© May 2026 | IJIRT | Volume 12 Issue 12 | ISSN: 2349-6002

Predicting Environmental Changes Using Machine Learning

Rajendra Dilip Walave[1], Chaitanya Anil Sarje[2], Mohit Rooparam Sirvi[3], Sanket Vinayak Jagtap[4]
[1,2,3,4] S.Y.MSc(Computer Science), MIT Arts, Commerce & Science College, Alandi(D), Pune, Savitribai Phule Pune University

***Abstract*—Environmental variability poses far-reaching consequences for ecosystems, economic systems, and public welfare. This study investigates four widely adopted machine learning (ML) classifiers — K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Decision Trees (DT), and Naive Bayes (NB) — for predicting categorical climate states from meteorological measurements including temperature, humidity, and wind speed. The research further examines whether predictive performance can be elevated through stacking ensemble learning. The standalone Decision Tree achieved the highest accuracy at 92.6%, while the stacking ensemble reduced Decision Tree training time to 3.62 seconds — a significant gain for real-time forecasting scenarios. These findings demonstrate the practical value of ML-based ensemble techniques for environmental forecasting applications.**

***Index Terms*—Climate prediction, Decision Tree, ensemble learning, machine learning, stacking, Support Vector Machine.**

I. INTRODUCTION

Environmental variability poses far-reaching consequences for ecosystems, economic systems, and public welfare. The capacity to anticipate and respond to climatic shifts — whether manifested as extreme heat events, altered precipitation regimes, or shifting wind patterns — is central to robust resource management, disaster risk reduction, and evidence-based environmental policy. Conventional meteorological forecasting frameworks, grounded in deterministic physical equations, have served well for decades; however, they encounter inherent limitations when applied to high-dimensional, non-linear climatic systems where the volume of sensor data surpasses the capability of traditional analytical methods.

The advent of machine learning (ML) offers a compelling complementary paradigm. Rather than encoding domain knowledge as explicit equations, ML models extract predictive structure directly from historical observations, allowing them to capture subtle statistical relationships that classical models may overlook. This capability is particularly valuable in environmental science, where variable interactions are numerous and often non-linear.

The present study investigates the suitability of four widely adopted ML classifiers — K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Decision Trees (DT), and Naive Bayes (NB) — for predicting categorical climate states from meteorological measurements. The research further examines whether predictive performance can be elevated through stacking ensemble learning, a meta-learning strategy that synthesises the outputs of multiple base classifiers into a unified, more robust prediction.

A. Research Objectives

The study is guided by the following objectives: (1) To assess and compare the individual predictive performance of KNN, SVM, DT, and NB on a historical climate dataset characterised by temperature, humidity, and wind speed variables. (2) To implement and evaluate a stacking ensemble model combining SVM and Decision Tree as base learners with a logistic regression meta-learner. (3) To measure each model's performance across accuracy, precision, recall, F1-score, and training time. (4) To derive practical insights for the development of scalable, interpretable ML-based environmental forecasting systems. The analysis reveals that standalone Decision Trees achieve the highest accuracy (92.6%), while stacking ensemble application reduces training time to 3.62 seconds — a significant gain for real-time forecasting scenarios.

II. LITERATURE REVIEW

A growing body of research supports the integration of machine learning into environmental and

meteorological science. The following review synthesises key works that inform the present study.
Kadam et al. [22] developed a regression framework for estimating atmospheric visibility using climatic indicators such as temperature, humidity, wind speed, precipitation, and barometric pressure. Drawing on a high-resolution NOAA dataset with hourly records, the authors demonstrated that robust meteorological data combined with principled feature engineering can yield reliable visibility estimates.

Jakaria et al. [23] examined linear regression and functional regression approaches for short-horizon temperature forecasting. The review also synthesised complementary findings including Krasnopolsky's hybrid neural-physical framework, Radhika et al.'s application of SVM to weather classification, and Montori et al.'s crowdsensing platform (Sen-Square). These studies highlight the value of leveraging spatially distributed data to improve forecast skill.

Yang et al. [17] conducted a comprehensive survey of interpretability methods applied to ML weather and climate models. They found that explainability is a critical prerequisite for practitioner trust and identified several open challenges, including the need for standardised evaluation benchmarks.

Hamdan et al. [14] reviewed the expanding role of AI and ML in climate change research, noting that ML models have become indispensable for processing large-scale datasets — including satellite imagery and multi-decadal observational records — to generate forecasts of climate patterns and extreme weather phenomena.

Bochenek and Ustrnul [19] systematically reviewed over 500 research articles published since 2018 on ML applications in meteorology and climatology. They found that deep learning, random forests, artificial neural networks, SVM, and XGBoost dominate the literature.

Vázquez-Ramírez et al. [15] surveyed ML applications to climate change, emphasising the urgency created by projected global temperature increases of 0.15–0.6 °C per decade attributable to greenhouse gas emissions. Deep learning architectures — particularly convolutional neural networks — showed promise in predicting sea-ice extent and extreme weather events.

Chen et al. [18] presented an extensive review of interpretable ML techniques in weather and climate prediction, recommending that future research develop richer interpretability workflows that feed directly into the model development cycle.

III. DATASET DESCRIPTION

This study employs a Historical Climate Dataset comprising 203 annotated climate records. Each record captures three continuous meteorological attributes — temperature, humidity, and wind speed — along with a discrete categorical label indicating the prevailing climate class. The dataset is stored in CSV format with a total size of approximately 2.1 MB.

A. Feature Descriptions

Temperature (°C): Spans a range from approximately 11 °C to 130 °C, encompassing both temperate and extreme thermal conditions. Humidity (%): Relative humidity measurements ranging from 20% to 90%, reflecting a broad spectrum of moisture environments. Wind Speed (m/s): Wind speed values between 1 m/s and 30 m/s; elevated readings correspond to more turbulent or stormy conditions. Climate Category: A four-class nominal target variable — 0: Moderate, 1: Hot, 2: Cool, 3: Very Hot.

TABLE I: Sample entries from the Historical Climate Dataset

Temp (°C)	Humidity (%)	Wind (m/s)	Category
22	60	5	Moderate
30.5	45	10	Hot
15	80	2	Cool
25	50	7	Moderate
35	30	15	Hot

IV. EXPERIMENTAL METHODOLOGY

This section describes the step-by-step experimental pipeline employed to train, evaluate, and compare the individual and ensemble classifiers. The overarching aim was to classify climate records into one of four categories — Moderate, Hot, Cool, or Very Hot — and to quantify the trade-off between predictive accuracy and computational efficiency.

A. Dataset Preparation

The Historical Climate Dataset was partitioned into training and test subsets using a stratified split to preserve the proportional representation of all climate classes. Numerical features were standardised prior to training SVM models, as the algorithm is sensitive to feature scale. Decision Trees do not require normalisation and were trained on the raw feature values. Performance was assessed using accuracy, precision, recall, F1-score, and training time as primary evaluation metrics.

B. Base Classifiers

Support Vector Machines (SVM): SVM is a kernel-based supervised classifier that identifies an optimal decision boundary — a hyperplane — which maximises the separation margin between class clusters in a transformed feature space. SVM is particularly robust in situations where decision boundaries are non-linear, provided an appropriate kernel is selected through cross-validation. Its principal limitation lies in computational cost during training.

Decision Trees (DT): Decision Trees construct a hierarchical, tree-structured model in which each internal node represents a binary splitting rule on a single feature, and each leaf node corresponds to a predicted class. The most frequently adopted splitting criteria are information gain based on entropy and the Gini impurity index. Decision Trees offer a significant interpretability advantage but are prone to overfitting without regularisation.

C. Stacking Ensemble Learning

Stacking (stacked generalisation) is an ensemble technique that trains multiple diverse base learners on the same dataset and then employs a separate meta-learner to combine their predictions optimally. Unlike bagging or boosting, stacking learns the combination function from data [20]. In this study: (1) Base learners — SVM and Decision Tree — were trained independently on the training set. (2) Out-of-fold predictions were generated through k-fold cross-validation. (3) A logistic regression meta-learner was fitted on the out-of-fold predictions. (4) Final predictions on the held-out test set were obtained by passing base learner predictions through the trained meta-learner.

V. RESULTS AND ANALYSIS

The evaluation compared standalone SVM and Decision Tree classifiers, as well as their stacking ensemble variants. Performance was measured across precision, recall, F1-score, accuracy, and training time.

A. Standalone Model Performance

The standalone SVM model demonstrated strong performance across the Hot and Cool climate classes, achieving perfect recall (1.00) for the Hot category and an F1-score of 0.92 for both Hot and Cool. Performance was comparatively lower for the Moderate category, with a recall of 0.67 and F1-score of 0.77.

TABLE II: Classification Report — Standalone SVM

Climate Class	Precision	Recall	F1-Score
Moderate	0.89	0.67	0.77
Hot	0.86	1.00	0.92
Cool	0.88	0.96	0.92

The standalone Decision Tree achieved perfect precision for the Moderate class (1.00) and near-perfect recall for the Cool class (1.00), yielding an F1-score of 0.98 for Cool — the highest per-class F1 across all configurations.

TABLE III: Classification Report — Standalone Decision Tree

Climate Class	Precision	Recall	F1-Score
Moderate	1.00	0.83	0.91
Hot	0.71	0.83	0.77
Cool	0.96	1.00	0.98

TABLE IV: Overall Accuracy — Standalone Models

Algorithm	Accuracy (%)
SVM	87.8
Decision Tree	92.6

Decision Trees outperformed SVM by approximately 4.8 percentage points in standalone accuracy.

B. Stacking Ensemble Performance

TABLE V: Classification Report — SVM in Stacking Ensemble

Climate Class	Precision	Recall	F1-Score
Moderate	0.85	0.79	0.81

Hot	0.79	0.85	0.81
Cool	0.91	0.91	0.91
Very Hot	1.00	1.00	1.00

TABLE VI: Classification Report — Decision Tree in Stacking Ensemble

Climate Class	Precision	Recall	F1-Score
Moderate	0.92	0.79	0.85
Hot	0.80	0.92	0.86
Cool	0.91	0.91	0.91
Very Hot	1.00	1.00	1.00

TABLE VII: Overall Accuracy — Stacking Ensemble Models

Algorithm	Accuracy (%)
SVM (Stacking)	86.0
Decision Tree (Stacking)	88.0

TABLE VIII: Model Training Time Comparison

Algorithm	Training Time (s)
SVM (Stacking)	5.27
Decision Tree (Stacking)	3.62

The stacked Decision Tree attained 88% accuracy while requiring only 3.62 seconds of training time — approximately 31% faster than the stacked SVM (5.27 seconds).

VI. CONCLUSION

This research evaluated the predictive capabilities of Support Vector Machines and Decision Trees for classifying historical climate records into discrete environmental categories (Moderate, Hot, Cool, Very Hot), and further examined whether stacking ensemble learning could enhance these capabilities.
The standalone Decision Tree emerged as the most accurate individual classifier at 92.6%, demonstrating the algorithm's capacity to capture complex, non-linear decision boundaries in meteorological feature spaces. The standalone SVM achieved 87.8% accuracy, reflecting its strong generalisation properties, though at the cost of longer training times.

When stacking ensemble learning was applied, the Decision Tree variant within the ensemble attained 88% accuracy with a training time of only 3.62 seconds — a compelling trade-off between accuracy and speed relative to the stacked SVM, which required 5.27 seconds to achieve 86% accuracy. The slight accuracy reduction from standalone to stacked Decision Tree (92.6% to 88%) indicates that the current stacking configuration may not yet be fully optimised, presenting a direction for future refinement.

Overall, these findings underscore the practical value of ensemble techniques in environmental forecasting, particularly in scenarios demanding rapid model retraining. They also highlight the importance of configuration-aware model selection: the optimal choice between a standalone and an ensemble approach depends critically on the operational constraints of accuracy versus computation time.

VII. FUTURE SCOPE AND LIMITATIONS

A. Future Research Directions

(1) Advanced Ensemble Architectures: Investigating gradient boosting frameworks such as XGBoost or LightGBM may yield further accuracy gains. (2) Real-Time Data Assimilation: Integrating live data streams from IoT meteorological sensors and satellite imagery could substantially increase model responsiveness. (3) Deep Learning for Spatiotemporal Data: Recurrent architectures such as LSTMs and ConvLSTM are particularly suited to modelling sequential climatic dynamics. (4) Hybrid Physics-Informed ML Models: Coupling ML classifiers with numerical weather prediction frameworks could produce models that respect known physical constraints while benefiting from data-driven flexibility. (5) Explainable AI (XAI) Integration: Applying post-hoc explanation tools such as SHAP or LIME would improve interpretability. (6) Long-Horizon Climate Trend Modelling: Extending the classification approach to regression-based multi-decade trend forecasting. (7) Computational Optimisation: Hardware acceleration, model quantisation, and pruning techniques could further reduce inference latency.

B. Limitations

(1) Accuracy Trade-off in Ensemble Models: The application of stacking reduced Decision Tree accuracy from 92.6% to 88%. (2) Dataset Scale: With 203 records, the dataset is comparatively small; generalisation to geographically diverse or temporally extended datasets remains to be validated. (3) Data Quality Dependency: Both SVM and

Decision Trees are sensitive to noise and class imbalance. (4) Overfitting Risk: Standalone Decision Trees are particularly susceptible to overfitting on small datasets. (5) Interpretability of Stacking: The two-layer prediction architecture reduces interpretability relative to the standalone Decision Tree. (6) Short-Term Prediction Scope: The models were designed for short-horizon classification tasks; their applicability to multi-year climate projection has not been assessed.

REFERENCES

[1] A. Dumont, X. Liu, and L. Li, "A comprehensive survey of machine learning applications in weather and climate prediction," ArXiv, 2024.

[2] J. Pasquier and I. Diallo, "Evaluation of machine learning methods for weather and climate prediction," Geoscientific Model Development, vol. 16, no. 8, pp. 6433–6454, 2023.

[3] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," The American Statistician, vol. 46, no. 3, pp. 175–185, 1992.

[4] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Tech. Report, National Taiwan University, 2010.

[5] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[6] S. S. Keerthi et al., "Improvements to Platt's SMO algorithm for SVM classifier design," Neural Computation, vol. 13, no. 3, pp. 637–649, 2001.

[7] J. R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[8] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.

[9] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, Classification and Regression Trees. CRC Press, 1984.

[10] I. H. Sarker et al., "BehavDT: A behavioural decision tree for user-centric context-aware predictive modelling," Mobile Networks and Applications, vol. 25, pp. 1151–1161, 2019.

[11] I. H. Sarker et al., "IntrudTree: A machine learning-based cybersecurity intrusion detection model," Symmetry, vol. 12, no. 5, p. 754, 2020.

[12] L. Breiman, "Stacked regressions," Machine Learning, vol. 24, no. 1, pp. 49–64, 1996.

[13] Y. Chen, A. McGovern, and K. L. Elmore, "Enhancing weather prediction with ensemble machine learning," Journal of Applied Meteorology and Climatology, vol. 59, no. 1, pp. 19–33, 2020.

[14] A. Hamdan et al., "AI and machine learning in climate change research," Manuscript submitted for publication, 2024.

[15] S. Vázquez-Ramírez et al., "An analysis of climate change based on machine learning and an endoreversible model," Mathematics, vol. 11, no. 14, p. 3060, 2023.

[16] A. McGovern et al., "Using machine learning for real-time forecasting of severe weather events," Bulletin of the American Meteorological Society, vol. 98, no. 4, pp. 707–719, 2017.

[17] R. Yang et al., "Interpretable machine learning for weather and climate prediction: A survey," arXiv preprint, 2024.

[18] L. Chen et al., "Machine learning methods in climate prediction: A survey," Preprints, 2023.

[19] B. Bochenek and Z. Ustrnul, "Machine learning in weather prediction and climate analyses," Atmosphere, vol. 13, no. 2, p. 180, 2022.

[20] D. H. Wolpert, "Stacked generalisation," Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.

[21] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," Nature, vol. 521, no. 7553, pp. 452–459, 2015.

[22] G. Kadam et al., "Climate visibility prediction using machine learning," IRJET, vol. 10, no. 5, 2023.

[23] A. H. M. Jakaria et al., "Smart weather forecasting using machine learning: A case study in Tennessee," arXiv preprint, 2020.

[24] K. M. Ting and I. H. Witten, "Issues in stacked generalisation," Journal of Artificial Intelligence Research, vol. 10, pp. 271–289, 1999.

[25] A. J. Smola and V. N. Vapnik, "Support vector regression machines," Advances in Neural Information Processing Systems, vol. 9, pp. 155–161, 1997.