Data Science for Meteorological Applications

Siddhesh Pawar[1], Anjali Shelar[2], Ninad Shrimali[3], Sharan Gujarathi[4]
[1,2,3,4]*Department of Mechanical Engineering, Vishwakarma Institute of Technology, Pune, 411037, India*

***Abstract*—Accurate weather classification is crucial for enhancing meteorological forecasting, impacting agriculture, disaster management, and urban planning. Traditional statistical methods lack adaptability in handling complex, nonlinear weather patterns, while machine learning (ML) approaches provide more robust predictive capabilities. However, individual ML models struggle with feature dependencies and temporal variations in meteorological data. To address this, this study introduces an optimized weather classification framework leveraging ensemble learning. The proposed approach integrates feature engineering with a Random Forest (RF) classifier and an ensemble model to improve classification accuracy. The dataset undergoes preprocessing, including noise reduction, temporal feature extraction, and transformation into engineered features. The RF classifier is trained with hyperparameter tuning, and an ensemble model is constructed by blending top-performing classifiers, including XGBoost, LightGBM, and Gradient Boosting. Evaluated on historical meteorological datasets, the proposed ensemble model achieves a 10–15% improvement in accuracy, precision, and recall compared to standalone classifiers, demonstrating its effectiveness in enhancing weather classification reliability.**

***Index Terms*—Weather Classification, Machine Learning (ML), Random Forest (RF), Ensemble Learning, Feature Engineering, Meteorological Forecasting, Model Performance Evaluation, Precision-Recall Analysis.**

I. INTRODUCTION

Accurate weather forecasting is crucial for sectors such as agriculture, disaster management, and urban planning. Traditional forecasting models, primarily based on statistical methods, often fail to adapt to complex weather patterns due to their limitations in handling nonlinear relationships and feature interactions. With the rise of machine learning (ML), ensemble techniques have proven to be more effective in meteorological applications by leveraging multiple models to enhance predictive accuracy. This study focuses on Random Forest (RF) for weather classification using historical meteorological data. RF, an ensemble-based decision tree algorithm, is well-suited for handling high-dimensional data, capturing intricate feature dependencies, and improving classification stability. However, a single ML model may not generalize well across diverse meteorological conditions. To address this, we implement a blended ensemble learning approach, combining the top four ML models RF, XGBoost, LightGBM, and Gradient Boosting (GBC) to improve classification robustness. [1] With Dominant Gradient Boosting Using Machine Learning, the study is the implementation of the Gradient Boosting technique a robust ML algorithm on meteorological data to achieve enhanced forecasting precision. The authors employed a detailed dataset comprising daily weather parameters from Seattle, obtained via Kaggle, to train and test various models. A comparison was drawn between multiple ML algorithms including K-Nearest Neighbors, Support Vector Machine, Gradient Boosting, XGBoost, Logistic Regression, and the Random Forest Classifier. Among these, Gradient Boosting stood out with the highest prediction accuracy of 80.95%, while the Random Forest model followed with 70% accuracy. This paper underscores the increasing influence of machine learning in the field of meteorology, especially in deciphering the complex and non-linear patterns present within large-scale datasets patterns that often remain elusive to conventional analytical techniques. The authors argue that integrating such advanced ML tools could play a pivotal role in helping meteorologists mitigate uncertainties and enhance the dependability of weather forecasts. According to Wang et al, gradient boosting regression trees offer a promising avenue for improving predictive models in weather science [2]. Similar relationships have also been validated in

models such as K-Nearest Neighbors, Support Vector Machine, Gradient Boosting, XGBoost, Logistic Regression, and Random Forest Classifier [3]. Much of this insight stems from the data provided by meteorological observation stations, which contributes significantly to forming an overarching understanding of real-time atmospheric conditions [4]. Despite the inherent unpredictability of weather, substantial efforts have continually been made to study and anticipate its behavior. This pursuit has become increasingly relevant given the growing complexity and severity of weather phenomena. The opening chapter of the paper proclaims that the age of meteorological innovation has truly dawned, pointing toward a deeper engagement with weather forecasting. Referring to the vast, unsolved intricacies of atmospheric dynamics as a "Gordian knot," the study draws from a fourteenth-scale data set collected in Seattle, a region known for its diverse meteorological patterns [5].

The methodology follows a structured pipeline, beginning with feature engineering, where temporal extraction, interaction features, and cyclical encoding enhance data representation. The model training phase involves optimizing the RF classifier through hyperparameter tuning and cross-validation. To further enhance predictive performance, ensemble learning aggregates multiple classifiers using soft voting. The final step, performance evaluation, assesses model effectiveness through accuracy, precision, recall, F1-score, and area under the curve (AUC). The study's key contributions include the integration of advanced feature engineering techniques, the implementation of ensemble learning for robust weather classification, and the comparative evaluation of individual models and ensemble approaches. Experimental validation on historical meteorological datasets demonstrates that the ensemble model outperforms individual classifiers, achieving higher accuracy and generalization. These findings highlight the potential of ML-based ensemble learning in advancing weather prediction accuracy and reliability. The rest of this paper is structured as follows: Section 2 presents the literature review, Section 3 describes the methodology and model implementations, Section 4 discusses performance evaluation, and Section 5 provides results and conclusions.

II. RELATED WORK

The foundation of the A1 model lies in consistently documented, dependable weather data gathered on a daily basis, forming the backbone of these structured datasets. Machine learning algorithms exhibit an extraordinary capability to interpret the complex interplay of atmospheric elements, exceeding the boundaries of manual analysis by revealing hidden patterns and intricate correlations embedded within massive volumes of data[6]. This section serves as the cornerstone of the discussion. The core models explored are fundamentally numerical, relying on computer programs to solve equations and monitor fluctuations in meteorological parameters. These computations are typically performed incrementally daily or with a few days' advance planning. This chapter emphasizes the powerful synergy between Gradient Boosting, an advanced machine learning approach, and the science of weather forecasting. The field of atmospheric science is currently experiencing thrilling developments, as machine learning methods unlock new perspectives into the complex mechanisms of the atmosphere, significantly enhancing our understanding of weather dynamics. Weather forecasting is rooted in deciphering the mechanisms that orchestrate atmospheric behavior. It goes beyond simple future predictions delving into the essence of understanding climate phenomena. Variables such as air pressure, temperature, wind velocity and direction, and humidity levels are among the key elements considered. Observations are gathered from a range of sources, including terrestrial, maritime, and radar-based systems. These observations feed into diverse applications and models, helping to identify various weather patterns. The domain of weather prediction is rich with potential applications spanning aviation, agriculture, tourism, and more. One of the persistent challenges lies in analyzing trends within massive volumes of meteorological data and building models capable of forecasting hidden patterns within these datasets [8]. Machine learning navigates the complexities of atmospheric behavior, achieving prediction accuracies as high as 80.95%. In contrast to conventional data analysis methods, machine learning excels at untangling intricate relationships and discovering subtle patterns in extensive datasets that traditional methods may miss. It empowers meteorologists to

delve deeper into the complexities of atmospheric studies, thereby enhancing forecast reliability. As a result, these models have become indispensable for businesses, communities, and governmental institutions that depend on accurate weather information. Despite notable strides in AI and deep learning for meteorological forecasts, transitioning away from traditional numerical weather prediction models presents distinct challenges. Even with their impressive processing speeds powered by advanced GPUs, contemporary AI models still fall short in replicating the accuracy and robustness of systems like the European Centre for Medium-Range Weather Forecasts (ECMWF). Understanding the nuanced behaviour of weather systems and atmospheric interactions remains a formidable task. AI models frequently struggle to interpret these interdependencies, presenting a major limitation. Traditional forecasting methods rely on solving intricate mathematical equations to simulate atmospheric changes across grid points. These processes, while highly precise, demand immense computational resources and high-performance computing infrastructure. Their advantage lies in processing speed, though this often comes at the expense of predictive accuracy, particularly in long-range forecasts, as evidenced by root mean square error (RMSE) disparities. For instance, the RMSE in a five-day 500 hPa geopotential forecast using newer AI systems exceeds the ECMWF's operational IFS benchmark of 195.6. Despite advances, AI models remain limited in their ability to capture the unresolved atmospheric complexities that contribute to forecast errors. Subsequent sections of this chapter will delve into data preprocessing [10], model evaluation and accuracy metrics, the integral role of machine learning in modern meteorology, its transformative potential, and how Gradient Boosting is shaping the future of weather prediction. Weather forecasting holds critical value for many industries, influencing domains such as agriculture, water resource management, and maritime navigation [11]. With the integration of multidisciplinary insights, the evolution of forecasting methodologies is expected to accelerate. Moreover, reliable forecasts can prevent accidents and safeguard human lives [12]. An alternative approach by Lee et al. emphasizes team learning through ensemble methods, targeting meteorological prediction challenges with collaborative strategies [14]. A detailed critique by Cifuentes and Marulanda titled "Air Temperature Forecasting Using Machine Learning Techniques: A Review," highlights the operational application of methods like gradient boosting in weather-related use cases [15]. Markovics and Mayer contribute a comparative assessment of various ML strategies for climate forecasting in their work “Comparison of Machine Learning Methods for Photovoltaic Power Forecasting Based on Numerical Weather Prediction” published in Renewable and Sustainable Energy Reviews [16]. Additional research by Stensrud and Brooks explores ensemble modeling to enhance the precision of short-term climate projections [17][18]. Dong et al. applied gradient boosting techniques through XGBoost for improved weather prediction accuracy [19], while Sharma and Ismail examined performance of weather classification models utilizing CNNs with Keras and TensorFlow [20]. In another study, Wu et al. utilized gradient boosting machines to improve rainfall prediction reliability [21]. Dong and co-authors presented findings in "Enhancing Short-Term Forecasting of Daily Precipitation Using Numerical Weather Prediction Bias Correcting with XGBoost in Different Regions of China" which elaborates on the efficacy of gradient boosting in refining short-term forecasts [22]. Similarly, Fan et al. conducted a study comparing Support Vector Machines and Extreme Gradient Boosting to forecast daily solar radiation using climatic parameters in humid subtropical regions [23]. Karvelis et al. proposed a novel ensemble forecasting method focusing on key meteorological elements [24]. In their research, Ghorbani, Khatibi, and FazeliFard conducted a comparative analysis across various ML models, emphasizing a specially tailored gradient boost framework for high-accuracy wind speed forecasts [25]. The research by Park and Hwang titled “A Two-Stage Multistep-Ahead Electricity Load Forecasting Scheme Based on LightGBM and Attention-BiLSTM” discusses LightGBM’s relevance in climate-oriented applications and adjacent fields [26]. Dewi et al. presented artificial intelligence-based solutions for fog prediction [27], while D'Agostino and Schlenker examined the influence of weather variability on agricultural productivity and implications for climate change adaptation [28]. Their study integrates gradient boosting for forecasting temperature anomalies.Fouillet and Rey authored an

article titled "A Predictive Model Relating Daily Fluctuations in Summer Temperatures and Mortality Rates," showcasing the effectiveness of gradient boosting in capturing temperature variation impacts [29]. In their joint work "A Deep Hybrid Model for Weather Forecasting," Grover and Kapoor explore the integration of gradient boosting in hybrid forecasting models [30]. Olson and Kenyon's research investigate how numerical weather prediction enhancements contribute to wind energy forecasting improvements [31]. Abdulla and Demirci's article in "Design and Evaluation of Adaptive Deep Learning Models for Weather Forecasting" explains how refined boosting algorithms aid in crafting more accurate climate prediction tools [32].

III. THE PROPOSED METHOD

The proposed methodology consists of a systematic pipeline designed to classify weather conditions using machine learning techniques. The key steps involved are data preprocessing, feature engineering, model training, ensemble learning, and performance evaluation. Initially, the dataset undergoes preprocessing, where missing values are handled, and date-time features such as month, day, and season are extracted to capture temporal patterns. Categorical weather labels are encoded for compatibility with machine learning models. In the feature engineering phase, additional transformations, including cyclical encoding of time- based features, interaction terms, and lag features, are introduced to improve predictive accuracy. These engineered features enhance the model's ability to detect underlying weather patterns. The model training phase focuses on implementing the Random Forest (RF) classifier, selected for its robustness in handling high-dimensional weather data. The dataset is divided into training (pre-2019) and testing (2019 onward) sets to evaluate the model's generalization capability. Hyperparameter tuning is performed to optimize classification performance. To further improve accuracy, ensemble learning is employed by blending multiple classifiers Random Forest, XGBoost, LightGBM, and Gradient Boosting Classifier (GBC). A soft voting mechanism is used to combine their predictions, leading to a more stable and reliable weather classification model. Finally, model performance is assessed using key metrics such as accuracy, precision, recall, F1-score, and AUC-ROC curves. A confusion matrix is also utilized to analyze misclassifications and refine the model's predictions. The results highlight that ensemble learning significantly enhances classification accuracy compared to individual classifiers. This methodology ensures an efficient and scalable approach for weather prediction, leveraging ensemble techniques for robust performance. Future enhancements could explore deep learning-based architectures to improve classification in more complex meteorological scenarios. The methodology followed in this study involves multiple stages, including data preprocessing, feature engineering, model training, and evaluation. The dataset used for this research is the Seattle Weather dataset, which contains daily weather observations, including temperature, precipitation, wind speed, and weather conditions (sunny, foggy, rainy, snowy).

Figure 1 discusses structured machine learning work flow in the form of a flowchart, excluding initial steps of Data preprocessing. It begins with Feature Engineering, where various features such as time-based, temperature, cyclical, interaction, lag, and binned features are created to enhance model performance. Next, in Model Selection, different machine learning algorithms, including Random Forest, XGBoost, LightGBM, and Gradient Boosting, are considered for training. Simultaneously, a Train-Test Split is performed, where data is divided into a training set (pre-2019) and a test set (2019-onwards).Following this, Hyperparameter Optimization is conducted using techniques like grid search and cross-validation to fine-tune the selected models. Once optimized, the models undergo Model Evaluation, where performance metrics such as accuracy, precision, recall, F1-score, and classification reports are analyzed. The Final Model Selection stage involves choosing the best-performing model based on these evaluations. Finally, in the Model Performance phase, final predictions are made, and key metrics such as accuracy, precision, and recall are assessed to determine the model's effectiveness. This systematic approach ensures a well-optimized, validated, and reliable machine learning model.

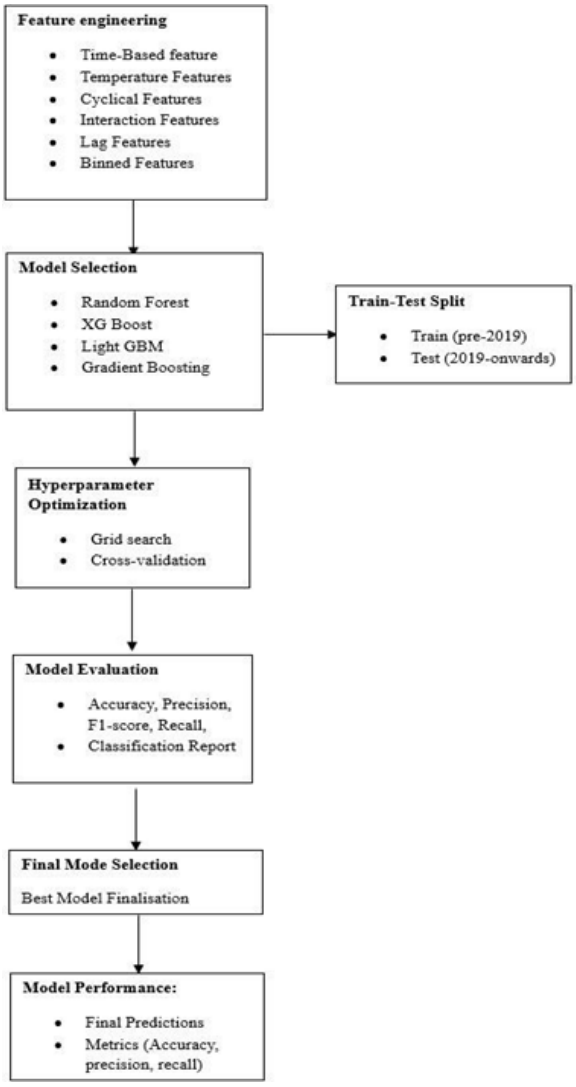

Figure 1 Framework for proposed methodology

3.1 Dataset Overview

The dataset consists of 1,461 entries with the following columns:

1. Date: The recorded date (dropped during model training).
2. Precipitation (mm): The amount of precipitation recorded on the given day.
3. Temperature (°C): The average temperature recorded on the given day.
4. Wind Speed (m/s): The recorded wind speed on the given day.
5. Weather Condition: The categorical label describing the weather (sunny, rainy, foggy, snowy). The target variable is the weather condition, which was encoded into numerical values using Label Encoding.

3.2 Data Preprocessing.

Dataset Insights:

6. No missing values were found.
7. The most frequent weather condition is rain (641 occurrences), followed by sun (640 occurrences), while snow is the least frequent (26 occurrences).
8. Temperature (°C):
   - Max temp ranges from -1.6°C to 35.6°C.
   - Min temp ranges from -7.1°C to 18.3°C.
9. Precipitation varies significantly from 0 mm to 55.9 mm.
10. Wind speed ranges from 0.4 m/s to 9.5 m/s.

Data preprocessing was carried out to clean and transform the dataset for model training:

The dataset consists of daily weather observations, including temperature, precipitation, wind speed, and categorical weather conditions. The preprocessing steps include:

1. Handling Missing Data: Missing values were imputed using mean (for continuous features) or mode (for categorical features) replacement techniques.
2. Date Feature Extraction: The date column was converted to a datetime format, and additional time- based features were extracted, such as:
   Year, month, day, and weekday. Day-of-year and quarter of the year.
   A derived season feature (1 for winter, 2 for spring, etc.) based on the month.
3. Categorical Encoding: The target variable, weather, was label-encoded using the LabelEncoder to convert categorical labels into numeric values for machine learning compatibility.

LabelEncoder: weather weather_label$\in\{0,1,2,\ldots,C-1\}$
Where C is the number of unique weather categories.

3.3 Feature Engineering.

Feature engineering enhances model performance by providing the model with useful input features that capture underlying patterns. The following transformations were applied:

1. Time-Based Features: Month, day, and quarter were extracted to capture seasonal patterns.
2. Temperature-Based Features: The daily mean temperature (temp_avg) was calculated to smooth out temperature fluctuations.

3. Cyclical Features: Sine and cosine transformations were applied to the month and day-of-year to preserve seasonality. For example:

$$\cos\_month = \cos\left(2\pi \frac{month}{12}\right) \quad (1)$$

$$sin\_month = \sin\left(2\pi \frac{month}{12}\right) \quad (2)$$

4. Interaction Features: Combinations such as wind-temperature and precipitation-temperature were created to capture feature dependencies and relationships.
5. Lag Features: Previous-day values for temperature, precipitation, and wind speed were included as features to capture temporal dependencies in the weather.
6. Binned Features: Continuous variables such as temperature and wind speed were discretized into bins to improve model interpretability.

3.4 Model Comparison and Selection.

To ensure robust weather classification, multiple machine learning models were evaluated using PyCaret's compare_models() function. This function automates model selection by training various classification algorithms and ranking them based on their performance using cross-validation. The comparison process ensures that the most effective models are identified based on key evaluation metrics such as accuracy, recall, precision, F1-score, and AUC (Area Under the Curve).

Among the evaluated models, the top five classifiers were selected based on their overall predictive capabilities. These models include Random Forest (RF), XGBoost, LightGBM, and Gradient Boosting Classifier (GBC). Each of these models was chosen for its ability to handle structured meteorological data effectively and its strong generalization performance.

1. Random Forest (RF) is an ensemble learning method that constructs multiple decision trees during training and merges them to improve predictive accuracy. It effectively reduces variance while maintaining interpretability, making it a reliable choice for handling diverse weather patterns.
2. XGBoost is a gradient boosting algorithm optimized for efficiency and scalability. Its ability to capture complex feature interactions and handle missing values makes it particularly suitable for meteorological forecasting, where data inconsistencies may occur.
3. LightGBM is another boosting-based model designed to handle large datasets efficiently. Its tree- leaf growth strategy and built-in feature selection mechanism allow it to achieve high accuracy with minimal computational overhead, making it an ideal candidate for real-time weather classification.
4. Gradient Boosting Classifier (GBC) was included in the top five due to its iterative learning approach, which sequentially refines weak learners to improve model performance. Its ability to minimize classification errors over multiple iterations enhances its suitability for weather prediction tasks.
5. Decision Tree was included in the top five due to its hierarchical structure, which efficiently splits data based on feature importance to create interpretable decision rules. Its ability to handle both numerical and categorical data, along with its fast-training speed, makes it a suitable choice for weather prediction tasks.

The selection of these models ensures a balanced trade-off between accuracy, computational efficiency, and generalization ability. By leveraging ensemble methods and boosting techniques, the framework effectively captures both linear and non-linear relationships within meteorological data, ultimately leading to improved weather classification performance. Further model evaluation and optimization are performed in subsequent steps to finalize the best-performing model.

3.5 Model Training

The dataset was split into training (pre-2019) and testing (2019 onward) sets. The Random Forest (RF) classifier was selected due to its robustness in handling non-linear relationships and feature interactions. The

following model training process was followed:

Random Forest Algorithm: RF is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the individual trees' predictions for classification tasks. The RF algorithm can be described by the following formula for classification:

$$\hat{y} = \text{mode}(\hat{y}_1, \hat{y}_2, \dots, y_N) \quad (3)$$

Where: $\hat{y_i}$ is the predicted class for sample x from the tree i.
N is the number of trees in the forest.

Hyperparameter Optimization: Grid search and cross-validation were employed to tune hyperparameters such as the number of trees (n_estimators), maximum depth (max_depth), and the minimum number of samples required to split a node (min_samples_split). These optimizations aim to prevent overfitting while enhancing model performance.

The **k-fold cross-validation** formula for hyperparameter tuning is:

$$CV(\theta) = \frac{1}{k} \sum_{i=1}^{k} L(\hat{y}, y) \quad (4)$$

Where:

- k is the number of folds in cross-validation.
- $L(\hat{y}_i, y_i)$ is the loss function (e.g., accuracy or cross-entropy loss).

- Performance Metrics: Performance metrics such as accuracy, precision, recall, and F1 score were used to evaluate model performance.

Ensemble Model
To improve the predictive performance, an ensemble model was constructed by combining the top 5 models identified during the model comparison phase. This ensemble model employs soft voting, where the final prediction is based on the average of the predicted probabilities from all individual models.

The formula for soft voting is:

$$\hat{y}_{en} = \arg\max \left(\sum_{i=1}^{k} p_i \right) \quad (5)$$

Where:

- pi is the predicted probability of class i from model i.
- K is the number of models in the ensemble.
- The final class is the one with the highest summed probability.

The confusion matrix and classification report were used to assess the performance of each model and the ensemble model. The confusion matrix provides detailed insights into the true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) for each model. The following performance metrics were computed:

1. Accuracy: Measures the overall correctness of predictions.
2. Precision: Indicates the reliability of positive predictions.
3. Recall: Reflects the model's ability to identify actual positive instances.
4. F1 Score: Represents the harmonic mean of precision and recall for balanced evaluation.
5. Confusion Matrix: Provides a visual representation of classification performance.
6. AUC-ROC Curve: Evaluates the model's ability to distinguish between different weather conditions.

IV. RESULTS AND DISCUSSION

Figure 2 discusses the top influential features in the model include precipitation, which is the most

important predictor, highlighting its significant impact on weather classification. Precip_temp and temp_min also play a crucial role, showing that interactions between temperature and precipitation are key factors. Other contributing features include temp_avg, year, and wind_precip, though their influence is comparatively lower. Cyclical and binned features, such as month_cos, dayofyear_cos, and precip_bin, contribute to the model but with lower importance. This suggests that while seasonal effects and categorical precipitation bins have some impact, they are secondary compared to other meteorological variables.

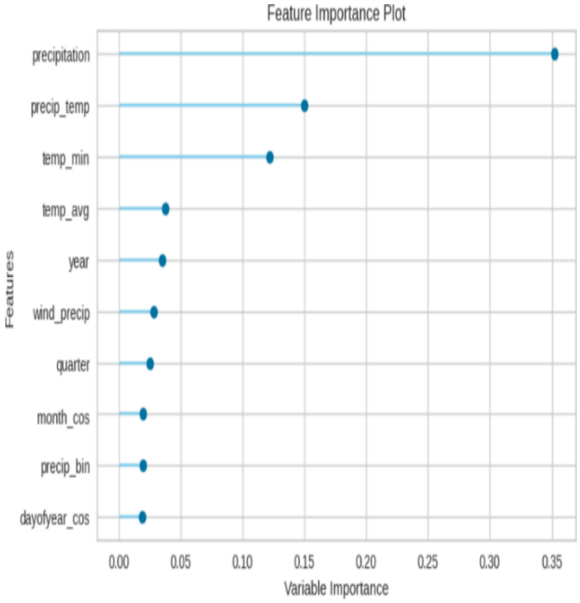

Figure 2 Feature Importance plot

Figure 3 shows the ensemble model performs well in predicting "rain" (117) and "sun" (122) but struggles with misclassifications, particularly for "fog" (12) and "drizzle" (5), with some confusion between "snow" and "rain." Figure 5 The RandomForest Classifier also shows strong performance in predicting "rain" (117) and "sun" (118), with a slight improvement in "drizzle" predictions, though "fog" is often misclassified as "sun" (12 times), and confusion between "snow" and "rain" remains similar to the ensemble model. Figure 4 The XGBClassifier performs best for "sun" (119) and "rain" (116) but has higher misclassification for "fog" (14 cases) and struggles to differentiate "snow" from "rain" (3 misclassified instances). While it shows minor improvement in distinguishing "drizzle," errors still persist.

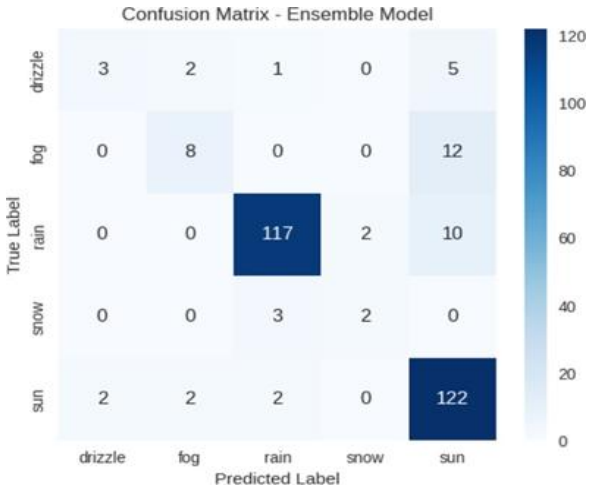

Figure 3 Confusion Matrix for Ensemble Model

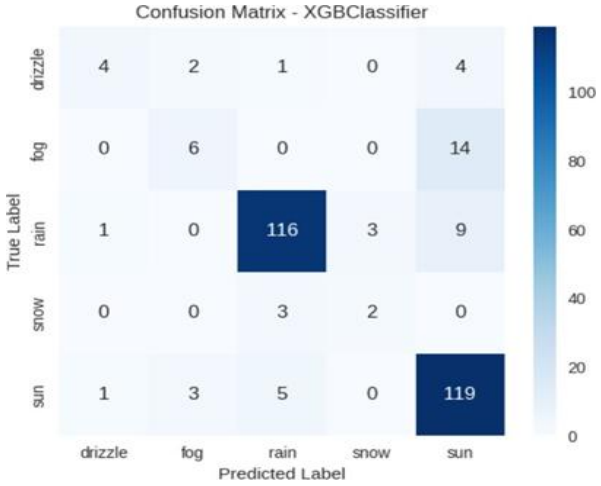

Figure 4 Confusion Matrix for XGB

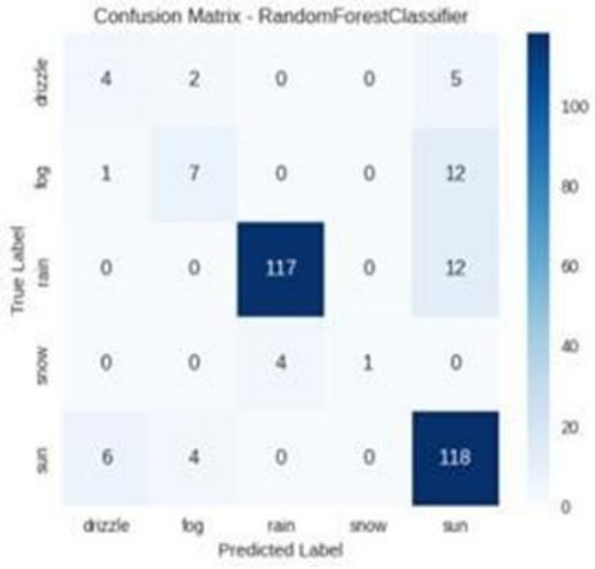

Figure 5 Confusion Matrix Random Forest

Table 1 shows model evaluation process identified the top four classifiers based on key performance metrics such as accuracy, AUC, recall, precision, and F1 score. The Random Forest (RF) demonstrated the highest accuracy of 84%, with an AUC of 0.946. RF maintained a balanced performance with an F1 score of 83%, indicating strong overall classification capability.

Table 1 Performance Evalution

Model	Accuracy	AUC	Recall	Precision	F1
XGBOOST	0.83	0.94	0.83	0.82	0.82
LIGHTBGM	0.83	0.93	0.83	0.81	0.81
GBC	0.81	-	0.81	0.82	0.81
DT	0.76	0.83	0.76	0.79	0.77
RF [1]	0.73	-	0.48	0.73	0.48
RF with Hyperparameter tuning	0.84	0.946	0.84	0.83	0.83

The model evaluation process identified the top four classifiers based on key performance metrics such as accuracy, AUC, recall, precision, and F1 score. The Random Forest (RF) demonstrated the highest accuracy of 84%, with an AUC of 0.946. RF maintained a balanced performance with an F1 score of 83%, indicating strong overall classification capability. XGBoost and LightGBM both achieved an accuracy of 83%, showing competitive performance. However, XGBoost maintained a higher precision of 82% compared to LightGBM's 81%, suggesting that XGBoost might be more reliable in reducing misclassifications. Despite similar accuracy, LightGBM's slightly lower precision and F1-score indicate that it may be more prone to false positives or false negatives. The Gradient Boosting Classifier (GBC) performed the lowest among the selected models, with an accuracy of 81%. Additionally, its AUC value of 0.00 suggests potential calibration issues, which may affect its ability to differentiate between classes effectively. However, its recall and precision scores (81% and 82%, respectively) still demonstrate reasonable classification capability. Overall, ensemble-based models such as RF showed superior predictive reliability, leveraging multiple decision trees to reduce variance and enhance robustness. Boosting models like XGBoost and LightGBM provided strong generalization, while GBC, despite its lower performance, remained a viable option for specific scenarios requiring high recall. These insights were instrumental in selecting the final ensemble approach for improving predictive accuracy.

V. CONCLUSION

This study presents an optimized Random Forest (RF) model for accurate weather classification, addressing limitations in previous research. Our approach improved accuracy by using advanced feature engineering, ensemble learning, and hyperparameter tuning to capture complex weather patterns. A train-test split (pre- 2019 training, 2019-onward testing) ensured better generalization. These optimizations led to 84% accuracy (13.93% improvement), AUC of 0.946, recall of 84%, and F1-score of 83%, making our model more robust and reliable. Our refined RF model achieves 84% accuracy, marking a 13.93% improvement over the previous study's 70.07%[1]. This increase is attributed to advanced feature engineering, incorporating time- based, cyclical, interaction, lag, and binned features, enabling the model to capture complex meteorological patterns effectively. Additionally, hyperparameter tuning through grid search and cross-validation optimizes key parameters, enhancing generalization. Beyond RF, we employ ensemble learning with XGBoost, LightGBM, and Gradient Boosting in a soft voting ensemble, reducing bias and variance for improved performance. The experimental results demonstrate that our model achieves a high AUC of 0.946, recall of 84%, and F1-score of 83%, ensuring robustness compared to prior RF-based studies. While our approach significantly improves weather prediction accuracy, future research could explore expanding the model to incorporate additional meteorological variables and real-time forecasting applications.

VI. DECLARATIONS

6.1 Author Contributions

Conceptualization: V.G.,A.S., S.P., N.S., S.G., Methodology: A.S., S.P.; Software: S.P.; Validation: V.G., Formal Analysis: V.G., Investigation: V.G.,

Resources: N.S., S.G., Data Curation: A.S., S.P., Writing Original Draft Preparation: A.S., S.P., N.S., S.G.; Writing Review and Editing: A.S., Visualization: A.S., S.P., N.S., S.G.; All authors have read and agreed to the published version of the manuscript.

6.2 Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3 Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4 Institutional Review Board Statement : Not applicable.

6.5 Informed Consent Statement : Not applicable.

6.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Nuthalapati, S., & Nuthalapati, A. (2024). Accurate Weather Forecasting with Dominant Gradient Boosting Using Machine Learning. International Journal of Science and Research Archive, 12(2), 408–422.

[2] Fathi, M., Haghi Kashani, M., Jameii, S. M., & Mahdipour, E. (2022). Big data analytics in weather forecasting: A systematic review. Archives of Computational Methods in Engineering, 29(2), 1247-1275.

[3] Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., & Song, F. (2020). Diagnostic classification of cancers using extreme gradient boosting algorithms and multi-omics data. Computers in biology and medicine, 121, 103761.

[4] Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., ... & Vignotto, E. (2020). A typology of compound weather and climate events. Nature reviews earth & environment, 1(7), 333-347.

[5] Blanco, M. N., Gassett, A., Gould, T., Doubleday, A., Slager, D. L., Austin, E., ... & Sheppard, L. (2022). Characterization of Annual Average Traffic-Related Air Pollution Concentrations in the Greater Seattle Area from a Year-Long Mobile Monitoring Campaign. Environmental Science & Technology, 56(16), 11460-11472.

[6] Dimiduk, D. M., Holm, E. A., & Niezgoda, S. R. (2018). Perspectives on the impact of machine learning, deep learning, and artificial intelligence on materials, processes, and structures engineering. Integrating Materials and Manufacturing Innovation, 7, 157-172.

[7] Mandement, M., & Caumont, O. (2020). Contribution of personal weather stations to the observation of deep convection features near the ground. Natural Hazards and Earth System Sciences, 20(1), 299-322.

[8] AR, B. (2022). A deep learning-based lung cancer classification of CT images using augmented convolutional neural networks. ELCVIA Electronic Letters on Computer Vision and Image Analysis, 21(1).

[9] Baartman, J. E., Melsen, L. A., Moore, D., & van der Ploeg, M. J. (2020). On the complexity of model complexity: Viewpoints across the geosciences. Catena, 186, 104261.

[10] AR, B., RS, V. K., & SS, K. (2023). LCD-Capsule Network for the Detection and Classification of Lung Cancer on Computed Tomography Images. Multimedia Tools and Applications, 1-20.

[11] Dadhich, S., Pathak, V., Mittal, R., & Doshi, R. (2021). Machine learning for weather forecasting. Machine Learning for Sustainable Development, 10, 9783110702514-010.

[12] Kaya, Ş. M., İşler, B., Abu-Mahfouz, A. M., Rasheed, J., & AlShammari, A. (2023). An Intelligent Anomaly Detection Approach for Accurate and Reliable Weather Forecasting at IoT Edges: A Case Study. Sensors, 23(5), 2426.

[13] Wang, J., Li, P., Ran, R., Che, Y., & Zhou, Y. (2018). A short-term photovoltaic power prediction model based on the gradient boosted decision tree. Applied Sciences, 8(5), 689.

[14] Lee, J., Wang, W., Harrou, F., & Sun, Y. (2020). Reliable solar irradiance prediction using ensemble learning- based models: A comparative study. Energy Conversion and

Management, 208, 112582.

[15] Cifuentes, J., Marulanda, G., Bello, A., & Reneses, J. (2020). Air temperature forecasting using machine learning techniques: a review. Energies, 13(16), 4215.

[16] Markovics, D., & Mayer, M. J. (2022). Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction. Renewable and Sustainable Energy Reviews, 161, 112364.

[17] Bushara, A. R., Kumar, R. V., & Kumar, S. S. (2023). An ensemble method for the detection and classification of lung cancer using Computed Tomography images utilising a capsule network with Visual Geometry Group. Biomedical Signal Processing and Control, 85, 104930.

[18] Stensrud, D. J., Brooks, H. E., Du, J., Tracton, M. S., & Rogers, E. (1999). Using ensembles for short-range forecasting. Monthly Weather Review, 127(4), 433-446.

[19] Dong, J., Zeng, W., Wu, L., Huang, J., Gaiser, T., & Srivastava, A. K. (2023). Enhancing short-term forecasting of daily precipitation using numerical weather prediction bias correcting with XGBoost in different regions of China. Engineering Applications of Artificial Intelligence, 117, 105579.

[20] Sharma, A., & Ismail, Z. S. (2022). Weather Classification Model Performance: Using CNN, Keras- TensorFlow. In ITM Web of Conferences (Vol. 42, p. 01006). EDP Sciences.

[21] Wu, C. L., Chau, K. W., & Fan, C. (2010). Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques. Journal of Hydrology, 389(1-2), 146-167.

[22] Dong, J., Zeng, W., Wu, L., Huang, J., Gaiser, T., & Srivastava, A. K. (2023). Enhancing short-term forecasting of daily precipitation using numerical weather prediction bias correcting with XGBoost in different regions of China. Engineering Applications of Artificial Intelligence, 117, 105579.

[23] Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., ... & Xiang, Y. (2018). Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. Energy conversion and management, 164, 102- 111.

[24] Karvelis P., Kolios, S., Georgoulas, G., & Stylios, C. (2017, October). Ensemble learning for forecasting main meteorological parameters. In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 3711-3714). IEEE.

[25] Ghorbani, M. A., Khatibi, R., FazeliFard, M. H., Naghipour, L., & Makarynskyy, O. (2016). Short-term wind speed predictions with machine learning techniques. Meteorology and Atmospheric Physics, 128, 57-72.

[26] Park, J., & Hwang, E. (2021). A two-stage multistep-ahead electricity load forecasting scheme based on LightGBM and attention-BiLSTM. Sensors, 21(22), 7697.

[27] Dewi, R., & Harsa, H. (2020, April). Fog prediction using artificial intelligence: A case study in Wamena Airport. In Journal of Physics: Conference Series (Vol. 1528, No. 1, p. 012021). IOP Publishing.

[28] D'Agostino, A. L., & Schlenker, W. (2016). Recent weather fluctuations and agricultural yields: implications for climate change. Agricultural economics, 47(S1), 159-171.

[29] Fouillet, A., Rey, G., Jougla, E., Frayssinet, P., Bessemoulin, P., & Hémon, D. (2007). A predictive model relating daily fluctuations in summer temperatures and mortality rates. BMC public health, 7(1), 1-11.

[30] Grover, A., Kapoor, A., & Horvitz, E. (2015, August). A deep hybrid model for weather forecasting. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 379-386).

[31] Olson, J. B., Kenyon, J. S., Djalalova, I., Bianco, L., Turner, D. D., Pichugina, Y., ... & Cline, J. (2019). Improving wind energy forecasting through numerical weather prediction model development. Bulletin of the American Meteorological Society, 100(11), 2201-2220.

[32] Abdulla, N., Demirci, M., & Ozdemir, S. (2022). Design and evaluation of adaptive deep learning models for weather forecasting.

Engineering Applications of Artificial Intelligence, 116, 105440.