Available online at www.sciencedirect.com

ScienceDirect

Advances in Space Research 60 (2017) 1271–1282

ADVANCES IN SPACE RESEARCH (a COSPAR publication)

www.elsevier.com/locate/asr

A random forest algorithm for nowcasting of intense precipitation events

Saurabh Das [a,*], Rohit Chakraborty [b], Animesh Maitra [b]

[a] Center for Soft Computing Research, Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata 700108, India
[b] Institute of Radiophysics and Electronics, University of Calcutta, 92 A. P. C. Road, Kolkata 700009, India

Received 7 April 2016; received in revised form 8 March 2017; accepted 19 March 2017
Available online 24 March 2017

Abstract

Automatic nowcasting of convective initiation and thunderstorms has potential applications in several sectors including aviation planning and disaster management. In this paper, random forest based machine learning algorithm is tested for nowcasting of convective rain with a ground based radiometer. Brightness temperatures measured at 14 frequencies (7 frequencies in 22–31 GHz band and 7 frequencies in 51–58 GHz bands) are utilized as the inputs of the model. The lower frequency band is associated to the water vapor absorption whereas the upper frequency band relates to the oxygen absorption and hence, provide information on the temperature and humidity of the atmosphere. Synthetic minority over-sampling technique is used to balance the data set and 10-fold cross validation is used to assess the performance of the model. Results indicate that random forest algorithm with fixed alarm generation time of 30 min and 60 min performs quite well (probability of detection of all types of weather condition ~90%) with low false alarms. It is, however, also observed that reducing the alarm generation time improves the threat score significantly and also decreases false alarms. The proposed model is found to be very sensitive to the boundary layer instability as indicated by the variable importance measure. The study shows the suitability of a random forest algorithm for nowcasting application utilizing a large number of input parameters from diverse sources and can be utilized in other forecasting problems.

Keywords: Nowcasting; Machine learning; Random forest; Convective rain; Microwave radiometer

1. Introduction

Short term weather forecasting (or nowcasting) is a challenging task, particularly in case of convective or thunderstorm system. Due to its inherent stochastic nature, the development, growth and decay of such system is a very short time scale phenomenon, thus making it difficult to predict well in advance (Wilson et al., 1998). Nowcasting of thunderstorms and convective initiation has important application in disaster management, aviation flight planning and air traffic managements. Recent studies have focused on machine learning techniques in such problems utilizing a combination of observations from different sources, such as Numerical Weather Prediction (NWP) models, radar, satellite and ground based sensors to nowcast in a few hour time frames (Williams, 2013; Li et al., 2012; Marzano et al., 2007; Rivolta et al., 2006). However, identification of relative importance of the atmospheric parameters is one of the important aspect of such machine learning methods due to the availability of a large number of input variables (or features) and considering the computational cost.

* Corresponding author at: Center for Soft Computing Research, Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata 700108, WB, India.
E-mail address: das.saurabh01@gmail.com (S. Das).

http://dx.doi.org/10.1016/j.asr.2017.03.026
0273-1177/

The recent advances in machine learning techniques offer a variety of algorithms, particularly the development of Artificial Neural Network (ANN), Fuzzy information System (FIS), Support Vector Machine (SVM), decision trees, ensemble learning, etc. which provide an excellent opportunity for modelling such highly non-linear systems (Ortiz-García et al., 2015; Raghavendra and Deka, 2014; Wei, 2012, 2013; Asklany et al., 2011; Pankiewicz, 1995). The additional benefit of machine learning methods beside automated response in case of nowcasting application is that it can provide the relationships among various features which ultimately help to understand the underlying physical processes. Although ANN or SVM in such non-linear problems are widely applied (Babel et al., 2015; Grimes et al., 2003; Tapiador et al., 2004; Wei, 2012; Hsu et al., 1997), direct interpretation of the rules are difficult. On the other hand, classification using decision tree approaches is useful in modelling of complex relationships among variables with the added advantages of identifying the importance of each variable (Zhao and Zhang, 2008). Decision trees are computationally fast, easily understandable and do not require a prior knowledge of the data. Due to these reasons, the application of decision tree in case of nowcasting of convective precipitation is gaining popularity (Williams, 2013; Williams et al., 2008; Wei, 2013; Cai et al., 2009; Colquhoun, 1987). Further, due to its capability to identify the influential role of different features, it's rather advantageous to use decision trees in forecasting applications.

Ground based radiometer, which measures the brightness temperatures at different frequencies, has recently been demonstrated to be useful in nowcasting applications (Chakraborty et al., 2014; Madhulatha et al., 2013; Das et al., 2012; Chan, 2009; Chan and Lee, 2011). The radiometric measurements are directly related to the atmospheric temperature and humidity profile, which in turns points to the instability of the atmosphere. There are various instability indices derived from humidity and temperature profiles, which are frequently used in various combinations to predict the convective initiation (Koffi et al., 2007; Showalter, 1953; Wilson et al., 1998; Chakraborty et al., 2016). In a recent study, Das et al. (2012) and consequently, Chakraborty et al. (2014) showed that the brightness temperatures measured by a ground based multi frequency radiometer related to the oxygen and water vapor line can also be used directly to nowcast the convective rain. They reported accuracy of the model was 90% with an over prediction of 10%. However, in their model, brightness temperature of only two frequencies are utilized and hence some useful information from other frequencies were ignored which may help to reduce the false alarm rate. The alarm generation time was also not fixed in their model and it varies from a few minutes up to 40 min for different events. The present work is focused on these two issues.

In this present paper, a Random Forest (RF) based model utilizing all available frequencies has been attempted to reduce the over prediction and improve the lead time for alarm generation with the same experimental data set used by Chakraborty et al.(2014). A fixed time interval of 30 min and 60 min are chosen as alarm generation time. In addition, the relative importance of the predictors are also assessed using the RF model.

2. Methodology

2.1. Random forest technique

The decision tree is a non-linear and non-parametric supervised classification algorithm where the non-terminal nodes indicate the features and terminal nodes are outcomes. Random forest is a type of meta classifier which is based on an ensemble of unpruned trees (Breiman, 2001). The trees are generated by n random feature selection. The value of n is taken as the square root of the number of features. The random forest is generated by bootstrapping of training samples (bagging). A "bagged" training sample is initially obtained by selecting m elements (including duplicates) randomly from the training set consisting of m elements and replaced after each selection. On an average, each tree is trained on $\sim 2/3$ of the dataset. The out-of-bag (OOB) samples are then used to evaluate the performance of the trees. This also provides a measure of importance for each features. To calculate the feature importance, each feature is permuted across the out-of-bag observations. This is done for every tree and changes in the prediction error is estimated. If the new model's accuracy changes significantly, it would indicate that feature is important for the accuracy of the original model. To get a normalize measure of the variable importance, the ensemble average of this measure is then divided by the standard deviation value of overall ensemble.

The classification is finally done based on the majority vote of each generated tree in the ensemble. The possibility of over fitting of training data sets can be minimized in an ensemble method which is a significant drawback of a single decision tree. It has superior performance over single tree algorithm and less affected by the presence of noise in the data set (Breiman, 2001).

2.2. Experimental details

Data from a multi-frequency microwave radiometer (RPG-HATPRO) is used to study the capability of the RF model in nowcasting rain. The instrument is operating at Institute of Radio Physics and Electronics, University of Calcutta since 2009. The radiometer operates in two frequency bands, with seven frequencies in 22–31.4 GHz and seven frequencies in 51.26–58 GHz. The raw data are logged to a computer in binary format which is then converted to ASCII data through the radiometer software. The choices of these frequency bands are due to sensitivity of these frequencies to water vapor and temperature variations of the atmosphere. The 22–31.4 GHz band relates to

the pressure broadened weak water vapor line which can be used to get the water vapor profile. On the other hand, the atmosphere becomes transparent with frequencies away from the oxygen absorption line and can measure brightness temperatures (BT) originated from different layers of the atmosphere. Since the mixing ratio and temperature dependence of oxygen absorption is known, the well defined weighting function of the 51.26–58 GHz band can be utilized to get the temperature profile of the atmosphere.

Three types of calibration are used for this radiometer. While absolute calibration was done once in a year by plotting the calibration curve between antenna temperature and receiver voltages, noise injection calibration is performed in every measurement cycle. Further, the sky tipping calibration was used to remove the cosmic noises from the measurements. The BTs are measured continuously in every 3 s with a systematic gap of a few minutes after every 5 min interval. The time gap is used for internal calibration of the system. The range of the BTs measured by radiometer is 0–800 K with an accuracy of 0.5 K (Rose and Czekala, 2009). In addition, a rain sensor and a disdrometer are used to measure the rain occurrences. The brightness temperatures measurements can then be converted to temperature and humidity profiles using suitable inversion method (Ajil et al., 2010; Cimini et al., 2006; Haobo et al., 2011; Bleisch et al., 2011).

Direct performance assessment of radiometer is difficult and, therefore, the performance of the radiometer is assessed by comparing the radiometer derived profiles of humidity and temperature with that obtained from a local radiosonde and MODIS observations. Detailed comparisons of this radiometer with other instruments were provided in earlier studies by Chakraborty et al. (2014, 2016), and Majumder et al. (2015) and only the major observations are briefly discussed here. The results shows good performance of the radiometer against the radiosonde profiles with a correlation coefficient of 0.94 for temperature profiles and 0.83 for relative humidity profiles. It is observed that both the profiles are in good agreement, but, the difference between retrieves and actual humidity profile is relatively more than the temperature profiles. The retrieved temperature profiles have a small bias only above 2 km, whereas humidity profiles have a wet bias up to 5 km and a dry bias above that height (Chakraborty et al., 2014). The quadratic regression retrieval model, which was developed based on simulated brightness temperature from a large number of radiosonde observations, may be responsible for such biases in retrieving profiles. Comparison of precipitable water vapor measured by radiometer and MODIS on-board Aqua satellite have been made which also indicates a very good correlation coefficient of 0.9 (Majumder et al., 2015). In a recent study (Chakraborty et al., 2016), the atmospheric instability estimated from the same radiometer is compared with the values obtained from radiosonde measurements. The results also show close agreement with correlation coefficient of 0.8 for convective available potential energy, and, 0.9 for both K-index and precipitable water vapor.

Since every retrieval algorithm relies on some mathematical relation between BT and temperature or humidity profiles, each has its own limitations. For this reason, the raw brightness temperature data is used in the present model without any prior assumption of the relationship between brightness temperature and atmospheric parameters. Further detailed description of this radiometer, working principles and its performance assessment in different geographical regions may be obtained from Rose and Czekala (2009).

Another important consideration is that radiometer measurements saturates during heavy rain and may not be reliable in such conditions. An example of radiometric measurement of temperature and humidity profiles of 30 March 2011 is shown in Fig. 1. Two rain events occurred on this day with maximum rain rate of 20 mm/h. The temperature and humidity profile measured by the radiometer are shown in Fig. 1(a) and (b). The occurrences of the rain

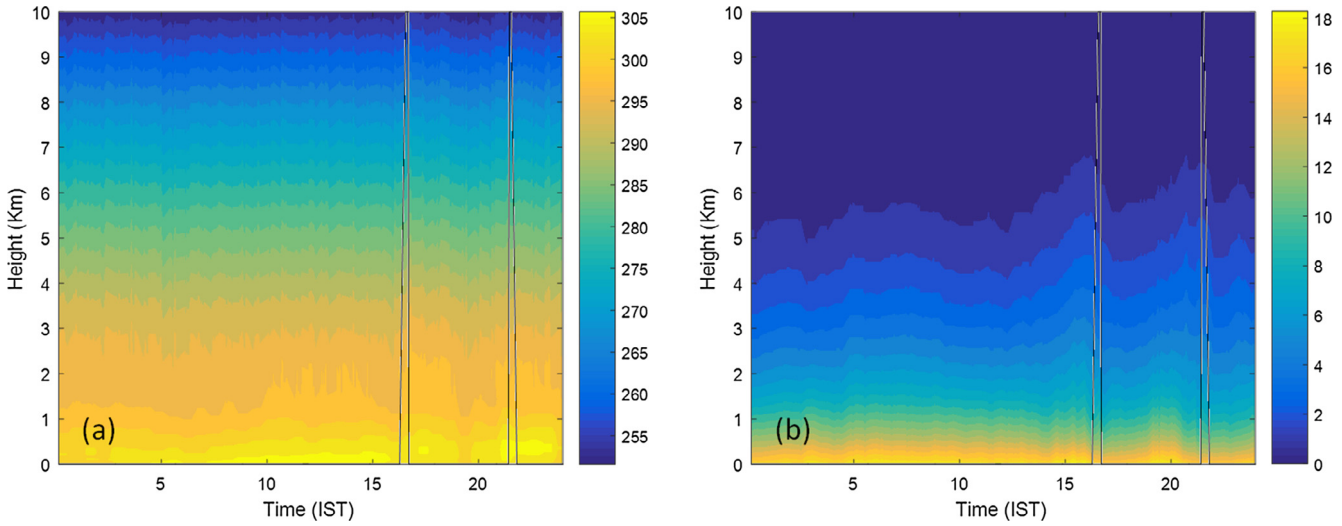

Fig. 1. (a) Temperature (in K) and (b) absolute humidity (in g/m^3) profiles as measured by radiometer. Vertical lines indicate the occurrence of rain.

Table 1
Attributes of the RF model.

Index number	Attribute
1	Average BT at 22.24 GHz
2	Average BT at 23.04 GHz
3	Average BT at 23.84 GHz
4	Average BT at 25.44 GHz
5	Average BT at 26.24 GHz
6	Average BT at 27.84 GHz
7	Average BT at 31.40 GHz
8	Average BT at 51.26 GHz
9	Average BT at 52.28 GHz
10	Average BT at 53.86 GHz
11	Average BT at 54.94 GHz
12	Average BT at 56.66 GHz
13	Average BT at 57.30 GHz
14	Average BT at 58.00 GHz
15	Standard deviation of BT at 22.24 GHz
16	Standard deviation of BT at 23.04 GHz
17	Standard deviation of BT at 23.84 GHz
18	Standard deviation of BT at 25.44 GHz
19	Standard deviation of BT at 26.24 GHz
20	Standard deviation of BT at 27.84 GHz
21	Standard deviation of BT at 31.40 GHz
22	Standard deviation of BT at 51.26 GHz
23	Standard deviation of BT at 52.28 GHz
24	Standard deviation of BT at 53.86 GHz
25	Standard deviation of BT at 54.94 GHz
26	Standard deviation of BT at 56.66 GHz
27	Standard deviation of BT at 57.30 GHz
28	Standard deviation of BT at 58.00 GHz
29	Range of BT at 22.24 GHz
30	Range of BT at 23.04 GHz
31	Range of BT at 23.84 GHz
32	Range of BT at 25.44 GHz
33	Range of BT at 26.24 GHz
34	Range of BT at 27.84 GHz
35	Range of BT at 31.40 GHz
36	Range of BT at 51.26 GHz
37	Range of BT at 52.28 GHz
38	Range of BT at 53.86 GHz
39	Range of BT at 54.94 GHz
40	Range of BT at 56.66 GHz
41	Range of BT at 57.30 GHz
42	Range of BT at 58.00 GHz

events are indicated by the vertical lines for reference. It can be seen that before the start of the event, atmospheric temperature and relative humidity changes significantly. However, during the rain event, the performance of the radiometer degraded. This means that the radiometer can forecast the rain with non-rainy period data, but may give false prediction of the future state with measurements taken during already raining condition. To avoid this problem, the present RF model is designed in such a way that it can be able to detect the already raining instances and does not use that measurement in nowcasting.

2.3. Training and test data set

The present study is based on the experimental measurements for the same rain events used earlier by Chakraborty et al. (2014). The training and test data set for the RF model is developed from the observations of ground based radiometer and a co-located rain sensor for the pre-monsoon (March–May) months of years 2011–2013 at Kolkata, India. Rain in the pre-monsoon season is mostly due to convective activity. The data set contains 62 rainy days, i.e. rain occurs on that day, at least for 5 min continuously and maximum rain rate exceeds 10 mm/h (Chakraborty et al., 2014). In total 107 distinct rain events are present in the data sets.

For the purpose of the present study, these data are grouped in a predefined time period. We study the performance of RF with two different lead time periods, namely, one hour and half an hour. Since the atmospheric condition prior to a convective system is expected to vary significantly, in addition to the average values of the parameters, we also consider their range and standard deviation for the said time periods. Thus, measurements of BT at 14 frequencies yields 42 input variables for the model with one output element of rain prediction. The input variables are summarized in Table 1.

Further, as we want to predict the possibility of rain in the future, there can be two possibilities of present states, whether it's already raining or no rain. So we assign three labels (0, 1, and 2) to the instances corresponds to three situations as follows:

1. Present → No Rain, Future → No Rain, Label → 0
2. Present → No Rain, Future → Rain, Label → 1
3. Present → Rain, Future → Any, Label → 2

In the third class, we do not bother whether it will rain or not in the future since radiometer is not a reliable instrument for nowcasting application in raining condition. So the prediction of future rain status from the measurements taken in already raining situation is out of scope. However, we need the 3rd class to make sure that the model can identify the already raining situations automatically. A schematic diagram of the proposed detection scheme is shown in Fig. 2. It is observed that about 71%, 8% and 21% of the total samples belongs to class label 0, 1 and 2, respectively in case of one hour lead time.

It is to be emphasized here that the non-rainy days are not included explicitly in the data set for model development. This is because the data set already contains large number of non-rainy samples and including non-rainy days will decrease the proportion of label 1 data (positive events) to a great extent. Hence, there will be a serious concern of over-fitting for majority class with such highly imbalanced data and class balancing will be difficult in such scenario. However, the non-rainy days are also used to assess the model performance as discussed in later sections.

2.4. Class imbalance problem

The radiometric data pertaining the pre-monsoon months, contains a very large number of non-rainy samples

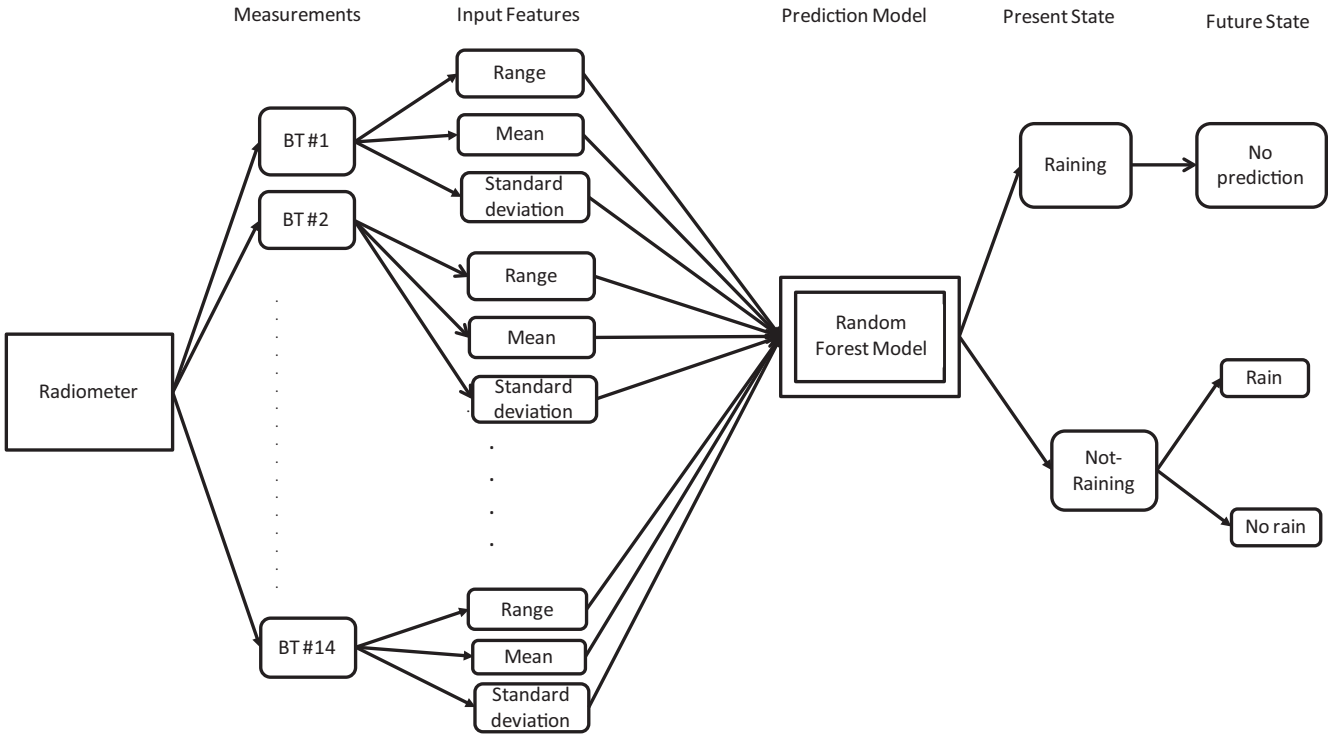

Fig. 2. Schematic diagram of the proposed detection process.

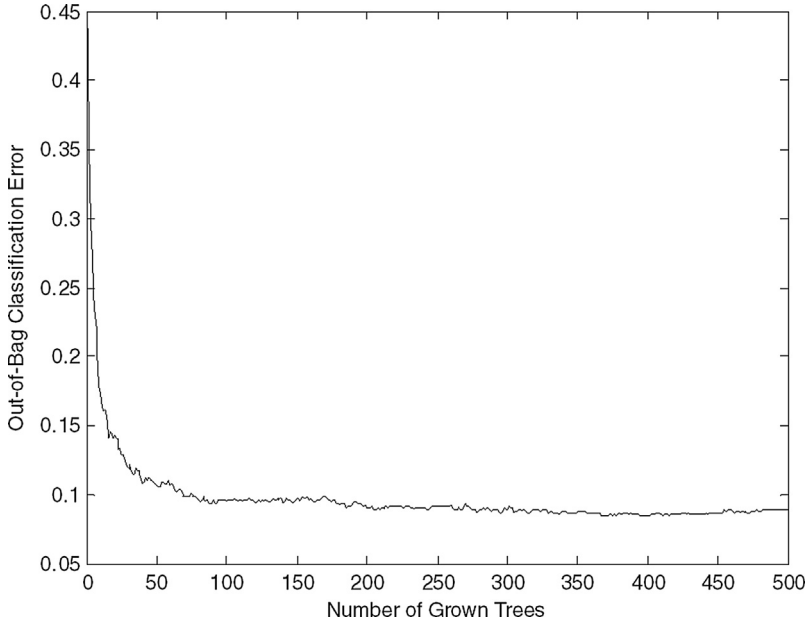

Fig. 3. Out-of-bag classification error with number of grown trees.

than the raining sample as pointed out in the previous section. Particularly, the class with label 1, which is of primary interest, is extremely low. The model developed based on this original data set can therefore have over fitting problems with class label 0. To address this problem, the data set is needed to be balanced, either by under-sampling of majority class or by over-sampling of a minority class. Since in under-sampling, a significant amount of information is discarded, over-sampling of data set is chosen. In the present case, Synthetic Minority Over-sampling Technique (SMOTE) is used for over-sampling the minority class which synthetically generate data using seven nearest neighbor points (Chawla et al., 2002). New minority class data is generated by interpolating several nearest

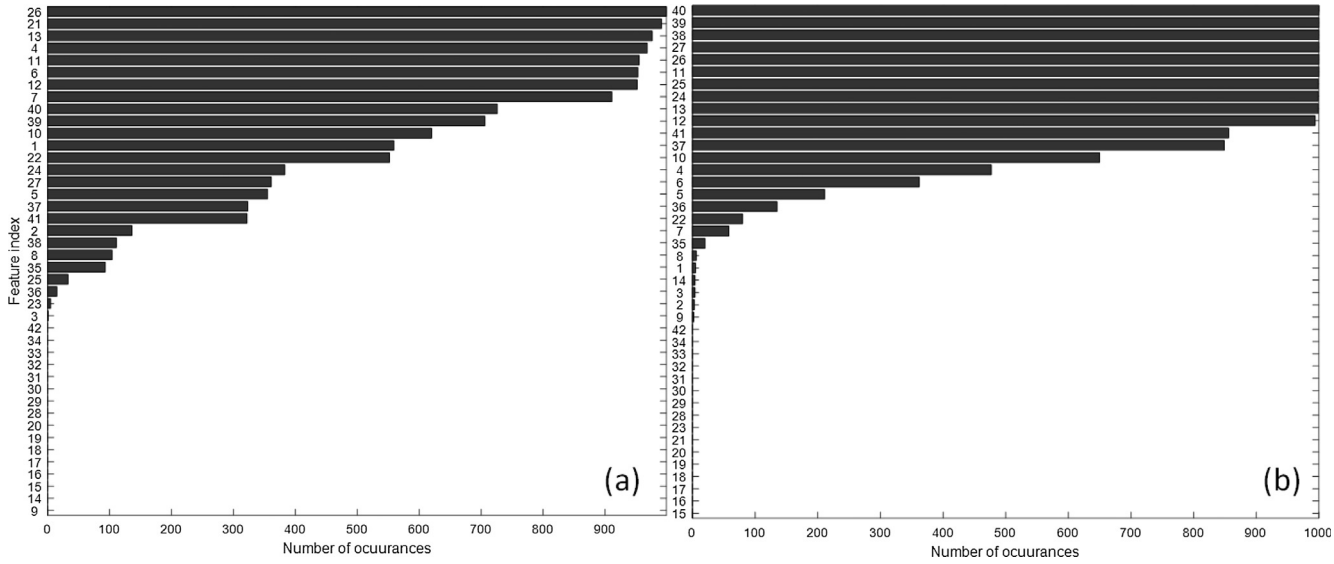

Fig. 4. Number of occurrences of different features in top 15 position for RF models in 100 repetitions of 10-fold cross validations with lead time of (a) 60 min and (b) 30 min.

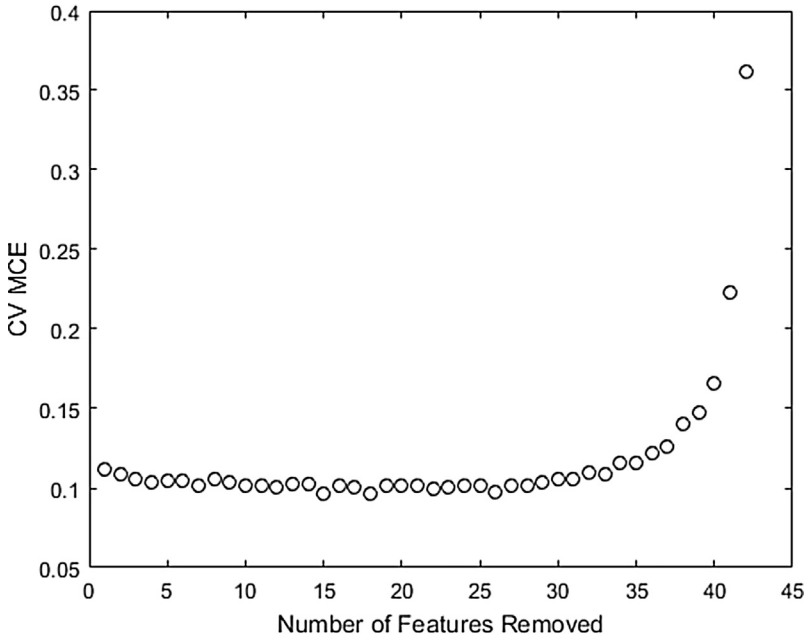

Fig. 5. Variation of mis-classification error with reduced numbers of features.

neighbor minority class data and hence, the over fitting problem is avoided.

2.5. Feature selection

The 42 feature used to develop the models are not of similar importance. The inherent feature ranking capability of RF provide a understanding of the relative importance of different features. However, selecting a subset of the features with similar model performance for prediction of the rain event is computationally advantageous. Hence, a wrapper based backward feature selection technique is used to obtain a optimal subset of the features which can also provide satisfactory performance.

2.6. Validation

Since the aim of the present study is to develop a model which can forecast the rain with a fixed lead time, the correct forecast or hit indicate the instances which is predicted as well as actually rains (i.e. label 1 is predicted as label 1). Similarly the misses are when it actually rains, but our model either say it's non-raining or already raining instances (i.e. label 1 is predicted as label 0 or label 2).

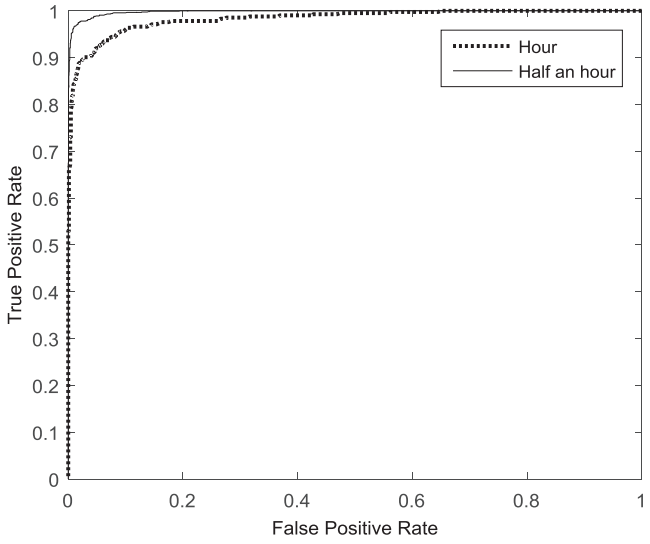

Fig. 6. Receiver operating characteristic (ROC) curve for RF model with two different lead times.

The false alarm generates when the model says it will rain, but actually it's non-raining or already raining (i.e. label 0 or label 2 is predicted as label 1).

A standard 10-fold cross validation technique is used to assess the performance of the models for lead times of 30 min and 60 min and the experiments are repeated 100 times. The data is stratified so that each fold has all class labels in original proportion. Hence, the model is tested with ~1/10 of the each sample class in each cross validation. To evaluate the model performance in our case, several performance criteria are used, namely the probability of detection (POD), proportion correct (PC), false alarm rate (FAR) and threat score (TS). The POD is defined as the number of events correctly forecasted divided by the total number of events observed. The PC is defined as the correct forecasts of all types divided by the total sample size. FAR is defined as the number of false alarms divided by the total number of hits and false alarms. TS is defined as the number of hits divided by the total number of hits, false alarms and misses.

As already mentioned, the non-rainy days are not included in development of the RF models and hence, the cross-validation of the model is conditioned to rainy days only. However, the models are also tested with 174 non-rainy days separately. Since these days have no positive cases, the performance of the models are evaluated

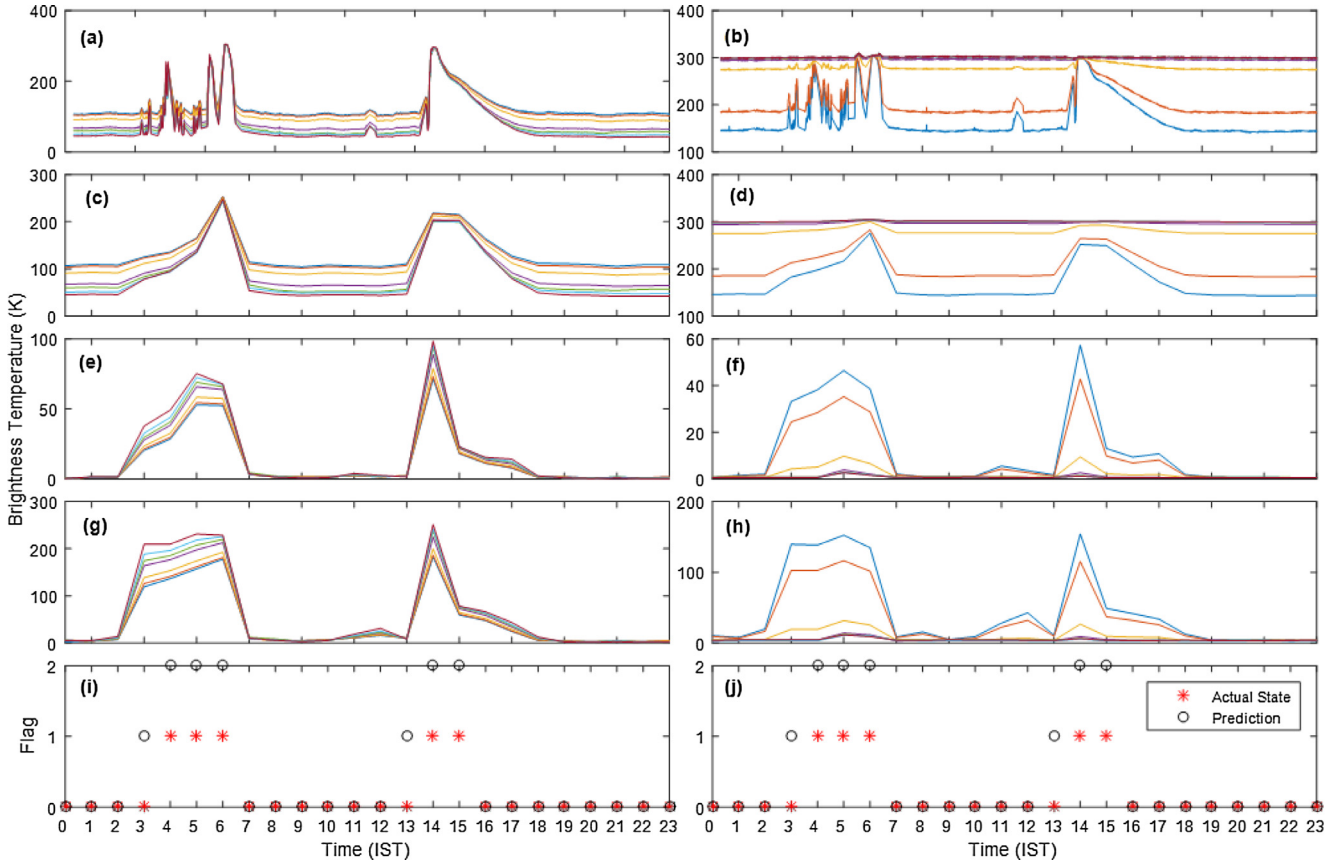

Fig. 7. An example of successful prediction by the RF model with 60 min lead time. Times series of BT are shown for (a) 22–31 GHz and (b) 51–58 GHz. Average BT values during 60 min periods are shown for (c) 22–31 GHz and (d) 51–58 GHz. Standard deviation of BTs during 60 min periods are shown for (c) 22–31 GHz and (d) 51–58 GHz. Range of BTs during 60 min periods are shown for (c) 22–31 GHz and (d) 51–58 GHz. The actual and predicted states are shown in (i and j).

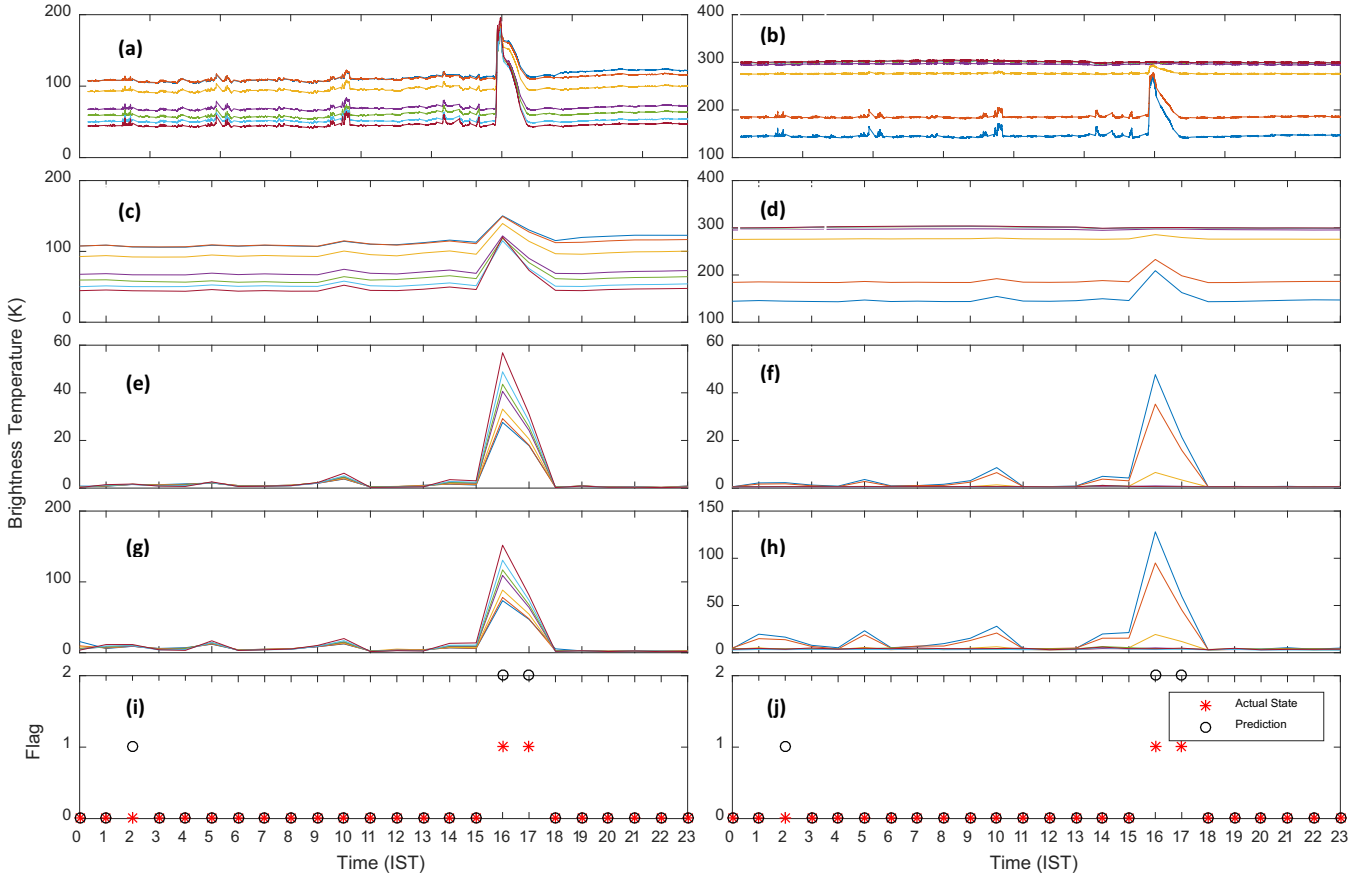

Fig. 8. An example of unsuccessful prediction by the RF model with 60 min lead time. Times series of BT are shown for (a) 22–31 GHz and (b) 51–58 GHz. Average BT values during 60 min periods are shown for (c) 22–31 GHz and (d) 51–58 GHz. Standard deviation of BTs during 60 min periods are shown for (c) 22–31 GHz and (d) 51–58 GHz. Range of BTs during 60 min periods are shown for (c) 22–31 GHz and (d) 51–58 GHz. The actual and predicted states are shown in (i–j).

on the basis of false alarm with respect to non-rainy samples.

3. Results

The present RF model is developed with 500 trees which are generated with random feature selection. The fixed alarm times of 30 min and 60 min are tested with these specifications. Fig. 3 indicates the OOB error variation with the number of trees in the model. It can be noted that, initially the OOB error falls fast with an increase in the number of trees and almost saturates after ~100. As the number of trees increased further, the performance of the model doesn't improve significantly.

The variable importance is estimated using the out-of-bag samples and an inherent testing of the model performance. It is observed that all 42 features are not of similar importance to the model. It is to be noted here that the variable importance measured by the RF model is not fixed and the ranking of the individual variable may vary as we repeat the experiment many times. This is because every time a new model is developed with a different sample set. So to study the stability of the features' rankings, the frequency distribution of each parameters are studied. In Fig. 4, the frequency of occurrence of different features in top 15 positions are shown. It is observed that only a few parameters out of 42 parameters are in top 15 position, however, the number of parameters are more than 15. This may be because there exist correlation between some of the parameters, and hence these parameters also appears as influential parameters.

In any models, reducing the number of input features while maintaining the performance is often computationally advantageous. As the variable importance ranking indicate that there exist some features which are correlated, an optimal subsets of feature can be selected for developing the model. To select a subset of the features, a wrapper based backward feature selection method is used. In Fig. 5, the variation of misclassification error (MCE) in 10-fold cross-validation is plotted against the number of features removed for RF model with 60 min lead time. It is seen that the model performance degraded drastically for input numbers of features less than 10 in the RF model. The optimal number of features is taken as 10 and the feature indexes are 1,4,6,13,17,26,35,38,39,41.

It can be seen that the average values of BTs for frequencies 22.24 GHz, 25.44 GHz, 27.84 GHz and 57.30 GHz appears in the list whereas the standard deviation of BTs

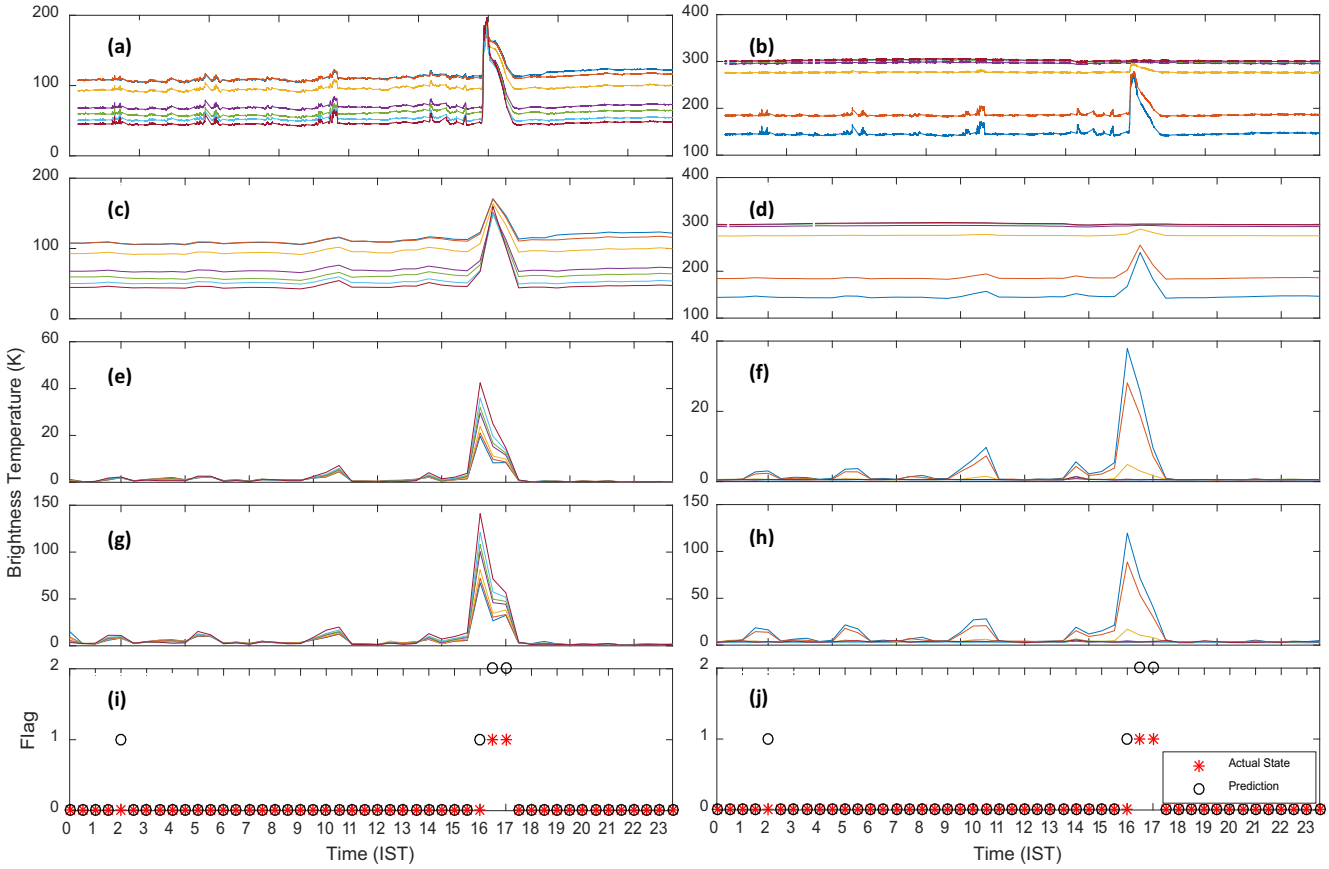

Fig. 9. The improvement in prediction of the same event by the RF model with 30 min lead time. Time series of BT are shown for (a) 22–31 GHz and (b) 51–58 GHz. Average BT values during 30 min periods are shown for (c) 22–31 GHz and (d) 51–58 GHz. Standard deviation of BTs during 30 min periods are shown for (c) 22–31 GHz and (d) 51–58 GHz. Range of BTs during 30 min periods are shown for (c) 22–31 GHz and (d) 51–58 GHz. The actual and predicted states are shown in (i–j).

for frequencies 23.84 GHz and 56.66 GHz are present in the model. The range values of BTs of frequencies 31.40 GHz, 53.86 GHz, 54.94 GHz and 57.30 GHz are in the list. Since the frequency range 51–58 GHz is used for the atmospheric temperature profiling, the changes in the brightness temperatures of these frequencies are directly related to the atmospheric temperature, and hence to the instability. Further, radiative transfer theory indicates that different frequencies can have different weighting functions which peaks for different layers of the atmosphere (Ullaby et al., 1982). As we move from 51 GHz to 58 GHz, the maximum weight shifts to the boundary layer region and 57.30 GHz has the peak normalized difference weights in the lowest region of the boundary layer (Ullaby et al., 1982; Rose and Czekala, 2009). Changes in temperature profiles of the boundary layer are primarily captured by the range and standard deviation values of the BT at 57.30 GHz and 56.66 GHz. Thus the instability developed in the boundary layer region is captured by the proposed technique, which is one of the important region of convective initiation (Fabry, 2006; Bennett et al., 2006). On the other hand, 22–31 GHz are sensitive to the water vapor and enables these frequencies to react sharply to the changes in atmospheric water content. Hence, average values of the BT for frequencies in 22–31 GHz, which are identified as important variables in this model, actually relate to the amount of total water vapor in the atmosphere. Changes in the liquid water content during the specified period is considered in the model by the range values of BT at 31.40 GHz. The variable importance figure also indicates that both the absolute value as well their relative changes during the pre-defined time period is important is nowcast the convective phenomena. It's worth mentioning here that the model previously proposed by Chakraborty et al. (2014) also utilized the BT values of 22.24 GHz and 58 GHz to predict the rain occurrence.

The receiver operating characteristics (ROC) curves for 30 min and 60 min lead times are shown in Fig. 6. ROC curve is generated by plotting the true positive rate against the false positive rate for the different possible cut points of a model. ROC curve is rather more meaningful only in the case of binary classifier, however, in case of three class problems like ours it gives only an indication of the model performance. In the present case, ROC curve is generated considering the label 1 as 'true' and both label 0 and label 2 as 'false'. We can note that the model with half an hour lead time performs better than the model with one hour lead time. This is not unexpected considering the nature of convective rain which sometimes can develop very fast.

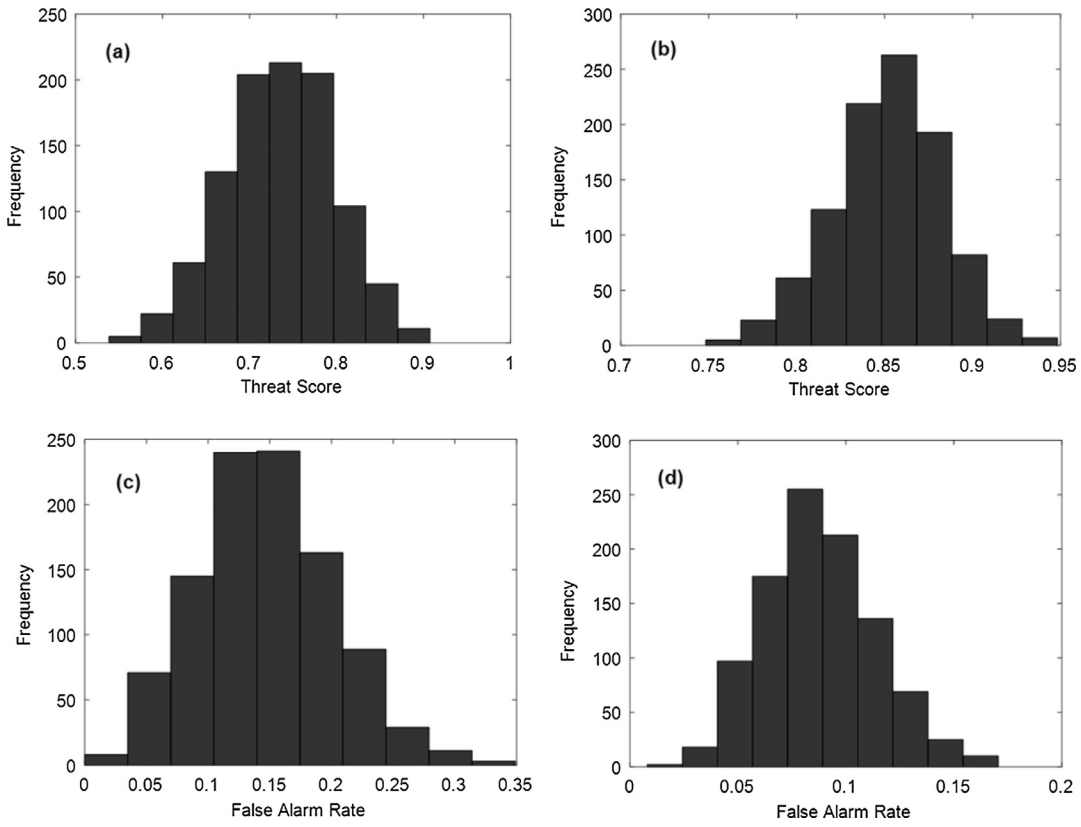

Fig. 10. Distribution of threat score of obtained in 100 repetition of 10-fold cross-validation with a lead time of (a) 60 min and (b) 30 min. Respective FAR distributions of the experiments are shown in (c) and (d).

To illustrate the rationale behind the performance of the proposed technique, Figs. 7 and 8 are shown. In Fig. 7, the radiometric measurements are shown for 30 May, 2011 in the form of time series of the input features. The brightness temperature values of 22–31 GHz band are plotted in Fig. 7(a) whereas in Fig. 7(b) the same values for 51–58 GHz are shown. One can note the signal has many small and large fluctuations and also have systematic gaps due to the calibration processes. In, Fig. 7 (c) and (d), the average value of the parameters for each one hour period is shown. Similarly in Fig. 7(e) and (f) shows the standard deviation and in Fig. 7(g) and (h) indicates the range of the BTs during each one hour periods. Fig. 7(i) and (j) shows the actual state of the atmosphere (*) and the predicted state (o). The non-rainy periods are indicated by 0 value whereas the raining periods are indicated by value 2. The proposed technique generates alarm on 3 and 13 h by changing its state to 1 from 0 indicating that there will be rain in next hour. The proposed model also successfully detects the already raining state during 4–6 h and 14–15 h by changing the value from 1 to 2. The rest of the times it shows 0 values, indicating no rain probability in next one hour.

In Fig. 8, another example of 4 June, 2011 is shown where the proposed model with one hour time gap does not able to predict the state of next hour successfully for every instants. It generates a false alarm during 2 h and also does not generate the alarm during 15 h as expected. A close look to the input parameters indicate that there are some changes in the parameters during 2 h which may trigger the false alarm. Similarly, the changes during 15 h does not seem sufficient to trigger the alarm in present model.

In Fig. 9, the same day is shown with RF model of 30 min lead time. In this case also the false alarm is generated at 2 h, but the model now able to predict the rain successfully during 16 h. This is because the signature of rain is weak due to longer time frame in earlier case which is now detected due to small time window. It also indicate that reducing alarm generation time keeping other model configuration same will further improve the threat score, but may not able to eliminate all false alarm effectively due to natural fluctuations in BT time series.

The frequency distribution of threat score and false alarm rate for 100 repetition of 10-fold cross validation is shown in Fig. 10. It can be seen that there exist some variability in the performances of the models in different experiments, but the overall performance of 30 min lead time is higher than the 60 min lead time. The frequency distribution of FAR of non-rainy days for these two model are also shown in Fig. 11. Though the non-rainy days are not considered explicitly in developing the models, the FAR of non-rainy days are also comparable with that of rainy days.

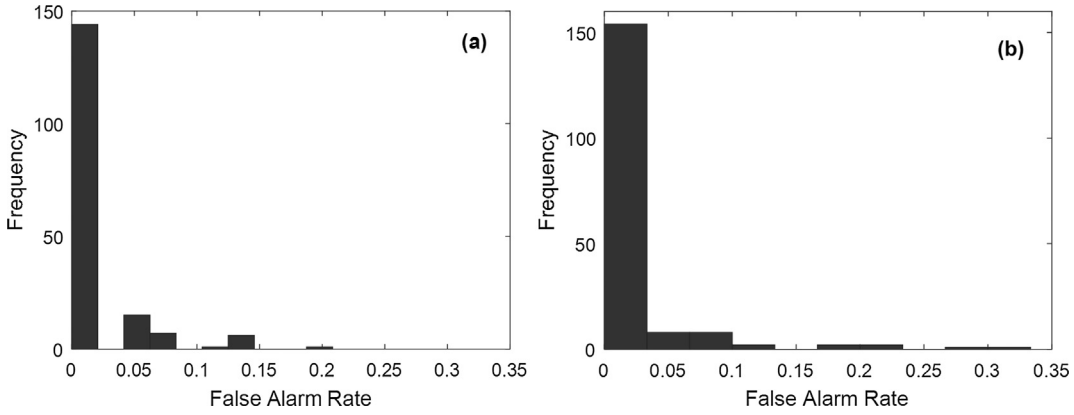

Fig. 11. FAR for non-rainy days in RF models with lead time of (a) 60 min, and (b) 30 min.

Table 2
Performances of different models in rainy days.

Model	FAR	TS	POD	PC
Hourly RF	0.15	0.74	0.86	0.90
Half hourly RF	0.09	0.85	0.91	0.95
Chakraborty et al. (2014)	0.10	–	0.90	–

'–' indicates no result.

The average performance parameters of the models are summarized in Table 2. This also indicates that the performance of the model is comparatively better for half an hour lead time than an hour lead time. In both the cases POD of rain events is ~90%. The correct prediction of all weather conditions is 90% and 95% for an hour and half an hour lead time, respectively. The threat score is 74% and 85% for these two lead times, respectively. The false alarm rates are also very low with both the models in rainy as well in non-rainy days. Chakraborty et al. (2014) obtained similar POD of 90%, but the lead time was varying from a few minutes to ~40 min. The results indicate although half an hour lead time gives an optimal prediction of the convective rain, the RF model for an hour lead time also performs reasonably well.

4. Conclusion

Machine learning algorithms are now-a-days used in various real life problems including atmospheric sciences. The application of the random forest algorithm has been tested for nowcasting application with a ground based radiometer. RF provides an additional advantage to identify most useful features from a large number of input features. 10 features are identified as the optimal set for predicting the rain instances. The model is tested with two different lead times, namely half an hour and one hour. The present models indicate reasonable good performances with low false alarm probability in both the configurations. It is observed that the POD of rain events are ~90% for both the models with ~10% false alarms. The threat score are 85% and 74% for half an hour and one hour lead time, respectively. The threat score improves with reducing lead time as convective initiation is a relatively short time scale phenomena. Integrating satellite and weather radar data in this model is expected to further improve the performance of this model and will be attempted in the future.

Acknowledgement

The work has been supported by the Department of Science and Technology, India under INSPIRE Faculty scheme (DST/INSPIRE/04/2014/002492). The radiometer was procured under the Space Science Promotion Scheme (E 33013/3/2009-V) at the University of Calcutta funded by Indian Space Research Organization. The authors are also thankful to Prof. Animesh Maitra, University of Calcutta for providing the ground based radiometer data and the two anonymous reviewers for their constructive suggestion for the betterment of the study.

References

Ajil, K.S., Thapliyal, P.K., Shukla, M.V., Pal, P.K., Joshi, P.C., Navalgund, R.R., 2010. A new technique for temperature and humidity profile retrieval from infrared-sounder observations using the adaptive neuro-fuzzy inference system. IEEE Trans. Geosci. Rem. 48, 1650–1659.

Asklany, S.A., Elhelow, K., Youssef, I.K., El-Wahab, M.A., 2011. Rainfall events prediction using rule-based fuzzy inference system. Atmos. Res. 101, 228–236.

Babel, Mukand S., Badgujar, Girish B., Shinde, Victor R., 2015. Using the mutual information technique to select explanatory variables in artificial neural networks for rainfall forecasting. Meteorol. Appl. 22 (3), 610–616.

Bennett, L.J., Browning, K.A., Blyth, A.M., Parker, D.J., Clark, P.A., 2006. A review of the initiation of precipitating convection in the United Kingdom. Q. J. R. Meteorol. Soc. 132, 1001–1020.

Bleisch, R., Kämpfer, N., Haefele, A., 2011. Retrieval of tropospheric water vapour by using spectra of a 22 GHz radiometer. Atmos. Meas. Tech. 4, 1891–1903.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Cai, H., Kessinger, C.J., Ahijevych, D.A., et al., 2009. Nowcasting oceanic convection for aviation using random forest classification. In: 16th Conference on Satellite Meteorology and Oceanography. American Meteorological Society, Phoenix, AZ.

Chakraborty, R., Das, S., Jana, S., Maitra, A., 2014. Nowcasting of rain events using multi-frequency radiometric observations. J. Hydrol. 513, 467–474.

Chakraborty, R., Das, S., Maitra, A., 2016. Prediction of convective events using multi-frequency radiometric observations at Kolkata. Atmos. Res. 169, 24–31.

Chan, P.W., 2009. Performance and application of a multi-wavelength, ground-based microwave radiometer in intense convective weather. Meteorol. Z. 18, 253–265.

Chan, P.W., Lee, Y.F., 2011. Application of ground based multi-channel microwave radiometer to the alerting low level wind shear. Meteorol. Z. 20 (4), 423–429.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

Cimini, D., Hewison, T.J., Martin, L., Güldner, J., Gaffard, C., Marzano, F., 2006. Temperature and humidity profile retrievals from ground based microwave radiometers during TUC. Meteorol. Z. 15, 45–56.

Colquhoun, J.R., 1987. A decision tree method of forecasting thunderstorms, severe thunderstorms and tornadoes. Wea. Forecast. 2, 337–345.

Das, S., Chakraborty, R., Talukdar, S., Maitra, A., 2012. Nowcasting of tropical rain using dual frequency atmospheric brightness temperatures at Kolkata. In: 5th International Conference on Computers and Devices for Communication (CODEC), vol. 1(4), pp. 17–19. http://dx.doi.org/10.1109/CODEC.2012.6509338.

Fabry, F., 2006. The spatial variability of moisture in the boundary layer and its effect on convection initiation: project-long characterization. Mon. Wea. Rev. 134, 79–91.

Grimes, D.I.F., Coppola, E., Verdecchia, M., Visconti, G., 2003. A neural network approach to real time rainfall estimation for Africa using satellite data. J. Hydrometeorol. 4 (6), 1119–1133.

Haobo, T., Mao, J., Chen, H., Chan, P.W., Wu, D., Li, F., Deng, T., 2011. A study of a retrieval method for temperature and humidity profiles from microwave radiometer observations based on principal component analysis and stepwise regression. J. Atmos. Ocean. Technol. 28, 378–389.

Hsu, K.L., Gao, X., Sorooshian, S., Gupta, H.V., 1997. Precipitation estimation from remotely sensed information using artificial neural networks. J. Appl. Meteor. 36, 1176–1190.

Koffi, E.N., Schneebeli, M., Brocard, E., Mätzler, C., 2007. The Use of Radiometer Derived Convective Indices in Thunderstorm Nowcasting. Mätzler Research Report Nr. 2007-02-MW, March 2007. Bern University.

Li, N., Wei, M., Niu, B., Mu, X., 2012. A new radar-based storm identification and warning technique. Meteorol. Appl. 19 (1), 17–25.

Madhulatha, A., Rajeevan, M., Venkat Ratnam, M., Bahte, J., Naidu, C. V., 2013. Nowcasting severe convective activity over southeast India using ground based microwave radiometer observations. J. Geophys. Res. 118, 1–13.

Majumder, S., Das, S., Maitra, A., 2015. Study of tropospheric delay over Indian region from MODIS, NCEP/NCAR data and ground based water vapor measurements at Kolkata. Adv. Space Res. 56 (6), 1115–1124.

Marzano, F.S., Rivolta, G., Coppola, E., Tomassetti, B., Verdecchia, M., 2007. Rainfall nowcasting from multisatellite passive-sensor images using a recurrent neural network. IEEE Trans. Geosci. Rem. Sens. 45 (11), 3800–3812.

Ortiz-García, E.G., Salcedo-Sanz, S., Casanova-Mateo, C., 2014. Accurate precipitation prediction with support vector classifiers: a study including novel predictive variables and observational data. Atmos. Res. 139, 128–136.

Pankiewicz, G.S., 1995. Pattern recognition techniques for the identification of cloud and cloud systems. Meteorol. Appl. 2 (3), 257–271.

Rose, Th., Czekala, H., 2009. RPG-HATPRO Radiometer Operating Manual, Radiometer Physics GmbH, Version 7.99.

Raghavendra, S.N., Deka, P.C., 2014. Support vector machine applications in the field of hydrology: a review. Appl. Soft Comput. 19, 372–386.

Rivolta, G., Marzano, F.S., Coppola, E., Verdecchia, M., 2006. Artificial neural-network technique for precipitation nowcasting from satellite imagery. Adv. Geosci. 7, 97–103.

Showalter, A.K., 1953. Stability index for forecasting thunderstorms. Bull. Am. Meteorol. Soc. 34, 250–252.

Tapiador, FJ., Kidd, Levizzani, CV., Marzano, FS., 2004. A neural networks-based PMW–IR fusion technique to derive half hourly rainfall estimates at 0.1 resolution. J. Appl. Meteorol. 43 (4), 576–594.

Ullaby, F.T., Moore, R.K., Fung, A.K., 1982. Microwave remote sensing: active and passive. In: Volume Scattering and Emission Theory, Advanced Systems and Applications, vol. III. Artech House Inc., Dedham, Massachusetts.

Wei, Chih-Chiang, 2012. Wavelet support vector machines for forecasting precipitation in tropical cyclones: comparisons with GSVM, regression, and MM5. Wea. Forecast. 27, 438–450.

Wei, Chih-Chiang, 2013. Soft computing techniques in ensemble precipitation nowcast. Appl. Soft Comput. 13 (2), 793–805.

Wilson, J.W., Crook, A.N., Mueller, C.K., Sun, J., Dixon, M., 1998. Nowcasting thunderstorms: a status report. Bull. Am. Meteorol. Soc. 79, 2079–2099.

Williams, J.K., Ahijevych, D.A., Kessinger, C.J., et al., 2008. A machine-learning approach to finding weather regimes and skillful predictor combinations for short-term storm forecasting. In: 13th Conference on Aviation, Range and Aerospace Meteorology. American Meteorological Society, New Orleans, LA.

Williams, J.K., 2013. Using random forests to diagnose aviation turbulence. Mach. Learn. 95, 51–70.

Zhao, Y., Zhang, Y., 2008. Comparison of decision tree methods for finding active objects. Adv. Space Res. 41 (12), 1955–1959.