deep-learning model to predict thunderstorms within 400 km 2 south Texas domains. Meteorol Appl. https://doi.org/10.1002/met.1905

Koteshwaram MP, Srinivasan V (1958) Thunderstorms over Gangetic West Bengal in the pre-monsoon season and the synoptic factors favourable for their formation. Indian J Meteorol Geophys 9:301–312

Krishna Rao PR (1966) Thunderstorm studies in India- A review. Indian J Meteorol Geophys 12:3–13

Kunz M (2007) The skill of convective parameters and indices to predict isolated and severe thunderstorms. Nat Hazards Earth Syst Sci 7(2):327–342

Kunz M, Sander J, Kottmeier C (2009) Recent trends of thunderstorm and hailstorm frequency and their relation to atmospheric characteristics in southwest Germany. Int J Climatol 29(15):2283–2297. https://doi.org/10.1002/joc.1865

Lee J-G, Min K-H, Park H, Kim Y, Chung C-Y, Chang E-C (2020) Improvement of the rapid-development thunderstorm (RDT) algorithm for use with the GK2A satellite. Asia Pac J Atmos Sci 56:307–319. https://doi.org/10.1007/s13143-020-00182-6

Leinonen J, Hamann U, Germann U, Mecikalski JR (2022) Nowcasting thunderstorm hazards using machine learning: the impact of data sources on performance. Nat Hazard 22:577–597. https://doi.org/10.5194/nhess-22-577-2022

Lin Y, Kumjian MR (2022) Influences of CAPE on hail production in simulated supercell storms. J Atmos Sci 79:179–204. https://doi.org/10.1175/jas-d-21-0054.1

Litta AJ, Mary Idicula S, Mohanty UC (2013) Artificial neural network model in prediction of Meteorological parameters During Premonsoon thunderstorms. Int J Atmos Sci 2013:1–14. https://doi.org/10.1155/2013/525383

Lohar D, Pal B (1995) The effect of irrigation on Premonsoon season precipitation over South West Bengal, India. J Clim 8:2567–2570. https://doi.org/10.1175/1520-0442(1995)008%3c2567:teoiop%3e2.0.co;2

Lundberg SM, Lee S (2017a) A Unified Approach to Interpreting Model Predictions. Adv Neural Inf Process Syst, pp 4765–4774

Lundberg SM, Lee S-I (2017b) Consistent feature attribution for tree ensembles. arXiv:1706.06060

Maddalena L, Gori M, Pal SK (2020) Pattern recognition and beyond: Alfredo Petrosino's scientific results. Pattern Recognit Lett 138:659–669. https://doi.org/10.1016/j.patrec.2020.07.032

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Struct 405:442–451. https://doi.org/10.1016/0005-2795(75)90109-9

Metz CE (1978) Basic principles of ROC analysis. Semin Nucl Med 8:283–298. https://doi.org/10.1016/s0001-2998(78)80014-2

McCann DW (1992) A neural network short-term forecast of significant thunderstorms. Wea Forecasting 7(3):525–534. https://doi.org/10.1175/1520-0434(1992)007%3c0525:annstf%3e2.0.co;2

McCullagh P, Nelder JA (1989) An outline of generalized linear models. Generalized Linear Models 21–47. https://doi.org/10.1007/978-1-4899-3242-6_2

McGovern A, Elmore KL, Gagne DJ et al (2017) Using artificial intelligence to improve real-time decision-making for high-impact weather. Bull Am Meteorol Soc 98:2073–2090. https://doi.org/10.1175/bams-d-16-0123.1

Miller RC (1972) Notes on the analysis of severe storm forecasting procedures of the air force global weather center AFGWC Tech Rep 200. Air Weather Service, Scott AFB, IL

Mostajabi A, Finney DL, Rubinstein M, Rachidi F (2019) Nowcasting lightning occurrence from commonly available Meteorological parameters using machine learning techniques. npj Clim Atmos Sci. https://doi.org/10.1038/s41612-019-0098-0

Mostajabi A, Finney D, Rubinstein M, Rachidi F (2020) Nowcasting lightning occurrence using machine learning Techniques: The challenge of identifying outliers. EGU General Assembly 2020. https://doi.org/10.5194/egusphere-egu2020-22302

Nie F, Hu Z, Li X (2018) An investigation for loss functions widely used in machine learning. Commun Inf Syst 18(1):37–52. https://doi.org/10.4310/cis.2018.v18.n1.a2

Pal A, Pal SK (2017) Pattern recognition: evolution, mining and big data. In: Pal A, Pal SK (eds) Pattern recognition and big data. World Scientific, Singapore, pp 1–36

Pal SK, Meher SK, Skowron A (2015) Data science, big data and granular mining. Pattern Recognit Lett 67:109–112. https://doi.org/10.1016/j.patrec.2015.08.001

Pathak A, Pal SK (1986) Fuzzy grammars in syntactic recognition of skeletal maturity from X-rays. IEEE Trans Syst Man Cybern 16:657–667. https://doi.org/10.1109/tsmc.1986.289310

Pradhan D, DE UK, Singh UV (2012) Development of nowcasting technique and evaluation of convective indices for thunderstorm prediction in Gangetic West Bengal (India) using Doppler Weather Radar and upper air data. Mausam 63(2):299–318

Rabbani G, Kardani-Yazd N, Mansouri Daneshvar MR (2020) Factors affecting severe weather threat index in urban areas of Turkey and Iran. Environ Syst Res. https://doi.org/10.1186/s40068-020-00173-6

Raman, Raghvan K (1961) Diurnal variations of thunderstorms in India during different seasons. Indian J Meteorol Geophys 12:115–124

Rao KN, Daniel CEJ, Balasubramanian LV (1971) Thunderstorms over India, IMD published Scientific Report No. 153

Rasp S, Lerch S (2018) Neural networks for postprocessing Ensemble weather forecasts. Mon Weather Rev 146:3885–3900. https://doi.org/10.1175/mwr-d-18-0187.1

Riemann-Campe K, Fraedrich K, Lunkeit F (2009) Global climatology of Convective available potential Energy (cape) and convective Inhibition (cin) In ERA-40 reanalysis. Atmos Res 93:534–545. https://doi.org/10.1016/j.atmosres.2008.09.037

Sahu RK, Dadich J, Tyagi B et al (2020) Evaluating the impact of climate change in threshold values of thermodynamic indices during pre-monsoon thunderstorm season over Eastern India. Nat Hazards 102:1541–1569. https://doi.org/10.1007/s11069-020-03978-x

Schultz MG, Betancourt C, Gong B, Kleinert F, Langguth M, Leufen LH, Mozaffari A, Stadtler S (2021) Can deep learning beat numerical weather prediction? Philos Trans A Math Phys Eng Sci 379:20200097. https://doi.org/10.1098/rsta.2020.0097

Sen D, Pal SK (2010) Gradient histogram: Thresholding in a region of interest for edge detection. Image Vis Comput 28:677–695. https://doi.org/10.1016/j.imavis.2009.10.010

Sánchez JL, Marcos JL, de la Fuente MT, Castro A (1998) A logistic regression model applied to short term forecast of hail risk. Phys Chem Earth 23:645–648. https://doi.org/10.1016/s0079-1946(98)00102-5

Shapley LS (1953) A value For N-PERSON GAMES. Contributions to the Theory of Games (AM-28), Volume II 307–318. https://doi.org/10.1515/9781400881970-018

Shrestha Y, Zhang YR, Doviak R, Chan PW (2021) Lightning flash rate nowcasting based on polarimetric radar data and machine learning. Int J Remote Sens 42:6762–6780. https://doi.org/10.1080/01431161.2021.1933243

Smita M (2021) logistic regression model –A REVIEW. Int J Innov Sci Res Technol 6(5):1276–1280

Stankova E, Tokareva IO, Dyachenko NV (2021) On the possibility of using neural networks for the thunderstorm forecasting. Computational Science and Its Applications – ICCSA 2021. Springer International Publishing, Cham, pp 350–359

STORM (Severe Thunderstorms—Observations and Regional Modeling) Programme (2005) Science plan. Department of Science and Technology, Government of India

Strumbelj E, Kononenko I (2010) An efficient explanation of individual classifications using game theory. J Mach Learn Res 11:1–18. https://doi.org/10.1145/1756006.1756007

Tyagi B, Naresh Krishna V, Satyanarayana ANV (2011) Study of thermodynamic indices in forecasting pre-monsoon thunderstorms over Kolkata during STORM pilot phase 2006–2008. Nat Hazards (dordr) 56:681–698. https://doi.org/10.1007/s11069-010-9582-x

Ukkonen P, Manzato A, Mäkelä A (2017) Evaluation of Thunderstorm predictors for Finland Using Reanalyses and neural networks. J Appl Meteorol Climatol 56:2335–2352. https://doi.org/10.1175/jamc-d-16-0361.1

UmaKanth N, Satyanarayana GC, Simon B, Rao MC (2019) Some climatological aspects of convective systems at five major Cities of West Bengal, India. Earth Sci India 12(2):105–116

Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7(1):91. https://doi.org/10.1186/1471-2105-7-91

Zhou K, Zheng Y, Li B, Dong W, Zhang X (2019) Forecasting different types of convective weather: A deep learning approach. J Meteorol Res 33:797–809. https://doi.org/10.1007/s13351-019-8162-6

Zhou K, Zheng Y, Dong W, Wang T (2020) A deep learning network for cloud-to-ground lightning nowcasting with multisource data. J Atmos Ocean Technol 37:927–942. https://doi.org/10.1175/jtech-d-19-0146.1