scientific reports

Check for updates

OPEN

A comparative study of severe thunderstorm among statistical and ANN methodologies

Sonia Bhattacharya[1]✉ & Himadri Chakraborty Bhattacharyya [2]

Severe Thunderstorms are the extreme weather convective features. It causes local calamities in various ways. Proper prediction with lead time is an important factor to prevent such calamities from saving people. Here, both probabilistic and machine learning techniques are applied to weather data to obtain proper predictions. Traditional methodologies are already available for such prediction purposes. However, Naïve Bayes and RBFN (Radial Basis Function Network) methodology have been introduced here with some specific weather parameters that has not done before remarkably. A comparative study was performed on weather data including Naïve Bayes, Multilayer Perceptron (MLP), K-nearest neighbor (KNN) and Radial Basis Function Network (RBFN). All these data have been procured from Kolkata located in north-east India. The result obtained by applying the Radial Basis Function Network is better among the three methods, yielding a correct prediction of 95% for severe "squall-storms" and 94% for "no storm". The predictions have a sufficient lead time of 10- 12 h.

Thunderstorms are mesoscale convective processes that cause some extreme weather events. That may include suffering, heavy rainfall, hail and gusty winds[1]. Wind speed of at least 45 km/h having minimum duration of 1 min is called 'squall'. Generally, a thunder-squall can persist for a maximum of one minute with a spatial extent of 100 kms [2]. Since the thunderstorm is an extreme convective weather event, proper prediction is needed to alert the people who reside within the devastating region (100 km)[3]. Thunderstorms occur every year in pre monsoon season over North east India. The days that have a record of 'squall' wind has considered here as 'thunderstorm days'. Similarly, the days that have no such records of 'squall' wind has considered here as 'no thunderstorm' days[4,5]. The main aim of this study is to forecast severe thunderstorm with an enough lead time by comparing among various ANN (Artificial Neural Network) methodologies i.e. Multilayer Perceptron (MLP), K-Nearest Neighbor (KNN) Method, Radial Basis Function Network (RBFN) and statistical methodology i.e. e. Naive Bayes Method. The life cycle of a thunderstorm has three stages: cumulus (updraft persisting throughout the cell), mature (presence of both the updraft and the downdraft), and dissipating (manifested only by the downdraft all through the cell)[6]. It is the towering cumulus or the cumulonimbus clouds of the convective origin and high vertical extent that are capable of producing lightning and thunder. The study revealed that every cumulus tower is sheared at a much lower rate than if it drifted with the wind[7]. Outlying cumuli are frequently torn as under when subjected to strong vertical shear[8]. Thunderstorms can be categorized as single cells, multicells, squall lines and supercells [9].Byers and Batton (1949) performed a study with the help of radar data. The simulation of mesoscale model is helpful to justify the physics and dynamics of the severe thunderstorms[3,10]. Mathematical physics establishes that primary value problem is mainly to forecast the state of the atmosphere, whereas future weather is predicted by integrating the governing partial differential equations, starting from the observed current weather[11]. Most weather prediction systems use a combination of empirical and dynamical techniques[12]. Thunderstorms forecasting is a complicated task in weather prediction. The reason behind this is the small spatial and temporal extension of thunderstorms and the inherent nonlinearity of their dynamics and physics[13].The success and failure of predictions is exactly known and pathways to obtain better predictive skill can be efficiently tested[14,15]. Parameterizations play an important part in forecasting skill since they determine the main features of the simulated weather, such as clouds and precipitation[11]. Forecast consistency is determined by contrasting forecast circulations with the observed constancy of occurrence[16]. Convective Available Potential Energy (CAPE) value indicates the presence of updrafts, more the value more the possibility of severe thunderstorm. Convective Inhibition (CINE) is the energy that needs to be overcome in order for convection to occur[17]. The roles of CINE and CAPE have been studied for the forecasting purpose by many researchers[18].

[1]State Aided College Teacher, Department of Computer Science, Panihati Mahavidyalaya Barasat Road, Sodepur, Kolkata, India. [2]Principal, Bangabasi College, University of Calcutta, UGC Sponsored visiting Faculty, Institutue of Radio Physics and Electronics, University of Calcutta, Kolkata, India. ✉email: sonia.rpe2020@gmail.com

Some weather parameters which have been considered here in this work are very much related to generate the CAPE by overcoming CINE. Researches also show that vertical velocity, relative humidity and wind shear, plays very vital role to form the severe thunderstorm[4]. Therefore other weather parameters have been considered here in this study to analyse their relations if any with the formation of severe thunderstorm. Methodologies have been developed to offer vital information on the probability of severe weather[19]. The Numerical Weather Prediction (NWP) model is a very useful tool for diagnosing the structure of thunderstorms. The applications of different NWP models on different weather parameters (such as vertical velocity, relative humidity and wind shear) yields promising result. The application of the NN model in the field of meteorology has been increasingly applied in meteorological research[20–24]. The application of a neural network that learns rather than analysing such compound relationships has revealed an immense deal of assurance in accomplishing the objective of weather forecasting with elevated accuracy[25–27]. The weather prediction reports require some intelligent calculations that can deal with the nonlinear dataset. This creates some rules and patterns to learn from the experimental data to forecast the weather in the future[28]. ANNs (Artificial Neural Networks) have the benefit of their skill to learn and become accustomed[27]. Gyanesh Shrivastava et al., revealed that BPN (Back-Propagation Neural Network) and RBFN (Radial Basis Function Network) are competent model for predicting monsoon rainfall. The forecast of monsoon rainfall based on artificial neural network is a well-researched problem[29]. These models are also effective for small range weather forecast. BPN and RBFN give suitable solutions for the prediction of long-range weather forecasting[30]. Chaudhuri et al., has showed in their studies the use of multilayer perceptron logic and fuzzy logic to analyse the role of different weather parameters for thunderstorm prediction purpose[24,31,32].

The Multilayer Perceptron[33] and KNN (K-Nearest Neighbor) all have been applied on different weather parameters (such as moisture data) for severe thunderstorm prediction purpose previously. There are many studies where statistical and different machine learning techniques have been applied on different weather parameters to predict severe thunderstorm. But here in this study two different methodologies have been introduced on others weather parameters for prediction purpose which are innovative. The results obtained from these methodologies have been compared here with the conventional methodologies (such as Multilayer Perceptron, KNN) also. Here the RBFN and Naive Bayes classifier are introduced for the severe thunderstorm prediction purpose. RBFN has not previously been used for this purpose. The Naive Bayes classifier has been chosen for lightning storm detection purposes using lightning data[34]. Li et al. (2019) applied Naive Bayes for sandstorm prediction purposes. Wu et al. (2015) used RBFN to predict rainfall forecasts, which gave 88.49% correct prediction. Surface temperature prediction has been done using RBFN with a good accuracy of Litta et al. (2015). However there is no benchmark study that predicts severe thunderstorms using RBFN and Naive Bayes using the mentioned weather data with a high accuracy level. The main aspect of this study is as follows:

- A less used methodology has been applied here which gives a high accuracy rate.
- A comparative study was performed among conventional methodologies (Multilayer Perceptron, KNN, Naive Bayes) and RBFN
- A comparative study also reveals that RBFN gives much more promising results than the others.
- This study has a lead time of 10–12 h which is very much important so that the government can take proper precautions to save life and property.

In this study six different weather parameters were considered for severe thunder storm prediction. These six weather parameters are cloud coverage, sunshine hours, pressure at the freezing level and three different dry adiabatic lapse rates at different geopotential heights of the atmosphere. Here different methodologies (both statistical and ANNs) have been applied to these weather parameters for prediction purposes[4,35]. The Naive Bayes classifier has been applied here as a statistical methodology to these weather parameters. This yields more than 85% correct prediction for 'squall days' and 86.34% correct prediction for 'no squall days'. The application of the K Nearest Neighbour (KNN) method on the mentioned data set gives more than 88% correct prediction for 'squall days' and more than 87% correct prediction for 'no squall days'. Then Multilayer Perceptron (MLP) has been applied on the six mentioned weather parameters which produce 91.8% correct prediction for 'squall days' and 89.27% correct prediction for 'no squall days'. The most promising results emerge by the application of Radial Basis Function network (RBFN). This gives more than 95% correct prediction for 'squall days' and more than 94% correct prediction for 'no squall days' on the mentioned weather parameters. In case of a proper weather forecast correctness of a methodology is not the only one factor. A weather prediction without any lead time has no importance. Therefore in case of a proper weather forecast enough lead time is very much necessary. Here in this study a lot of importance has given to these matters. Lead time is the duration of time which is predicted before the onset of the occurrence of the event. The development of the thunderstorm generally begins from the early morning and occurs on the evening time. It can develop within the span of 10–12 h and then it occurs. The lead time is not only important to alert the people but also the Government for taking precautionary measures. All these predictions have a lead time of 10 to 12 h which is necessary to save life and property from damages. Accurate forecasts not only save lives but also support emergency management and mitigation. It also prevents economic losses from high impact weather. It can create major financial revenue in the energy, agriculture, transport and recreational sectors.

Plan of work.

- Weather parameters selection.
- Data collection and processing.
- Application of different methodologies (Naive Bayes, MLP, KNN, RBFN) on the processed data.

- Comparison among the results obtained from different methodologies, skill score calculation.
- Conclusion

Data

In this study, weather data of 33 years have been considered for prediction purposes. The data of three months March–April-May (MAM) for every year from 1969 to 2002 have been chosen. These three months are known as the pre-monsoon months in India.

Data collection. In this paper, real field meteorological data have been collected at the weather station Kolkata (22.3⁰ N/88.3⁰ E), North-East India at morning 0 GMT (6:00 am). These entire real field data are Radio-sonde observational data and collected from the meteorological station (Here Kolkata, Alipore) operated by Indian Meteorological Department, Government of India (IMD). The errors were corrected at the time of observation by IMD. So, all the real field data are here error free and normalized.

North-East India generally signifies Gangatic West Bengal, Coastal region of West Bengal and Assam. The days when thunderstorms take place denoted as thunderstorm days and the days when thunderstorms did take place denoted as no thunderstorm days here in this study. The numbers of 'thunderstorm' days are 161 and 'no thunderstorm' days are 2805. In this study 100 squall days and 2600 no squall days have been considered for training purposes. These training data has been arranged in 1:26 orders and the remaining 61 squall days and 205 days has been considered as the test data set.

Data description. In the current study different weather parameters has been considered for analysis purpose. These weather parameters are: Sun Shine Hour as X1, Pressure at freezing level (FRZ) as X2, Cloud coverage (Octa Nh) as X3 and three different dry adiabatic lapse rates at three different geo-potential heights of atmosphere as X4, X5 and X6. These parameters are essential data for the formation of thunder clouds. The main aim of this study is to predict the thunderstorms by analyzing the numerical data responsible for cloud generation. All these weather parameters are discussed in detail below.

Sunshine hour. The duration of the sun or the time of the sun is a climatic indicator that measures the duration of the sun in a certain phase (typically a day or a year) for a certain position on earth. It is usually articulated as an average of quite a few years. This measures the total energy delivered by sunlight over a period of time. As per the definition given by WMO in 2003 sunshine time is the period during which direct solar irradiance exceeds a threshold value of 120 watts per square meter (W/m^2). This value is equal to the degree of solar radiation shortly after sunrise or shortly before sunset in cloudless regions. This differential heating of the atmosphere near the earth's surface relative to the atmospheric column aloft is ultimately responsible for an instability or conditional instability. For ordinary gases such as atmospheric air that obeys the ideal gas law, parcel density at any altitude (or pressure) is determined by temperature and the buoyancy force is proportional to the temperature difference between the air parcel and its surroundings[36]. The measurement is performed by comparing the recording of the time of day using the Campbell-Stokes solar recorder with a real-time solar radiation[37].

The sun is the ultimate source of energy for thunderstorm convection and, at a larger scale, for the general circulation of the atmosphere. Because of the clear atmospheric transparency to solar radiation, more than half of the incoming sunlight is absorbed by the Earth's surface. A statistical data from World Meteorological Organization Standard Normal shows that mean values of Sun Shine Hour for the three months: March–April-May (MAM) of 1971–1990 are comparatively larger with the other months[38]. The most of the thunderstorm cases occur in these three months.

Solar heating drives convective currents, so thunderstorms tend to be most frequent when and where solar radiation is most intense. Hence, in most areas, thunderstorms are most frequent during the warmest hours of the day[39]. It has been observed that the phase change from water to ice or snow tends to accelerate the parcel or column upwards[40], but synoptic acceleration is important if and only if the accumulation rate is at a frozen level or higher[41]. At first, the phase change occurs in a semi-isobaric manner and the rising air can be warmed far above the surrounding temperature. It follows that heavy rains and rainstorms can be expected to occur when such levels of accumulation are present in the pre-emitting noise when the instability phenomenon begins[40]. Therefore, sunshine hours play a vital role in the formation of thunder clouds.

Pressure at freezing level (FRZ). The melting level in the troposphere where the water freezes is known as the FRZ (freezing point)[40]. It is situated at the intersection between the 0 °C isotherm and the temperature ratchet. An FRZ level with a pressure level of 650 mb or closer to the surface, in severe weather conditions, will usually carry large hailstones. This will have more time to grow in cold air and will have less time to melt as it falls to the surface[42]. As a result of the convection process the hot air rises, transferring temperature to the upper levels of the atmosphere from the Earth's surface. The water vapor they contain begins to cool, release heat, condense and form clouds[40]. The pressure at freezing level is measured using aneroid barometer. It is a device for measuring atmospheric pressure without the use of fluids[43].

Cloud coverage (Octa Nh). The measure of atmospheric moisture is indicated by the cloud content of the upper atmosphere. The most important ingredient to form the thunder cloud is this atmospheric moisture. The amount of moisture in the upper air increases with the increase in cloud content. The Cloud coverage is measured by Ceilometers[44].

3

Dry adiabatic lapse rates. The dry bulb temperature difference between two consecutive levels at different geopotential heights of the atmosphere is the measure of the dry adiabatic lapse rate (dT/dZ). In this study four heights (dZ) of the atmosphere have been considered. These heights are (a) 700 hpa and 600 hpa (approximately 3100 to 4500 m), denoted by X4 (b) 600 hpa and 400 hpa (approximately 4500 to 7500 m), denoted by X5, and (c) 400 hpa and 300 hpa (approximately 7500 to 9600 m), denoted by X6. The temperature differences (dT) between these consecutive heights have been taken into account. The change in temperature is measured by Thermistors. Thermistors are temperature-dependent resistors, changing resistance with changes in temperature. They are very sensitive and react to very small changes in temperature[45].

The dry adiabatic lapse rate of the atmosphere is the measure of the conditional instability[39]. The conditional instability in the atmosphere is the reason for presence of moisture which would carry out to the upper atmosphere from the surface level to form thunder clouds[36]. A statistics from world Meteorological Standard Organization reflects that mean value of the dry bulb temperature remains the maximum during the March–April-May. The thunderstorms occur in these three months. It can be observed from the statistical data that the mean values of dry bulb temperature during other months (except March–April-May) were comparatively lower[46].

Methodologies

Naïve Bayes classifier. Naïve Bayes classifier is a supervised learning algorithm and is utilized for the solution of classification problems[47]. It is based on the Bayes theorem[47]. The Bayes decision theorem is a fundamental statistical approach to recognize a pattern[47]. It is preferable for high dimensional training data sets and quick prediction. Bayes theorem states that,

$$P(AB) = \frac{P(BA)P(A)}{P(B)} \tag{1}$$

where,

P (A|B) is posterior probability: Probability of hypothesis A on the observed event B.
P (B|A) is likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.
P (A) is prior probability: Probability of hypothesis before observing the evidence.
P (B) is marginal probability: Probability of Evidence.

The expression P (A) refers to the probability that event A will occur. P (A|B) stands for the probability that event A will happen; given that event B has already happened. In other words, P(A|B) is the probability of the object belonging to class A i.e., the probability of the attribute values (predictors which are Sun Shine Hour, Pressure at freezing level, Cloud coverage and three different dry adiabatic lapse rates at three different geo-potential heights of atmosphere) B belonging to class A (squall or no squall days)[48].

Here is the algorithm for Naive Bayes procedure:

Convert the training dataset into corresponding frequency tables.
Generate likelihood table by finding the probabilities of the mentioned parameters.
Then the Bayes theorem is used to compute the posterior probability.

Naive Bayes is straightforward probabilistic classifier[49]. This often gives reasonable solution in many real-world problems[50]. Despite of its unrealistic independence hypothesis, the Naïve Bayes classifier is astonishingly successful in exercise[50]. The performance of Naïve Bayes classification is fairly good, as evidenced by the many experimental studies[51]. In this study Table 1 from the result Sect. 4 shows that Naïve Bayes classification yields 85.25% correct prediction for 'squall days' and 86.34% correct prediction for 'no squall days'.

K-nearest neighbor (K-NN). K Nearest Neighbor (K-NN) is one of the familiar names in the field of data classification[52]. The K-NN algorithm was successfully applied by Cover in1967. This is a straightforward algorithm that reserves all existing cases and classifies new cases created based on the amount of vicinity. The K-NN determines the way that which of the points from the training sets is similar enough to be considered[53]. The k value in the k-NN algorithm defines how many neighbors will be checked to determine the classification of a specific query point. For example, if k = 1, the instance will be assigned to the same class as its single nearest neighbor[54]. The principle of the algorithm is established on a comparison between a given testing data point and training data points[52]. This sorts out the training data points which are in close vicinity (neighbors) with test data points, and then predicts the corresponding class label of these neighbors[53]. It can be said that neighbors are measured by a distance or dissimilarity measure that can be computed between samples based on the independent variables[52]. KNN is a non-parametric procedure to classify items built on closest training instances in the feature space[53]. One of the best examples of instance-based learning or lazy learning is KNN[52]. Here, the function is estimated locally and all calculations are delayed until classification[8]. In the classification stage, K is a user-

Training data for squall days = 100	Training data for no squall days = 2600	
	Squall (total no. of test data = 61)	No-squall (total no. of test data = 205)
Six variables	52, 85.25%	177, 86.34%

Table 1. Correct prediction of 'squall days' and 'no squall' days using Naïve Bayes classification.

defined constant, and these are not previously labeled[53]. Here in this study K have been chosen as 1, 3, and 5. All training data vectors have a class label[53]. The training stage of the algorithm contains only loading the feature vectors and class labels of the training objects[55]. The similarity measure has been considered between each data vector of test data set with each data vector of training data set. Similarity between two vectors can be defined as,$p = (p1, p2,…, …., p\gamma)$, $q = (q1, q2,… , q\gamma)$ is defined as,

$$\frac{\sum_{i=1}^{\gamma} p_i q_i}{\sqrt{\left(\sum_{i=1}^{\gamma} p_i^2 \sum_{i=1}^{\gamma} q_i^2\right)}} \quad (2)$$

Here p corresponds to training data vector and q corresponds to test data vector. Here value of γ is 6 since numbers of parameters are six. Here p1 and q1 corresponds to Sun Shine Hour (variable X1), p2 and q2 corresponds to Pressure at freezing level (FRZ, variable X2), p3 and q3 corresponds to Cloud coverage (Octa Nh, variable X3), p4 and q4 (variable X4), p5 and q5 (variable X5), p6 and q6 (variable X6) corresponds to three different dry adiabatic lapse rates at three different geo-potential heights of atmosphere respectively. The flowchart for KNN has been depicted in the Fig. 1.

The cosine angle between two vectors indicates the similarity measure between them[52], which will be greater if the angle value is smaller. The similarity measure indicates the vicinity of each data vector of the test set with each data vector of the training set. These cosine angles are arranged in decreasing order. The result of Table 2 from section "Result" shows that 3NN gives the most promising result in comparison with 1NN and 5NN. The 88.52% correct prediction for 'squall days' and 87.8% correct prediction for 'no squall days' were obtained by applying 3NN.

Multilayer perceptron. One of the most widely used empirical approaches for weather prediction is artificial neural network[56]. A three-layered Multilayer Perceptron (MLP) network has been applied to the above mentioned six weather variables. It consists of an input layer, one hidden layer and an output layer. As such, neural networks are extremely complex[51]. The ANN (Artificial Neural Network) reduces the error using a variety of algorithms. This produces an approximated value that is close to the real value[57]. One of the most promising branches of artificial intelligence is neural network. It has many applications in the field of space weather prediction such as forecasting geomagnetic storms[58] and solar flairs[59]. A single layer perceptron with one input produces decision regions under the form of semi planes[51]. The addition of one layer causes every neuron to act as a standard perceptron for the outputs of the neurons in the anterior layer. Therefore, the output of the network can evaluate convex decision regions, which results from the intersection of the semi planes produced by the neurons[60]. Sequentially, a three-layer perceptron can create arbitrary decision areas[60].

Learning phase. In the learning phase of the Multilayer Perceptron, the 'occurrence' of storm days is represented by a value of 1 and 'no occurrence' of the storm days is represented by the value of 0. Every unit of every layer is associated with every unit of the next layer by the connection weights[4]. The sigmoid function is chosen as the transfer function which acts as a nonlinear activation function. Two different modes of learning the weights of an MLP exist. These are Batch mode learning and On-line learning. Here, On-line method of learning the weights is considered[51].

Feed forward stage. The multilayer perceptron is the neural network model that is commonly known and most frequently used in different types of applications. Generally, the signals are transferred within the network unidirectionally from input to output. The initial part of this architecture is called the feed forward stage of the network[60]. In this stage each node (say *i*) in layer α is joined to each node (say *j*) in the next layer ($\alpha + 1$), with a connection weight represented by $W_{ij}^{(\alpha)}$[60]. Let S_i be the i-th input node in the input layer. Then the activation unit for the hidden layer is Y_i, which is the output from the nodes of the input layer. Y_i is the total input received for the j-th node in the hidden layer.

$$Yi = \sum_{i=1}^{n} S_i W_{ij} \quad (3)$$

The output from the j-th node of the hidden layer is Y_j. A transfer function is used to obtain this[51].

$$Y_j = \frac{1}{1 + \exp(-Y_i)} \quad (4)$$

This is valid for every layer.

Connection weights. Connection weights (*W*'s) are adjusted to trivial random values in the range (− 0.5 to 0.5)[4]. A threshold value is correspondingly presumed. The weight values are altered in back propagation stage of the learning of the model until the error is reduced[4]. The test data is validated by these modified weights. The gradient descent technique is mainly used in back propagation process to modify the weights. It is used to minimize the chances of becoming trapped in local optimal points or saddle points of the network[51].

Error. The error function is measured by the mean square error. This is given as follows,
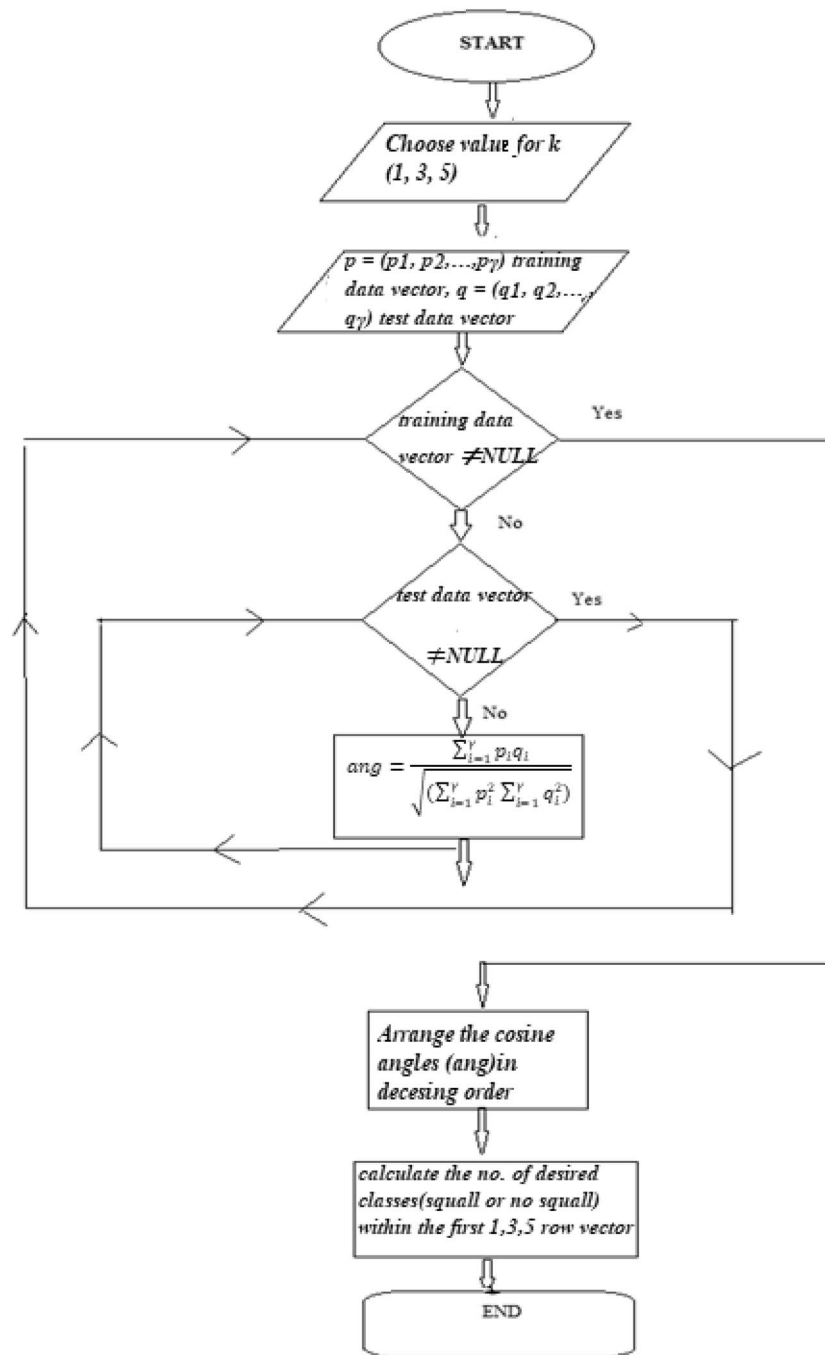
Figure 1. Flowchart for KNN methodology.

K = 1		K = 3		K = 5	
Training data for squall days = 100	Training data for no squall days = 2600	Training data for squall days = 100	Training data for no squall days = 2600	Training data for squall days = 100	Training data for no squall days = 2600
Squall (total no. of test data = 61)	No-squall (total no. of test data = 205)	Squall (total no. of test data = 61)	No-squall (total no. of test data = 205)	Squall (total no. of test data = 61)	No-squall (total no. of test data = 205)
52, 85.25%	174, 84.88%	54, 88.52%	180, 87.8%	50, 81.97%	175, 85.37%

Table 2. Correct prediction of 'squall days' and 'no squall' days using K Nearest Neighbor Method.

6

$$E = \frac{\sum_{i=1}^{2} \left(o_j - e_j\right)^2}{2} \tag{5}$$

The expected output (e_j) for each data point in the training set is recognized[51]. For a specific scenario the real output value for the j-th node in the output layer is o_j[51]. The error has to be reduced during the training time using back propagation. Iteration is continued until the error is reduced approximately 0.005 to 0.001[4].

Back propagation of error. In the present case, the back propagation rule is applied to the set of training patterns of data. This rule basically uses the gradient descent technique for changing the weights. The main aim is to arbitrate the modification of weight representation of an input–output pattern pair. Since given data can be used numerous times during training, let us use the index m to denote the presentation step for the training pair at step m[51]. For training a multilayer feed-forward neural network, the subsequent approximation is used by applying the gradient descent along the error surface[51] to determine the increase in the weight connecting units j and i:

$$\Delta w_{ij}(m) = -\eta \frac{\delta E(m)}{\delta W_{ij}} \tag{6}$$

where $\eta = 0.01$ is the learning rate parameter.

$E(m)$ denotes the measure of performance, the negative derivative of $E(m)$ with respect to the weight Wij can be defined as the negative gradient of $E(m)$.

Updation of weights. The weight update is given by,

$$W_{ij}(m+1) = W_{ij}(m) + \Delta W_{ij}(m) \tag{7}$$

The modified weights are used in the test dataset to validate the outputs[51]. Sometimes, if the number of iterations becomes too much large or if the classifications on the test set are insufficient, the error may not be minimized[51]. In such cases, the architecture of MLP is to be modified by modifying the number of nodes in the hidden layer or by changing the number of hidden layers[4]. MLP include too many parameters because it is fully connected. Each node is connected to another in a very dense web — resulting in redundancy and inefficiency[61]. Here in this study three layered MLP has been considered. These are 6–3-2, 6–4-2, and 6–5-2. Here the first layer represents input layer, second layer represents hidden layer and third layer represents output layer. Table 3 from section "Result" shows that applying MLP gives 91.8% correct prediction for 'squall days' and 89.27% correct prediction for 'no squall days' obtained. The flowchart for MLP has been depicted in the Fig. 2.

Radial basis function network. Artificial Neural Network (ANNs) offers a methodology for explaining different kinds of nonlinear problems that are complex to solve by conventional methodologies[62]. There are several types of ANN (Artificial Neural Network) and the Radial Basis function is one of them. Radial Basis Functional Networks (RBFNs) are non-linear layered feed forward networks[63]. It can implement arbitrary non-linear transformations of the input space. There are different applications of RBFNs[64]. The RBFNs are most effective for prediction purposes such as weather prediction, modeling, pattern recognition, and image compression[64,65]. It contains three different layers: input layer, hidden layer and output layer. The hidden layer is multidimensional and defined as radial counters[47].

Each hidden unit is defined as a radial center and every center represents one or some of the input patterns[66]. The network is known as a 'localized receptive field network'[64]. The hidden units in RBFN have Gaussian activation functions as follows:

$$\emptyset_i(X) = \varphi(||xt_i||) \tag{8}$$

where $||\boldsymbol{xt_i}||$ denotes the Euclidean norm function and φ is the RBF neuron activation function. The input vector is denoted by x i.e., the input weather data and t_i denote the neuron's prototype vector. The approximation of output, by an RBF will be denoted by $\hat{y}_t$.

Design of the Network	No. of correctly classified and % of accurate points for 'squall storm' days in test data sheet	No. of correctly classified and % of accurate points for 'no squall storm' days in test data sheet	No. of correctly classified and % of accurate points for in data sheet. Data set for both 'squall storm 'days and 'no squall storm ' days are considered
	Total no. of squall storm days = 61	Total no. of no squall days = 205	Total no. of days = 266
6–3-2	53, 86.88%	177, 86.34%	230, 86.47%
6–4-2	56, 91.8%	183, 89.27%	239, 89.85%
6–5-2	50, 90.16%	183, 89.27%	233, 87.59%

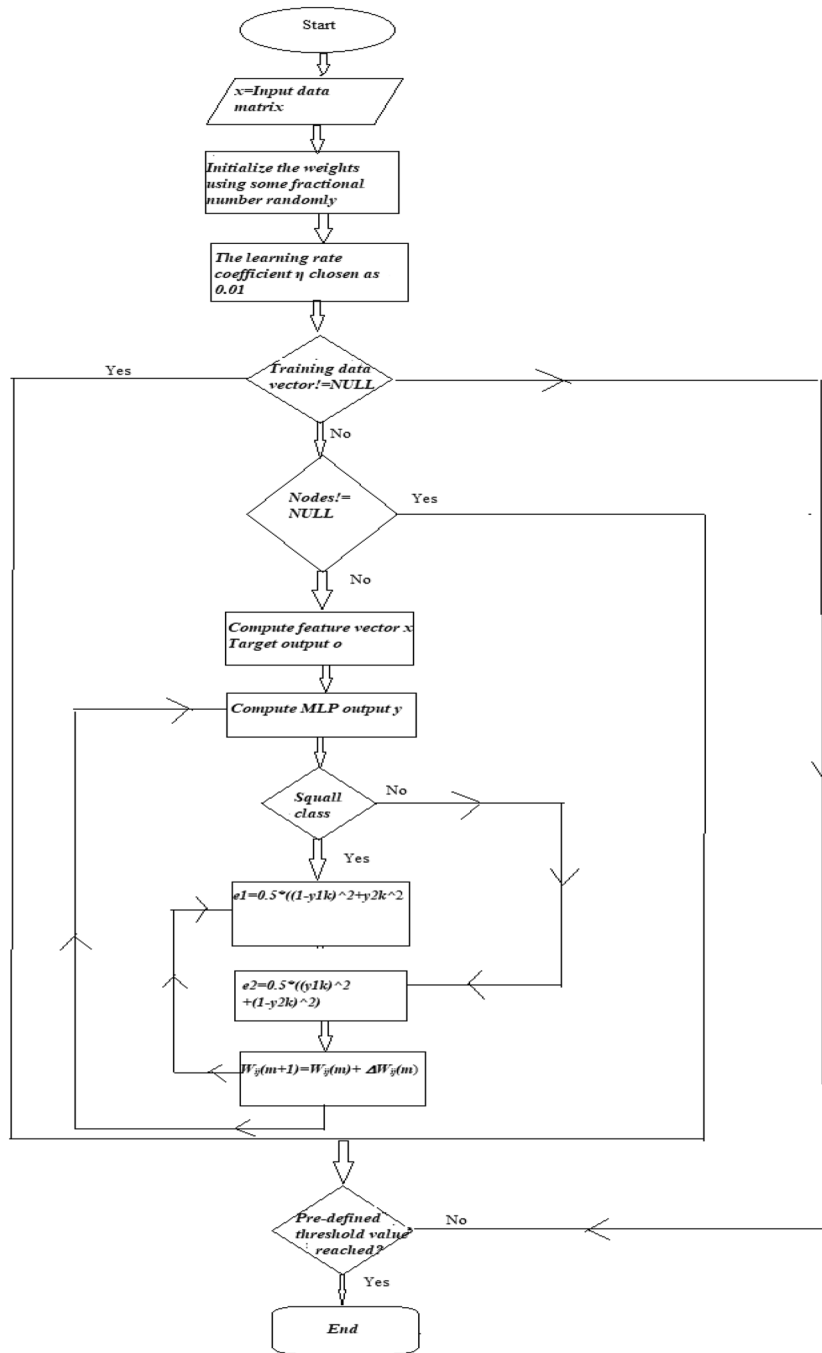**Table 3.** Correct prediction of 'squall days' and 'no squall' days using MLP.

Figure 2. Flowchart for MLP methodology.

$$\overset{\wedge}{\underset{t}{Y}} = \sum_{i=1}^{m} \lambda_i \emptyset(x_i, \ C_i, \ \sigma_i) \tag{9}$$

This approximation will be the weighted sum of m Gaussian kernels $\emptyset$:

$$\emptyset(x_i, \ C_i, \ \sigma_i) = \exp\left(\frac{-x - x_t}{\sqrt{2}\sigma_t}\right)^2 \tag{10}$$

Gaussian kernels are used to determine the complexity of RBFN. The various parameters to specify are the positions of the Gaussian kernels (Ci)[66]. The second parameter to be chosen is the standard deviation (or width) of the different Gaussian kernels σi. The last parameter is denoted by the multiplicative factor λi[66].

The hidden layer in RBF is of high dimension, which has a different purpose than in a multilayer feed forward network[66].The radial distance di, between the input vector x and the center of basis function Ci is computed for each unit i in the hidden layer as follows:

$$d_i = x - C_i \tag{11}$$

$$y = f(x) = \sum_{i=1}^{k} w_i \varphi_i (x - C_i) \tag{12}$$

Here, f denotes nonlinear activation function, x denotes input, φ_1, φ_2, …, …, φ_m denotes RBF centers in the input vector space[63]; every neuron in the hidden layer has its adjoining center, X denotes the input vector, k denotes the total number of hidden layer neurons and i denotes the j-th node in the hidden layer[63]. Although the training is faster in RBF network but classification is slow in comparison to Multi layer Perceptron due to fact that
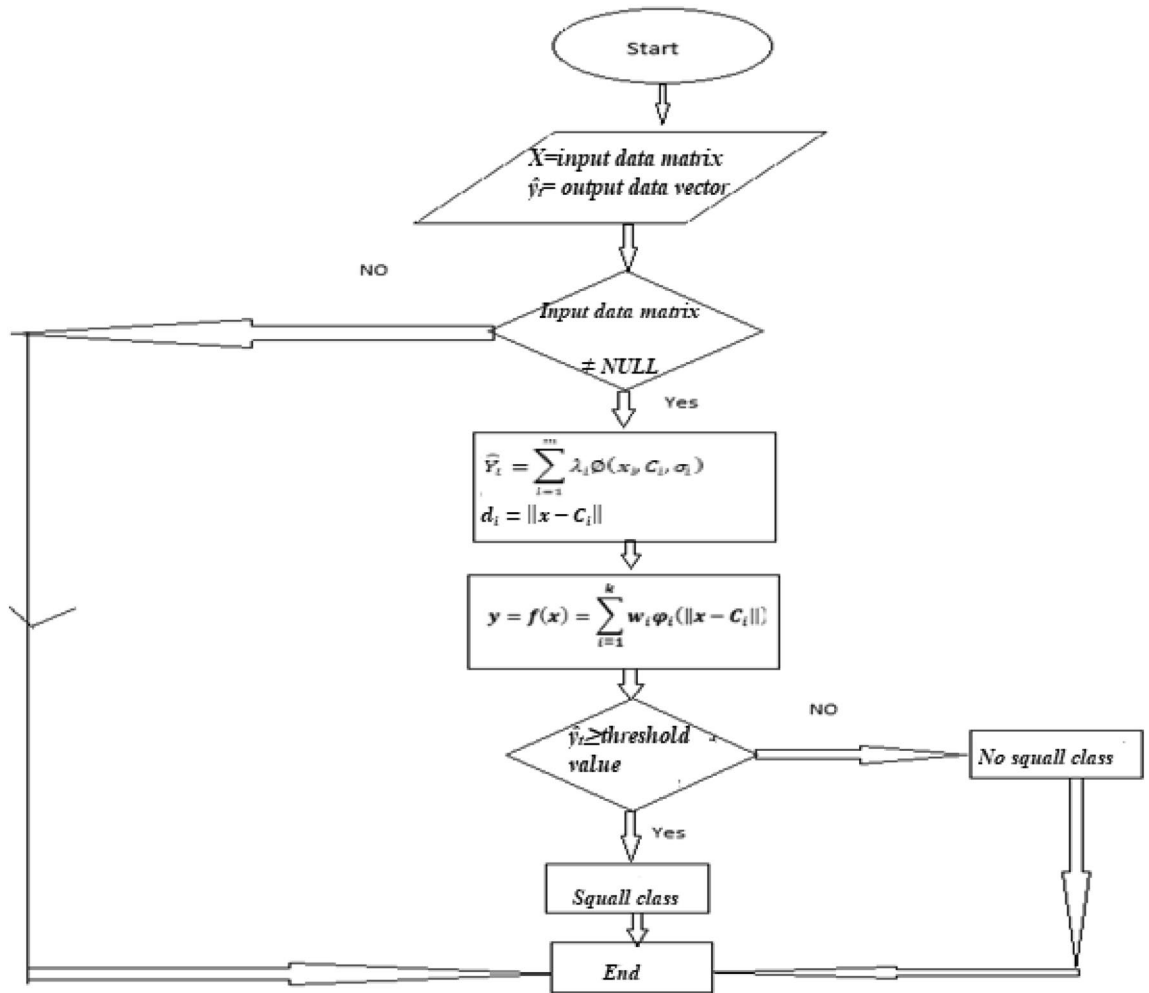

Figure 3. Flowchart for RBFN methodology.

Design of the Network	No. of correctly classified and % of accurate points for 'squall storm' days in test data sheet	No. of correctly classified and % of accurate points for 'no squall storm' days in test data sheet	No. of correctly classified and % of accurate points for in data sheet. Data set for both 'squall storm 'days and 'no squall storm ' days are considered
	Total no. of squall storm days = 61	Total no. of no squall days = 205	Total no. of days = 266
6–7-1	56, 91.8%	190, 92.68%	246, 92.48%
6–8-1	58, 95.08%	193, 94.15%	251, 94.36%
6–9-1	58, 95.08%	190, 92.68%	248, 93.23%

Table 4. Correct prediction of 'squall days' and 'no squall' days using six weather variables using RBFN.

every node in hidden layer have to compute the RBF function for the input sample vector during classification[67]. Here in this study three layered RBFN has been considered. These are 6-7-1, 6-8-1, and 6-9-1. Here the first layer represents input layer, second layer represents hidden layer and third layer represents output layer. The flowchart for RBFN has been depicted in the Fig. 3.

Table 4 from section "Result" shows that RBFN gives 95.08% correct prediction for squall days and 94.15% correct prediction for no squall days.

Result

Here results of four different methodologies have been represented. A total of 61 squall and 205 no squall days were chosen as test data randomly from 1969 to 2002 from the three months of March–April-May (MAM). There was a strong squall line over the sky of Kolkata (22.3°N/88.3°E) on these 61 squall days and severe thunderstorm occurred. There was no thunderstorm activity observed during these 205 no squall days over Kolkata (22.3°N/88.3°E). Here in this study WEKA 3.8.5 has been used as a common package tool to perform Naïve Bayes, K-NN, MLP and RBFN. This is free software and the operating platform is Windows 7.

The result of Table 1 shows that the application of Naïve Bayes methodology on the above-mentioned sample days produces 85.25% correct prediction for 'squall days' and 86.34% correct prediction for 'no squall days'.

Table 2 shows that KNN yields better results on these six weather variables in comparison with Naïve Bayes methodology. The result of Table 2 shows that 3NN gives the most promising result in comparison with 1NN and 5NN. The 88.52% correct prediction for 'squall days' and 87.8% correct prediction for 'no squall days' were obtained by applying 3NN. Table 2 shows that KNN yields better results on these six weather variables in comparison with Naïve Bayes methodology.

Table 3 shows that applying MLP 91.8% correct prediction for 'squall days' and 89.27% correct prediction for 'no squall days' has been obtained.

The most promising results yield from the application of RBFN (Table 4) on these six weather variables. RBFN gives 95.08% correct prediction for squall days and 94.15% correct prediction for no squall days.

It can be concluded from Table 5 that among these four methodologies RBFN gives the lowest misclassification rate for squall days.

The Heidke Skill Score (HSS) has also been applied here for the purpose of forecast which is a measure of skill. The Heidke Skill Score (HSS) is a skill score for categorical forecasts[68]. It is defined as follows,

$$ HSS = \frac{\left(\left(Hit + Correct\ negatives\right) - \left(Expected\ correct\right)_{random}\right)}{N - \left(Expected\ correct\right)_{random}} $$

where, $(expected\ correct)_{random} = \frac{1}{N}\left[(hit+misses)(hit+false\ alarms)+(correct\ negatives+misses)(correct\ negatives+false\ alarms)\right]$

Naïve Bayes		MLP		KNN		RBFN	
Squall class	No squall class	Squall class	No squall class	Squall class	No squall class	Squall class	No squall class
0.15	0.14	0.08	0.11	0.11	0.12	0.05	0.06

Table 5. Misclassification rate comparison among four methodologies applied to six weather variables.

Forecast	Observed		
	Yes	No	Total
Yes	Hits	False alarms	Forecast yes
No	Misses	Correct negatives	forecast no
Total	Observed yes	Observed no	Total

Table 6. Contingency Table.

Forecast	Observed		
	Yes	no	Total
Yes	52	28	80
No	9	177	186
Total	61	205	266

Table 7. The contingency table for Naïve Bayes.

Here, N denotes total number of test data; the term hit represents event forecast to occur, and did occur; miss denotes event forecast not to occur, but did occur; false alarm represents event forecast to occur, but did not occur and correct negative represents event forecast not to occur, and did not occur.

It can be analyzed from Table 6 (contingency table) what types of errors are being made.

Here 'yes' indicates squall days and 'no' indicates no squall days.

The HSS for the different methodologies has been obtained from the following contingency tables.

Therefore it can be obtained from the Table 7 that the Heidke Skill Score (HSS) for Naïve Bayes is 0.66.

Therefore it can be obtained from the Table 8 that the Heidke Skill Score (HSS) is 0.62 for 1 NN, 0.69 for 3NN and 0.61 for 5NN respectively.

Therefore it can be obtained from the Table 9 that the Heidke Skill Score (HSS) for MLP is 0.74.

Therefore it can be obtained from the Table 10 that the Heidke Skill Score (HSS) for RBFN is 0.85. The HSS measures the fractional improvement of the forecast over the standard forecast. HSS 0 means no skill, and a perfect forecast obtains a HSS of 1. Here RBFN exhibits the HSS value as 0.85 which is close to 1. Therefore it can be said that RBFN gives the best result among the other three methodologies here.

Conclusion

The study here predicts severe thunderstorms using both statistical and ANN methodologies on numerical weather data. The numerical simulation depends on the volume of the input data set[69]. Neural network classifiers have been attractive alternatives to conventional classifiers by numerous researchers[7]. The methodologies that are considered here have advantages and disadvantage both. The ANN methodologies produce output even with incomplete information. The ANN methodologies have much more fault tolerant capability[70]. The MLP and RBFN methodology both work well for large amount of data. In case of MLP there is loss of non convex function when there is more than one local minimum[70]. Although the training is faster in RBF network but classification is slow in comparison to Multi layer Perceptron due to fact that every node in hidden layer have to compute the RBF function for the input sample vector during classification[71]. RBF network works more effectively on noised input data set[71]. The KNN on the other hand gives better classification on rare events; it performs well for multi-classification issues[72]. The KNN shows poor result if the sample size is not properly balanced[72]. The choice of the value of K is one of the most crucial factors for correct prediction. The Naive Bayes methodology is easy to implement and the training is fast. The main disadvantage of Naive Bayes methodology is conditional independence assumption which does not always hold. In most situations, the feature show some form of dependency[47]. Different previous studies has showed that the application of MLP, KNN on weather parameters like moisture difference and wind shear can produce very effective result for thunderstorm prediction purpose[52,52]. Therefore, here in this study some different kind of weather parameters has been considered for thunderstorm prediction

	1NN			3NN				5NN			
	Observed			Observed				Observed			
Forecast	Yes	No	Total	Forecast	yes	no	Total	Forecast	Yes	No	Total
Yes	52	31	83	Yes	54	25	79	Yes	50	30	80
No	9	174	183	No	7	180	187	No	11	175	186
Total	61	205	266	Total	61	205	266	Total	61	205	266

Table 8. The contingency table for 1NN, 3NN and 5NN.

	Observed		
Forecast	Yes	No	Total
Yes	56	22	78
No	5	183	188
Total	61	205	266

Table 9. The contingency table for MLP.

	Observed		
Forecast	Yes	No	Total
Yes	58	12	70
No	3	193	196
Total	61	205	266

Table 10. The contingency table for RBFN.

purpose. There are many studies that used both statistical and ANN methodologies to predict severe thunderstorm. But there is no notable study where RBFN and Naive Bayes methodologies have been used for severe thunderstorm prediction successfully. RBFN gives more accuracy and builds the model faster than MLP. The aim of this study is not only to predict severe thunderstorm correctly but also to establish an effective comparative findings among ANN and statistical methodology. The present study can be extended in future by the analysis of cloud imageries for thunderstorm prediction purpose. Table 5 shows that among the four methodologies, RBFN exhibits the minimum misclassification rate. In this work the best result have been obtained by applying RBFN (ANN methodology) among the other methodology that have been used on the weather data. It can be concluded that the Naïve Bayes methodology yields less promising results for 'squall' days in comparison with the other three methodologies. Overall both the statistical and ANN methodologies give more than 80% correct prediction for severe thunderstorm in this study. Generally, thunderstorms occur in the North-East India during the evening. Lead time is the period between the time of prediction and occurrence of the event. Thunderstorm is a catastrophic event, generating in the early morning and occurring in the evening time. So, accurate prediction with enough lead time is very pertinent to protect the social life. Sufficient lead time is also helpful for local Government to make the people alert and to take safety measures for the people. Therefore, in this study 10–12 h as the lead time has been considered.

Data availability
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

References
1. Brooks, H. E. & Wilhelmson, R. B. Numerical simulation of a low-precipitation supercell thunderstorm. *Meteorol. Atmos. Phys.* **49**(1–4), 3–17 (1992).
2. Farley, R. D., Wang, S. & Orville, H. D. A comparison of 3D model results with observations for an isolated CCOPE thunderstorm. *Meteorol. Atmos. Phys.* **49**(1–4), 187–207 (1992).
3. Byers, H. R. and Braham, R. R. *The Thunderstorms, U.S. Govt. Printing Office*, 287 pp (1949)
4. Chakrabarty H, Bhattacharya S, Prediction of Severe Thunderstorms applying Neural Network using RSRW Data. Int. J Comput. Appl. 89: 1–5 DOI: https://doi.org/10.5120/15712-4362 (2014)
5. Jenamani, R. K., Vashisth, R. C. & Bhan, S. C. Characteristics of thunderstorms and squalls over Indira Gandhi International (IGI) airport, New Delhi Impact on environment especially on summer's day temperatures and use in forecasting. *Mausam* **60**(4), 4569 (2009).
6. Rajeevan, M., Kesarkar, A., Thampi, S. B., Rao, K. N. & Radhakrishna, B. Rajsekhar, M Sensitivity of WRF cloud microphysics to simulations of a severe thunderstorm event over Southeast India. *Ann. Geophys.* **28**, 603–619 (2010).
7. Newton, C. W., Dynamics of Severe Convective Storms. Meteorological Monographs. *Am. Meteorol. Soc.* 5: 33–58 (1963) DOI: https://doi.org/10.1007/978-1-940033-56-3_2
8. Cover, T. M., Hart, P. E., Nearest neighbor pattern classification. *IEEE Trans Inf Theor. IT* **13**(1), 1053964 (1967) DOI: https://doi.org/10.1109/TIT.1967.1053964
9. Browning, K. A. General circulation of middle latitude thunderstorms; In: *Thunderstorm Morphology and Dynamics*; (ed.) Kessler E; Univ. Oklahoma Press, 133–152 (1986)
10. Lal, A. Forecasting of thunderstorm around Delhi and Jodhpur. *Mausam* **40**, 267–268 (1989).
11. Bauer, P., Thorpe, A., Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55 (2015)
12. Balsamo, Gianpaolo, Rui Salgado, Emanuel Dutra, S. Boussetta, T. Stockdale, and Miguel Potes. On the contribution of lakes in predicting near-surface temperature in a global weather forecasting model. Tellus A: Dynamic Meteorology and Oceanography 64, no. 1 (2012)
13. P. Schultz. Relationships of several stability indices to convective weather events in Northeast Colorado. *Weather Forecast.* **4**(1), 73–80 (1989) DOI: https://doi.org/10.1175/1520-0434(1989)004<0073:ROSSIT>2.0.CO;2
14. Jung, T., Miller, M. J. & Palmer, T. N. Diagnosing the origin of extended-rangeforecasterrors. *Mon. Weath. Rev.* **138**, 2434–2446 (2010)
15. Duc, L., Saito, K. & Seko, H. Spatial-temporal fractions verification for high-resolution ensemble *Forecasts. Tellus* **A65**, 18171 (2013).
16. Erro, C. A. T. & Stephenson, D. B. Extremal Dependence Indices: improvedverification measures for deterministic forecasts of rare binary events. *Weather Forecast.* **26**, 699–713 (2011).
17. Williams, E. R. *et al.* A radar and electrical study of tropical hot towers. *J. Atmos. Sci.* **49**, 1386–1395 (1992).
18. Chaudhuri, S. & Chattopadhyay, S. Measure of CINE—A relevant parameter for forecasting pre-monsoon thunderstorms over GWB. *MAUSAM* **52**(4), 679–684 (2001).
19. Dole, R.et al.The making of an extreme event: putting the pieces together. *Bull. Am. Meteorol. Soc.* **95**, 427–440 (2014).
20. Pozzi, M., Malmgren, B. A. & Monechi, S. Sea surface temperature and isotopic construction from nano plankton data using artificial neural networks. *Palaeontol. Electron* **3**, 4–14 (2000).
21. Richaume, P., Badran, F., Crepon, M., Mejia, C. & Roquet, H. Neural networkwind retrieval from ERS-1 scatterometer data. *J. Geophys. Res* **105**, 8737–8751 (2000).
22. Bourras, D., Liu, W. T., Eymard, L. & Tang, W. Evaluation of latent heat flux fields from satellites and models during SEMAPHORE. *Journal of Applied Meteorology and Climatology* **42**(2), 227–239 (2003).
23. Mitra, A. K., Kundu, P. K., Sharma, A. K. & Roy Bhowmik, S. K. A neural network approach for temperature retrieval from AMSU-A measurements onboard NOAA-15 and NOAA-16 satellites and a case study during Gonu cyclone. *Atmósfera* **23**(3), 225–239 (2010).
24. Chaudhuri, S. & Chattopadhyay, S. Multi layer perceptron model in pattern recognition of surface parameters during pre-monsoon thunderstorm. *MAUSAM* **53**(4), 417–424 (2002).
25. Devi, C., Reddy, B., Kumar, K., Reddy, B., Nayak, N., An approach for weather prediction using back propagation. *Int. J. Eng. Trends Technol.* **3**, 19–23 (2012)
26. Hayati, M. & Mohebi, Z. Application of artificial neural networks for temperature forecasting. *Int. J. Electr. Comput. Eng.* **1**(4), 662–666 (2007).

27. Moro, Q.I., Alonso, L., Vivaracho, C.E., Application of neural networks to weather forecasting with local data. *Proc. of the 12th IASTED international conference on applied informatics.* Annecy, 68–70 (1994)

28. Priyanka, M., Chhaya, N., Siddheshwar, Kini., Krishnanjali, S., Weather Forecasting using Neural Network. *Int. J. Eng. Res. Technol. ICIATE Conf. Proc.* 5(1), 4259 (2017)

29. Chaudhury, S., Goswami, S. & Das, D. Meta-heuristic ant colony optimization technique to forecast the amount of summer monsoon rainfall: Skill comparison with Markov chain model. *Theor. Appl. Climatol.* **116**, 585–595 (2014).

30. Gyanesh, S., Sanjeev Karmakar Manoj Kumar, K. & Pulak, G. Application of artificial neural networks in weather forecasting: A comprehensive literature review. *Int. J. Comput. Appl.* **51**, 0975–8887 (2012).

31. Chaudhuri, S. & Chattopadhyay, S. Consequences of pre-monsoon thunderstorm -A fuzzy logic approach. *Mausam* **55**(1), 119–122. https://doi.org/10.54302/mausam.v55i1.938 (2022).

32. Chaudhuri, S., Khan, F., Das, D. & Mondal, P. Dey S Probing for overshooting as extreme event of thunderstorms. *Nat. Hazards* **102**, 1571–1588. https://doi.org/10.1007/s11069-020-03977-y (2020).

33. Litta, A. J., Idicula, S. M. & Francis, C. N. Artificial neural network model for the prediction of thunderstorms over Kolkata. *Int. J. Comput. Appl.* **50**(11), 1135. https://doi.org/10.5120/7819-1135 (2012).

34. Tiancheng L, Qing-dao-er-ji R, Ying Q *Application of Improved Naive Bayesian-CNN Classification Algorithm in Sandstorm Prediction in Inner Mongolia. Advances in Meteorology*. Hindwai. https://doi.org/10.1155/2019/5176576. (2019)

35. Chakrabarty, H. & Bhattacharya, S. Forecasting of severe thunderstorms using upper air data. *Int. J. Sci. Eng. Res.* **6**(7), 45628 (2015).

36. Volland, H. *Handbook of Atmospheric Electrodynamics*. CRC Press. **1**, 28 (1995)

37. Sanchez Romero, A., González, J.-A., Calbó, J. & Sanchez Lorenzo, A. Characterization of the Campbell-stokes sunshine duration recorder and its ability to derive direct solar radiation by using digital image processing. *Geophys. Res. Abstr.* **16**, 2014–4998 (2014).

38. http://data.un.org/Data.aspx?d=CLINO&f=ElementCode%3A15%3BCountryCode%3AKO

39. Moran J. M., Moran M. D., *Meteorology The Atmosphere and the Science of Weather.* Prentice Hall. Fifth Edition. 316–318 (1986)

40. Yang, Y., Na, Z., Yukun, H. & Xinyao, Z. Effect of wind speed on sunshine hours in three cities in Northern China. *Clim. Res.* **39**(2), 149–157. https://doi.org/10.3354/cr00820 (2009).

41. Wilk, K. E., *Research Concerning Analysis Of Severe Thunderstorms. Geophysics Research Directorate Air Force Cambridge Research Laboratories Office Of Aerospace Research United States Air Force Bedford, Massachusetts*. 19(604), 4940 (1961)

42. Norman, R. B., (1946) Thunderstorms and the freezing level. *Bulletin of the American Meteorological Society. Am. Meteorol. Soc.* 27(2):54–58

43. Garratt, J. R., Bird, I. G. & Stevenson, J. An electrical-readout, oven-controlled, aneroid barometer for meteorological application. *J. Atmos. Ocean. Technol.* **3**(4), 605–613 (1968).

44. Sharma, S. *et al.* Evaluation of cloud base height measurements from ceilometer CL31 and MODIS satellite over Ahmedabad. *India. Atmos. Meas. Tech. Discuss.* **8**, 11729–11752. https://doi.org/10.5194/amtd-8-11729 (2015).

45. Rajanish, K. K. & Gourish, M. N. Thermistors—In search of new applications, manufacturers cultivate advanced NTC techniques. *Sens. Rev.* **22**(4), 334–340. https://doi.org/10.1108/02602280210444654 (2002).

46. http://data.un.org/Data.aspx?d=CLINO&f=ElementCode%3a02

47. Lan, L. & Vucetic, S. Improving accuracy of microarray classification by a simple multi-task feature selection filter. *Int. J. data Min.* **5**(2), 189–208 (2011).

48. Gupta, G. K. Introduction to data mining with case studies. 3rd Edition ISBN: 978–81–203–5002–1 (2019)

49. Taheri, S. O. N. A. *Learning the Naive Bayes Classifier with Optimization Models.* **23**(4), 787–795 (2013)

50. Pouria, K., Sunita, D., Short survey on naive bayes algorithm. *Int. J. Adv. Eng. Res. Dev.* (2017)

51. Yegnanarayana, B. *Artificial neural networks*. Prentice Hall of India Pvt Ltd. (1999)

52. Chakrabarty, H. Application of K-nearest neighbor technique to predict severe thunderstorms. *Int. J. Comput. Appl.* **110**(10), 0975–8887 (2015).

53. Moradian, M. & Baraani, A. KNNBA: K-nearest-neighbor-based-association algorithm. *J. Theor. Appl. Inf. Technol.* **6**(1), 123–129 (2009).

54. https://www.ibm.com/in-en/topics/knn

55. Wu, J., Long, J. & Liu, M. Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm. *Neurocomputing* **148**, 136–142 (2015).

56. Litta, A. J., Idicula, S. M. & Francis, C. N. Artificial neural network model for the prediction of thunderstorms over Kolkata. *Int. J. Comput. Appl.* **50**(11), 45896 (2012).

57. Abhishek, K., Singh, M. P., Ghosh, S. & Anand, A. Weather forecasting model using artificial neural network. *Proc. Technol.* **4**, 311–318. https://doi.org/10.1016/j.protcy.2012.05.047 (2012).

58. Lundstedt, H., *Magnetic Storm. Geophysical Monograph Series.* pp. 98 (1997) https://doi.org/10.1029/GM098.

59. Rong, L., Wang, H., He, H., Cui, Y., and Du, Z., Support vector mahine combined with K-nearest Neighbors for solar flare forecasting. *Chin. J. Astron. Astrophys*. **7**, 441–447 DOI:https://doi.org/10.1088/1009-9271/7/3/15 (2007)

60. Popescu, M. C., Balas, V. E., Perescu-Popescu, L. & Mastorakis, N. Multilayer perceptron and neural networks. *WSEAS Trans. Circ. Syst.* **8**(7), 579–588 (2009).

61. Lavine B.K., Blank T.R. 3.18—Feed-forward neural networks. comprehensive chemometrics. Elsevier. ISBN 9780444527011 (2009) DOI: https://doi.org/10.1016/B978-044452701-1.00026-0.

62. Shereef, I.K., 7 Baboo, S.S., An efficient weather forecasting system using artificial neural network. Int. J. Environ. Sci. Dev. **1**, 321–326 (2010) http://www.ijesd.org/papers/63-D472.pdf

63. El-Feghi, I., Zubia, Z. & Abozgaya, S. Efficient weather forecasting using artificial neural network as function approximator. *Int. J. Neural Netw. Adv. Appl.* **1**, 49–55 (2014).

64. Boopathi, G., Arockiasamy, S., Image compression: Wavelet transform using radial basis function (RBF) neural Network. *INDICON* **5241**, 340–344 (2012)

65. Virginia, E-D., Biometric identification system using a radial basis network. *Proc. 34th Annual IEEE International .Carnahan Conf. on Security Technology* pp. 47–51 DOI:https://doi.org/10.1109/CCST.2000.891165 (2000)

66. Haykin, S. *Neural Networks* (Macmillan Publishing, 1994).

67. Sharkawy, A.-N. Principle of neural network and its main types: Review. *J. Adv. Appl. Comput. Math.* **7**, 8–19. https://doi.org/10.15377/2409-5761.2020.07.2 (2020).

68. Hyvärinen, O. A probabilistic derivation of heidke skill score. *Weather Forecast.* **29**(1), 177–181. https://doi.org/10.1175/WAF-D-13-00103.1 (2014).

69. Banik, J. J., Hwang. H. S., Tropical cyclone intensity prediction using regression method andneural network. J. Meteorol. Soc. Jpn. **76**(5):711–717 (1998) DOI: https://doi.org/10.2151/jmsj1965.76.5_711

70. Chester, D. L. Why two hidden layers are better than one. *Int. Jt. Conf. Neural Netw.* **456**, 265–268 (1990).

71. Sharkawy, A.-N. Principle of neural network and its main types: Review. *J. Adv. Appl. Comput. Math.* **7**(1), 8–19. https://doi.org/10.15377/2409-5761.2020.07.2 (2020).

72. Sun, J., Weixing, D. & Niancai, S. A survey of KNN algorithm. *Inf. Eng. Appl. Comput.* https://doi.org/10.18063/ieac.v1i1.770 (2018).

73. Shadiq, M. A., Keoptimalan Naïve Bayes DalamKlasifikasi. **1**, 31, (2009)

74. Byers, H. R., & Battan, L. J. Some effects of vertical wind shear on thunderstorm structure. *Bull. Am. Meteorol. Soc.* **30**(5), 168–175 (1949).

75. Chung, C. Y. C., Kumar, V. R., *Knowledge Acquisition using a Neural Network for Weather Forecasting Knowledge-Based System Neural Computing and Applications*. Springer (1993) **1**, 215–223 DOI: https://doi.org/10.1007/BF01414951

76. Litta, A. J., Naveen, C. Francis radial basis function network for hourly surface temperature prediction. *Int. J. Eng. Res. Technol. IJERT RTPPTDM-2015 Conf. Proc.* (2015)

77. Zhou, C., Li, L., Wang, H., & Liu, S. Modeling thunderstorm based on paralleled and improved naive bayes. *2nd International Conf. on Computer Modeling, Simulation and Algorithm*. J. Phys. Conf. Ser. (2020) doi: https://doi.org/10.1088/1742-6596/1624/2/022025

78. Chengdong, Z., Leixiao, L., Hui, W. & Shuang, L. Modeling thunderstorm based on paralleled and improved Naïve Bayes. *J. Phys. Conf. Ser. Comput. Model. Simul. Technol.* **5892**, 1624. https://doi.org/10.1088/1742-6596/1624/2/022025 (2020).

Author contributions

S.B. implemented the methodologies, did the formal analysis and investigation, wrote original draft and H.C(B). designed the concept, reviewed and edited the manuscript, collected the fund and supervised all over.

Funding

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.