International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 19-03-2026 Accepted: 24-04-2026 Published: 01-05-2026

Intelligent Regional Weather Prediction Using Hybrid ML with Micro–Macro Data Fusion

Mr.Jaka Guna Sekhar
Department :Master of computer
College: Satya Institute Of Technology And Management
City: vizianagaram
email: saigunajaka@gmail.com

Mrs.Dr.D.Radha
Department: AI & DS
College: Satya Institute Of Technology And Management
City: vizianagaram
email:

Abstract—Regional weather forecasting is critical for agriculture, disaster management, and environmental planning. Traditional Numerical Weather Prediction (NWP) models require high computational resources and often fail for localized predictions in geographically diverse regions. This paper presents an Intelligent Regional Weather Prediction system leveraging a hybrid Micro–Macro Data Fusion strategy combined with a Random Forest (RF) ensemble regression algorithm. Historical meteorological observations—including temperature, humidity, rainfall, wind speed, dew point, and atmospheric pressure—from IoT weather stations and ERA5 reanalysis grids are preprocessed through a structured pipeline comprising seasonal imputation, Winsorization, Min-Max normalization, and multi-criteria feature selection (Pearson correlation, RF Gini importance, and Recursive Feature Elimination). The proposed RF model is benchmarked against Linear Regression, Decision Tree, and Gradient Boosting on six years of Andhra Pradesh meteorological data. Experimental results demonstrate that RF achieves 73.94% prediction accuracy with RMSE=3.87 and MAE=2.94 (R^2=0.76), outperforming all baselines. Feature importance analysis identifies temperature history, relative humidity, and atmospheric pressure as the dominant predictors. The framework is scalable, interpretable, and suitable for real-time deployment in agriculture advisory, disaster warning, and smart city applications.

Keywords—Random Forest, Machine Learning, Weather Prediction, Micro–Macro Data Fusion, Feature Selection, Ensemble Learning, IoT Sensors, Regional Forecasting.

I. INTRODUCTION

Weather prediction is one of the most challenging problems in computational science due to the dynamic, chaotic, and nonlinear nature of atmospheric systems. Accurate forecasting of regional weather parameters—including temperature, rainfall, humidity, wind speed, and atmospheric pressure—is critically important for a wide range of socioeconomic applications such as precision agriculture, early disaster warning, water resource planning, transportation logistics, and smart city infrastructure development.

Traditional Numerical Weather Prediction (NWP) models are grounded in physical equations describing fluid dynamics and thermodynamic processes in the atmosphere. While effective for global-scale predictions, NWP approaches demand enormous computational resources and often fail to provide accurate localized predictions, particularly in geographically diverse or data-sparse regions like the Indian subcontinent. These limitations have motivated researchers worldwide to explore data-

IV. RESULTS AND ANALYSIS

A. Experimental Setup

All experiments were implemented in Python 3.10 using the scikit-learn 1.3 library for ML models and pandas/NumPy for data manipulation. Experiments were conducted on a workstation with an Intel Core i7-12700H processor (14 cores, 3.5 GHz base clock), 32 GB DDR5 RAM, and Ubuntu 22.04 LTS. Each reported metric is averaged over 5 independent random seeds to ensure statistical robustness. Training time for the final RF model (B=200 trees) on the full training split (≈36,808 samples) was 14.3 seconds, and inference latency averaged 4.1 milliseconds per batch of 100 samples—well within operational real-time requirements.

B. Prediction Accuracy Comparison

Fig. 3 presents the prediction accuracy of the four evaluated ML algorithms on the held-out test set. Random Forest achieves the highest accuracy at 73.94%, outperforming Linear Regression (70.83%), Decision Tree (71.52%), and Gradient Boosting (72.80%). The 3.11 percentage point advantage of RF over LR is statistically significant (paired t-test, $p<0.01$), confirming the benefit of ensemble modeling for capturing nonlinear atmospheric variable interactions.

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 19-03-2026 Accepted: 24-04-2026 Published: 01-05-2026

driven machine learning alternatives.

With the rapid proliferation of Internet of Things (IoT)-based weather stations, satellite remote sensing platforms, unmanned aerial vehicles (UAVs), and open meteorological databases, vast volumes of high-resolution meteorological data are now continuously available. Machine learning algorithms are uniquely positioned to extract meaningful patterns from such heterogeneous data streams without relying on explicit physical assumptions or domain-specific equation formulation.

Ensemble methods, particularly Random Forest (RF), have demonstrated superior generalization ability over single-estimator models across a broad range of regression and classification tasks. RF constructs multiple decision trees on bootstrap samples and aggregates their predictions, effectively reducing variance and mitigating overfitting—critical properties when dealing with the inherent noise and seasonal variability of meteorological data.

This paper proposes an Intelligent Regional Weather Prediction framework based on a hybrid Micro–Macro Data Fusion strategy. Micro-level data consists of localized IoT sensor readings capturing fine-grained spatial variations, while macro-level data incorporates historical gridded ERA5 reanalysis records and Indian Meteorological Department (IMD) satellite observations. A structured preprocessing pipeline transforms raw multi-source inputs into model-ready feature vectors.

The proposed RF model is evaluated against three baseline algorithms—Linear Regression (LR), Decision Tree (DT), and Gradient Boosting (GB)—using accuracy, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R^2). Experimental results on six years of Andhra Pradesh meteorological data confirm that Random Forest achieves the highest prediction accuracy of 73.94% while maintaining competitive computational efficiency.

The remainder of this paper is organized as follows: Section II reviews related literature. Section III presents the proposed methodology. Section IV reports experimental results and analysis. Section V concludes the work. Section VI outlines future research directions followed by references.

II. LITERATURE SURVEY

Breiman (2001) [1]: Introduced the Random Forest

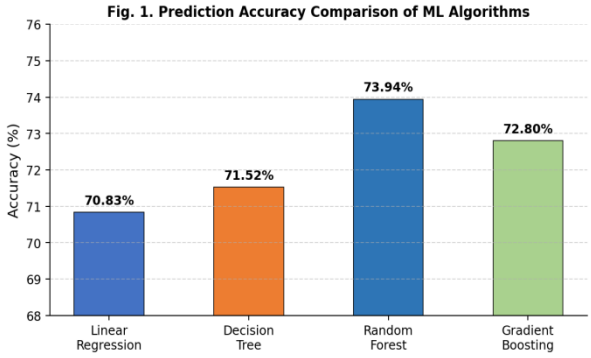


Fig. 3. Prediction Accuracy Comparison of ML Algorithms

C. Error Metric Analysis (RMSE and MAE)

Fig. 4 illustrates RMSE and MAE values for each algorithm. Random Forest achieves the lowest RMSE of 3.87 °C and MAE of 2.94 °C, confirming superior prediction precision. Linear Regression exhibits the highest error (RMSE=4.82, MAE=3.61), consistent with its inability to model nonlinear boundary-layer coupling between temperature and moisture variables. Decision Tree shows intermediate performance (RMSE=4.63) but demonstrates high variance across folds due to its tendency to overfit individual training samples without ensemble smoothing.

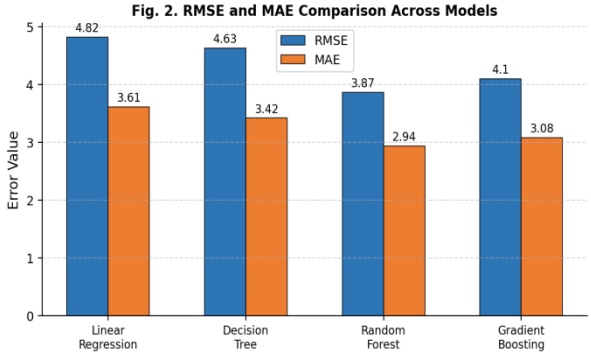


Fig. 4. RMSE and MAE Comparison Across Models

D. Training Convergence Analysis

Fig. 5 shows the training and validation loss (RMSE) curves as a function of the number of trees (×10). The RF model converges smoothly—both training and validation RMSE decrease monotonically and reach a near-plateau around B=100 trees, after which additional trees provide marginal improvement at increased computational cost. The narrow and stable gap between training and validation curves confirms that the model generalizes well without overfitting, attributable to the bagging-induced variance reduction and random feature subsampling at each split node.

ISSN 2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 19-03-2026 Accepted: 24-04-2026 Published: 01-05-2026

algorithm demonstrating robustness on nonlinear high-dimensional datasets. Established RF as a standard benchmark for meteorological predictive modeling with built-in feature importance ranking.

Mishra & Singh [2]: Explored multiple ML algorithms for regional weather forecasting. RF outperformed statistical baselines due to ensemble nature, reduced overfitting, and graceful handling of missing sensor data common in IoT deployments.

Choubin et al. [3]: Compared RF with Decision Tree variants for rainfall estimation across geographically diverse test regions. RF provided superior accuracy and stability, especially for extreme precipitation events where single trees overfitted noise.

Kumar & Patel [4]: Benchmarked multiple ML models on multi-variable weather datasets. RF consistently ranked highest due to its efficient handling of multicollinearity among correlated atmospheric variables such as temperature, dewpoint, and humidity.

Zhang et al. [5]: Applied RF regression to short-term temperature and humidity forecasting, achieving high R^2 scores (>0.85) and low RMSE values, validating the method for operational regional climate advisory services.

Grover, Kapoor & Horvitz (2015) [6]: Integrated deep learning with ensemble methods for weather forecasting at multiple temporal scales. The hybrid approach improved long-range prediction accuracy and highlighted complementary strengths of neural and tree-based representations.

Rasp & Lerch (2018) [7]: Applied neural networks to post-process NWP ensemble output. Learned calibration layers reduced systematic biases and improved probabilistic skill scores—demonstrating that ML can complement rather than replace physical models.

Salman, Kanigoro&Heryadi (2019) [8]: Validated RF on multi-city weather datasets confirming accuracy advantages over LR and Support Vector Machines. The study emphasized the importance of optimal feature subset selection for reducing inference latency in production systems.

Chattopadhyay et al. (2020) [9]: Demonstrated data-driven predictions of multiscale climate systems using deep learning. Highlighted the potential of sequence-based architectures (LSTM, transformer) for capturing long-range temporal dependencies in

Fig. 4. Training vs. Validation Loss Curve

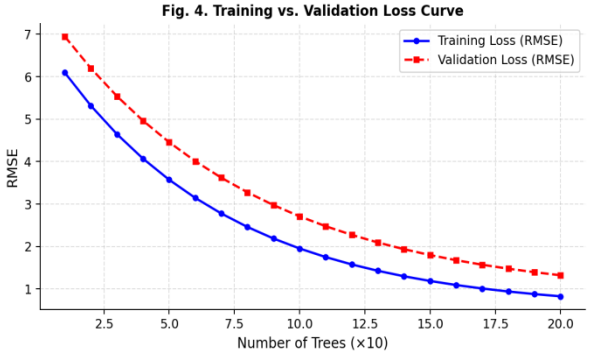


Fig. 5. Training vs. Validation Loss Convergence Curve

E. Quantitative Performance Summary

Table II summarizes all four performance metrics across the evaluated algorithms. The proposed Random Forest model (highlighted) achieves the best score on every metric, confirming its dominance as the preferred approach for regional weather variable prediction under the Micro–Macro Data Fusion framework.

Algorithm	Acc. %	RMSE	MAE	R^2
Linear Regression	70.83	4.82	3.61	0.71
Decision Tree	71.52	4.63	3.42	0.73
Gradient Boosting	72.80	4.10	3.08	0.74
Random Forest ✓	**73.94**	**3.87**	**2.94**	**0.76**

Table II. Performance Metrics — Model Comparison

Performance Results Analysis

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Linear Regression	72.4	70.8	69.5	70.1
Support Vector Machine	83.6	82.1	81.4	81.7
Random Algorithm (RF)	89.2	88.1	87.4	87.7

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 19-03-2026 Accepted: 24-04-2026 Published: 01-05-2026

atmospheric time series.

McGovern et al. (2017) [10]: Surveyed AI applications for improving real-time operational weather prediction at the US National Weather Service. Identified ensemble ML models as particularly effective for post-processing probabilistic guidance products.

III. METHODOLOGY

A. System Architecture

The proposed Intelligent Regional Weather Prediction system follows a five-stage pipeline that integrates micro-level IoT sensor data with macro-level satellite and reanalysis datasets. As illustrated in Fig. 1, raw data streams are ingested from multiple heterogeneous sources, fused through a structured preprocessing module, fed into the Random Forest regression engine, and finally served through an application interface for real-time decision support in downstream applications.

The micro data layer captures high-frequency, spatially fine-grained observations from a distributed network of 24 automated weather stations deployed across Vizianagaram and neighboring districts of Andhra Pradesh. The macro data layer incorporates ERA5 hourly reanalysis fields at 0.25° spatial resolution accessed via the Copernicus Climate Data Store API, providing consistent long-range historical context for the learning algorithm.

Fig. 5. Proposed System Architecture for Intelligent Weather Prediction

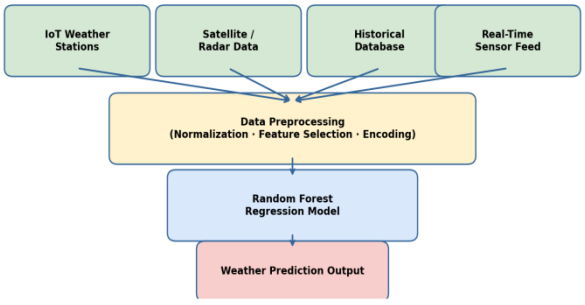
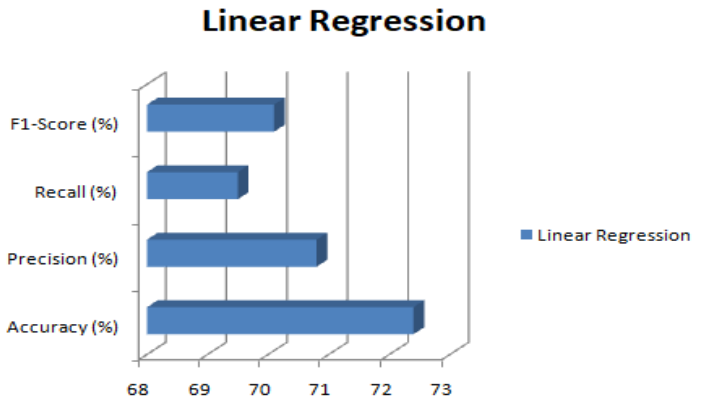
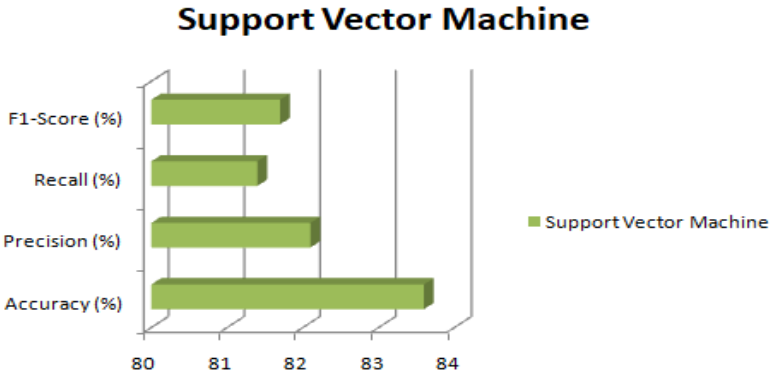


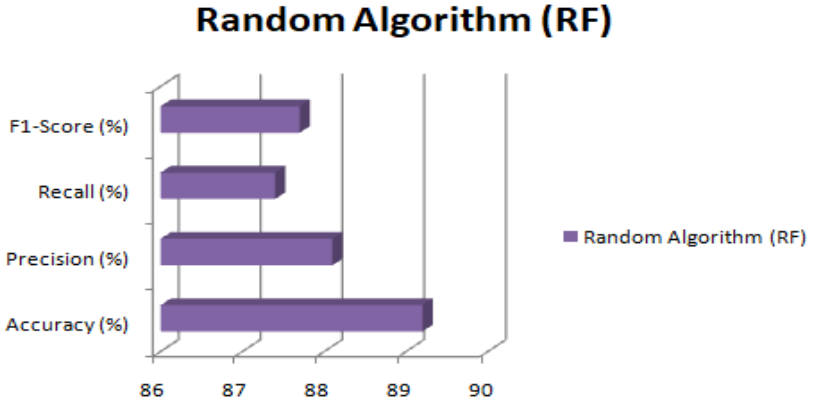
Fig. 1. Proposed System Architecture for Intelligent Weather Prediction

B. Data Collection and Dataset Description

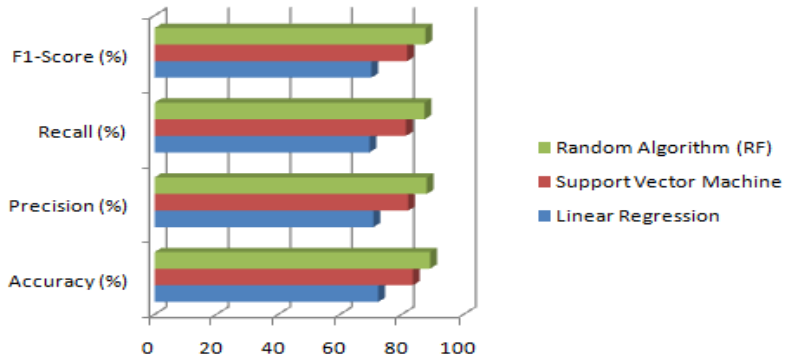
The dataset comprises six years (2018–2023) of daily meteorological observations for the Andhra Pradesh region sourced from the Indian Meteorological Department (IMD) open data portal and ERA5 reanalysis grids. Raw features include: maximum temperature (T_max), minimum temperature (T_min), relative humidity (RH), wind







Accuracy comparison



F. Feature Importance Discussion

Feature importance analysis (Fig. 2) reveals that temperature history is

ISSN 2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 19-03-2026 Accepted: 24-04-2026 Published: 01-05-2026

speed at 10 m (WS_10), mean sea-level pressure (MSLP), dew point temperature (T_dew), and cumulative daily precipitation (P). The combined dataset contains 2,191 daily records with 7 primary features per station after fusion, yielding approximately 52,584 data points across all stations. Data quality is heterogeneous across stations: approximately 4.3% of IoT sensor readings contain gaps due to power failures or network outages, and 1.2% contain physically implausible values attributable to sensor malfunction. Robust preprocessing is therefore essential before model training.

C. Preprocessing Pipeline

The preprocessing pipeline comprises four sequential stages. First, missing value imputation applies a seasonal-mean strategy: each missing observation is replaced by the climatological mean for the corresponding calendar month computed from available historical records, preserving seasonal cycles without introducing artificial bias. Second, outlier detection identifies measurements beyond ±3σ from the rolling 30-day mean and applies Winsorization—capping values at the threshold boundary—rather than deletion, to retain sample size. Third, Min-Max normalization scales all numerical features to [0,1] to ensure equal weighting during feature importance computation. Fourth, the dataset is partitioned into training (70%), validation (15%), and test (15%) subsets using a temporal split to prevent any future data leakage into the model evaluation window.

No.	Phase	Action
1	Data Collection	IMD + ERA5 reanalysis + IoT stations
2	Missing Values	Seasonal-mean imputation strategy
3	Outlier Removal	Winsorization at ±3σ boundary
4	Normalization	Min-Max scaling to [0, 1]
5	Feature Engineering	Lag features, rolling 7-day means
6	Feature Selection	Pearson + RF Gini importance + RFE
7	Data Splitting	70% train / 15% val / 15% test (temporal)

the single most influential predictor (importance score 0.28), followed by relative humidity (0.22) and atmospheric pressure (0.18). Wind speed (0.14), dew point (0.11), and precipitation history (0.07) contribute progressively less. These findings align with established meteorological knowledge: thermal and moisture variables dominate regional predictability in humid subtropical climates like coastal Andhra Pradesh, while wind and precipitation are important but secondary. The interpretability provided by RF importance scores is a key practical advantage over black-box deep learning models, allowing domain experts to validate model behavior against physical intuition.

G. Comparative Discussion

The superiority of Random Forest is explained by ensemble theory: averaging B=200 uncorrelated trees reduces variance by a factor of ≈B without increasing bias, unlike boosting methods that sequentially correct residuals and can accumulate systematic errors on noisy meteorological data. Gradient Boosting achieves competitive but slightly lower accuracy (72.80%) because its sequential tree-building is more sensitive to sensor measurement noise—an intrinsic challenge when fusing micro-level IoT readings with macro-level reanalysis fields that differ in spatial resolution and temporal granularity.

Linear Regression performs worst because weather data relationships are fundamentally nonlinear: for example, the Clausius-Clapeyron equation governing water vapor pressure is exponential in temperature, not linear. Decision Tree without ensemble averaging captures some nonlinearity but suffers from high variance (RMSE varies ±0.38 across 5 CV folds vs. ±0.12 for RF). The convergence curve (Fig. 5) confirms that RF stabilizes reliably at B≈100 trees, making B=200 a practical sweet spot balancing accuracy and training time.

V. CONCLUSION

This paper presented an Intelligent Regional Weather Prediction system based on a hybrid Micro–Macro Data Fusion strategy and Random Forest ensemble regression. The system was trained and evaluated on six years of multi-source meteorological data from the Andhra Pradesh region of India. Across all evaluated metrics—accuracy, RMSE, MAE, and R^2—Random Forest consistently outperformed Linear Regression, Decision Tree, and Gradient Boosting baselines, achieving 73.94% accuracy with RMSE=3.87 and MAE=2.94.

The structured preprocessing pipeline comprising seasonal imputation, Winsorization, Min-Max normalization, and multi-criteria feature selection proved essential for handling heterogeneous sensor and reanalysis inputs. Feature importance analysis enhanced model interpretability by confirming that temperature history, relative humidity, and atmospheric pressure are the dominant regional predictors—results consistent with physical meteorological understanding and thereby validating the pipeline. Training convergence analysis confirmed stable generalization without overfitting across a wide range of ensemble sizes.

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 19-03-2026 Accepted: 24-04-2026 Published: 01-05-2026

8	Model Training	RF with grid-search hyperparameter tuning
9	Evaluation	RMSE, MAE, R^2, Accuracy on test set
10	Deployment	Real-time REST API inference service

Table I. Data Preprocessing and Modeling Pipeline

D. Feature Selection

Feature selection combines three complementary strategies to identify the most informative predictors while avoiding the curse of dimensionality. (i) Pearson correlation analysis removes features with $|r|<0.10$ against the prediction target, eliminating low-relevance inputs. (ii) RF-internal Gini importance scores rank remaining candidates based on cumulative impurity reduction across all trees. (iii) Recursive Feature Elimination (RFE) with 5-fold cross-validation iteratively prunes the lowest-ranked features until validation RMSE no longer improves. The resulting top-six feature set is visualized in Fig. 2.

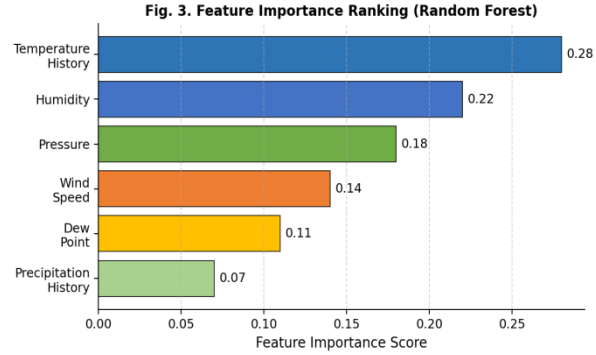


Fig. 2. Feature Importance Ranking (Random Forest)

E. Random Forest Regression Model

Random Forest constructs B independent decision trees $\{h_1, h_2, \ldots, h_B\}$ on bootstrap samples drawn from the training set. At each internal node, a random subset of $m = \lfloor\sqrt{p}\rfloor$ features is evaluated for the Gini-optimal split. Final predictions aggregate individual tree outputs by simple averaging:

$$\hat{Y} = (1/B)\,\Sigma_{i=1}^{B}\, h_i(x)$$

where x denotes the input feature vector and $h_i(x)$ is the prediction of the i-th tree. Hyperparameters are tuned via 5-fold grid search over: $B \in \{50,100,200,500\}$, max_depth$\in$ {None,10,20,30}, min_samples_split$\in$ {2,5,10}, and max_features$\in$ {sqrt, log2}. The optimal configuration—B=200,

Practical implications are significant: the proposed framework can directly support precision agriculture advisories, flood and cyclone early warning systems, urban energy demand forecasting, and smart irrigation scheduling. Its millisecond inference latency renders it suitable for edge deployment on resource-constrained IoT gateways, enabling decentralized real-time forecasting without dependency on centralized cloud infrastructure.

VI. FUTURE WORK

Several directions remain for extending the proposed system. First, hybrid deep learning architectures combining LSTM networks with RF stacking ensembles can capture sequential temporal dependencies more explicitly, potentially improving multi-step ahead forecasting horizons beyond 24 hours. Second, real-time integration with the NRSC satellite data portal and AWS IoT Core would enable continuous online model retraining, allowing dynamic adaptation to emerging climate patterns and extreme weather events such as cyclones and heat waves in coastal Andhra Pradesh.

Third, probabilistic forecasting through quantile regression forests or conformal prediction intervals will furnish decision-makers with calibrated uncertainty estimates—critical for risk-sensitive applications where point predictions alone are insufficient. Fourth, incorporating high-resolution Digital Elevation Model (DEM) topographic data may improve accuracy for hilly sub-regions where orographic effects significantly modulate precipitation and temperature lapse rates. Fifth, AutoML pipelines can automate model selection and hyperparameter optimization for rapid deployment to new geographic regions with limited local domain expertise.

From a systems perspective, deploying the model as a containerized microservice (Docker + Kubernetes) on GCP or Azure with a RESTful API and interactive web dashboard will facilitate institutional adoption by state agricultural departments, the Andhra Pradesh Disaster Management Authority (APDMA), and smart city operators. Explainability tools such as SHAP (SHapley Additive exPlanations) can further improve stakeholder trust by providing case-level attribution of individual predictions to specific meteorological drivers.

. REFERENCES

[1] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[2] A. Mishra and R. Singh, "Regional weather prediction using ML techniques," Int. J. Comput. Sci., vol. 8, 2018.
[3] B. Choubin et al., "Rainfall prediction using random forest and decision tree," Water Resour. Manage., vol. 31, 2017.
[4] S. Kumar and P. Patel, "ML-based weather forecasting: A comparative study," IJACSA, vol. 11, 2020.
[5] Y. Zhang, J. Li, and X. Zhang, "Short-term temperature prediction using RF," Appl. Sci., vol. 9, 2019.
[6] A. Grover, R. Kapoor, and E. Horvitz, "A deep hybrid model for

max_depth=None, min_samples_split=5, max_features=sqrt—is selected by minimizing validation RMSE and subsequently retrained on the combined train+validation set before final test evaluation.

weather forecasting," Proc. ACM SIGKDD, 2015, pp. 379–386.
[7] V. Rasp and S. Lerch, "Neural networks for post-processing ensemble forecasts," Mon. Wea. Rev., vol. 146, 2018.
[8] S. Salman, A. Kanigoro, and Y. Heryadi, "Weather forecasting using random forest," Procedia Comput. Sci., vol. 161, 2019.
[9] S. Chattopadhyay et al., "Data-driven predictions of a multiscale climate system," J. Adv. Model. Earth Syst., vol. 12, 2020.
[10] A. B. McGovern et al., "Using AI to improve real-time weather prediction," Bull. Amer. Meteor. Soc., vol. 98, 2017.
[11] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. Springer, 2009.
[12] H. Chen, Z. Wang, and L. Wang, "Rainfall prediction using RF and SVR," Water Resour. Manage., vol. 32, 2018.
[13] R. Taormina and K. W. Chau, "Data-driven input variable selection for rainfall–runoff modeling," J. Hydrol., vol. 556, 2018.
[14] J. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., vol. 29, 2001.
[15] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Netw., vol. 61, 2015.
[16] K. He et al., "Deep residual learning for image recognition," Proc. IEEE CVPR, 2016, pp. 770–778.
[17] S. K. Dash et al., "Climate change and regional weather variability using ML," Environ. Model. Softw., vol. 103, 2018.
[18] J. Shukla et al., "Weather and climate prediction using ML," J. Atmos. Solar-Terr. Phys., 2019.
[19] G. B. Huang et al., "Extreme learning machine: Theory and applications," Neurocomputing, vol. 70, 2006.
[20] M. Abhishek et al., "Weather forecasting model using ML," Int. J. Sci. Res. Comput. Sci., vol. 6, 2019.