International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 06 June-2026 | Impact Factor: 3.5

Weather Forecasting using Hybrid Machine Learning Framework

Shrawani Parkale[1]
Department of Information Technology,
Vidya Pratishthan's Kamalnayan Bajaj Institute of Information and Technology, Baramati, India
shrawaniparkale@gmail.com

Maithili Huske[2]
Department of Information Technology,
Vidya Pratishthan's Kamalnayan Bajaj Institute of Information and Technology, Baramati, India
maithilihuske@gmail.com

Deepika Dombe[3]
Department of Information Technology,
Vidya Pratishthan's Kamalnayan Bajaj Institute of Information and Technology, Baramati, India
dombedeepika@gmail.com

Pradip Paithane[4]
Department of Information Technology
Vidya Pratishthan's Kamalnayan Bajaj Institute of Information and Technology, Baramati, India
paithanepradip@gmail.com

Abstract—Machine learning (ML) has played a major role in enhancing weather forecasting systems in terms of predictive performance and reliability. The traditional numerical weather prediction techniques are limited to incomplete information, nonlinear atmospheric interactions and the challenges of modeling complicated meteorological patterns. These constraints have led to studies of hybrid ML systems that integrate preprocessing, feature engineering and ensemble learning algorithms [1], [4]. This review will discuss weather prediction hybrid models such as Linear Regression, K-Nearest Neighbors (KNN), XGBoost, CatBoost, AdaBoost, and LightGBM. Wavelet packet denoising and ensemble are used to increase the accuracy and robustness of prediction [1]. The existing body of literature shows that there has been significant advancement in ML-based forecasting [18], but there are issues related to scalability, interpretability, and real-time implementation. Hybrid ML models always achieve better accuracy, stability, and flexibility compared to single ones. Recent advancements in machine learning and the increasing availability of meteorological data have enabled the development of intelligent forecasting systems that support accurate, timely, and reliable decision-making across weather-sensitive domains. This work has critically examined the literature available, compared the methods, found deficiencies and future directions on how to come up with scalable and intelligent weather forecasting systems.

Index Terms—Weather Forecasting, Machine Learning, En-semble Learning, CatBoost, Gradient Boosting, Hybrid Models, Weather Prediction.

I. INTRODUCTION

The weather forecast has assumed a very significant role in the modern society as it directly affects the agricultural industry, disaster preparedness and aviation as well as transport and climate management. Proper forecasting of the weather patterns allows making a better plan, minimizing losses, and enhancing the security of the population in case of extreme events in the nature. Nevertheless, predicting weather is not an easy process due to the high nonlinear, dynamic, and interconnection of multiple and interdependent parameters in the atmosphere. The classical models of numerical weather prediction methods are based on mathematical solutions and physical models that calculate the atmospheric conditions. These models have been extensively used but have been found to have problems in form of data noise, poor spatial resolution as well as inability to accommodate quickly changing weather patterns. Moreover, wrong initial conditions may give erroneous predictions and unreliable results. Machine learning procedures have become influential as far as creating weather forecasts are concerned, due to the accessibility of vast meteorological data, and the progress in computational intelligence. Machine learning models can also detect latent patterns and relationships on past data which makes them better fitted to address complex environmental systems [4], [19]. However, single machine learning models can be associated with overfitting, sensitivity to noise, and low capacity to generalize. In order to eliminate these shortcomings, hybrid machine learning systems have been proposed, which involves the use of preprocessing methods, feature selection and ensemble methods to increase the accuracy of forecasting. It has been demonstrated that a combination of the use of wavelet-based denoising alongside boosting and ensemble learning methods greatly enhances predictive performance by minimizing noise, as well as, improving feature representation [1], [4], [18]. Recent studies put the focus on the role of ensemble-based methods and gradient boosting algorithms that include XGBoost, CatBoost, and LightGBM in the creation of a robust forecasting system [12], [22], [8]. These algorithms enhance the reliability of the predictions though the combination of several weak learners and the optimization of the performance through an iterative method. The real-time forecasting and adaptive learning can also be achieved using hybrid models that is why they can be used in the present-day monitoring of the environment. This work aims to review the literature on the use of hybrid machine learning frameworks in weather prediction, to study methodology of recent literature, and to define research gaps that can be used to enhance future research [3], [23].Traditional forecasting approaches often face challenges in capturing complex atmospheric interactions, motivating researchers to explore data-driven methodologies capable of learning hidden patterns from large-scale meteorological datasets. As a result, machine learning has emerged as a promising paradigm for enhancing forecasting performance and supporting intelligent weather prediction applications.Recent advancements in computational resources, data availability, and artificial intelligence have accelerated the adoption of machine learning techniques across various environmental applications. Numerous studies have investigated the effectiveness of different predictive models, preprocessing strategies, and ensemble approaches for weather forecasting. However, identifying suitable methodologies that balance forecasting accuracy, interpretability, and practical applicability remains an important research challenge. This review aims to examine these developments and provide a comprehensive overview of current trends in machine learning-based weather forecasting.

II. Literature Review

Weather forecasting is an area where machine learning methods are being applied and increasingly so because of their capacity to model nonlinear and complex relationships in the weather data. The initial research works were mostly based on regression models, decision trees, and neural networks to perform prediction tasks. The recent studies have moved towards the ensemble and boosting methods, which enhance the prediction performance based on the combination of multiple models. Combination of preprocessing methods, feature engineering, and machine learning algorithms have become significant as well. Specifically, there has been a long-term adoption of wavelet-based denoising techniques to improve the quality of data and enhance the level of forecasting reliability.

There is a lot of literature on the application of machine learning in weather forecasting to enhance the performance of predictions and address nonlinear weather patterns. To improve the quality of data and the performance of the model, Niu et al. suggested a hybrid forecasting method combining wavelet packet denoising and CatBoost algorithm. Their experiment showed that a preprocessing of the meteorological signals prior to training the model has great effects on the reliability of the prediction and the errors generated by noise [1].

Kair et al. discussed the application of machine learning methods in smart weather forecasting systems and the necessity to create scalable and data-driven forecasting models. Their work emphasized in the way that machine learning models could be useful in decision-making through learning relationships among the atmospheric variables and creating more adaptive forecasting systems [2].

Saravana et al. concentrated on the next-generation artificial intelligence-based climate and weather prediction models and mentioned the efficiency of the combination of various machine learning methods. Significance of combination of hybrid and ensemble learning approaches was emphasized in the study to enhance robustness in forecasting and effective handling of massive meteorological data [3].

Mienye and Sun have provided an in-depth overview of algorithms of ensemble learning used in weather prediction and have shown that by using a combination of different models, predictive ability and model bias are improved. They found that ensemble methods are more effective especially when one is working with large and complicated environmental data [4]. Comparative study of various machine learning-based rainfall prediction models in urban areas done by Kumar et al. found that the ensemble-based models are more effective than single algorithms in consistency of predictions and model forecast. It was also found in the research that feature selection and preprocessing of data are critical in obtaining reliable prediction results [5].

A comparative review of existing studies shows that ensemble and hybrid learning approaches generally perform better than standalone models in weather prediction tasks. Most research demonstrates improved stability and robustness when multiple algorithms are combined. However, there are still typical problems, along with data quality problems, the overfitting of the model and high computational requirements. These limitations highlight the need for more effective and scalable forecasting systems.

A. Dataset Characteristics

The effectiveness of machine learning–based weather forecasting largely depends on the quality, diversity, and temporal depth of meteorological datasets. Weather prediction uses multivariate time-series data from meteorological stations, satellites, and environmental sensors, including parameters such as temperature, humidity, pressure, rainfall, wind speed, solar radiation, and dew point [5], [17].

Studies show that large and high-resolution datasets improve forecasting accuracy by helping models capture nonlinear

TABLE I: Literature Survey Summary

Sr. No.	Authors and Year	Methodology	Dataset	Evaluation Metrics Used	Advantages	Limitations
1	Niu et al. (2021)	CatBoost + Wavelet Packet Denoising	Meteorological dataset	MAE, RMSE, MSE, R^2	Higher accuracy than LSTM/Seq2Seq; Faster	Manual tuning needed; Complex preprocessing
2	Kair et al. (2025)	CatBoost, XGBoost, Linear/Bayesian Regression	Atmospheric data	MSE, MAE, Interpretable (SHAP)	Simple preprocessing and scaling; Feature importance via SHAP	Short-term; No hybrid/deep learning; Small dataset; High CPU/GPU demand
3	Saravana et al. (2025)	Hybrid AI Models (LSTM, CNN, XGBoost, etc.)	Climate datasets	MAE, RMSE, R^2 (discussed)	Covers hybrid models for accuracy; Explains wavelet/EMD; Identifies future trends (XAI, AutoML)	High computation cost; Overfitting and low interpretability
4	Mienye & Sun (2022)	Explains ensemble concepts (Bagging, Boosting, Stacking)	Various benchmark datasets	Accuracy, RMSE, MAE	Improves accuracy & reduces bias; Detailed math & applications	High training time; Overfitting risk; Hard to interpret results; Lacks dataset proof
5	Kumar et al. (2023)	CatBoost, XGBoost, LGBM, Linear Models	Urban rainfall dataset	MAE, MSE, RMSE, R^2	Useful for flood management; Compared daily vs weekly forecasts	Regional dependency; Some prediction errors
6	Jakaria et al. (2020)	Random Forest Regression (RFR)	Tennessee local weather data	RMSE	Fast, simple, accurate, low cost	Small dataset; Predicts only temperature
7	Singh et al. (2021)	Recurrent Neural Network (RNN)	Indian meteorological dataset	RMSE	Best accuracy; Captures time-dependence	Larger time window increases error
8	Shaiba et al. (2022)	Ensemble Learning (RF + GBDT + KNN + Naive Bayes)	Large-scale weather big data	MAE, MSE, Accuracy	High accuracy (95%), faster prediction	Complex, heavy preprocessing
9	Shimaila & Siddiqi (2023)	Multiple Linear Regression (Hybrid Regression Models)	Meteorological datasets	MSE, MAE, MSLE, R^2	Real-time, continuous prediction; Low MSE	Small dataset; Limited validation
10	Raghuwanshi et al. (2024)	Random Forest + Optimization (PSO, GA)	Regional weather datasets	Accuracy, Precision, Recall, RMSE	High accuracy; Optimized performance	Computationally heavy; Tuning required

relationships, seasonal trends, and long-term climate patterns [2], [18]. However, these datasets often contain missing values, noise, and inconsistent recording intervals, which affect model performance. Therefore, preprocessing steps like normalization, denoising, interpolation, and feature scaling are necessary before training [1], [15].Hybrid and ensemble forecasting approaches further enhance prediction performance by combining data from multiple sources, enabling models to learn complex atmospheric interactions and improve robustness [4], [22].

Studies indicate that large-scale meteorological datasets improve forecasting performance by enabling models to capture seasonal variability and non-linear atmospheric interactions [2], [18]. However, increasing dataset size also introduces challenges such as computational overhead, memory constraints, and data inconsistency, which may degrade model performance if preprocessing is inadequate.

The primary dataset used in this study is the Jena Climate Dataset (Weather Long-Term Time Series Forecasting), obtained from Kaggle. The dataset contains 52,696 meteorological observations recorded at 10-minute intervals by a weather monitoring station located in Jena, Germany. It comprises 21 atmospheric and environmental variables, including temperature, humidity, atmospheric pressure, dew point, wind speed, wind direction, rainfall, radiation-related measurements, and other weather-related indicators.

To facilitate forecasting analysis, four key meteorological variables—temperature, relative humidity, rainfall, and wind speed—were considered as target variables. The remaining weather attributes, together with temporal information extracted from timestamps, were utilized to characterize atmospheric conditions and capture temporal weather patterns. The dataset spans diverse environmental conditions and seasonal variations, making it suitable for investigating machine learning-based weather forecasting approaches and evaluating their ability to model complex meteorological relationships.

B. Analysis of Existing Work

Existing research in weather forecasting demonstrates a strong transition from traditional statistical techniques to machine learning and hybrid forecasting approaches. Early models primarily relied on regression and numerical weather prediction techniques, which were limited in capturing non-linear atmospheric behavior. With the emergence of machine learning, models such as decision trees, neural networks, and support vector machines improved prediction capabilities but still faced challenges related to data noise, overfitting, and limited scalability [5], [7]

Recent studies emphasize ensemble learning and boosting algorithms, which combine multiple learners to improve prediction accuracy and reduce variance. Techniques such as XGBoost, CatBoost, and LightGBM have shown superior performance in modeling complex meteorological relationships compared to standalone models [12], [22]. Furthermore, hybrid frameworks integrating preprocessing techniques such as wavelet denoising and feature engineering have demonstrated improved robustness and noise handling in time-series weather data [1], [15].

Despite these advancements, variability in datasets, regional dependency, and computational complexity remain major concerns in large-scale weather prediction systems [18].

C. *Research Gaps Identified*

Although significant progress has been made in machine learning-based weather forecasting, several research gaps still exist. Many studies rely on region-specific datasets, which limit the generalization of forecasting models to other climatic conditions. Data quality issues such as missing values, noise, and imbalance continue to affect prediction reliability.

Additionally, standalone machine learning models often fail to capture dynamic atmospheric interactions due to their limited adaptability. While ensemble and boosting techniques improve performance, they increase computational complexity and require extensive training time. Real-time forecasting and deployment in resource-constrained environments remain challenging.

Another critical gap is the lack of integration between domain knowledge from meteorology and data-driven machine learning approaches. Most existing studies focus purely on algorithmic performance without incorporating climate dynamics and environmental context [18], [22].Another important limitation noted in existing studies is the lack of real-time deployment validation. Most forecasting models are tested with offline datasets, which may not accurately reflect changing atmospheric conditions. Additionally, boosting and ensemble models often have lower interpretability, even though they are very accurate in predictions. Finding the right balance between accuracy, scalability, and explainability is still a challenge in research.

III. Methodology

Weather forecasting hybrid machine learning practices generally embrace an apparent multi-step procedure. This architecture comprises pre processing, feature engineering, integration of models and evaluation. The recent research explains the necessity to use multiple methods rather than use individual predictive models. This approach improves robustness and stability [1], [4], [18].

The literature analysis shows that previous researchers depended mostly upon individual models like Linear Regression, Random Forest, and Recurrent Neural Networks to predict the weather independently without any combination algorithms [6], [7]. Although these models had a reasonable predictive capability, they were also prone to failure in nonlinear interactions in the atmosphere and sensitivity to noise.

Recent trends in research have been shifting towards hybrid and ensemble-based approaches, in which one or more algorithms are used in a single framework. It has been found that studies that combine preprocessing methods with boosting algorithms or stacking of multiple learners have shown to be more robust, stable, and better predictors of the future results than those that do not combine these approaches to forecasting future outcomes [1], [4], [18]. This is a change

Fig. 1: Workflow diagram

that can be attributed to the realization that a combination of complementary models improves the overall predictability of the weather forecasting systems.

A. *Preprocessing and Denoising*

An important observation that is found throughout the literature is the importance of preprocessing to enhance model reliability. Weather statistics are usually noisy and inconsistent and they have negative impacts on learning algorithms. To overcome this problem, a number of studies use signal processing prior to model training.

Wavelet Packet Denoising (WPD) is a well known popular preprocessing technique in meteorological time-series data. To illustrate, Niu et al. [5] have shown that the combination of the wavelet-based denoising mechanism with the boosting algorithms leads to the enhancement of signal transparency, as well as to the increase of forecasting accuracy. These example preprocessing steps allow machine learning models to better represent relevant atmospheric patterns and they do not have as much sensitivity to noise [15].

B. *Feature Selection and Representation*

In hybrid forecasting structures, feature engineering is very important. Numerous investigations use correlation-based analysis and model-based importance ranking to identify significant atmospheric variables to be utilized. Definite selection of features minimizes redundancy, increases the computational efficiency, as well as the generalization ability of forecasting models [5], [12]. Hybrid methods focus on the structured feature representation to provide more realistic nonlinear interactions of the atmosphere.

C. Ensembling of Models

Another significant trend discovered in the literature is the replacement of the traditional model of regression by the ensemble and gradient boosting models. Linear models have the advantage of offering explainable baselines, but they are not always able to capture intricate atmospheric relationships.

Recent research emphasizes the success of increasing the algorithms like the XGBoost, LightGBM, AdaBoost, and Cat-Boost in weather prediction activities as they can improve prediction errors in an iterative manner and can make predictions of nonlinear relationships [12], [22]. The additional learning methods include ensemble learning including bagging, boosting, stacking, which increase the stability through combination of several predictive models [4], [8]. The hybrid structure comprising of preprocessing, feature engineering, and boosting algorithms is always robust and adaptable, compared to the individual model [1], [18].

D. Theoretical Foundations of Machine Learning Models

The proposed hybrid weather forecasting framework integrates advanced preprocessing techniques, ensemble machine learning algorithms, and explainable artificial intelligence to improve forecasting accuracy and reliability. Unlike traditional forecasting systems that rely on a single predictive model, the proposed architecture combines multiple machine learning algorithms through a hybrid stacking mechanism. The framework consists of five major stages: data acquisition, data preprocessing, feature engineering, ensemble model training, and explainable prediction generation. Historical meteorological observations and real-time weather data obtained from weather APIs are utilized as input sources. Weather parameters including temperature, humidity, atmospheric pressure, rainfall, wind speed, visibility, and dew point are collected and transformed into structured features suitable for machine learning algorithms.

To improve data quality, Wavelet Packet Denoising (WPD) is applied to remove unwanted noise.The processed data is then used to train multiple machine learning models including Linear Regression, K-Nearest Neighbors, Random Forest, AdaBoost, XGBoost, LightGBM, CatBoost, and Support Vector Regression. Predictions generated by these models are combined using a ensemble strategy to generate the final weather forecast. Finally, AI based explainability is integrated to interpret model decisions and identify the most influential weather parameters contributing to the forecast. This hybrid architecture aims to improve prediction performance, robustness, scalability, and interpretability simultaneously.

E. Theoretical Foundations of Machine Learning Models

Several machine learning algorithms have been widely used for weather and climate predictions. These models include simple linear methods, as well as ensemble and gradient boosting techniques. Each offers its own strengths and considerations.

1. Linear Regression

Linear Regression models the relationship between meteorological variables and the target variable as a linear combination:

$$\hat{y} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \tag{1}$$

where x_i represent input weather parameters (temperature, humidity, pressure, etc.), and β_i denote regression coefficients.

The parameters are estimated by minimizing the Mean Squared Error (MSE):

$$J(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

Although Linear Regression provides a simple and interpretable baseline model, it is limited in capturing nonlinear atmospheric interactions.

2. K-Nearest Neighbors (KNN)

KNN is a non-parametric model that predicts values based on similarity between samples. For a test sample x, the prediction is computed as:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i \tag{3}$$

where the k nearest neighbors are selected based on Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{m} (x_{il} - x_{jl})^2} \tag{4}$$

KNN is effective in capturing local weather patterns but suffers from high computational cost for large-scale meteorological datasets.

3. AdaBoost

AdaBoost constructs a strong learner by combining multiple weak learners sequentially. At iteration t, a weak learner $h_t(x)$ is trained to minimize weighted error:

$$e_t = \sum w_i \, I(y_i \neq h_t(x_i)) \tag{5}$$

The model weight is computed as:

$$\alpha_t = \frac{1}{2} \ln \frac{1 - e_t}{e_t} \tag{6}$$

The final prediction is:

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{7}$$

AdaBoost reduces bias and improves prediction accuracy but is sensitive to noisy meteorological data.

4. XGBoost

XGBoost is an optimized gradient boosting framework that minimizes the objective function:

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{8}$$

where L represents the differentiable loss function and $\Omega(f_k)$ denotes the regularization term to control model complexity.

The model uses second-order Taylor approximation of the loss function and sequentially fits decision trees to residual errors. Regularization enhances generalization and reduces overfitting, making XGBoost highly suitable for nonlinear weather prediction tasks.

5. LightGBM

LightGBM is a gradient boosting framework that improves efficiency using histogram-based learning and leaf-wise tree growth.

$$y\hat{}^{(t)} = y\hat{}^{(t-1)} + \eta f_t(x) \tag{9}$$

where η represents the learning rate.

LightGBM introduces Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), which reduce training time and memory usage, making it suitable for large-scale meteorological datasets.

6. CatBoost

CatBoost is a gradient boosting algorithm designed to effectively handle categorical features. The objective function is:

$$Obj = \sum L(y_i, y\hat{}_i) + \Omega(f) \tag{10}$$

It employs ordered target encoding, symmetric decision trees, and ordered boosting to prevent prediction shift. CatBoost demonstrates strong performance in structured weather datasets containing mixed numerical and categorical variables.

IV. System Design and UML Diagrams

This section presents the system design representations and Unified Modeling Language (UML) diagrams associated with the proposed weather forecasting framework. UML diagrams provide a visual understanding of the interactions, workflow, structural organization, and operational behavior of the forecasting system. These diagrams help illustrate how different system components collaborate to support weather prediction, data management, and result visualization. The presented models offer a high-level representation of the forecasting environment and serve as a conceptual blueprint for understanding the overall system functionality.

A. Use Case Diagram

The use case diagram illustrates the interaction between users and the weather forecasting system. It identifies the primary functionalities supported by the framework, including weather data acquisition, forecast generation, result visualization, and system management. The diagram provides an overview of the services offered by the forecasting platform and highlights the roles of different actors involved in the forecasting process.

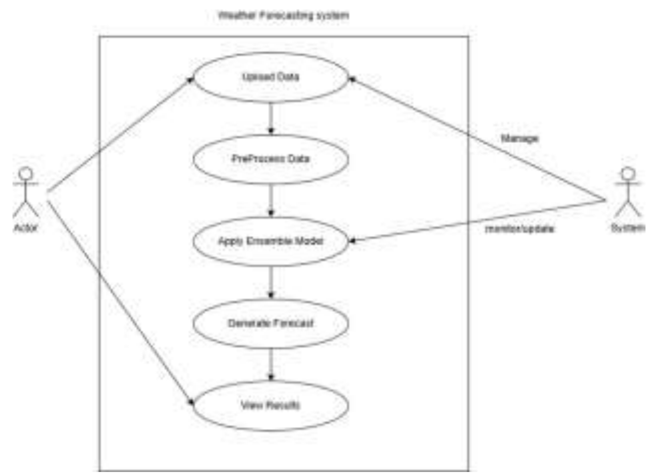

Fig. 2: Use case diagram of system

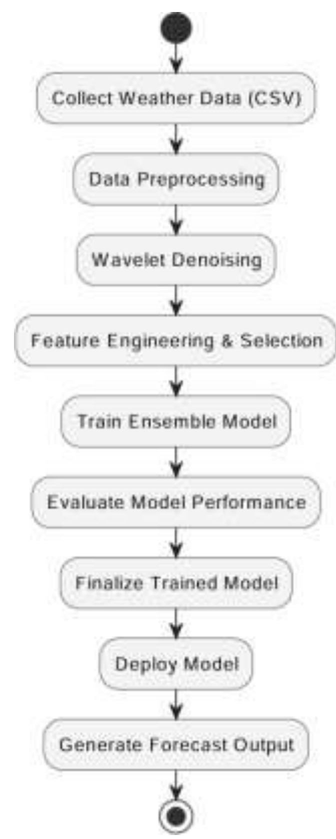

Fig. 3: Activity diagram of system

B. Activity Diagram

The activity diagram represents the overall workflow of the weather forecasting process. It describes the sequence of operations beginning with data collection and preparation, followed by analytical processing and forecast generation. The diagram demonstrates how information flows through different stages of the system and provides a simplified view of the forecasting lifecycle.

C. Sequence Diagram

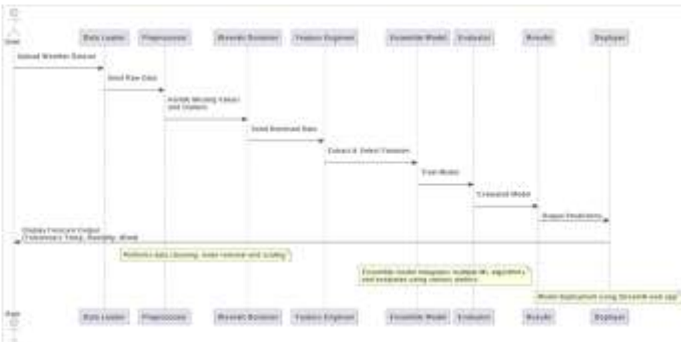

Fig. 4: Sequence diagram for System Developement

The sequence diagram illustrates the communication between various system components during the execution of forecasting tasks. It shows the exchange of information among users, processing modules, forecasting components, and result presentation interfaces. The diagram helps in understanding the chronological order of interactions involved in generating weather predictions.

D. Class Diagram

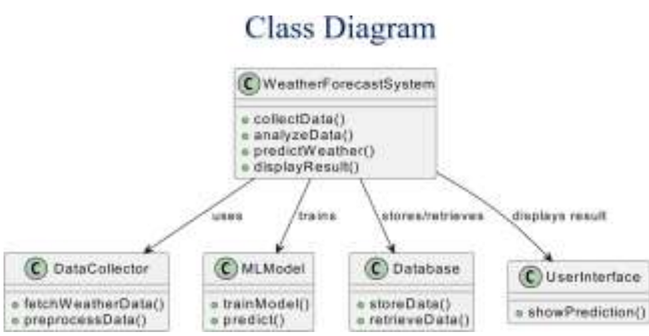

Fig. 5: Class diagram

The class diagram provides a structural representation of the forecasting framework by identifying the major entities and their relationships. It illustrates how different components responsible for data handling, forecasting operations, and user interaction are organized within the system. The diagram offers a conceptual view of the system architecture and supports a better understanding of component-level organization.

V. Explainable Artificial Intelligence in Weather Forecasting

The growing use of machine learning in weather forecasting has increased the need for transparent and interpretable prediction systems. Although advanced models can achieve high forecasting accuracy, understanding their decision-making processes remains a challenge. Explainable Artificial Intelligence (XAI) addresses this issue by providing insights into model behavior and identifying the factors that influence forecasting outcomes.

By improving transparency and user trust, XAI supports more reliable decision-making in weather-sensitive domains such as agriculture, disaster management, transportation, and environmental monitoring. As intelligent forecasting systems continue to evolve, explainability is expected to play an important role in enhancing their practical applicability and adoption.

VI. Dataset Description

Weather forecasting models rely heavily on the availability of diverse and high-quality meteorological data. Such datasets are typically collected from weather stations, satellite observations, environmental sensors, and publicly available weather repositories. These data sources provide valuable information regarding atmospheric conditions and enable machine learning algorithms to identify patterns associated with weather variations over time.

The datasets used in weather forecasting generally consist of multiple meteorological parameters, including temperature, humidity, atmospheric pressure, rainfall, wind speed, visibility, and other environmental indicators. The availability of historical observations allows predictive models to learn temporal trends and seasonal characteristics, thereby improving forecasting effectiveness. Furthermore, the integration of data collected from different geographical regions enhances the ability of forecasting systems to generalize across diverse climatic conditions.

VII. Performance Analysis

Weather forecasting aptitude of machine learning models is assessed in terms of conventional regression indicators including Mean Absolute Error (MAE), Root mean Square error (RMSE), and Mean Squared error (MSE). They are metrics used to quantify the difference between the predicted and actual values of weather and are commonly used in the research on forecasting [5], [12]. Boosting algorithms achieve better forecasting results because of their sequential error-correction method. This method allows for improved modeling of nonlinear weather relationships. However, these models can be more complex and less transparent than simpler regression techniques. XGBoost, CatBoost, AdaBoost and LightGBM boosting based models usually have a higher accuracy than conventional regression based models as they can adapt non linear relationships and high dimensional data. Ensemble learning enhances robustness through lessening variation and consolidating various predictions [4], [8]. Hybrid models that include preprocessing and several learning algorithms are always better than independent models in the accuracy of prediction, noise tolerance and ability to generalize. Researchers have found that wavelet denoising has been useful in machine learning models to improve forecasting accuracy in a significant way [1], [15]. Although these benefits are present, issues like computational complexity, need to have large datasets, and model interpretability remain. The upcoming studies should be conducted on the real-time implementation, scalable systems, and their integration with climate simulation models to enhance the reliability of forecasts [18], [22].

VIII. Comparative Evaluation of Forecasting Techniques

A comparison of existing forecasting techniques indicates that different machine learning models perform differently depending on data characteristics, prediction horizon, and feature availability. Regression models provide baseline prediction capability but struggle with nonlinear patterns. Neural networks and deep learning approaches capture temporal dependencies effectively but require large datasets and significant computational resources. Ensemble learning models improve stability and accuracy by combining multiple predictions, whereas boosting algorithms enhance performance by sequentially reducing prediction errors [4], [8], [12].

Hybrid frameworks that integrate preprocessing, feature selection, and ensemble modeling demonstrate the highest performance across multiple studies. These models effectively address limitations such as noise sensitivity and overfitting while improving generalization capability [1], [18]

IX. Conclusion and future work

This work demonstrates the increased significance of hybrid machine learning systems in weather prediction. The review of the literature shows that the combination of preprocessing methods and boosting and ensemble algorithms provides a high degree of predictive robustness and accuracy as compared to single models. Hybrid techniques are effective in nonlinear atmospheric interactions and also noise sensitivity, which enables them to be used in complex meteorology forecasting tasks.

Although considerable progress has been achieved, challenges related to computational complexity, model interpretability, and regional generalization remain active areas of research. Future studies should focus on the development of scalable and computationally efficient forecasting frameworks capable of supporting real-time applications. The integration of explainable artificial intelligence techniques and meteorological domain knowledge may further enhance transparency, reliability, and practical applicability of intelligent forecasting systems.Overall, the findings highlight the growing potential of data-driven forecasting approaches in addressing modern weather prediction challenges. Continued advancements in machine learning and intelligent analytical techniques are expected to contribute significantly to the development of more accurate, adaptable, and user-centric weather forecasting solutions across diverse environmental conditions.

X. ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Prof. Pradip Paithane for his continuous guidance and motivation. The authors also give gratitude to the Department of Information Technology, Baramati, for providing the infras structure required for this project. We also acknowledge the valuable contributions of the research community and developers of open-source meterological information, datasets and machine learning libraries.

References

[1] D. Niu, L. Diao, Z. Zang, H. Che, T. Zhang, and X. Chen, "A Machine-Learning Approach Combining Wavelet Packet Denoising with CatBoost for Weather Forecasting," MDPI, vol. 12, p. 1618, 2021.
[2] D. Kair, A. Baktygaliyev, A. Kassymova, A. Joldybayev, N. Ramazan, and A. Kartbayev, "Advancing Information Systems for Smart Decision Making Using Machine Learning-Based Weather Prediction," 2025.
[3] Saravana M. K., Dhanush B., Harish Potadar, Lakshman S., and Roshan Zameer, "Next Generation Weather Forecasting: A Survey on Integrating AI Models for Accurate and Scalable Climate Predictions," IJSRST, vol. 12, no. 2, pp. 1251–1266, 2025.
[4] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," IEEE, vol. 10, pp. 473–480, 2022.
[5] V. Kumar, N. Kedam, K. V. Sharma, K. M. Khedher, and A. E. Alluqmani, "A Comparison of Machine Learning Models for Predicting Rainfall in Urban Metropolitan Cities," MDPI, vol. 15, pp. 1–24, 2023.
[6] A. H. M. Jakaria, M. Hossain, and M. A. Rahman, "Smart Weather Forecasting Using Machine Learning: A Case Study in Tennessee," arXiv:2008.10789, 2020.
[7] S. Singh, M. Kaushik, A. Gupta, and A. K. Malviya, "Weather Forecasting Using Machine Learning Techniques," 2021.
[8] H. Shaiba et al., "Weather Forecasting Prediction Using Ensemble Machine Learning for Big Data Applications," Computers, Materials Continua, 2022.
[9] Shimaila and S. Siddiqi, "Enhancing the Effectiveness of Weather Forecasting Using Ensemble Machine Learning Techniques," 2023.
[10] M. M. Raghuwanshi et al., "Weather Prediction with Machine Learning," International Journal of Innovative Science and Research Technology, vol. 9, no. 9, 2024.
[11] A. Ansari, H. Kumar, and I. Mubarak, "Weather Prediction Using Machine Learning," International Journal of Scientific Development and Research, vol. 10, no. 4, 2025.
[12] S. B. Nuthalapati and A. Nuthalapati, "Accurate Weather Forecasting with Dominant Gradient Boosting Using Machine Learning," International Journal of Science and Research Archive, vol. 12, no. 2, pp. 408–422, 2024.
[13] M. Azad, T. H. Nehal, and M. Moshkov, "A Novel Ensemble Learning Method Using Majority-Based Voting of Multiple Selective Decision Trees," Springer Nature Computer Science, 2024.
[14] C. B. Pande et al., "Forecasting of Meteorological Drought Using Ensemble and Machine Learning Models," Environmental Sciences Europe, 2024.
[15] A. Altunkaynak and M. Ozger, "Comparison of Discrete and Continuous Wavelet–Multilayer Perceptron Methods for Daily Precipitation Prediction," ASCE Journal of Hydrologic Engineering, 2023.
[16] S. Ghosh and M. K. Gourisaria, "A Pragmatic Ensemble Learning Approach for Rainfall Prediction," Discover Internet of Things, Springer, 2023.
[17] V. Kumar et al., "A Comparative Study of Machine Learning Models for Daily and Weekly Rainfall Forecasting," 2024.
[18] L. J. Slater et al., "Hybrid Forecasting: Blending Climate Predictions with AI Models," Hydrology and Earth System Sciences, vol. 27, no. 5, pp. 1865–1889, 2023.
[19] K. Kashinath et al., "Physics-Informed Machine Learning: Case Studies for Weather and Climate Modeling," Nature Reviews Physics, 2023.
[20] S. Ravishankar and N. Ravi, "A Combined System for Predicting Weather Using Machine Learning and Optimization Techniques," International Journal of Engineering and Science, vol. 11, no. 3, 2025.
[21] S. Z. Hussain and K. Rajakumari, "Weather Forecasting Using Machine Learning Techniques," International Journal of Advanced Research in Science, Communication and Technology, 2025.
[22] H. Al-Sukhni et al., "Data-Driven Weather Prediction: Integrating Deep Learning and Ensemble Models for Robust Weather Forecasting," Journal of Artificial Intelligence Research, 2025.
[23] K. Nalini et al., "Real-Time Weather Estimation and Forecasting Using Hybrid Machine Learning Approaches," Asian Journal of Research in Computer Science, 2025.
[24] L. Evans et al., "Developing a Hybrid Weather Forecasting Model Using Machine Learning and Traditional Approaches," 2024.
[25] I. Price et al., "Probabilistic Weather Forecasting with Machine Learning," Nature, vol. 625, pp. 123–140, 2024.