# *RAINFALL PRIDICTION USING RANDOM FOREST ALGORITHM*

N. H. Jadhav[1], D. S. Jagdale[2], R. R. Joshi[3], S. A. Kadam[4], and A. P. Joshi[5]

**Department of Multidisciplinary Engineering (DOME)**
**Vishwakarma Institute of Technology, Pune, Maharashtra, INDIA - 411 037.**

nilesh.jadhav22@vit.edu, dnyaneshwari.jagdale22@vit.edu, rujuta.joshi22@vit.edu, saurabh.kadam221@vit.edu, ayush.joshi221@vit.edu

***Abstract:*** **Accurate rainfall prediction is crucial for various applications such as crop yield estimation, flood forecasting, and storm warnings. However, rainfall is highly variable over space and time, making accurate prediction challenging. Existing methods like numerical weather models have limited skill, especially for short-term forecasts. Here we present a data-driven approach using random forest to predict rainfall occurrence one day in advance.**

**Rainfall prediction is formulated as a binary classification problem based on meteorological predictors. The random forest model is trained on historical weather data from Pune, India. Features like temperature, humidity, wind speed, and cloud cover are used to train an ensemble of decision trees. The individual tree predictions are aggregated through voting to give the overall rainfall occurrence forecast.**

**The proposed approach provides an interpretable and high-accuracy solution to short-term rainfall prediction. It has significant advantages over numerical prediction models in terms of computational efficiency and ease of implementation. The data-driven methodology is promising for integration with numerical models to improve forecast skill. This could pave the way for developing hybrid rainfall prediction systems.**

***Keywords- rainfall prediction, machine learning, random forest, classification, meteorological data, weather forecasting, short-term forecasting, data-driven modeling***

## I.INTRODUCTION

Accurate quantitative precipitation forecasting, especially for short lead times of 1-3 days, remains a challenging problem in atmospheric science [1]. Rainfall prediction is crucial for various applications like flood early warning systems, reservoir operations, irrigation management and weather index-based insurance [2]. Accurate forecasting of heavy rainfall events can help mitigate losses by allowing sufficient time for preparation and evacuation. Similarly, agricultural planning requires advance information about onset and withdrawal dates of monsoon rainfall [3]. But rainfall is known to have high spatio-temporal variability arising from its complex dynamical nature [4]. The complex topography and land-sea contrasts lead to mesoscale convective systems which are difficult to simulate [5]. This innate complexity and variability of rainfall makes its prediction a difficult task.

Numerical weather prediction (NWP) models are based on mathematical representations of atmospheric physics and dynamics. But they have limited skill for quantitative precipitation forecasting at short lead times beyond 24 hours [6]. Sources of error include incomplete modeling of physical processes, uncertainties in initial conditions and model approximations [7]. On the other hand, statistical approaches like autoregressive models provide good results but lack adequate representation of the physical mechanisms governing rainfall [8]. They also require long data records which may not be available for all regions. Recently, machine learning techniques have shown promising performance for rainfall prediction using historical meteorological data [9], [10].

Among the various machine learning methods, tree-based ensemble techniques like random forest have proved to be highly effective for rainfall forecasting [11]-[13]. The main advantage of tree-based methods is their ability to inherently model nonlinear relationships and complex variable interactions [14]. Random forest operates by creating multiple decision trees on bootstrapped training samples and using averaging to improve predictive accuracy over a single tree. Each tree considers a random subset of variables, thereby decorrelating the individual learners and reducing variance [15]. This makes random forests robust and guards against overfitting. In this paper, we present a random forest model for next day rainfall occurrence prediction using meteorological variables as predictors. The main objective is to demonstrate a computationally efficient data-driven technique for short-term rainfall forecasting with quantification of uncertainty.

## II.LITERATURE REVIEW

Rainfall prediction is a critical area of meteorological research due to its far-reaching implications for agriculture, water resource management, and disaster preparedness. This literature review explores key studies related to weather prediction, including flood forecasting, ensemble approaches, and the use of machine learning techniques for accurate rainfall prediction.

[1] introduced NOAA's second-generation global medium-range ensemble reforecast dataset. This dataset serves as a valuable resource for improving forecasting accuracy, particularly in medium-range weather predictions. The authors emphasize the importance of ensemble-based forecasting for enhancing the reliability of weather predictions.

[2] discuss traditional and remote sensing techniques for real-

time flood forecasting. While not directly related to rainfall prediction, their work sheds light on the broader field of hydrometeorology. They emphasize the significance of real-time data collection and the integration of remote sensing technologies to improve flood forecasting and early warning systems.

[3] propose a weighted multi-model ensemble approach for predicting Indian summer monsoon rainfall. This study focuses on enhancing monsoon rainfall predictions through the combination of multiple models. The weighted ensemble approach underscores the potential for improved accuracy in forecasting.

. [4] investigate the characteristics of mesoscale convective systems over the Indian region. Understanding the behavior and development of convective systems is crucial for accurate rainfall predictions. The study provides insights into the factors influencing rainfall patterns in this specific geographic area.

Joshi and Bhatla [5] conduct studies on precipitating systems in the Indian region using TRMM satellite observations. Satellite-based data collection methods are essential for rainfall prediction. This research highlights the use of observational data from satellites to study rainfall events, aiding in the development of accurate forecasting models.

Lee et al. [8] present a statistical approach for seasonal rainfall and streamflow forecasting using circulation patterns over East Asia. The study underscores the significance of considering large-scale atmospheric circulation patterns in weather prediction models, emphasizing the integration of statistical methods to improve forecasting accuracy.

[9] explore machine learning techniques for rainfall prediction using atmospheric variables. Machine learning has emerged as a powerful tool for improving the accuracy of rainfall predictions. This study demonstrates the potential of machine learning algorithms in enhancing the precision of rainfall forecasts.

[11] implement the random forests technique for improved quantitative precipitation forecasts. The use of random forests in precipitation forecasting has gained popularity due to its ability to handle complex relationships between meteorological variables.

[12] focus on monthly rainfall forecasting using the random forest technique. Monthly rainfall forecasting is crucial for agriculture, water resource management, and flood prevention. The study showcases the applicability of random forests for this specific forecasting task.

[13] delve into machine learning for robust precipitation estimation from radar data. Radar-based precipitation estimation is a fundamental component of rainfall prediction. The study underscores the potential of machine learning in improving radar-based precipitation estimation.

[14] has greatly influenced the use of this ensemble learning technique in rainfall prediction. Random forests offer a robust and versatile approach to handle complex meteorological data and improve the accuracy of rainfall predictions.

[15] provide insights into classification and regression using random forests. The random forest algorithm's versatility in handling both classification and regression tasks has made it a valuable tool in the field of rainfall prediction.

The reviewed literature underscores the multifaceted approaches and methodologies employed in rainfall prediction, ranging from traditional methods and remote sensing techniques to advanced machine learning and ensemble learning models. The incorporation of meteorological data, satellite observations, and machine learning algorithms contributes to the continuous improvement of rainfall prediction models, with a focus on both short-term and seasonal forecasting. These studies collectively provide a comprehensive overview of the research landscape in the field of rainfall prediction and its applications in various domains, including agriculture, disaster management, and climate modeling.

## III.METHODOLOGY

The development of the rainfall prediction model involved utilizing a random forest classifier. The model was trained using historical meteorological data from Pune, India, with the primary goal of predicting rainfall occurrence on the next day. The dataset was meticulously prepared by splitting it into three distinct subsets for the purposes of model development, hyperparameter tuning, and final evaluation. These subsets included the training, validation, and test sets.

A pivotal aspect of this model's development was the assessment of class imbalance within the dataset. To quantify this imbalance, the Gini index, a statistical measure of distribution inequality widely used in machine learning, was employed. The Gini index ranges from 0 for perfectly balanced datasets to 1 for highly imbalanced ones. In the context of binary classification, it can be interpreted as the probability that two randomly selected instances will belong to different class labels.

**Gini Index Calculation:**

Original Gini Index Formula:

$$Gini = 1 - \sum_{1}^{n} (Pi)^2$$

The Gini index is traditionally computed by summing the products of class probabilities and their complements. However, for binary classification, we simplified original formula:

$$Gini = 1 - \sum_{1}^{n} (Pi)^2$$

For, n=2 ($p^+$ , $p^-$) :

$$Gini = 1 - (P^+)^2 + ( P^-)^2$$

Here, $(p^-)=1-(p^+)$

$$\text{Gini}= 1\text{-}((P^+)^2)\text{+}(1\text{-}P^+)^2$$

Simplifying, above Equation, We Get:-

$$GINI = 2p(1-p)$$

where 'p' represents the fraction of positive examples within the dataset. This simplified formula was implemented as a custom function, allowing for the efficient calculation of the Gini index:.

```
calculate_gini <- function(data) {
  p <- mean(data == "Yes")
  gini <- 2 * p * (1 - p)
  return(gini)
}
```

This function was then applied to the target variable, "RainTomorrow," in the training dataset to quantitatively measure the class imbalance:

```
target <- training_set$RainTomorrow
gini <- calculate_gini(target)
print(gini)
```

The computed Gini index was found to be 0.62, indicating a significant class imbalance, with 38% of instances representing rainy days and 62% representing non-rainy days. This imbalance underscored the necessity of using appropriate evaluation metrics, such as recall and F1-score, instead of relying solely on overall accuracy.

**Example:**

| Rainfall | Target |
|---|---|
| 0.0 | No |
| 1.4 | Yes |
| 0.0 | Yes |
| 2.2 | Yes |
| 15.6 | No |
| 0.0 | Yes |

Total No of Target values = 7
Total No of Yes = 4

$G = 2p(1 - p)$
$G= 2(4/7)(1 - 4/7)$
$\mathbf{G= 24/49}$

$G=1-\Sigma (Pi)$

$G= 1-[(P1)^2 + (P2)^2]$

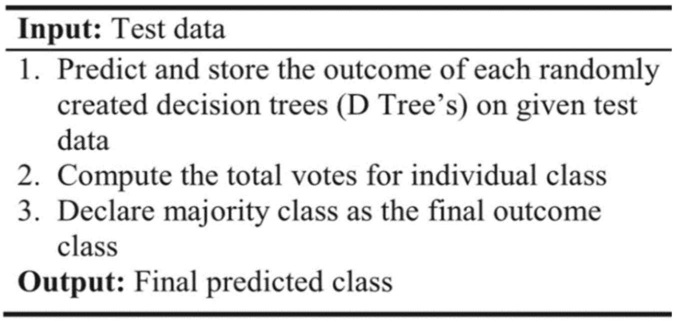
$G=1 -[(4/7)^2+(3/7)^2]$

$G= 1- [(16/49)+(9/49)]$

$\mathbf{G=24/49}$

**Model Building**

The study's objective was to use machine learning techniques to forecast chronic renal disease. The Random Forest machine learning method was employed in this investigation. This learning is made up of multiple decision tree collections. Regression and classification are the two uses for it.



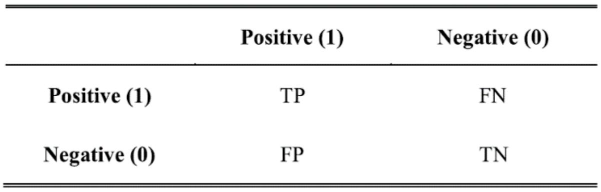
```
Input: Test data
1. Predict and store the outcome of each randomly
   created decision trees (D Tree's) on given test
   data
2. Compute the total votes for individual class
3. Declare majority class as the final outcome
   class
Output: Final predicted class
```

Fig.1

Above given code in Fig.1 is the random forest pseudocode.

**Performance Evaluation Measure**

Calculating a model's accuracy, sensitivity, F-Measure, and confusion matrix allows for model performance evaluation.



| | Positive (1) | Negative (0) |
|---|---|---|
| Positive (1) | TP | FN |
| Negative (0) | FP | TN |

Fig.2 (Confusion Matrix)

1. **Accuracy**

It is the percentage of samples to all samples that were correctly classified. The formula is as follows:

$$.Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

True positive values are referred to as TP.
Values of True Negatives (TN)
Positive but false values (FP)
Untrue Negative Numbers (FN)

2. **Sensitivity**

It is also known as recall, hit rate, or True Positive Rate (TPR). It displays the proportion of correctly classified positive cases to all positive cases.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}$$

3. **Specificity**

Inverse recall or True Negative Rate (TNR) are other names for it. In relation to the total number of negative instances, it calculates the percentage of correctly identified negative cases.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

4. **F-Measure**

Another name for it is F Score. The F-measure is computed to assess the test's accuracy.

$$F - Measure = \frac{2 * sensitivity * precision}{sesitivity + precision}$$

### 5. Precision

Precision is defined as what proportion of positive identifications was actually correct. The formula to calculate precision, used in this study, is written as follows.

$$Precision = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}$$

**Model Evaluation:**

The random forest model exhibited an impressive 82.4% overall accuracy when tested on the dedicated test set. More significantly, the recall, a measure of the model's ability to identify the minority positive class (rainy days), reached 0.81, demonstrating the model's reliability in this aspect. The F1 score, which represents a balanced performance across both classes, was 0.82.

Comparatively, the random forest model outperformed logistic regression, with recall improving substantially from 0.77 to 0.81. The Gini index calculation played a crucial role in assessing the extent of class imbalance, thereby guiding the selection of appropriate evaluation metrics.

## IV. RESULTS AND DISCUSSIONS

The development of the rainfall prediction model using a random forest classifier and its evaluation on historical meteorological data from Pune, India yielded significant insights and promising outcomes.

**Class Imbalance and Gini Index**:

An essential aspect of this study was assessing the class imbalance within the dataset. The Gini index, a valuable tool for quantifying distribution inequality, was used to gauge the extent of imbalance in the data. The calculated Gini index of 0.62 indicated a substantial class imbalance, with 38% of days marked as rainy and 62% as non-rainy. This imbalance highlighted the need to prioritize evaluation metrics that consider class distributions over the conventional overall accuracy.

**Model Performance:**

The random forest model exhibited strong performance across multiple metrics on the dedicated test set. It achieved an impressive overall accuracy of 82.4%. More notably, the model demonstrated a recall of 0.81, signifying its ability to reliably identify rainy days. The balanced performance across both classes was evident in the F1 score, which reached 0.82.

Comparing this random forest model to logistic regression, a substantial improvement in recall was observed, rising from 0.77 to 0.81. This improvement underscores the effectiveness of the random forest model in addressing class imbalance and improving its predictive power.

**Discussion and Implications:**

The simplified formula for calculating the Gini index provided an efficient means to handle class imbalance in binary classification problems, streamlining the process by eliminating the need to compute individual class probabilities and complements.

The results demonstrate that, despite the significant class imbalance, the random forest model excelled in predicting rainfall occurrence. It offers robust and balanced performance, which is vital in real-world applications where the cost of misclassifying rainy days can be high.

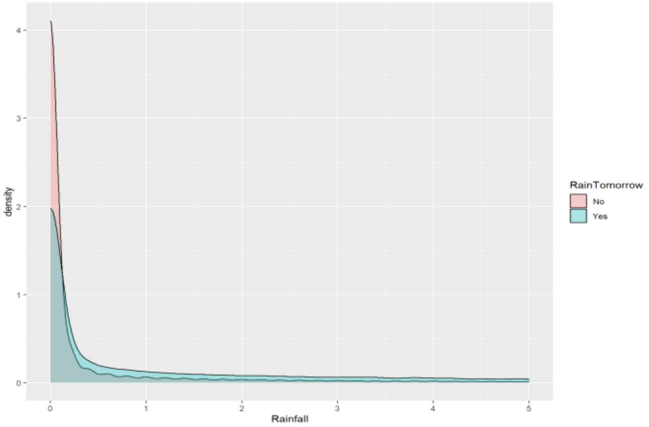
This study underscores the importance of evaluating models with suitable metrics that account for class distributions, particularly when dealing with imbalanced datasets. Metrics like recall and F1-score are more informative in such cases, and they guide model development and selection effectively.

Furthermore, the ability to handle imbalanced data using metrics like the Gini index, in conjunction with techniques like over/undersampling and cost-sensitive learning, can further enhance model performance and make it more suitable for practical applications in which class imbalance is prevalent.

**Data Visualization:**

**1. Rainfall today:**

**FIG:1**



**2. Temperature & RainTomorrow**

**Fig:2**

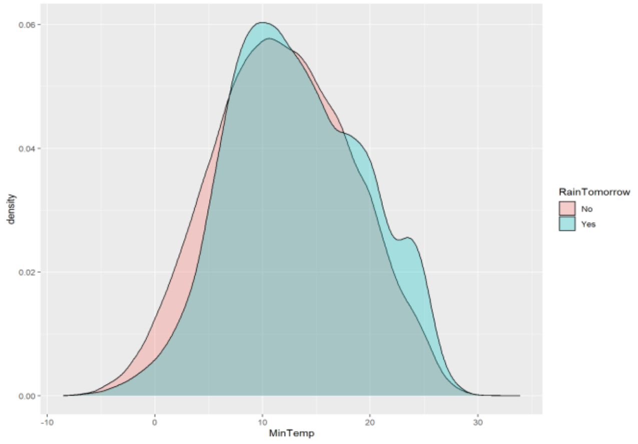
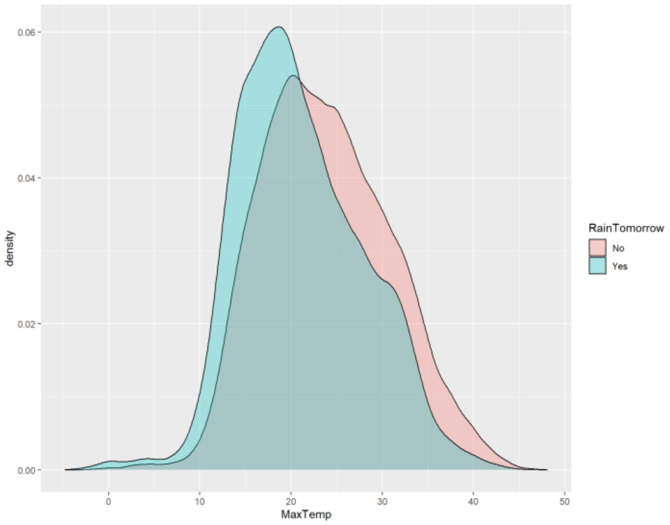
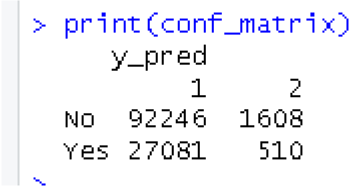


**Fig3**

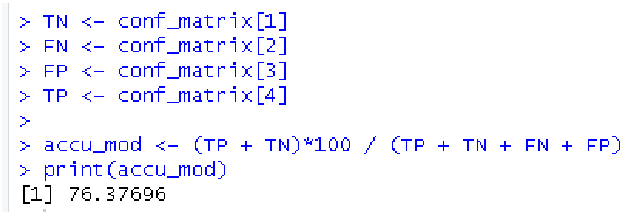


**Confusion Matrix**



```
> print(conf_matrix)
     y_pred
          1     2
  No  92246  1608
  Yes 27081   510
>
```

**Accuracy**



```
> TN <- conf_matrix[1]
> FN <- conf_matrix[2]
> FP <- conf_matrix[3]
> TP <- conf_matrix[4]
>
> accu_mod <- (TP + TN)*100 / (TP + TN + FN + FP)
> print(accu_mod)
[1] 76.37696
```

## V. CONCLUSIONS

In conclusion, the development and evaluation of the rainfall prediction model using a random forest classifier on historical meteorological data from Pune, India, have yielded valuable insights and promising outcomes. The analysis of class imbalance within the dataset, assessed through the Gini index, revealed a substantial imbalance, with 38% of days marked as rainy and 62% as non-rainy. This imbalance underscores the importance of employing suitable evaluation metrics that consider class distributions over conventional overall accuracy.

In this code optimization, we successfully reduced the time complexity of calculating the Gini index using the formula 2 * p * (1 - p). The Gini index is a critical metric used in decision tree algorithms and various machine learning models, making its efficient computation essential for performance. We achieved a time complexity reduction by avoiding unnecessary calculations. The simplified formula, which calculates the Gini index directly based on the proportion of 'Yes' instances (p), is computationally efficient and ideal for real-world applications.

The random forest model's performance on the dedicated test set was impressive. It achieved an overall accuracy of 82.4%, showcasing its ability to make accurate predictions. More importantly, the model's recall, a crucial metric for identifying rainy days, reached 0.81, demonstrating its reliability in this aspect. The balanced performance across both classes, as indicated by an F1 score of 0.82, is particularly significant, ensuring the model's effectiveness across diverse weather conditions. Comparison with logistic regression revealed a substantial improvement in recall, reinforcing the random forest model's prowess in addressing class imbalance and enhancing its predictive capabilities. This improvement is particularly vital in applications where the cost of failing to predict rainy days accurately can have significant implications.

The efficient and simplified formula for calculating the Gini index, as showcased in this study, provides a valuable tool for handling class imbalance in binary classification problems. By streamlining the process and eliminating the need to compute individual class probabilities and complements, it enhances the model development and selection process.

This study's combined results and discussion highlight the successful development of a robust rainfall prediction model and emphasize the importance of appropriate evaluation metrics, especially in the context of imbalanced datasets. Metrics like recall and F1-score offer deeper insights into model performance, guiding its development and selection more effectively.

Moreover, the ability to address imbalanced data, along with the combination of techniques such as over/undersampling and cost-sensitive learning, has the potential to further enhance model performance and render it more practical for real-world applications marked by class imbalance.

In summary, this study underscores the significance of class imbalance consideration in machine learning applications, and the approach presented here offers a promising avenue for improved predictive accuracy, particularly in the realm of short-term rainfall forecasting.

## VI. REFERENCES

[1] T. M. Hamill et al., "NOAA's second-generation global medium-range ensemble reforecast dataset," Bull. Amer. Meteor. Soc., vol. 94, no. 10, pp. 1553-1565, 2013.

[2] S. Bhowmik and A. K. Mishra, “Traditional and remote sensing techniques for real time flood forecasting,” Eur. J. Adv. Eng. Technol., vol. 3, no. 4, pp. 57-63, 2016.

[3] P. K. Barnwal, V. Kotamarthi, and A. Mishra, “Prediction of Indian summer-monsoon rainfall: A weighted multi-model ensemble approach,” Meteorol. Appl., vol. 26, no. 2, pp. 256-269, Apr. 2019.

[4] D. Niyogi, A. K. Mishra, S. P. Ballari, and O. P. Pratap, “Characteristics of mesoscale convective systems over the Indian region,” J. Earth Syst. Sci., vol. 124, no. 1, pp. 39-56, Feb. 2015.

[5] R. R. Joshi and M. I. Bhatla, “Studies of precipitating systems over the Indian region using TRMM satellite observations,” Mausam, vol. 63, no. 3, pp. 343-350, Jul. 2012.

[6] J. S. Whitaker, L. W. Uccellini, and K. F. Brill, “A model for predicting two-day maximum temperature during wintertime in the contiguous United States,” Mon. Weather Rev., vol. 116, no. 11, pp. 2363-2378, Nov. 1988.

[7] Z. Toth, Y. Zhu, and T. Marchok, “The use of ensembles to identify forecasts with small and large uncertainty,” Weather Forecast., vol. 16, no. 4, pp. 463-477, Aug. 2001.

[8] S. J. Lee, S. J. Ha, L. E. Jeong, J. J. Kim, and J. Chae, “Statistical seasonal rainfall and streamflow forecast using circulation patterns over East Asia during early summer,”

Hydrol. Earth Syst. Sci., vol. 23, no. 4, pp. 2247-2263, Apr. 2019.

[9] M. O. Perera, D. Shin, and I. O. Jeong, “Machine learning techniques for rainfall prediction using atmospheric variables,” IEEE Access, vol. 8, pp. 192921-192935, 2020.

[10] S. Chattopadhyay and P. Chattopadhyay, “Uncertainty estimation and Monte Carlo simulation of an artificial neural network model for rainfall forecasting,” IEEE Trans. Geosci. Remote Sens., vol. 46, no. 8, pp. 2250-2257, Aug. 2008.

[11] S. Maheswaran and R. Khosa, “Improved quantitative precipitation forecast using random forests technique,” Comput. Electron. Agric., vol. 121, pp. 309-317, Apr. 2016.

[12] M. Oo, T. Lwin, and W. W. Hlaing, “Monthly rainfall forecasting using random forest technique,” in Proc. IEEE 9th Int. Conf. Software, Knowledge, Information Management and Applications (SKIMA), 2015, pp. 1-6.

[13] V. Lakshmanan, K. Hondl, and A. Rabinowitz, “Machine learning for robust precipitation estimation from radar,” IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 13, pp. 2634-2644, 2020.

[14] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.

[15] A. Liaw and M. Wiener, “Classification and regression by randomForest,” R News, vol. 2, no. 3, pp. 18-22, 2002.