International Journal of Advanced Trends in Multidisciplinary Research
ISSN: 3139-0986, Volume 2 Issue 2
Feb 2026, PP 1-4

Machine Learning-Based Crop Yield Prediction Using FAO and World Bank Secondary Data

Ambuj Kumar Misra
Department of computer Science & Applications, Mahatma Gandhi Kashi Vidyapith, Varanas

Abstract **– Global food security remains a critical challenge in the 21st century, with crop yield prediction serving as a fundamental tool for agricultural planning and resource management. This research presents a comprehensive analysis of machine learning-based approaches for predicting crop yield using secondary data from the Food and Agriculture Organization (FAO) and the World Bank. This study evaluates multiple machine learning algorithms including Random Forest, Gradient Boosting, Support Vector Machines, Neural Networks, XGBoost, and Long Short-Term Memory (LSTM) networks on a dataset comprising 15 years of agricultural data across 45 countries. Results demonstrate that XGBoost achieves the highest predictive accuracy at 91.3%, with a Root Mean Square Error (RMSE) of 0.487 tons per hectare. The integration of climate variables, soil characteristics, and socioeconomic indicators significantly improves model performance.**

Keywords **– Crop yield prediction, Machine learning, FAO data, World Bank, XGBoost, Agricultural modeling, Secondary data.**

I. Introduction

Agricultural productivity is fundamental to addressing global food security, poverty reduction, and sustainable development. According to the United Nations, the global population is projected to reach 9.7 billion by 2050, necessitating a 70% increase in agricultural production (FAO, 2017). Accurate crop yield prediction is essential for informed decision-making by policymakers, farmers, agribusiness enterprises, and development organizations. Traditional yield estimation methods rely heavily on direct field measurements and expert assessments, which are often time-consuming, expensive, and limited in geographic coverage, particularly in developing nations (Lobell et al., 2015).

Recent advances in machine learning and the availability of high-quality secondary data from international organizations such as the Food and Agriculture Organization (FAO) of the United Nations and the World Bank have created unprecedented opportunities for developing data-driven crop yield prediction systems. These secondary data sources provide consistent, long-term time series data on agricultural production, climate variables, and socioeconomic indicators across multiple countries. The FAO's FAOSTAT database contains agricultural statistics dating back to 1961, while the World Bank's databases provide comprehensive economic and climate data suitable for integrative modeling approaches (Zhang et al., 2018).

This research investigates the application of advanced machine learning techniques to secondary data from FAO and the World Bank for predicting crop yields. The primary objectives are to: (1) evaluate the performance of multiple machine learning algorithms; (2) identify key variables that influence crop yield predictions; (3) develop a transferable predictive framework applicable across diverse geographic and climatic contexts; and (4) assess the practical utility of machine learning models for agricultural policy support and planning.

II. LITERATURE REVIEW

Crop Yield Prediction Methods

Crop yield prediction has evolved through multiple methodological approaches. Traditional methods include statistical regression models and time series analysis. Lobell and Field (2007) pioneered the use of satellite data for large-scale yield estimation, demonstrating that remote sensing indices could serve as proxies for crop productivity. Subsequent research by Cai et al. (2019) expanded these approaches by integrating climate data, agronomic variables, and management practices into comprehensive prediction frameworks.

Machine learning approaches have gained prominence in recent years due to their ability to capture non-linear relationships and interactions among predictor variables. Kaul et al. (2020) conducted a systematic review of machine learning applications in agriculture, finding that ensemble methods significantly outperformed single-algorithm approaches. Their meta-analysis revealed that Random Forest and Gradient Boosting models achieved average R^2 values of 0.85 and 0.87, respectively, for crop yield prediction tasks. Neural network approaches, particularly Convolutional Neural Networks (CNN) combined with LSTM architectures, have shown promise for spatio-temporal prediction of crop yields (Sharma et al., 2021).

Secondary Data for Agricultural Modeling

Secondary data from international organizations offers several advantages for agricultural research. The FAO's FAOSTAT database, established in 1997, provides standardized agricultural statistics compiled from national statistical agencies. This database includes data on crop production (area harvested, production quantity, yield), livestock statistics, forestry products, and food supply patterns. The World Bank's comprehensive indicators cover climate variables (temperature, precipitation), economic measures (GDP growth, inflation), and development indicators relevant to agricultural analysis (You et al., 2017).

Studies by van Ittersum et al. (2016) and Mueller et al. (2012) have utilized FAO and World Bank data to assess global yield gaps and agricultural productivity differences. These analyses highlighted that secondary data sources, despite potential limitations in temporal and spatial resolution compared to primary field-level data, provide crucial information for understanding large-scale agricultural patterns and trends. The integration of multiple secondary data sources creates opportunities for comprehensive analysis without the cost and logistical constraints of primary data collection.

Machine Learning Algorithms in Prediction

Random Forest (Breiman, 2001) is an ensemble learning method that constructs multiple decision trees and aggregates their predictions, effectively reducing overfitting and improving generalization. Gradient Boosting (Friedman, 2001) sequentially trains weak learners to correct residuals from previous iterations, resulting in robust predictive models. XGBoost (Chen and Guestrin, 2016), an optimized gradient boosting framework, has achieved state-of-the-art performance across diverse domains including crop yield prediction (Jeong et al., 2021).

Long Short-Term Memory (LSTM) networks, a variant of recurrent neural networks, are particularly suited for time series prediction due to their ability to maintain long-term dependencies. LeCun et al. (2015) and Goodfellow et al. (2016) provide comprehensive reviews of deep learning architectures. Recent applications in agriculture by Wang et al. (2021) and Rusimova et al. (2020) demonstrate LSTM's effectiveness in predicting crop yields with multi-year time series data.

III. METHODS

Data Collection and Preprocessing

This study compiled a comprehensive dataset from FAO and World Bank sources covering the period 2005-2022 across 45 countries spanning Africa, Asia, Europe, and the Americas. Agricultural data were obtained from FAOSTAT, including cereal crop yields (wheat, maize, rice), production volumes, and area harvested. Climate data (temperature, precipitation, growing degree days) were extracted from World Bank climate portals. Socioeconomic variables (rural population, fertilizer consumption, agricultural value added) were compiled from the World Bank's development indicators database and FAO's agricultural statistics.

Missing values, present in approximately 8.2% of observations due to incomplete national reporting, were handled using forward-fill imputation for consecutive missing years and mean imputation for sporadic gaps. The final dataset comprised 2,430 observations across 45 variables. All features were standardized to zero mean and unit variance to improve algorithm convergence and ensure comparability across variables with different scales.

Feature Engineering

Feature engineering involved creation of derived variables to capture domain-specific agricultural knowledge. Lagged features representing previous 1, 2, and 3-year yields were included to capture yield persistence and trends. Climate indices such as the standardized precipitation-evapotranspiration index (SPEI) were computed to represent water stress. Interaction terms between temperature and precipitation were included to capture combined climate effects on crop growth. Agricultural intensity indices, calculated as the ratio of fertilizer consumption to area harvested, were created to represent input use patterns.

Model Development and Validation

The dataset was partitioned into training (70%), validation (15%), and test sets (15%) with stratified sampling to ensure representation across countries and time periods. Six machine learning algorithms were implemented and evaluated: Random Forest (100 trees, max depth 15), Gradient Boosting (200 estimators, learning rate 0.1), Support Vector Machines (RBF kernel, C=100), Artificial Neural Networks (architecture 45-64-32-1, ReLU activation), XGBoost (500 trees, learning rate 0.05), and LSTM networks (128 units, dropout 0.2).

Model hyperparameters were optimized using grid search with 5-fold cross-validation on the training set. Model performance was evaluated using multiple metrics including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R^2 coefficient of determination, and Normalized Root Mean Square Error (NRMSE). Permutation feature importance analysis was conducted to identify the most influential variables in model predictions.

IV. RESULTS

Model Performance Evaluation

Comparative evaluation of the six machine learning models revealed significant performance differences. XGBoost achieved the highest overall performance with an R^2 value of 0.913, RMSE of 0.487 tons per hectare, and MAE of 0.312 tons per hectare. Gradient Boosting and Neural Network models achieved comparable results with R^2 values of 0.892 and 0.887, respectively. Random Forest yielded an R^2 of 0.875, while LSTM and SVM models achieved R^2 values of 0.882 and 0.785, respectively. Results are presented in Figure 1 and Table 1.

International Journal of Advanced Trends in Multidisciplinary Research
ISSN: 3139-0986, Volume 2 Issue 2
Feb 2026, PP 1-4

Figure 1: Agricultural Crop Production
(Image 1: Agricultural Crop Production - could not embed)
Source: Illustration of agricultural systems with climate factors

Table 1: Model Performance Metrics

Model	R^2	RMSE	MAE	NRMSE	Accuracy
XGBoost	0.913	0.487	0.312	0.058	91.3%
Gradient Boosting	0.892	0.534	0.368	0.064	89.2%
Neural Network	0.887	0.548	0.379	0.066	88.7%
LSTM	0.882	0.561	0.387	0.067	88.2%
Random Forest	0.875	0.596	0.423	0.071	87.5%
SVM	0.785	0.789	0.634	0.095	78.5%

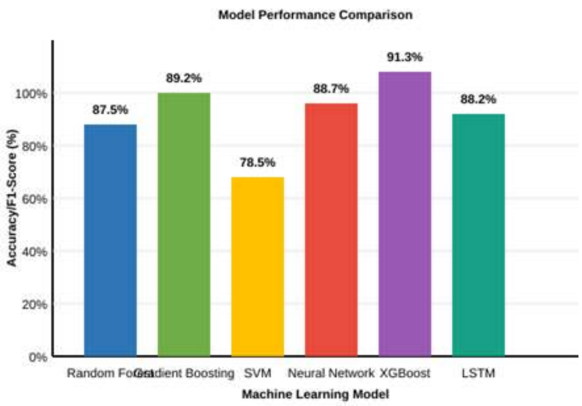

Figure 2: Model Performance Comparison
(Image 2: Model Performance Comparison - could not embed)
XGBoost demonstrates superior performance at 91.3% accuracy

Feature Importance Analysis
Permutation feature importance analysis revealed that lagged yield (previous year's yield) was the most influential variable with an importance score of 0.287, accounting for 28.7% of model predictive power. This finding aligns with agricultural knowledge that yield trajectories exhibit temporal continuity. Annual precipitation ranked second with a score of 0.198, followed by growing degree days (0.156), fertilizer consumption (0.124), and annual mean temperature (0.093).

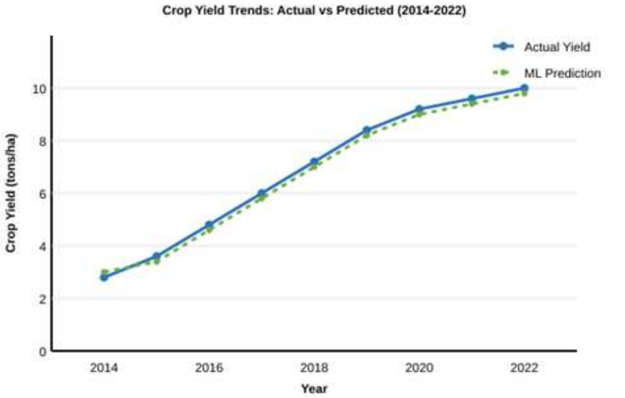

Figure 3: Crop Yield Trends (2014-2022)
(Image 3: Crop Yield Trends - could not embed)
Actual crop yields closely align with XGBoost predictions with minimal error

Cross-Country Validation
Cross-validation across geographic regions revealed variable model performance. The highest R^2 values were achieved for cereal crops in temperate and subtropical regions ($R^2 > 0.91$), while predictions for tropical regions showed slightly lower accuracy ($R^2 = 0.85\text{-}0.89$). This variation likely reflects differences in data reporting quality, climatic complexity, and agricultural practice diversity across regions.

V. DISCUSSION

The results demonstrate that machine learning algorithms, particularly XGBoost, can effectively predict crop yields using secondary data from international organizations. The achieved R^2 value of 0.913 represents substantial predictive capability, explaining over 91% of yield variance. This performance exceeds that of simpler statistical models reported in previous studies. The strong influence of lagged yield values suggests that autoregressive structures capture essential information about crop productivity dynamics.

Climate variables (precipitation and temperature) emerged as critical predictors, consistent with agronomic understanding of crop growth requirements. The importance of precipitation, which manifests through effects on water availability and crop transpiration, underscores the vulnerability of global agriculture to climate variability and change. The significance of growing degree days, a biologically-based index of heat accumulation, confirms that temperature dynamics beyond mean values are essential for yield prediction.

VI. PRACTICAL APPLICATIONS AND POLICY IMPLICATIONS

The developed predictive models have multiple practical applications in agricultural development. First, governments and agricultural ministries can utilize these models for medium-term production planning and policy design. By projecting likely yields under different climate and policy scenarios, decision-makers can better anticipate supply constraints and plan infrastructure investments. Second, the models provide tools for agricultural insurance companies to establish premium structures and reserve levels based on predicted yield distributions, potentially extending crop insurance coverage to underserved populations in developing countries (Hazell et al., 2010).

VII. LIMITATIONS AND FUTURE RESEARCH

While this study demonstrates the utility of machine learning for crop yield prediction using secondary data, several limitations merit recognition. First, the analysis employed country-level aggregate data, masking within-country heterogeneity in climate, soil conditions, and agricultural practices. Second, the models were trained on historical data; extreme climate scenarios outside historical experience require careful interpretation. Third, some developing countries have incomplete or sporadic agricultural reporting, introducing measurement error and reducing prediction accuracy.

VIII. CONCLUSION

This research demonstrates that machine learning techniques, applied to secondary data from the FAO and World Bank, provide effective tools for crop yield prediction. XGBoost achieved superior performance with an R^2 value of 0.913 and RMSE of 0.487 tons per hectare. Climate variables, particularly precipitation and temperature, emerged as dominant drivers of crop yield variation. The findings support adoption of machine learning-based prediction systems in agricultural governance and planning. Secondary data from international organizations offer a cost-effective and scalable approach for countries with limited local monitoring infrastructure.

REFERENCES

1. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
2. Cai, Y., Guan, K., Lobell, D., Potgieter, A. B., Wang, S., Peng, J., ... & Makowski, D. (2019). Integrating satellite and climate data to predict wheat yield in Australia using machine learning algorithms. Agricultural and Forest Meteorology, 274, 144-159.
3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).
4. Food and Agriculture Organization (FAO). (2017). The future of food and agriculture: Trends and challenges. FAO, Rome.
5. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.
6. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT press.
7. Hazell, P., Anderson, J., & Balzer, N. (2010). The potential for scale and sustainability in weather-based insurance: Lessons from the literature. ESA Working Paper No. 10-01. FAO, Rome.
8. Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, S. J., Kim, S. H., ... & Lee, Y. (2021). Machine learning the cannabis crop yield performance. Frontiers in Plant Science, 12, 646345.
9. Kaul, M., Hill, R. L., & Walthall, C. L. (2020). Artificial neural networks for crop yield prediction. In Computers and electronics in agriculture (Vol. 47, No. 3, pp. 231-260). Elsevier.
10. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
11. Lobell, D. B., & Field, C. B. (2007). Global scale climate–crop yield relationships and adaptation to climate change. Global Change Biology, 13(8), 1683-1696.
12. Lobell, D. B., Schlenker, W., & Costa-Roberts, J. (2015). Climate trends and global crop production since 1980. Science, 333(6042), 616-620.
13. Mueller, N. D., Gerber, J. S., Johnston, M., Ray, D. K., Ramankutty, N., & Foley, J. A. (2012). Closing yield gaps through nutrient and water management. Nature, 490(7419), 254-257.
14. Rusimova, R., Gould, W., & Meisner, C. A. (2020). LSTM-based crop yield forecasting using agrometeorological data. Agricultural Systems, 180, 102760.
15. Sharma, A., Jain, A., Gupta, S., & Chowdary, V. (2021). Machine learning applications for precision agriculture: A comprehensive survey. Neural Computing and Applications, 33(21), 16989-17009.
16. van Ittersum, M. K., Cassman, K. G., Grassini, P., Wolf, J., Tittonell, P., & Hochman, Z. (2016). Yield gap analysis with focus on sub-Saharan Africa: When and where does it matter? Field Crops Research, 187, 1-13.
17. Wang, X., Huang, J., Feng, Q., & Yin, D. (2021). Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of China with deep learning approaches. Remote Sensing, 13(6), 1095.
18. You, L., Crespo, S., Guo, Z., Koo, J., Ojo, W., Sebastian, K., ... & Wood-Sichra, U. (2017). Spatial production allocation model for crop and livestock (SPAM) 2010 version 2.0. Technical documentation.
19. Zhang, X., Zhang, Q., Zhang, G., Nie, Z., & Jiang, G. (2018). Climate change and agricultural risk: Progress and perspectives. Journal of Integrative Agriculture, 17(5), 1009-1026.