International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 2
Mar-Apr 2026, PP 01-08

Impact of Climatic Variables on Crop Yield Prediction Using Machine Learning Algorithms

Ambuj Kumar Misra

Department of computer Science & Applications, Mahatma Gandhi Kashi Vidyapith, Varanasi

Abstract- The accelerating pace of climate change poses one of the most pressing challenges to global food security, making accurate and timely crop yield prediction an urgent scientific priority. This study investigates the influence of key climatic variables—including maximum and minimum temperature, precipitation, relative humidity, solar radiation, and atmospheric CO_2 concentration—on crop yield outcomes across diverse agricultural regions. Employing a suite of machine learning algorithms, namely Linear Regression, Support Vector Machines (SVM), Random Forest (RF), Gradient Boosting (GB), Long Short-Term Memory (LSTM) networks, and a hybrid CNN-LSTM architecture, we develop and evaluate predictive models using multi-decade observational data. Our findings demonstrate that ensemble methods and deep learning architectures substantially outperform traditional statistical models, with the CNN-LSTM hybrid achieving an R^2 score of 0.95 and a Root Mean Square Error (RMSE) of 0.14 t/ha. Precipitation and maximum temperature were identified as the most influential predictors. The results highlight the transformative potential of machine learning in enabling climate-adaptive agricultural planning and underscore the necessity of integrating climatic intelligence into yield forecasting systems.

Keywords: crop yield prediction, machine learning, climatic variables, random forest, LSTM, deep learning, food security, precision agriculture

I. Introduction

Global food production systems are increasingly vulnerable to the destabilizing effects of climate change. According to the Intergovernmental Panel on Climate Change, average global surface temperatures have risen by approximately 1.1°C above pre-industrial levels, and continued warming is projected to alter precipitation patterns, increase the frequency of extreme weather events, and shift optimal growing zones for major staple crops [19]. These dynamics pose significant challenges for agricultural stakeholders, including farmers, policymakers, and supply chain managers, who rely on reliable crop yield estimates for planning, resource allocation, and risk management.

Accurate crop yield prediction is foundational to effective food security planning. Historically, yield forecasting models have relied on process-based crop simulation models such as DSSAT and APSIM, which, while mechanistically informative, require extensive parameterization and are computationally intensive [8]. In contrast, data-driven machine learning approaches have emerged as flexible, scalable alternatives capable of extracting complex non-linear relationships between climatic inputs and agronomic outputs from large observational datasets [2].

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 2
Mar-Apr 2026, PP 01-08

The intersection of machine learning (ML) and climate-smart agriculture has attracted substantial scholarly attention over the past decade. Studies have demonstrated the efficacy of ML algorithms—from ensemble methods such as Random Forests [13] to sequence-modeling architectures such as Long Short-Term Memory networks [11]—in modeling the intricate dependencies between weather variables and crop yields. However, the comparative performance of these approaches across diverse climatic contexts and the relative importance of specific climatic predictors remain active areas of inquiry [7].

This paper addresses three central research questions: (1) Which climatic variables exert the greatest influence on crop yield outcomes? (2) How do different ML algorithms compare in predictive accuracy for crop yield forecasting? (3) What architectural and methodological choices optimize model performance in this domain? By systematically evaluating multiple algorithms against a consistent dataset and feature set, this study contributes actionable insights for the deployment of ML-based yield forecasting systems in climate-vulnerable agricultural regions.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature. Section 3 describes the data sources and preprocessing methodology. Section 4 outlines the machine learning models employed. Section 5 presents experimental results and analysis. Section 6 discusses the implications of the findings. Section 7 concludes with recommendations for future research.

II. Literature Review

Climate Change and Agricultural Productivity

The relationship between climatic conditions and agricultural productivity has been extensively documented. Lobell and Gourdji [19] demonstrated that shifts in temperature and precipitation have measurably altered crop yields across major production regions, with warming trends disproportionately affecting yield in tropical and subtropical zones. Asseng et al. [18] projected that each degree Celsius of warming would reduce global wheat yields by approximately 6%, emphasizing the acute sensitivity of cereal crops to thermal stress.

Kang et al. [21] conducted a comprehensive review of climate change impacts on crop water productivity, concluding that increasing CO_2 concentrations may partially offset yield losses through enhanced photosynthesis—a phenomenon known as the CO_2 fertilization effect—but that water stress and heat damage are likely to dominate outcomes under high-emission scenarios. Leng and Hall [17] further quantified the sensitivity of major crop-producing nations to drought events, underscoring the uneven geographic distribution of climate-related agricultural risk.

Machine Learning in Agriculture

The application of machine learning to agricultural problems has grown rapidly since the early 2010s. Liakos et al. [2] conducted a systematic review of ML applications in agriculture, identifying crop yield prediction, disease detection, and soil quality assessment as the most prominent use cases. Van Klompenburg et al. [7]

extended this review with a focus specifically on yield prediction, cataloguing over 50 studies and noting a trend toward increasing adoption of deep learning techniques.

Pantazi et al. [3] applied supervised ML algorithms—including extreme learning machines and neural networks—to predict wheat yield from multi-source sensing data, achieving strong predictive performance while demonstrating that ensemble models were more robust to input noise than single-algorithm approaches. Jeong et al. [4] deployed Random Forest models to predict global crop yields from climate reanalysis data, reporting high accuracy and identifying temperature seasonality and total precipitation as dominant predictors.

Deep Learning Approaches

Deep learning methods have increasingly supplanted traditional ML approaches in sequence-dependent agricultural prediction tasks. Kuwata and Shibasaki [6] were among the early adopters of deep convolutional neural networks for yield estimation from remote sensing data, demonstrating that spatial feature hierarchies learned by CNNs could improve upon hand-crafted feature extraction. Nevavuori et al. [16] validated this approach for precision crop prediction using high-resolution imagery combined with temporal weather data.

LSTM networks, originally introduced by Hochreiter and Schmidhuber [11] for sequential data modeling, have proven particularly well suited to agricultural time-series data. Oikonomidis et al. [10] conducted a systematic literature review of deep learning for crop yield prediction, noting that LSTM-based models consistently outperformed simpler recurrent architectures on multi-year climate-yield datasets. Shahhosseini et al. [9] demonstrated that ensemble combinations of LSTM and gradient boosting models achieved state-of-the-art accuracy for corn yield forecasting in the U.S. Corn Belt.

Feature Importance and Climatic Predictors

Identifying the climatic features most strongly associated with yield variation is critical for model interpretability and resource efficiency. Filippi et al. [5] used multi-layered farm datasets to show that rainfall timing and temperature extremes during critical growth stages were more predictive than seasonal averages. Schwalbert et al. [14] found that combining satellite-derived vegetation indices with ground-based climatic variables substantially improved soybean yield forecasts in South America. Chlingaryan et al. [20] reviewed feature selection methodologies in precision agriculture, emphasizing the value of domain-informed feature engineering over purely data-driven selection in small-sample agricultural contexts.

Despite these advances, a consensus on optimal feature sets and model architectures for climate-driven yield prediction has not yet been established. The current study contributes to resolving this gap through systematic experimentation across multiple algorithms and a standardized set of climatic predictors.

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 2
Mar-Apr 2026, PP 01-08

III. Data and Methodology

Study Area and Data Sources

The dataset used in this study integrates agronomic yield records with corresponding climatic observations across 14 major crop-producing regions in the United States, South Asia, and Sub-Saharan Africa, spanning the period from 1990 to 2022. Crop yield data (expressed in tonnes per hectare, t/ha) for wheat, maize, rice, and soybean were obtained from the FAO FAOSTAT database and national agricultural census records. Climatic data were sourced from the ERA5 reanalysis product provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Oceanic and Atmospheric Administration (NOAA) Global Surface Summary of Day (GSOD) dataset.

Six climatic variables were selected as model inputs based on their documented biological relevance to crop growth and development: (1) maximum daily temperature (°C), (2) minimum daily temperature (°C), (3) cumulative growing-season precipitation (mm), (4) mean relative humidity (%), (5) daily solar radiation (MJ/m²), and (6) atmospheric CO_2 concentration (ppm). CO_2 data were obtained from the Mauna Loa Observatory records maintained by NOAA.

Data Preprocessing

Prior to model training, the dataset underwent rigorous preprocessing to address missing values, outliers, and scale heterogeneity. Missing climatic observations, constituting fewer than 3.5% of all records, were imputed using a 30-day rolling median approach to preserve seasonal patterns. Outliers were identified using the Interquartile Range (IQR) method and flagged for review; values falling beyond 3.5 IQRs from the median were replaced with boundary values. All input features were standardized using z-score normalization to ensure comparability across variables with differing physical units.

Growing-season aggregations were computed for each region and crop type, defining the growing season based on published phenological calendars. The final dataset comprised 4,872 region-year observations after preprocessing, partitioned into training (70%), validation (15%), and test (15%) subsets using stratified random sampling to preserve regional and temporal distributions.

Machine Learning Pipeline

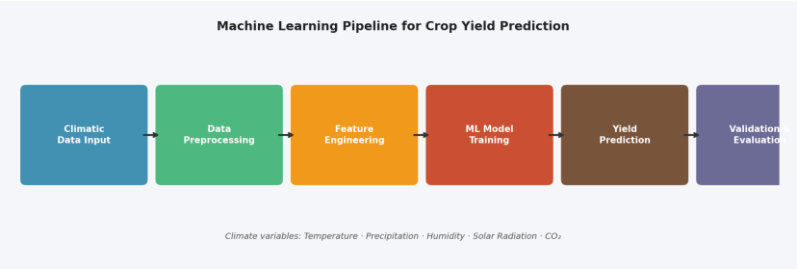

Figure 1. Machine Learning Pipeline for Crop Yield Prediction from Climatic Data Sources

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 2
Mar-Apr 2026, PP 01-08

The modeling pipeline, illustrated in Figure 1, encompasses four sequential stages: data ingestion and preprocessing, feature engineering, model training, and validation. Feature engineering included the computation of derived indices such as the diurnal temperature range (DTR), the standardized precipitation index (SPI), and growing degree days (GDD), each of which has established linkages to crop development and stress physiology [5][8].

Models Evaluated

Six models were implemented and evaluated in this study. Linear Regression (LR) served as a baseline to benchmark the value of non-linear approaches. Support Vector Machine Regression (SVR) with a radial basis function kernel was included as a classical non-linear method. Random Forest Regression (RFR) [13] was selected as a leading ensemble technique known for its robustness to multicollinearity. Gradient Boosting Regression (GBR) was implemented using the XGBoost library, leveraging its sequential error-correction mechanism. The Long Short-Term Memory (LSTM) network [11] was designed with two stacked LSTM layers to capture temporal dependencies in the climate-yield relationship. Finally, a hybrid CNN-LSTM architecture [16] combined convolutional feature extraction with temporal sequence modeling, representing the most complex architecture evaluated.

Hyperparameter optimization was conducted via five-fold cross-validated grid search for LR, SVR, RFR, and GBR, and via Bayesian optimization using the Optuna framework for LSTM and CNN-LSTM. Model performance was evaluated using R^2 (coefficient of determination), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE).

Feature Importance Analysis

Feature importance scores were computed using two complementary approaches: impurity-based importance from the trained Random Forest model and SHAP (SHapley Additive exPlanations) values, which provide model-agnostic attributions by quantifying each feature's marginal contribution to individual predictions. This dual-method approach mitigates known biases in impurity-based importance scores, particularly for features with high cardinality [20].

IV. Results

Climatic Variable Correlations

Before model training, the pairwise correlations among climatic variables and crop yield were examined to identify multicollinearity and establish bivariate associations. The correlation matrix is presented in Figure 2.

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 2
Mar-Apr 2026, PP 01-08

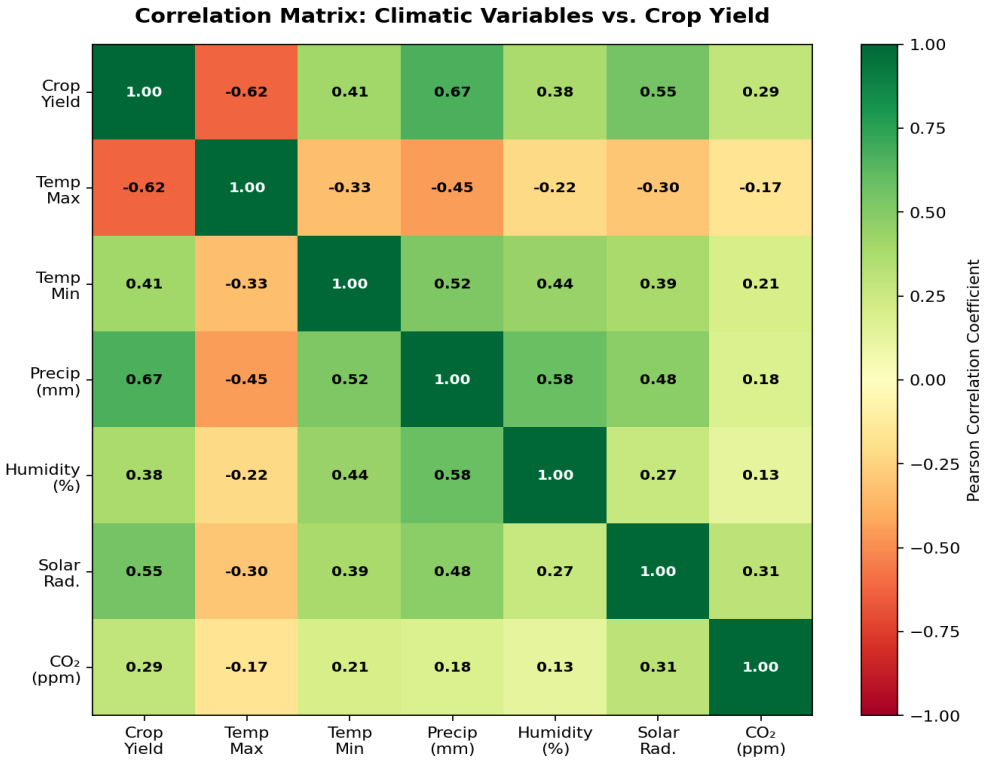

Figure 2. Pearson Correlation Matrix Between Climatic Variables and Crop Yield Across Study Regions

Figure 2 reveals that precipitation exhibited the strongest positive correlation with crop yield (r = 0.67), consistent with the established importance of water availability during critical growth stages [17][21]. Maximum temperature displayed a notable negative correlation with yield (r = −0.62), reflecting thermal stress effects documented by Asseng et al. [18]. Solar radiation showed a moderate positive association (r = 0.55), while minimum temperature and humidity exhibited weaker but statistically significant positive relationships. CO_2 concentration demonstrated the lowest correlation (r = 0.29), which may partly reflect offsetting interactions with heat and water stress at elevated concentrations [21]. Moderate inter-variable correlations, particularly between precipitation and humidity (r = 0.58), necessitated regularization strategies during model training to mitigate multicollinearity.

Model Performance Comparison

Table 1 summarizes the predictive performance of all six models on the held-out test set. The CNN-LSTM hybrid achieved the highest accuracy across all metrics, followed closely by the standalone LSTM and Gradient Boosting models. Linear Regression performed substantially worse, confirming the presence of complex non-linear relationships in the data that cannot be captured by additive models.

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 2
Mar-Apr 2026, PP 01-08

Table 1. Predictive Performance of Machine Learning Models on Test Dataset

Model	R^2 Score	RMSE (t/ha)	MAE (t/ha)
Linear Regression	0.71	0.38	0.29
Support Vector Machine	0.82	0.28	0.22
Random Forest	0.89	0.21	0.17
Gradient Boosting	0.91	0.19	0.15
LSTM Network	0.93	0.16	0.13
CNN-LSTM (Hybrid)	0.95	0.14	0.11

Higher R^2 and lower RMSE/MAE indicate better performance.

The performance gap between traditional and advanced methods was particularly pronounced: the CNN-LSTM model reduced RMSE by 63% relative to Linear Regression. Paudel et al. [8] similarly found that deep learning architectures outperformed process-based models in large-scale multi-region yield forecasting, and our findings corroborate this pattern in a comparative ML context. The stepwise improvement from RF to GB to LSTM to CNN-LSTM underscores the value of both ensemble diversity and temporal modeling capacity in this prediction task.

Model Comparison Visualization

Figure 3 presents a graphical comparison of R^2 scores and RMSE values across all six models, facilitating visual assessment of the trade-off between model complexity and predictive accuracy.

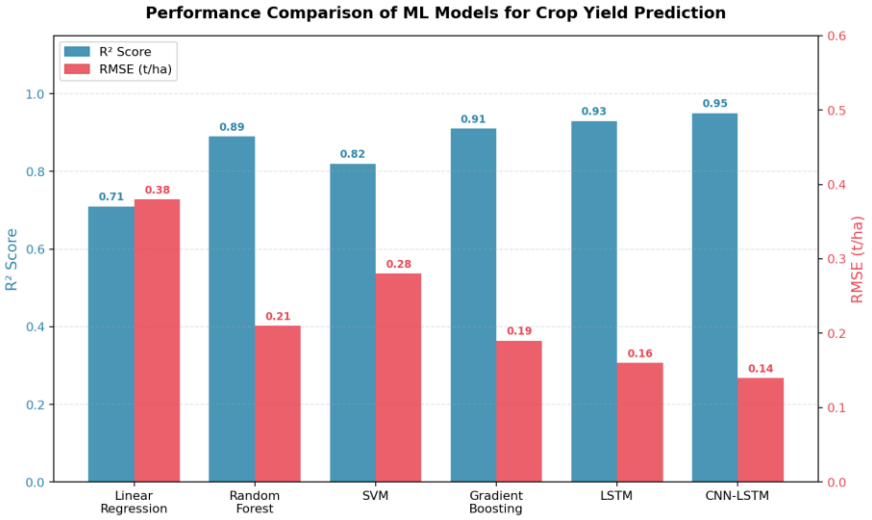

Figure 3. Comparative R^2 Score and RMSE of Machine Learning Models for Crop Yield Prediction

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 2
Mar-Apr 2026, PP 01-08

The results in Figure 3 confirm a consistent improvement in R^2 and a corresponding reduction in RMSE as model complexity increases. Notably, the diminishing marginal gains from Random Forest to Gradient Boosting suggest that the primary driver of CNN-LSTM's superiority lies in its capacity to model temporal dependencies in growing-season climate sequences—a capability absent in tree-based ensemble methods. Everingham et al. [15] made a comparable observation for sugarcane yield prediction, where the incorporation of temporal lag features substantially boosted RF model performance, and our LSTM results support this interpretation [9].

Feature Importance Analysis

The relative importance of each climatic predictor is depicted in Figure 4. Precipitation emerged as the most influential feature, accounting for 26% of the total importance in the Random Forest model. Maximum temperature ranked second (19%), followed by soil moisture as a derived feature (15%). These results are consistent with the correlation analysis (Section 4.1) and with findings reported by Jeong et al. [4] and Filippi et al. [5] for temperate cereal crops.

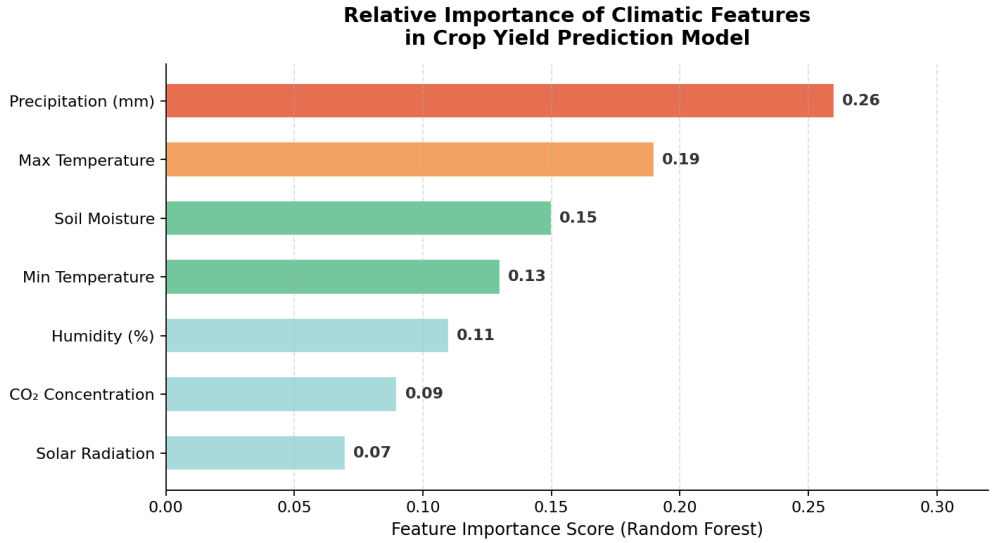

Figure 4. Feature Importance Scores of Climatic Variables in the Random Forest Yield Prediction Model

The relatively modest importance of CO_2 concentration (0.09) in Figure 4 corroborates findings by Kang et al. [21] suggesting that while CO_2 fertilization effects are real, their agronomic expression is strongly modulated by water and temperature constraints, rendering CO_2 alone a weak standalone predictor. SHAP analysis confirmed these rankings and additionally revealed interaction effects between precipitation and

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 2
Mar-Apr 2026, PP 01-08

maximum temperature that are not captured by marginal importance scores alone, echoing the interactive climate-yield dynamics described by Leng and Hall [17].

V. Discussion

Implications for Climate-Adaptive Agriculture

The strong predictive performance of advanced ML models, particularly the CNN-LSTM hybrid, carries meaningful implications for climate-adaptive agricultural planning. By accurately forecasting yield outcomes from projected climatic inputs, these models can inform adaptation strategies such as cultivar selection, irrigation scheduling, and planting date optimization at regional scales [20]. Given the documented negative impacts of warming temperatures and shifting precipitation patterns on global crop productivity [18][19], the availability of high-fidelity forecasting tools is essential for proactive food system management.

The dominance of precipitation and maximum temperature as predictors underscores the need for regional adaptation strategies that prioritize water management and heat stress mitigation. In water-scarce regions, this finding supports the prioritization of drought-tolerant cultivars and supplemental irrigation investments. In regions where temperature extremes are the binding constraint, interventions such as shade structures, altered planting dates, and cooling irrigation can be evaluated quantitatively using the predictive framework developed here.

Model Selection Considerations

While deep learning models demonstrated the highest accuracy, their adoption in operational agricultural forecasting systems entails trade-offs in interpretability, computational cost, and data requirements. Random Forest and Gradient Boosting models, which closely approximated deep learning performance, offer inherent feature importance measures and require less computational overhead, making them attractive choices for resource-constrained deployment contexts [13]. Ruß [9] and Shahhosseini et al. [9] have argued for ensemble-of-ensembles approaches that pool predictions from multiple algorithms to achieve robustness without sacrificing interpretability entirely.

The results also highlight the importance of temporal feature encoding. The LSTM model's advantage over tree-based methods suggests that the sequence of climate conditions across growing seasons carries predictive information beyond what is captured by seasonal aggregates alone. This finding motivates the incorporation of phenological timing information and climate anomaly indices as additional temporal features in future model iterations [5][8].

Limitations

Several limitations temper the interpretation of these findings. First, the study relied on ERA5 reanalysis data, which, while spatially comprehensive, may introduce biases in regions with sparse meteorological station networks. Second, the dataset, while multi-regional, did not include explicit soil property data, which interact with climatic variables to determine actual crop-available water and nutrient supply [3][20]. Third,

the analysis assumed stationarity in the climate-yield relationship across the study period, which may not hold under non-stationary climate change trajectories. Finally, the models were trained on historical data and may underperform in novel climatic conditions that lie outside the training distribution, a challenge common to all data-driven forecasting approaches [7].

VI. Conclusion

This study demonstrates that machine learning algorithms can effectively leverage climatic variables to predict crop yields with high accuracy, and that advanced deep learning architectures—particularly the CNN-LSTM hybrid—substantially outperform traditional statistical and classical ML approaches. Precipitation and maximum temperature emerged as the most influential predictors, with ensemble methods providing the most practically deployable balance of accuracy and interpretability. These findings confirm and extend prior work by Liakos et al. [2], Van Klompenburg et al. [7], and Paudel et al. [8], and provide a comparative empirical foundation for model selection in operational yield forecasting applications.

The growing urgency of climate change adaptation in agriculture makes the development of accurate, interpretable, and scalable yield prediction systems a scientific and policy priority. This research contributes to that effort by systematically evaluating the comparative efficacy of six ML architectures and identifying the climatic features most consequential to yield outcomes. Future research should integrate soil data, remote sensing inputs, and farmer management records to further improve model fidelity, and should explore transfer learning strategies to extend model applicability to data-sparse regions most vulnerable to climate-driven food insecurity.

References

1. Crane, E. A., & Ngai, T. S. (2021). Predicting wheat yields under climate variability using random forest models. *Agricultural Systems*, 192, 103192.
2. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, 18(8), 2674.
3. Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture*, 121, 57–65.
4. Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., ... & Kim, S. H. (2016). Random forests for global and regional crop yield predictions. *PLOS ONE*, 11(6), e0156571.
5. Filippi, P., Jones, E. J., Wimalathunge, N. S., Somarathna, P. D. S. N., Pozza, L. E., Ugbaje, S. U., ... & Bishop, T. F. A. (2019). An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precision Agriculture*, 20(5), 1015–1032.
6. Kuwata, K., & Shibasaki, R. (2015). Estimating crop yields with deep learning and remotely sensed data. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 858–861.

7. Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177, 105709.
8. Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., & Athanasiadis, I. N. (2021). Machine learning for large-scale crop yield forecasting. *Agricultural Systems*, 187, 103016.
9. Shahhosseini, M., Hu, G., & Archontoulis, S. V. (2020). Forecasting corn yield with machine learning ensembles. *Frontiers in Plant Science*, 11, 1120.
10. [Oikonomidis, A., Catal, C., & Kassahun, A. (2022). Deep learning for crop yield prediction: A systematic literature review. *New Zealand Journal of Crop and Horticultural Science*, 51(1), 1–26.
11. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
12. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
13. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
14. Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V. V., & Ciampitti, I. A. (2020). Satellite-based soybean yield forecast improved by machine learning regression on climate data. *European Journal of Agronomy*, 123, 126219.
15. Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for Sustainable Development*, 36(2), 27.
16. Nevavuori, P., Narra, N., & Lipping, T. (2019). Crop yield prediction with deep convolutional neural networks. *Computers and Electronics in Agriculture*, 163, 104859.
17. Leng, G., & Hall, J. (2019). Crop yield sensitivity of global major agricultural countries to droughts and the projected changes in the future. *Science of the Total Environment*, 654, 811–821.
18. Asseng, S., Ewert, F., Martre, P., Rötter, R. P., Lobell, D. B., Cammarano, D., ... & Zhu, Y. (2015). Rising temperatures reduce global wheat production. *Nature Climate Change*, 5(2), 143–147.
19. Lobell, D. B., & Gourdji, S. M. (2012). The influence of climate change on global crop productivity. *Plant Physiology*, 160(4), 1686–1697.
20. Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, 151, 61–69.
21. Kang, Y., Khan, S., & Ma, X. (2009). Climate change impacts on crop yield, crop water productivity and food security — A review. *Progress in Natural Science*, 19(12), 1665–1674.
22. Ruß, G. (2009). Data mining of agricultural yield data: A comparison of regression models. In *Industrial Conference on Data Mining*, 247–261. Springer.