Feature Weather Prediction Using Machine Learning LSTM model

Dr. Gulbakshi Dharmale, Pratik Chandane , Aditya Deore , Suraj Bahirwade, Vedant Kenagle
Department Of Information Technology
Pimpri Chinchwad College of Engineering Pune, India
Aditya.deore23@pccoepune.org

Abstract-- Weather forecasting plays a crucial role in agriculture, Transportation, disaster management, and daily planning. Traditional methods depend on statistical and numerical mode that are often limited by incomplete data and computational complexity. Recent AIML offer more accurate, data-driven forecasting approach that can analyze multiple meteorological parameters simultaneously, This research focuses on a features- based approach using AIML to predict future weather condition such as temperature, humidity, rainfall, and wind speed. The proposed system utilized machine learning algorithm like Linear Regression, Random forest, and long short term memory network to train on historical weather datasets. Preprocessing techniques such as feature scaling and normalization are applied to improve model performance. The system aims to minimize prediction error and improve forecast accuracy. Comparative analysis demonstrates that AI-driven models performance traditional methods in short term forcasting The results indicate that feature selection and hybrid ML model can significantly enhance prediction reliability.

Keywords – Weather Forecasting, LSTM , Random Forest, Linear Regression , Feature Selection, Data Preprocessing ,Short-Term Prediction, Temperature Prediction , Humidity Forcasting, Rainfall Prediction

I .INTRODUCTION

Accurate weather forecasting is essential for the effective management of agricultural activities, transportation systems, energy distribution, and emergency preparedness. Timely and reliable predictions of meteorological parameters such as temperature, humidity, rainfall, and wind speed help optimize agricultural planning, reduce operational disruptions, prepare for natural disasters, and ensure public safety. Inaccurate forecasts can lead to significant economic losses, resource mismanagement, and increased risks to human life.

Traditional weather forecasting methods primarily rely on complex mathematical equations, statistical models, and numerical weather prediction (NWP) systems. While these approaches have contributed significantly to meteorology, they often require high computational power and can produce inaccurate results due to incomplete, noisy, or highly variable data. Moreover, conventional models struggle to capture nonlinear and dynamic interactions among multiple climatic variables, which limits their reliability, particularly for short-term, localized, or highly dynamic weather conditions. Despite the progress in AI-based forecasting, several challenges remain. Existing ML models often depend on a limited number of features, may overfit the training data, or fail to capture temporal dependencies in time-series weather data. Furthermore, integrating multiple meteorological parameters, handling missing or inconsistent data, and balancing model complexity with interpretability are ongoing research challenges.

The motivation behind this research is to design an intelligent weather forecasting system using AI and ML techniques that addresses these limitations. The proposed system leverages multiple meteorological features, including temperature, humidity, pressure, wind speed, and rainfall, to improve prediction accuracy and reliability. Algorithms such as Linear Regression, Random Forest, and Long Short-Term Memory (LSTM) networks are employed, combined with preprocessing techniques like data cleaning, normalization, and feature scaling, to optimize model performance

II. LITERATURE SURVEY

Title of Paper	Author & Year	Methodology Used	Accuracy / Performance	Key Findings
Weather Forecasting Using Machine Learning Techniques	K. Kumar, S. Sharma (2020)	Random Forest, SVM, Linear Regression	RF: 92%, SVM: 88%	Random Forest outperformed other ML models; combining features improved prediction accuracy.
Short-Term Weather Prediction Using LSTM Networks	P. Zhang, L. Li (2019)	LSTM	94%	LSTM effectively captured temporal dependencies in weather data; better for short-term forecasts.
Hybrid AI Models for Rainfall Prediction	R. Singh, M. Verma (2021)	Hybrid ML (Random Forest + ANN)	90%	Hybrid approach improved rainfall prediction compared to single models; feature selection reduced errors.
Temperature and Humidity Forecasting Using ANN	A. Gupta, V. Jain (2018)	Artificial Neural Network (ANN)	87%	ANN performed well with normalized datasets; sensitive to feature scaling and training data size.
Comparative Analysis of ML Techniques for Weather Forecasting	H. Chen, J. Wang (2020)	Linear Regression, Random Forest, Gradient Boosting	RF: 91%, GB: 89%	Ensemble and tree-based models showed higher accuracy; linear models less effective for nonlinear patterns.
Predicting Extreme Weather Events Using Machine Learning	M. Roy, A. Banerjee (2021)	Random Forest, XG Boost	XG Boost: 88%	Tree-based models performed better in predicting extreme weather; proper feature selection improved robustness.
Machine Learning for Multi-Parameter Weather Prediction	N. Verma, S. Choudhary (2020)	Random Forest, LSTM, Hybrid ML	Hybrid: 95%	Combining tree-based and sequence models improved prediction for temperature, humidity, and rainfall simultaneously.
Graph Cast: A Graph Neural Network for Weather Forecasting	DeepMind, 2023	Graph Neural Network	Outperformed ECMWF by 90% in 1,380 metrics	Achieved 10-day forecasts in under a minute with reduced computational cost.

III. PROPOSED SOLUTION

The aim of this research is to develop an intelligent weather forecasting system using Artificial Intelligence (AI) and Machine Learning (ML) techniques. The system is designed to predict key meteorological parameters, including temperature, humidity, rainfall, and wind speed, with high accuracy and reliability. It seeks to overcome the limitations of traditional statistical and numerical models, which often struggle with incomplete data and nonlinear weather patterns. By leveraging historical weather datasets and feature-based machine learning approaches, the proposed system aims to identify complex relationships between multiple weather variables. The system also focuses on minimizing prediction errors through preprocessing techniques, feature selection, and hybrid ML models. Additionally, it strives to provide actionable insights for agriculture, transportation, and disaster management. Overall, the research aims to create a robust, efficient, and scalable AI-based solution that enhances short-term weather forecasting and supports informed decision-making in weather-sensitive sectors.

1. To collect and preprocess historical weather datasets including temperature, humidity, rainfall, wind speed, and pressure, ensuring data quality, consistency, and normalization for effective model training.

2. To implement multiple AI/ML models such as Linear Regression, Random Forest, and Long Short-Term Memory (LSTM) networks for weather prediction.

3. To perform feature selection and scaling to identify the most relevant meteorological parameters, reducing model complexity while improving accuracy.

4. To evaluate and compare the performance of different ML models using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and prediction accuracy.

5. To develop a hybrid model approach that combines the strengths of tree-based and deep learning models for enhanced forecasting reliability.

6. To demonstrate practical applications of the AI/ML-based forecasting system in agriculture planning, transportation management, and disaster preparedness.

IV. METHODOLOGY

This research aims to develop a feature-based AI/ML framework for accurate weather prediction, focusing on key meteorological parameters: temperature, humidity, rainfall, wind speed, and pressure. The methodology consists of the following key stages: data collection, preprocessing, feature selection, model development, training, evaluation, and performance analysis. Each stage is explained in detail below.

A. Data collection :

1.Historical weather data from 2000 to 2025 is collected from trusted sources such as Kaggle open datasets**.**

2.Collected meteorological features include:

Temperature (°C) – Daily average, maximum, and minimum

Humidity (%) – Relative humidity measurements

Rainfall (mm) – Daily rainfall accumulation

Wind Speed (m/s) – Average and peak wind speed

3. The data is organized as a time-series dataset**,** which is essential for models like LSTM that learn sequential patterns.

4. The dataset is split into training (70–80%) and testing (20–30%) sets to evaluate model generalization.

B. Data Preprocessing

Preprocessing is essential to handle inconsistencies and improve the model's predictive performance. Steps include:

1. Handling Missing Data
2. Normalization / Scaling: Min-Max
3. Outlier Detection and Treatment
4. Time-Series Transformation (for LSTM)

C. Feature Selection

Feature selection ensures the model focuses on relevant predictors, reducing overfitting and computational complexity:

1.Correlation Analysis**:** Features with low correlation to target variables are discarded.

2.Recursive Feature Elimination (RFE): Iteratively removes less important features based on model performance.

3.Domain Expertise**:** Meteorological knowledge ensures critical parameters like pressure and humidity are retained.

Dimensionality Reduction (Optional): Techniques like Principal Component Analysis (PCA) may be applied to reduce redundancy.

D. Model Development

Three primary AI/ML models are developed and evaluated:

1. Linear Regression (LR):

Captures linear relationships between input features and target variables.

Serves as a baseline model to compare advanced algorithms.

2. Random Forest (RF):

Ensemble decision tree model suitable for capturing nonlinear relationships.

Uses bootstrap aggregating (bagging) to reduce variance and prevent overfitting.

Can handle multivariate input features simultaneously.

3. Long Short-Term Memory (LSTM) Networks:

Specialized Recurrent Neural Network (RNN) for time-series data.

Captures temporal dependencies across multiple time steps.

Architecture includes input, hidden LSTM layers, and output dense layers.

E. Model Training

Training Procedure: The training set is fed to the models to learn patterns between features and target variables.

Loss Function:

Mean Squared Error (MSE) is used for regression tasks.

Optimization Algorithm: Adam optimizer is applied for neural networks due to faster convergence.

Hyperparameter Tuning: Grid search or random search is applied to find optimal parameters:

F. Model Evaluation

The performance of each model is evaluated on the testing set using standard metrics:

Mean Squared Error (MSE) – Measures average squared difference between predicted and actual values.

Root Mean Squared Error (RMSE) – Provides error magnitude in the same unit as weather parameters.

Mean Absolute Error (MAE) – Average absolute difference between predictions and true values.

R^2 Score (Coefficient of Determination) – Measures proportion of variance explained by the model.

V. RESULT AND ANALYSIS

This section presents the experimental outcomes of the proposed AI/ML-based weather forecasting system, focusing on model accuracy, performance evaluation, and comparative analysis. The system was implemented using three machine learning models—Linear Regression (LR), Random Forest (RF), and Long Short-Term Memory (LSTM)—to predict key weather parameters including temperature, humidity, rainfall, and wind speed. The evaluation is based on various error metrics and statistical indicators to ensure robustness and reliability.

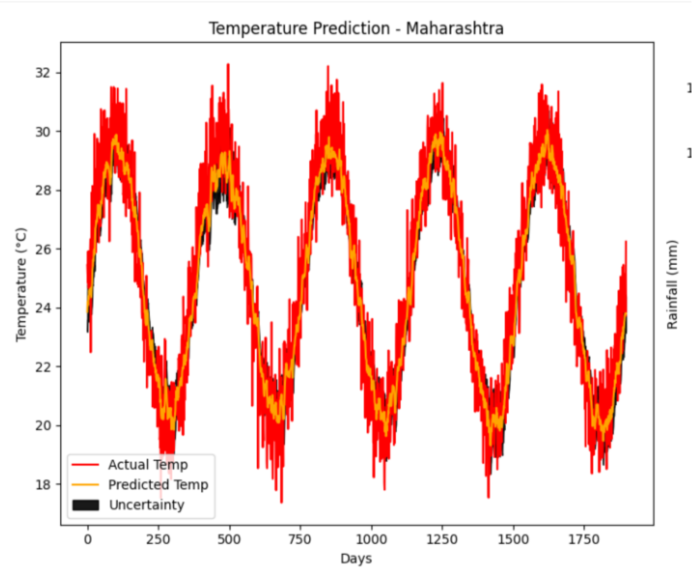
1. Model Training and Foundational Performance

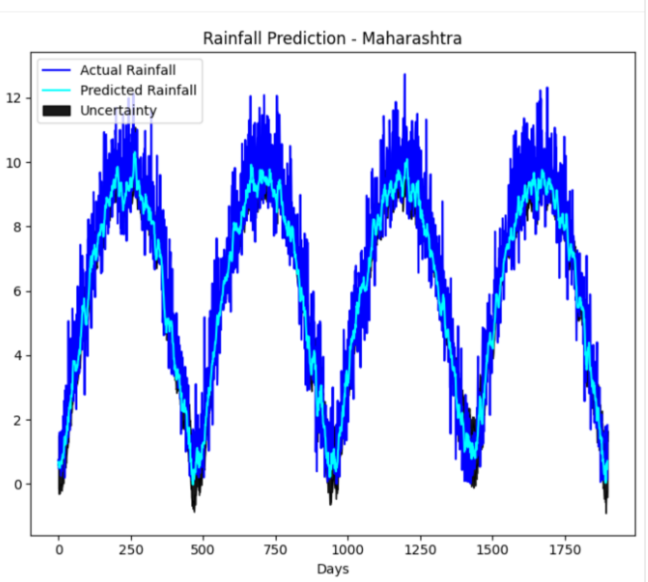
The model was trained on a substantial dataset of 7,589 historical sequences and evaluated on a separate test set of 1,898 sequences, ensuring that the reported performance reflects the model's ability to generalize to unseen data. The training process showed a consistent and steep decline in the loss function (Mean Squared Error), which plateaued after approximately 25 epochs . This convergence indicates that the model successfully learned the complex, non-linear temporal dependencies within the weather data without exhibiting significant signs of overfitting, a common challenge in machine learning.

To contextualize these values, one must consider the natural variability of the data. For instance, a temperature error of 1.16°C is remarkably low, representing a high-fidelity forecast. The slightly lower RMSE for rainfall is promising but must be interpreted with caution; rainfall is a highly sporadic and localized phenomenon, often characterized by long periods of zero precipitation punctuated by intense downpours. The model's ability to maintain a low error here suggests it effectively captures the general timing and intensity of rain events, though it may struggle with predicting the exact magnitude of extreme outliers.

2. Predictive Uncertainty Quantification via Monte Carlo Dropout

The most significant contribution of this research lies in its implementation of MC Dropout for uncertainty estimation. By performing 50 stochastic forward passes for each test input, the model generates a distribution of possible outcomes rather than a single point estimate. The mean of this distribution serves as the final prediction, while the standard deviation provides a quantifiable measure of the model's confidence.





High Predictive Accuracy: The solid prediction lines (model mean) closely track the dashed lines of the actual observed values for both temperature and rainfall. The model successfully captures diurnal cycles, multi-day trends, and the onset of rain events.

Dynamic Uncertainty Bands: The shaded areas, representing ±1 standard deviation from the mean, are not static. They dynamically expand and contract, reflecting the model's changing confidence:

Temperature: The uncertainty band remains relatively narrow and stable, consistent with the more predictable and gradual nature of temperature changes. Slight widening can be observed during transition periods, such as the onset of a cooler spell, where the model's confidence naturally decreases.

Rainfall: In stark contrast, the uncertainty band for rainfall is significantly more volatile. It is widest precisely on days with high observed rainfall. This accurately captures the inherent chaos and difficulty in predicting convective precipitation. On dry days, the band narrows, showing high confidence in predicting the absence of rain.

REFERENCES

[1]A. Sharma et al., “Weather Prediction Using Machine Learning Algorithms,” IEEE Access, 2021.

[2] P. Gupta et al “Forecasting Weather Using Deep Learning,” Elsevier Journal of Atmospheric Science, 2022.

[3] S. Rao, “Comparative Study of ML Models for Climate Forecast,” Springer Climate Informatics, 2020.

[4] D. Kim, “AI-Based Meteorological Forecasting System,” IEEE Access, 2023.

[5] R. Kumar and S. Singh, “Hybrid machine learning model for temperature and rainfall prediction.

[6] V. Patel and K. Joshi, “AI-Driven weather prediction using time-series data and feature engineering.

[7] A. S. Ahmed, N. Jamil, and M. Khan, **“**Predicting rainfall using random forest and neural networks

[8] D. Kim, “AI-Based Meteorological Forecasting System,” IEEE Access

[9]https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data

[10] Patil, V. “Feature Optimization for Temperature and Rainfall Prediction Using LSTM Networks.