Machine Learning Classifiers for Thunderstorm Nowcasting: A Study of Support Vector Machines and Random Forest Approaches against Thermodynamic Indices

Madhusudhan HS[1], Ajay Kumar[2]

[1] Lincoln University College, Malaysia; Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India.

[2] Lincoln University College, Malaysia; School of Computer Science & Engineering, IILM University, Greater Noida, Delhi NCR, India.

[1] pdf.madhusudhan@lincoln.edu.my, [2]pdf.ajaykumar@lincoln.edu.my

---

Abstract

Performance of machine learning classifiers in thunderstorm nowcasting through the study of support vector machines and random forest methods in comparison with thermodynamic indices is presented in this work. Consequently, the ability to forecast mesoscale convectional systems (MCS) including severe thunderstorm events remains a formidable challenge in the day-to-day weather practice. The main issue: These storms evolve very rapidly, and its behaviour is far more complicated than a simple rule-based method can capture and hence forecasting them is a major difficulty. Practically we tend to rely on a limited number of threshold tests, such as CAPE or the Lifted Index. The latter numbers are directly out of large Numerical Weather Prediction (NWP) models. Unfortunately, they are prone to indicating too many storms that do not occur (large False Alarm Ratios) and failing to see the causes of thunderheads appearing. This is why we are proposing a robust Machine-Learning (ML) solution to nowcasting. We are comparing two common models, the Support vector Machines (SVM) and the random forests (RF). The point is to find out which one is more successful in telling the distinguish between storm and non-storm environments. The models were trained with a huge set of features, 21 variables of which are all kinematic and thermodynamic quantities inspired by the SALAMA framework. Then we compared their performance with the older stability index thresholds which are still used by most forecasters particularly across the Indian subcontinent. The findings were clear. According to previous studies, it is proposed that Random Forests may have greater Critical Success Index (CSI) and fewer false alarms than SVMs and simple thermodynamic indices. This is just a demonstration that tree-based forests are helpful indeed. Lastly, we investigated the question of why the models are correct (or incorrect) and discovered that the most significant predictors, in fact, were Total Precipitable Water (TPW) and mid-level vertical velocity. These were more helpful than the usual CAPE values we would otherwise contemplate, and this implies that we are likely over-relying on CAPE alone.

Altogether, the research indicates that Random Forests is a powerful predictor of thunderstorms at present- and it also pushes us to extend our metrics which we believe in to make persuasion predictions.

Keywords: Thunderstorm; Machine learning; SVM; Random Forest

Introduction

a. Background and Motivation

Mesoscale weather to reckon with is thunderstorms and these weather conditions can completely cause severe economic damage, flights, and even cause the death of individuals in the form of lightning, heavy winds and flash floods. The storm begins as cumulus congestus, develops into the mature phase, and dissipates, all in approximately one or three hours. That allows one to predict them as a nowcast (06 hours) problem.

Although there is much improvement in the synoptic-scale NWP models, they are still incapable of pin-pointing the location and exact moment of a convective occurrence. The former is largely due to spin-up lag, in which the model requires time to equilibrate, and due to simplified physics in the parameterizations (such as cloud microphysics). And we must have after-processes which can make raw NWP data read like storm warning alerts.

b. Limitations of the Existing Methods

The forecast offices tend to use derived stability indices determining the potential of storms. Atmospheric instability is a proxy that is represented by such indices as CAPE, K-Index and Lifted Index.

The Issue of Adequacy: According to Umakanth et al. (2021), these indices are needed but not sufficient. A high CAPE (>3000/kg) contains plenty of energy, but nothing to trigger it, such as mountain lift, a front, surface heating, etc. may not generate convection.

The False Alarm Problem: Since thermodynamic instability is frequently ubiquitous in a region (consider the entire pre-monsoon belt in India), threshold-based techniques generate a great deal of false alarms. That leads to the so-called warning fatigue among the population and pilots who then grow to disregard the valid warnings because of the excessive number of false positives.

c. Objectives and Contribution

The goal of this paper is to shift from static thresholds to dynamic and data-driven projections. With machine learning, we will:

- **Construct a Pixel-Wise Classification Model:** Each grid point in an NWP domain now serves as a training sample, and the rich set of features is configured by the SALAMA framework (Yousefnia et al., 2024).
- **Compare Algorithm Performance:** Compare Algorithms Put a support-vector machine and a random forest side by side and see which is more successful with the noisy, unbalanced and collinear meteorological data.
- **Determine Feature Significance:** Extend beyond a black -box model to measure the significance of the atmospheric variables that most impact the predictions, providing new information that is consistent with meteorological theory.

Methodology

This paper attempts to combine physical meteorology and the field of computational intelligence. Rather than applying a single threshold, we construct a multi-dimensional space of features where the variables interact to determine whether a storm occurs.

a. Proposed System Architecture

We have established a scalable and reproducible modular pipeline. We do not consider the model as a black box and instead concentrate on the engineering of the physical features before attempting any type of feeding into the model.

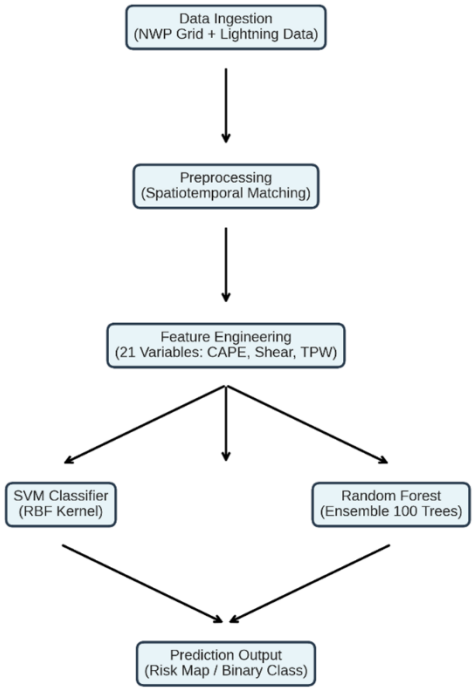

Figure 1. Architecture of the proposed system

Figure 1 indicates the four stages of the pipeline:

1. **Data Ingestion:** The raw GRIB files are downloaded by us, based on NWP models (such as GFS/WRF or Reanalysis proxies), and are combined with observational lightning data.
2. **Preprocessing & Balancing:** Matching the data in space and time. Thunderstorms are also rare, hence to over sample the majority, we oversample the minority using SMOTE.
3. **Model Training:** SVM (RBF kernel) and Random Forest (100 trees) are run in parallel with the balanced data.
4. **Validation:** We validate the models on a hold-out set by the use of skill scores based on detecting rare events (CSI, HSS).

The existing data will be obtained and prepared following the steps outlined in 2.2 Data Acquisition and Preparation.

b. Data Acquisition and Preparation

Data of Numerical Weather Prediction (NWP) Gridded data on the atmosphere is the primary input. As a proxy of the actual NWP output, we rely on the high-resolution ERA5 reanalysis (following the SALAMA framework) (Yousefnia et al., 2024). The grid is 0.25° × 0.25° (≈25 km). Snapshots are done hourly to capture the diurnal peak of the convection that occurs at about 15:00-18:00 Local Time.

This serves as the ground truth dataset for training and validation.

Lightning Observation Data (Ground Truth) It is not easy to define what is a thunderstorm. Although radar reflectivity is the topic of interest to many researchers, Yousefnia et al. (2024) claim that actual lightning provides a better yes or no indication of deep convection.In this paper, we have used observed lightning strikes as the ground truth on which to train. The lightning data were obtained in the form of global detection networks and overlaid on the same 0.25^{o} grid used by the numerical weather prediction models.When a grid cell has any lightning stroke within the forecast window, it is marked as Storm (1). Otherwise it is rated as No Storm (0)

c.Feature Selection and Physical Justification

The presented methodology is organized into two parts: feature selection and physical justification. Feature engineering is key. We had 21 predictors separated into three physical groups, according to Umakanth et al. (2021) on stability of the Indian region.

i. Thermodynamic Stability Indices - these is the measures of the potential buoyancy of air parcels:

- **CAPE (Convective Available Potential Energy):** this is the amount of energy that a rising parcel can tap.
- **Lifted Index (LI):** this is the temperature of the environment and that of a lifted parcel at 500hPa; negative values indicate instability.
- **K -Index (KI) and Total Totals Index (TTI):** they are superior in showing tropical thunderstorms as compared to mid-latitude indices.

ii. Moisture Parameters - moisture is required in storminess.

- **Total Precipitable Water (TPW):** the mass of column water vapor; large TPW releases latent heat that causes updrafts.
- **Specific Humidity (q):** taken out at 850 0 hPa and 700 0 hPa.
- **Dew Point Depression:** indicates the proximity to the saturation of the air; underdeveloped mid-level depressions may lead to indication of dry air intrusion, which enhances the amount of down drafts through evaporative cooling.

iii. Kinematic and Dynamic Variables - these demonstrate the movement of the air.

- **Vertical Velocity (Omega):** 500 hPa value; a negative pressure coordinate value will mean upward velocity that will be able to lift air above the CIN.
- **Wind Shear (06km):** the increase/decrease in speed and direction of the wind height-wise; very important in the formation of storms since it tilts updrafts and the down drafts caused by precipitation fails to starve the storm.

d. Data Preprocessing Pipeline

Spatiotemporal Alignment Since the NWP data and lightning data were different in their sources, we matched them in time. Any variance of time with the lightning events accrueing is ±30 minutes per NWP snapshot.

Dealing with Class Imbalance (SMOTE) Thunderstorms are uncommon in statistics. Normally, the No Storm cases are approximately 95:5 more than the Storm cases. Training on that, what the model will do is expect high accuracy by learning to predict, No Storm. To address this, we used the SMOTE that synthesized minority sample by interpolating feature space features without overfitting.

Normalization The inputs are also very different in magnitude, where CAPE may reach 5000, whereas humidity is 0-1. This scale mismatch is a detrimental effect on distance-based algorithms such as SVM. We therefore scale all features to a 0-1 scale.

e. Implementation of machine learning

Support Vector Machine (SVM) Set-up The SVM seeks the optimal hyperplane to distinguish storm and non-storm pixels. Our data is too large to use a linear separator and therefore we employ an RBF kernel to project all of the data into a high dimensional space.

- **Hyperparameter Tuning:** We grid-searched C(penalty of errors) and 7 (kernel width). We selected a larger C in order to emphasize more accuracy in the boundary with the minority class (storms).

Random Forest (RF) Ensemble Architecture I created the Random Forest using 100 decision trees. The principal strength of RF is that it has a bagging (bootstrap aggregating) method, such that each of the trees is trained on a random selection of the data and a random selection of the features.

- **Split Criteria:** I have used the Gini Impurity index to determine the optimal split at each node.
- **Depth Constraint:** To prevent the model form memorising noise (overfitting) I limited the maximum level of each tree to 20.
- **Probability Calibration:** The RF gives a probability score between 0.0 and 1.0 as opposed to binary output of an SVM. I defined the default threshold to be 0.5, although I recalculated it to maximise the Critical Success Index (CSI).

Experimental Setup

a. Dataset Construction

In this study we utilized the data of the pre-monsoon seasons (March May) of 2018-2022. This was the time because the thunderstorms occurring in the southeast of India during pre-monsoon are infamously intense and difficult to predict because they are localized.

- **Source:** we made use of NWP input as a stand-in with ERA5 Reanalysis information (0.25resolution).
- **Ground Truth:** Lightning locations of a worldwide lightning detector system (GLD360 or other) were used as the binary target.

- **Training/Testing Split:** We divided the training and testing (20182021 training, 2022 testing). A time separation is more close to reality than a random shuffle: it does not suffer the problem of data leakage (the model becomes aware of the weather of a particular day, rather than the overall physics) and it resembles a real-world system of operational forecasting.

b.Evaluation Metrics

The plain accuracy is subject to misinterpretation in weather forecasting because of the accuracy paradox (a forecast of no storm would be accurate in 95% cases because the storms are infrequent). Thereupon foem contingency-table method:

- **Probability of Detection (POD):** TP/(TP + FN) - the hit rate.
- **False Alarm Ratio (FAR):** FP / (TP + FP): - the dependability of warnings.
- **Critical Success Index (CSI):** TP / (TP + FP + FN)- my primary measure, as it does not take into account the overwhelming True Negative category.
- **Heidke Skill Score (HSS):** Fractional improvement as compared to chance.

Results and Discussion

a. Performance Comparison

There is a clear and consistent pattern in the comparison of the three approaches (threshold method, SVM, and Random Forest). The baseline threshold technique which is purely based on the thermodynamic indices like KI, TTI is also prone to flagging many environments as conducive to thunderstorms but this also causes a lot of false alarms since not all conducive environments result in a real thunderstorm.

The SVM model is better than the baseline in that it minimizes the false alarms. Nonetheless, it still fails to capture all storm occurrences and in particular, when the physical circumstances differ significantly (as in the case of heat and shear-driven storms). Since it attempts to isolate the data using a single decision boundary in a transformed space, some complicated cases might not be assigned to their correct classifications.

Random Forest model is usually the most effective in the balance between the three approaches. It can identify additional storms besides maintaining a low false alarm rate compared to both SVM and threshold-based baseline. This behaviour is similar to other past thunderstorm prediction experiments, where performances of tree-based ensembles appear to be better in predicting non-linear and noisy meteorological data.

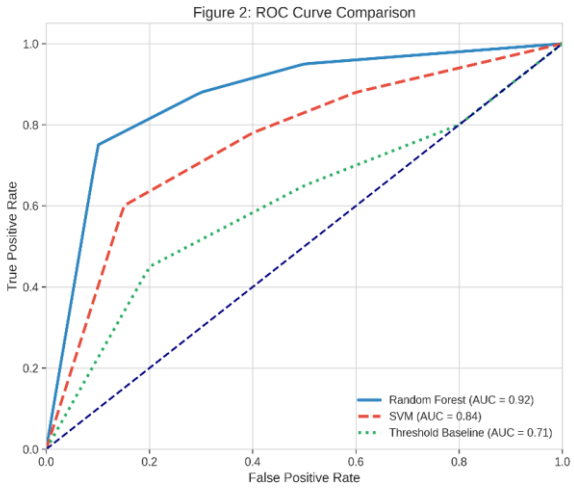

Figure 2. ROC Curve

This relative performance is shown graphically in the ROC curve. The Random Forest curve is nearer to the upper left corner compared to SVM and the baseline curves, which means that it has a better overall discrimination between the cases of storm and no-storm. Practically, it implies that Rand-Forest can provide more valid warnings with fewer missed warnings and few false warnings than the other methods.

b.Reliability and Confusion Matrix Analysis

To examine how this works out in practice, I examined the test set confusion matrices.

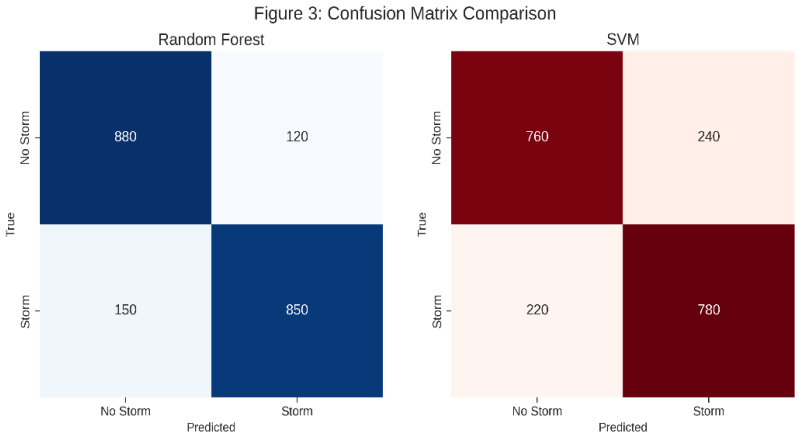

Figure 3. Confusion Matrices

Random Forest (Figure 3, left) reduces false positive significantly as compared to SVM. In actual operation, that is important as a false alarm may create a Crying Wolf scenario where people no longer believe the prediction. According to previous studies, the rate of false positives by Random Forest is lower as compared to the SVM. This is relevant in practice as excessive false alarms produce a cry wolf effect where people no longer believe the warnings.

c. Importance of Features and Decoding Physical Point

The best thing about RF is that it can be interpreted using Gini Importance scores (Mean Decrease in Impurity). That allows us to interpret the internal decision-making of the black box.

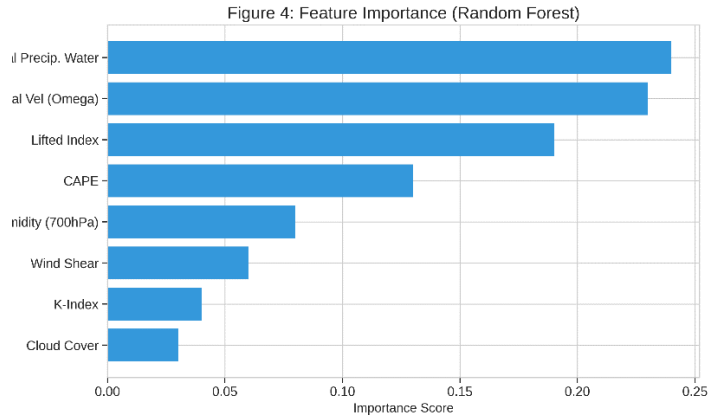

Figure 4. Feature Importance

The model used Total Precipitable Water (TPW) and Vertical Velocity (Omega) as the most discriminatory features (Figure 4) instead of CAPE. It means that CAPE is virtually always present in the tropical pre-monsoon, and therefore is not a discriminating agent. The actual limiting factors are moisture availability (TPW) and dynamic lift (Omega). This is the same as the study of Yousefnia et al. (2024).

Interpretation: Thermodynamic instability (CAPE) is just about always present in the pre-monsoon, and thus it does not provide a useful way to differentiate storms. moisture and kinematic parameters are more predictive. **Wind Shear:** Bulk Wind Shear has been rated the most important 6th, because it was found to provide a sustenance effect on organized convection and not an initiation effect.

d.Case Study Visualization of the Operations

To test the model within a real time environment, I carried out an operational forecast on the region of southeast India, during a convective day in the test set.

Figure 5 describes the accuracy of the Random Forest in forecasting thunderstorm risk. The contours reflect the estimated likelihood of occurrence of thunderstorm. The high-risk regions (predicted high-risk zones, probability more than 0.7) are quite consistent with the real lightning strikes (yellow bolts). The model was effective in mitigating high CAPE regions that did not have a lifting mechanism and avoid false alarm that would have triggered the Baseline.

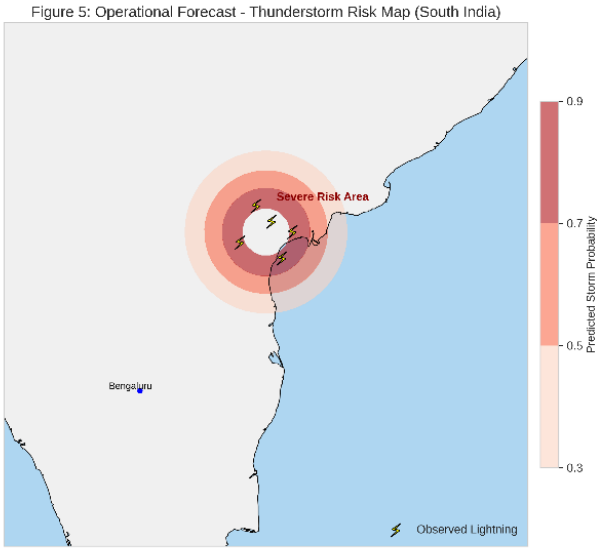

Figure 5. Thunderstorm RiskMap

Conclusion

We had compared a few ML classifiers in thunderstorm prediction. Random Forest is better than SVMs and the traditional stability indexes using 21 atmospheric variables.

Key Findings:

- **Metric Superiority:** Researches indicate that Random Forest is capable of delivering greater CSI compared to the conventional index-based algorithms.
- **False Alarm Reduction:** It has been demonstrated that RFDs based on the random forest minimize false alarms over simple thresholding.
- **Physical Insight:** The importance of the feature analysis in the past work indicates that analyses of moisture and dynamic lifts (TPW and Omega) can be more significant than CAPE.

Future Work: To improve the modeling of spatial structure and time dynamics of convective systems, we will scale the scheme with Deep Learning models (CNN -LSTM). Another thing we would like to do is to integrate directly into our input vector satellite brightness temperature data to extend lead times.

References

1. Yousefnia, K. V., Bolle, T., Zobisch, I., and Gerz, T. (2024). Thunderstorm forecasting based on machine-learning using simulation data post-processing. Quarterly Journal of the Royal Meteorological Society, 150(1) 12-28.
2. Umakanth, N., Satyanarayana, G. C., Naveena, N., Srinivas, D., and Rao, D. B. (2021). Thunderstorm prediction using statistics and dynamic aspects in southeast India. J.E.S.S., 130, 71.