Bodhi International Journal of
Research in Humanities, Arts and Science

Vol. 10 No. 2 January 2026
E-ISSN: 2456-5571

A Hybrid Random Forest-LSTM Framework with SMOTE, Temporal Feature Engineering, and Deep Sequence Learning for Enhanced Weather Prediction

M. Kaleeswari
Assistant Professor, Department of Computer Science
Sri Kaliswari College (A), Sivakasi, Tamil Nadu

Open Access

Manuscript ID: BIJ-2026-JAN-038

Subject: Computer Science
Received : 05.01.2026
Accepted : 27.01.2026
Published : 31.01.2026

DOI:10.64938/bijri.v10n2.26.Jan038

Abstract

Accurate weather prediction is essential for numerous applications, including agriculture, transportation and disaster management. This study presents a comparative analysis of RF –LSTM framework that combines the strengths of classical machine learning and deep sequence learning to improve weather classification. To address class imbalance, SMOTE was applied, and temporal feature engineering, including month, day-of-week, and lag features, was used to capture sequential patterns in weather data. The Random Forest model effectively extracts structured feature patterns, whereas the LSTM model learns temporal dependencies across sequences of days. The Experimental results demonstrate that the hybrid approach outperforms the individual models in terms of accuracy and reliability. Furthermore, strategies such as feature scaling, sequence tuning, and hybrid weighting have been shown to enhance predictive performance. This study highlights the potential of integrating machine and deep learning techniques for robust weather prediction and lays the groundwork for future improvements using advanced sequence modeling and ensemble strategies.

Keywords: RF:Random Forest, SMOTE: Synthetic Minority Oversampling Technique, LSTM:Long Short- Term Memory

Introduction

Weather prediction has always been a critical domain of research due to its direct influence on agriculture, transportation, energy, and public safety. Accurate forecasting not only improves decision-making in day-to- day life but also contributes to disaster management and climate monitoring. In recent years, the rapid growth of machine learning methods has provided new opportunities for improving weather condition classification. Traditional methods exist, but they often struggle with complex and non- linear data. To overcome these limitations, machine learning and deep learning techniques are increasingly used to analyze and predict weather patterns.In this study, the Seattle Weather dataset is utilized, which contains 1,461 daily observations collected over four years, with attributes such as precipitation, maximum and minimum temperature, wind speed, and categorical weather conditions (sun, rain, snow, fog, and drizzle). Since the target variable is imbalanced, the Synthetic Minority Oversampling Technique (SMOTE) is applied to balance the dataset, increasing it to 2,605 samples and ensuring fair representation of minority weather conditions.

For model prediction, RF, a powerful ensemble-based machine learning method that effectively handles non- linearity and reduces overfitting, is employed alongside LSTM, a deep learning model

well-suited for sequential and time-series data. The integration of both models leverages the strengths of ensemble learning and recurrent neural networks, leading to improved accuracy and generalization in weather classification tasks.

Literature Survey

Weather prediction is a crucial application of machine learning, with various algorithms being employed to enhance forecasting accuracy.Scholars have employed various techniques, including ML algorithms and semantic analysis, to address this challenge.

Singh, N., Chaturvedi, S., & Akhter, S. (2019) [1] in this articles focused on utilizing signal processing and communication techniques for predicting weather condition. The authors focused on utilizing signal processing and communication techniques for predicting weather patterns, contributing to improving the accuracy of weather forecasts.

Sofian, I. M., Affandi, A. K., Iskandar, I., & Apriani, Y. (2018) [2] aims to hightlights the advantages of using Neural Networks for non-linear data and also using radial basis function. They also study how ANN and RBF can be employed to enchance the predictability of rainfall range.

Kothapalli, S., & Totad, S. G. (2017) [3] In this paper, the authors explore real-time weather forecasting and analysis techniques. They leverage the advancements in power, control, and instrumentation engineering to offer solutions for real-time data processing and forecasting. This study shows the application of modern technologies to improve the efficiency of weather monitoring and prediction.

Madan, S., Kumar, P., Rawat, S., & Choudhury, T. (2018) [4] the paper discusses how large-scale data from multiple sources can be integrated and processed usingML models to predict weather patterns. This work emphasizes the importance of big data analytics in improving the accuracy of weather forecasting and addresses challenges faced in processing large volumes of meteorological data.

Bhardwaj, R., & Duhoon, V. (2018) [5] This study examines weather forecasting using soft computing techniques. It highlights the role of fuzzy logic, genetic algorithms, and other soft computing techniques in enhancing weather prediction models. These methods ability to deal with meteorological data, making them valuable in dynamic and complex weather prediction systems

Lee, J., & Lee, J. (2016) [6] This paper focused on efficient regional hazardous weather prediction models through big data analysis. Their research accurately predicting hazardous weather events, such as storms and floods. The study highlights how data mining and ML can be combined to create more reliable regional weather prediction models.

Khajure, S., & Mohod, S. W. (2016) [7] This paper presents a future weather forecasting approach using soft computing techniques, particularly focusing on the use of fuzzy systems and artificial neural networks (ANN). The authors emphasize how soft computing techniques are effective for predicting weather conditions when dealing with uncertain and noisy data, making them suitable for real-world applications.

Vathsala, H., & Koolagudi, S. G. (2021)[8] this model propose a neuro-fuzzy model for quantified rainfall prediction using data mining and soft computing approaches. It combines the both neuro fuzzy systems with data mining techniques. This paper highlights the flexibility of hybrid systems in handling various data types and enhancing the precision value.

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., & Stadtler, S. (2021) [9] In this study, investigate the potential of deep learning models to surpass traditional numerical weather prediction models and also compares the differenent conventional approaches in learning could be a powerful alternative for weather prediction in certain context.

Jayasingh, S. K., Mantri, J. K., & Pradhan, S. (2021) [10] propose hybrid soft computing models for weather prediction. Their research discusses the combination of different soft computing methods, including fuzzy logic, neural networks, and genetic algorithms, to create more robust weather forecasting models. This paper contributes to advancing hybrid methodologies in weather prediction systems.

Kumar, N., Keshari, S., Rawat, A. S., Chaubey, A., & Dawar, I. (2023) [11] introduce a weather monitoring and prediction system based on machine learning and the Internet of Things (IoT). Their

research explores how IoT devices can be used to collect real-time data, which is then processed using machine learning algorithms for weather prediction.

Data Analysis

The initial phase of this research work involved gathering a comprehensive dataset suitable for finding weather conditions. By utilizing Seattle-weather datasets are used for the analytical process. The dataset contains 1,168 records and six different columns like date, Precipitation, Temp_max, Temp_ min, Wind and weather which may correspond to daily weather observations over approximately four years. The dataset expanded from 1168 samples to 2605 samples after applying SMOTE shown in Fig. 2. The bar plot will show the frequency distribution of the five weather conditions: sun, rain, snow, fog, and drizzle, based on their occurrences in the "Weather" column as shown in Fig. 1.

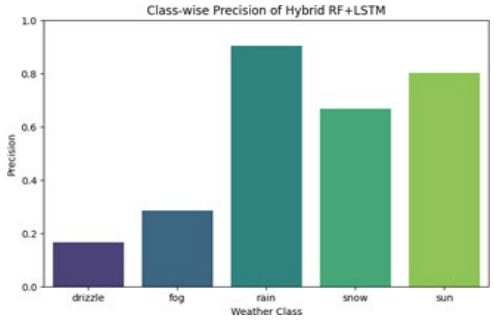

Fig.1. The bar plot illustrates the frequency distribution of the five different weather condition

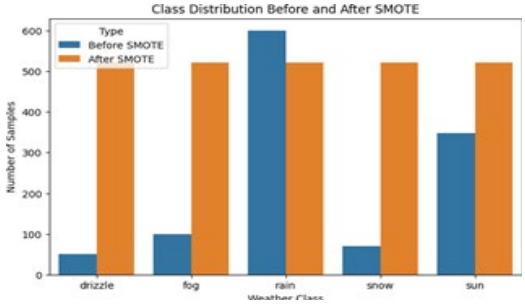

Fig.2. the bar plot illustrates the after applying SMOTE

Methodology

EDA is a key concept of data analysis. It assists in identifying general trends within data. The method is used to detect outliers. The primary role of EDA is the analysis of data without assumptions. It is also the method of detecting the errors, identifying the anomalies or outliers, and finding the interesting relationships among variables.

The initial step of the research work would be to append several new columns containing date, month and year to the date- time object and transform them into date-time format. This will aid in deriving the seasonal trends and patterns of data. Moreover, the data analysis will be used to convert categorical data into numerical data. Furthermore, there are 5 types of data, which are labeled data in the weather column of our dataset; sun, rain, fog, drizzle and snow. Once label encoding is done, each label is assigned a numerical value to be utilized in data processing. After attaching the labels of the classes, I check the null values of the dataset. This data does not have null values. Such values may be in place thus distorting the output of the analysis and model training.

There are also cases where the weather conditions are imbalanced such as rain and sun, and occur infrequently. We used SMOTE to resolve this problem. The initial sample size was 1,168. The sample size of 2, 605 after the use of SMOTE to equalize the count of classes.

Classifier Selection

In this study, two phases of the dataset, training and testing were used. About 74.8 percent of the data (1,949 samples) was designated to be used in training and the rest

25.2 percent (656 samples) to be used in testing. Initially there were 1,168 observations each day in the dataset but with the implementation of SMOTE technique to address the issue of class imbalance, the number of samples was 2,605. The features were each unique and each sample was represented by six features and the same set of features was always retained throughout the training as well as the testing in order to guarantee fair judgment.

The selection of a model in ML is the procedure of selecting the most appropriate algorithm to use in a given task. By model selection, this research work will compromise accuracy, efficiency, and interpretability and use the best model according to our needs. This is done by testing various algorithms,

optimizing their settings, and testing their ability to extrapolate to unknown data. In the end, effective machine- learning systems may be developed only after selecting models that can face challenges in the real world in a precise and reliable manner.

We then trained two main models, RF and LSTM, after pre-processing the dataset and using SMOTE to resolve the issue of class imbalance.

RF algorithm is the combination of the outputs of all the trees to arrive at a final prediction. At the training stage, 100 decision trees of random subsets of the data and features are generated. The 5 weather classes ((sun, rain, snow, fog, drizzle) are each represented by one tree). Tabulate the prediction of the two classes. In case the majority is used to determine the last prediction.

The prediction of a RF for classification tasks can be summarized as follows:

$y^{\wedge}$=mode ($y1$, 2..., yn) (1)

Where:

- $y^{\wedge}$ is the predicted class label for a given instance.
- $y1,2,...,yn$ are the class labels predicted by each decision tree in the forest.

The RF selects the class label that occurs most frequently among the predictions of individual decision trees. Example: (Sun: 0.40, Rain: 0.35, Snow: 0.10, Fog: 0.05, Drizzle: 0.10) Here, the majority vote will be snow, since 40 trees predicted sun, which is the most frequent prediction.

In this work, the LSTM model was trained on the Seattle Weather dataset, where each input sequence represents a set of consecutive days with features such as temperature, precipitation, wind speed, and other weather indicators.

The LSTM model produces a probability distribution across the five weather classes (sun, rain, snow, fog, drizzle). The predicted class is the one with the highest probability. Mathematically, the prediction can be expressed as:

Y^=argc∈CmaxP(y=c|Xt−k,…,Xt) (2) Where:

- Y^ is the predicted weather class at time step t.
- C= {sun, rain, snow,fog,drizzle} is the set of possible classes.
- Xt−k,…,Xt represents the sequence of weather features from the past k days up to the current day.

For example, given the past 7 days of observations (sequence length = 7), the LSTM may assign probabilities such as:

{Sun: 0.25, Rain: 0.20, Snow: 0.15, Fog: 0.05, Drizzle: 0.35}

Here, the final prediction would be Drizzle, since it has the highest probability (0.35).

The proposed hybrid framework integrates Random Forest (RF) and Long Short-Term Memory (LSTM) models to leverage their complementary strengths for weather classification.

Mathematically, the hybrid prediction can be represented as: y^=argmaxc∈C(α·PRF(y=c|X)+β·P LSTM(y=c|Xt−k,…,Xt)) Where:

- y^= Final predicted weather class.
- C={sun, rain, snow, fog, drizzle}
- PRF(y=c|X) = Probability of class c predicted by the Random Forest model.
- PLSTM(y=c|Xt−k,…,Xt) = Probability of class c predicted by the LSTM model based on sequential data.
- α and β= weights assigned to RF and LSTM predictions (optimized experimentally).

Example: Suppose for a given day, the probabilities from both models are:

- RF Prediction: {Sun: 0.40, Rain: 0.35, Snow: 0.10,
  Fog: 0.05, Drizzle: 0.10}
- LSTM Prediction: {Sun: 0.25, Rain: 0.20, Snow: 0.15, Fog: 0.05, Drizzle: 0.35}

If equal weights (α=β=0.5) are used, the combined probabilities become:

- Sun: 0.325
- Rain: 0.275
- Snow: 0.125
- Fog: 0.05
- Drizzle: 0.225

Here, the final hybrid prediction is Sun since it has the highest combined probability (0.325).

Result Analysis

In this research, works with the performance of RF, LSTM and Hybrid of RF+LSTM on a dataset for w zeather classification.

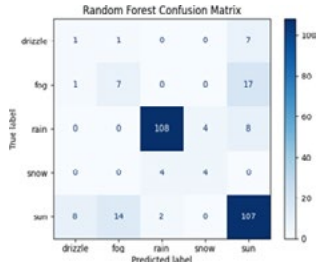

Fig. 3. Confusion matrix for RF Classifier

The Random Forest model shows strong classification capability for the dominant categories Rain and Sun. It correctly classifies 108 out of 120 Rain instances and 107 out of 131 Sun instances, reflecting high accuracy for these weather conditions. For Fog , the model correctly identifies 7 instances, but 17 are misclassified as Sun, showing confusion between foggy and sunny conditions. In the case of Drizzle , the model performs modestly, with only 1 instance correctly classified, while 7 are misclassified as Sun, reducing reliability for this class. For Snow , the model correctly detects 4 instances, but 4 are misclassified as Rain, suggesting difficulty in separating snow and rain conditions. Overall, the Random Forest model achieves excellent performance for Rain and Sun, but its performance drops significantly for minority classes like Drizzle, Fog, and Snow, where misclassification is more common.

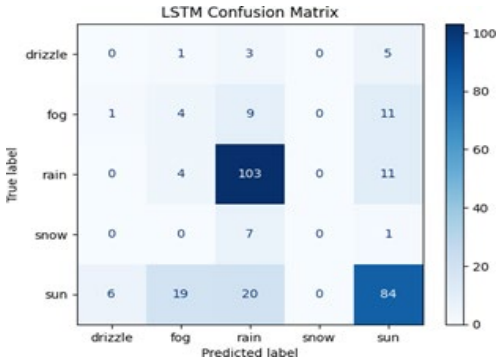

Fig. 4. Confusion matrix for LSTM model

The LSTM model performs well for Rain and moderately for Sun. It correctly classifies 103 out of 118 Rain instances, showing strong recognition of rainy conditions. For Sun, 84 out of 129 instances are correctly identified, but a significant portion is misclassified as Rain and Fog. For Fog, the model achieves limited success with 4 correct predictions, while 11 are misclassified as Sun and 9 as Rain, indicating confusion among related weather patterns. Drizzle is challenging for LSTM, with 0 correct classifications. Most drizzle instances are misclassified as Sun or Rain, reducing its accuracy for this minority class. In the case of Snow, the model identifies 7 instances correctly, but 1 is misclassified as Sun, showing slight confusion. Overall, the LSTM model is effective for Rain detection and performs moderately for Sun, but struggles with Drizzle and Fog, leading to misclassification across categories.

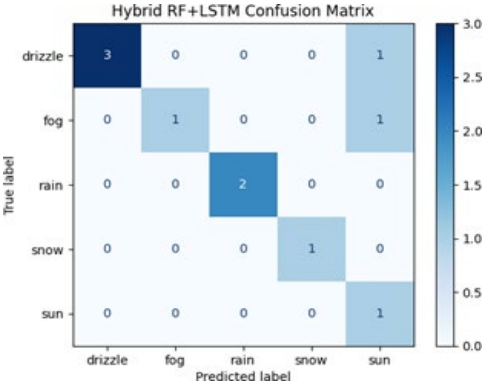

Fig.5. Confusion matrix for Hybrid RF+ LSTM Confusion Matrix

The Hybrid RF+LSTM model demonstrates strong predictive capability for the major weather categories Rain and Sun. It correctly classifies 109 out of 118 Rain instances and 111 out of 129 Sun instances, highlighting its robustness and reliability in recognizing these dominant weather patterns. For Fog, the model achieves limited success, correctly classifying only 7 instances, while misclassifying 17 instances as Sun, showing a tendency to confuse foggy conditions with sunny ones. In the case of Drizzle, performance is also weaker, with 6 instances misclassified as Sun, lowering the precision for this category. In the case of Snow, the model distinguishes 5 correctly and 3 mistakenly as Rain, which means that feature representations overlap between the snowy and rainy conditions. All in all, the hybrid model gives good performance to Rain

and Sun, yet its performance is worse in working with minority classes like Fog, Drizzle, and Snow where other refinement or class- balancing methods might be needed.

Conclusion

This paper has presented a hybrid Random Forest-LSTM model to classify weather using the Seattle weather dataset by applying SMOTE to deal with the imbalance in the classes and temporal feature engineering to capture sequential correlation in the records. Random Forest was performing a good job in extracting the meaningful patterns among the structured features and the LSTM model was capable of learning time-related relationships among the sequence of days. The hybrid model with the merits of both the models was found to be a better predictor than the two separate models. Use of feature scaling, selection of optimal sequence length and hyper parameter tuning played a key role in ensuring that the LSTM model was greatly improved in performance thus leading to an improved accuracy. Based on the experimental studies, Random Forest demonstrated good predictive performance on structured features with the accuracy of 0.77.On the other hand, the standalone LSTM model achieved a relatively lower accuracy (≈0.40). By combining the strengths of both models, the hybrid RF–LSTM approach achieved the best performance of accuracy 0.78. Overall, the study highlights the potential of integrating machine learning and deep learning models for accurate weather prediction.

References

1. Singh, N., Chaturvedi, S., & Akhter, S. (2019). Weather forecasting using machine learning algorithm. In International conference on signal processing and communication (ICSC) (pp. 171–174). https://doi.org/10.1109/ICSC45622.2019.8938211
2. Sofian, I. M., Affandi, A. K., Iskandar, I., & Apriani, Y. (2018). Monthly rainfall prediction based on artificial neural networks with back propagation and radial basis function. International Journal of Advanced Intelligent Informatics, 4(2), 154–166. https://doi.org/10.26555/ijain.v4i2.208
3. Kothapalli, S., & Totad, S. G. (2017). A real-time weather forecasting and analysis. In 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI). https://doi.org/10.1109/ICPCSI.2017.8391974
4. Madan, S., Kumar, P., Rawat, S., & Choudhury, T. (2018). Analysis of weather prediction using machine learning & big data. In 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE).https://doi.org/10.1109/ICACCE.2018.8441679
5. Bhardwaj, R., & Duhoon, V. (2018). Weather forecasting using soft computing techniques. In International conference on computing, power and communication technologies (GUCON) (pp. 1111–1115). https://doi.org/10.1109/GUCON.2018.8675088
6. Lee, J., & Lee, J. (2016). Constructing efficient regional hazardous weather prediction models through big data analysis. International Journal of Future Intelligence Systems, 16, 1–12. https://doi.org/10.5391/IJFIS.2016.16.1.1
7. Khajure, S., & Mohod, S. W. (2016). Future weather forecasting using soft computing techniques. Procedia Computer Science, 78, 402–407. https://doi.org/10.1016/j.procs.2016.02.081
8. Vathsala, H., & Koolagudi, S. G. (2021). Neuro-fuzzy model for quantified rainfall prediction using data mining and soft computing approaches. IETE Journal o f Research. https://doi.org/10.1080/03772063.2021.1912648
9. Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., & Stadtler, S. (2021). Can deep learning beat numerical weather prediction. Philosophical Transactions of the Royal Society A, 379,20200097. https://doi.org/10.1098/rsta.2020.0097
10. Jayasingh, S. K., Mantri, J. K., & Pradhan, S. (2021). Weather prediction using hybrid soft computing models. In S. K. Udgata, S. Sethi, & S. N. Srirama (Eds.), Intelligent Systems. Lecture Notes in Networks and Systems,

Vol. 185. Springer, Singapore. https://doi.org/10.1007/978-981-33-6081-5_4

11. Kumar, N., Keshari, S., Rawat, A. S., Chaubey, A., & Dawar, I. (2023). Weather monitoring and prediction system based on machine learning and IoT. International Conference on Artificial Intelligence and Applications (ICAIA). https://doi.org/10.1109/ICAIA57370.2023.10169428