

COMPARING MACHINE LEARNING ALGORITHMS
FOR LARGE-SCALE CROP YIELD PREDICTION USING
AGROECOLOGICAL PARAMETERS AND PESTICIDE
USAGE

Pavlo Lykhovyd^{1✉}, Liudmyla Hranovska¹, Oleksandr Averchev²,
Oleksandr Zhuikov², Gennadiy Karashchuk³, Dmytro Maksymov¹

Received on October 6, 2025

Presented by H. Najdenski, Member of BAS, on November 25, 2025

Abstract

Accurate crop yield prediction is essential to ensure global food security under changing climatic, environmental and agro-industrial conditions. This study presents a comprehensive comparative analysis of machine learning algorithms for large-scale yield prediction using agroecological parameters and pesticide usage as key explanatory variables. Crop yield prediction dataset, downloaded from Kaggle, comprising 28 242 records across multiple countries and crops was preprocessed and modelled with 17 algorithms, including tree-based, regression-based, support vector, neural network, boosting, and ensemble approaches. Model performance was assessed using root mean square error (RMSE) and coefficient of determination (R^2). Modelling and visualization were performed in Python 3 using corresponding external modules. Among global models, Extra Trees achieved the highest accuracy ($R^2 = 0.991$, RMSE = 8282.92 hg/ha), outperforming both gradient boosting and neural network approaches. Ensemble techniques, particularly stacking ensembles, provided comparable accuracy ($R^2 = 0.990$), confirming the robustness of tree-based methods. Feature importance analysis highlighted crop type (0.609) as the dominant predictor, while pesticides (0.110), average temperature (0.108), and rainfall (0.087) emerged as the most influential agro-environmental factors. Country-specific models achieved near-perfect predictive power ($R^2 \approx 1.0$) for

<https://doi.org/10.7546/CRABS.2026.01.17>

India, Brazil, Pakistan, Mexico, and Turkey, while Ukraine's best-performing model (XGBoost, $R^2 = 0.980$) revealed yields averaging 43.4% below the global mean. Crop-level analysis identified potatoes, cassava, and sweet potatoes as the highest-yielding crops globally. These results demonstrate the superiority of tree-based and ensemble models for yield forecasting and emphasize the value of localized modelling strategies. Findings provide actionable insights for optimizing agricultural practices and guiding sustainable intensification policies.

Key words: data-driven agriculture, feature importance analysis, regression, ensemble modelling, sustainability

Introduction. Large-scale crop yield prediction plays a vital role in modern agricultural science. It supports food security, informs policies to mitigate agroecological and agrotechnological risks, and contributes to the sustainable use of natural resources and long-term ecosystem stability [1]. Despite decades of research, yield prediction remains challenging due to the complex interplay of numerous environmental and human factors under uncertain natural conditions [2]. Addressing these complexities increasingly requires advanced computational methods, particularly those rooted in machine learning.

Current approaches to yield prediction vary widely by data sources, modelling techniques, and spatial or temporal scale. These include models based on historical data, remote sensing, crop growth simulations, and statistical or empirical methods – of which remote sensing and empirical models are now most prevalent [3]. Advances in machine learning have significantly enhanced empirical modelling, enabling more accurate capture of complex, nonlinear relationships in large, noisy datasets [4]. Machine learning also allows integration of diverse variables, including agrotechnological factors like pesticide use, alongside weather and remote sensing data, improving predictive power and decision-making relevance [5].

However, the performance of machine learning algorithms depends heavily on dataset characteristics and prediction goals. Comparative studies, such as DHALI WAL and WILLIAMS [6], have shown that ensemble methods like Random Forest can outperform traditional regressions in accuracy and reliability. Similarly, JABED and MURAD [7] noted that models such as Random Forest, Support Vector Machines, and neural networks consistently yield strong results, while deep learning methods like CNNs and LSTMs are becoming increasingly effective for more complex scenarios. Simpler regression-based models often fall short, particularly when compared to ensemble or deep learning approaches [8].

Yet, many studies focus on narrow datasets – single crops, regions, or limited features – making broad generalizations difficult. Few have systematically compared a wide array of algorithms using a unified, large-scale framework with impactful features. This study aims to fill that gap by providing a comprehensive comparison of established machine learning algorithms for large-scale crop yield prediction. Using a robust big data framework – integrating meteorological data

and pesticide usage – we evaluate algorithm performance to identify the most effective and efficient approaches for this critical task.

Methodology. Data source and structure. The analysis used a structured crop yield prediction dataset from Kaggle (<https://www.kaggle.com/datasets/mrigaankjaswal/crop-yield-prediction-dataset?resource=download>), containing variables such as country (Area), crop type (Item), year, yield (hg/ha), annual rainfall (mm), pesticide use (t/ha), and mean annual temperature ($^{\circ}\text{C}$). These features supported both agrotechnological and agroecological modelling.

Data preprocessing. The dataset was cleaned by imputing or removing missing values. Categorical features (e.g., crop type, country) were encoded, and numerical features scaled for distance-based algorithms. Stratified sampling ensured balanced train-test splits across crops and countries [9].

Machine learning models. Over 15 models were evaluated across global, country-specific, and crop-specific contexts, including: 1) Tree-based ensembles: Random Forest, Extra Trees, Gradient Boosting, XGBoost, LightGBM; 2) Linear models: Linear Regression, Ridge Regression, Lasso Regression, Elastic Net; 3) Kernel and distance-based methods: Support Vector Regression (SVR), k-Nearest Neighbours (KNN); 4) Other learners: Artificial Neural Networks, AdaBoost. Advanced methods such as Voting, Stacking, Bagging, and hyperparameter-optimized boosting (XGBoost, LightGBM) were also applied [10].

Model training and evaluation. Models were trained using Python ML libraries. Performance was evaluated using R^2 and RMSE to capture explained variance and prediction error [11]. Results were analyzed globally, by country, and by crop type.

Feature importance analysis. Tree-based models provided feature importance scores, assessing the impact of environmental (rainfall, temperature, pesticide use) and structural variables (crop, year) [12]. Rankings were generated globally and regionally.

Visualization and reporting. Results were summarized in high-resolution plots: model comparisons, country-level outcomes, feature importance, yield distributions, and trends. Ukraine was analyzed as a mandatory case. All visualizations were saved separately to ensure clarity.

Results and discussion. The comparative evaluation of machine learning models demonstrated a clear superiority of tree-based and ensemble approaches in predicting crop yields at a global scale. Among individual algorithms, Extra Trees achieved the best overall performance ($R^2 = 0.991$, RMSE = 8282.92 hg/ha), marginally outperforming ensemble strategies such as the Stacking Ensemble ($R^2 = 0.990$, RMSE = 8373.28 hg/ha). Random Forest and Bagged Trees also produced highly accurate predictions ($R^2 = 0.986$), whereas linear and regularized regression techniques exhibited weak explanatory power ($R^2 \approx 0.08$) and high error, indicating the nonlinearity and complexity of relationships between

the yields of the studied crops and explanatory variables. Similarly, support vector regression failed to generalize, returning a negative R^2 . Neural networks and boosting methods such as AdaBoost and Gradient Boosting achieved moderate results but consistently underperformed relative to tree-based ensembles (Table 1).

T a b l e 1

Comparison of machine learning models for yield prediction from temperature, rainfall, and pesticide use

Model	RMSE (hg/ha)	R^2
Random Forest	10188.48	0.986
Gradient Boosting	34773.82	0.833
Extra Trees	8282.92	0.991
Linear Regression	81501.76	0.084
Ridge Regression	81501.77	0.084
Lasso Regression	81501.78	0.084
Elastic Net	82012.86	0.073
Support Vector Regression	93406.09	-0.203
k-Nearest Neighbours	24599.55	0.917
AdaBoost	55102.77	0.581
Neural Network	54290.31	0.594
XGBoost	13469.57	0.975
LightGBM	17195.32	0.959
Voting Ensemble	13499.09	0.975
Stacking Ensemble	8373.28	0.990
Bagged Trees	10251.11	0.986
Advanced AdaBoost	51812.02	0.630
XGB Tuned	17780.06	0.956
LightGBM Tuned	19425.36	0.948

Feature importance analysis confirmed that crop type (0.609) was the most influential factor in determining yield outcomes, reflecting the inherent biological productivity differences among the crops, enrolled in the study. Environmental and pest management parameters – pesticide use (0.110), temperature (0.108), and rainfall (0.087) – were also identified as critical drivers, but with significantly lesser impact on the yields (Fig. 1). Notwithstanding the fact of somewhat higher feature importance value for pesticide usage, it has just a subtle difference with the importance of meteorological factors (air temperature and rainfall amounts), therefore, these three factors could be considered equally important critical factors after the crop type. This highlights the dual role of agronomic practices and climatic conditions in shaping yield variability, aligning with contemporary challenges in precision and climate-smart agriculture.

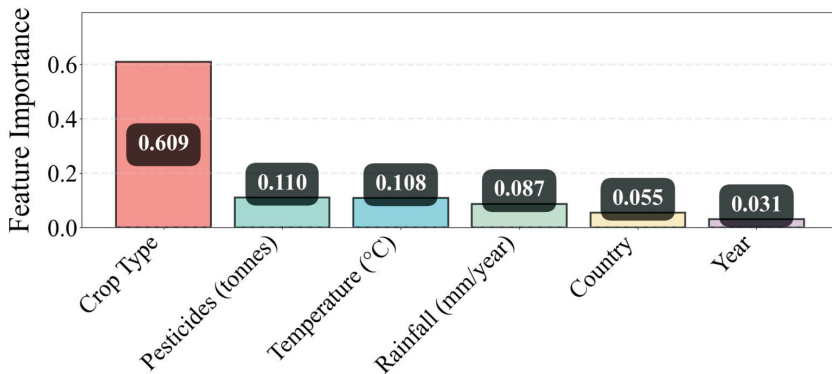


Fig. 1. Global feature importance analysis for yield prediction

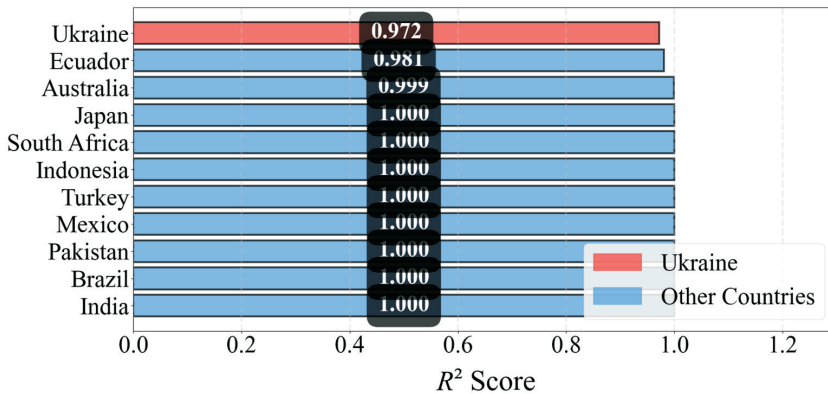


Fig. 2. Country-specific Random Forest model performance for top 10 countries and Ukraine

At the country level, localized models reached near-perfect accuracy ($R^2 \approx 1.000$) for India, Brazil, Pakistan, Mexico, and Turkey (Fig. 2). In contrast, the best model for Ukraine as the major European agrarian country was XGBoost ($R^2 = 0.980$), revealed average yields 43.4% below the global mean, underlining regional production gaps. Feature importance analysis showed an overwhelming dominance of crop type (0.904) as the primary determinant of yields in Ukraine, with only minor contributions from year (0.062), pesticides (0.022), and temperature (0.012). Rainfall and country-level context were negligible, emphasizing that yield prediction in Ukraine depends far more on biological crop characteristics than environmental variability. This suggests that crop-specific models offer greater utility for Ukrainian agriculture than generalized country-level approaches. Average yield analysis by crop confirmed potatoes as the highest-yielding crop ($> 120\,000$ hg/ha), nearly tripling the yields of staple cereals such as wheat ($\sim 28\,000$ hg/ha) and maize ($\sim 39\,000$ hg/ha). Rice (paddy) also performed

well ($\sim 42\,000$ hg/ha), while sorghum and soybeans produced significantly lower yields ($< 15\,000$ hg/ha). These disparities reflect both biological potential and management intensity, positioning potatoes and rice as highly productive crops, while cereals remain central due to dietary and market demand.

From a computational perspective, these results reinforce the value of interpretable, high-capacity machine learning models for agricultural decision-making, while also underscoring the risks of underperformance in heterogeneous regional datasets. Crop-specific analyses further identify potatoes, cassava, and sweet potatoes as globally dominant in yield potential. Grain yields remain well below the levels needed to ease food insecurity (Fig. 3). The largest fluctuations and uncertainty occur in potato and cassava yields, whereas the strongest stability is observed for soybeans, sorghum, and wheat.

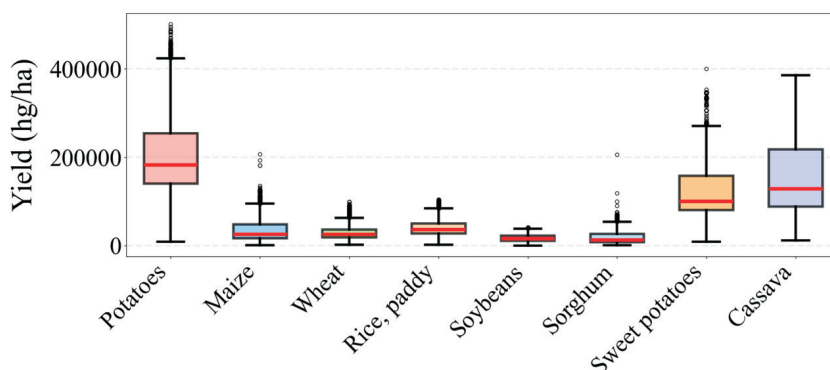


Fig. 3. Yield distribution of the major crops

The comparison of base and ensemble models for global crop-yield prediction revealed three patterns: (1) tree-based methods consistently delivered the highest accuracy; (2) ensemble models offered only slight improvements over the top single model (Extra Trees); and (3) linear and kernel methods proved unsuitable for this large-scale prediction task, likely due to their mismatch with the dataset rather than inherent algorithmic weakness. Recent studies report similar but less pronounced differences. The results show that while ensemble methods performed reliably, the Extra Trees model outperformed all combined ensembles, emphasizing the need to match algorithms to data complexity in yield prediction. Overall, the study demonstrates that robust, interpretable machine learning models can effectively support yield forecasting and sustainable agricultural planning.

Our study showed that for global crop-yield prediction using weather and pesticide data, the Extra Trees had the highest accuracy for global crop-yield prediction, followed by Stacking Ensemble, Bagged Trees, and Random Forest. Stacking was strong but slow and resource-heavy [13], while Random Forest was fastest and most efficient [14].

These findings are supported by previous research. JHAJHARIA et al. [15] reported Random Forest outperforming Support Vector Machine, Gradient Descent, Long Short-Term Memory, and Lasso regression in yield prediction for Rajasthan. Similarly, PANIGRAHI et al. [16] found it delivered the best accuracy among six regression models in predicting farm production. Another study comparing five ML models across six crop datasets showed Random Forest performed best for chickpea, while Gaussian Process led for wheat and pearl millets; K-Nearest Neighbours and AdaBoost were the weakest [17]. In our case, AdaBoost showed modest results, while KNN performed reasonably well. MAHESWARI and RAMANI [18] also found Random Forest superior to Logistic Regression, Naive Bayes, SVM, Decision Trees, KNN, and XGBoost. For rice yield prediction in India, Random Forest achieved R^2 scores of 0.85 and 0.81 for calibration and validation, respectively [19]. These studies confirm Random Forest’s consistent reliability across crops and regions, though model performance still varies by dataset and prediction task. Despite Random Forest’s efficiency and strong performance, Extra Trees was the top performer in our analysis. Therefore, Random Forest is recommended primarily for speed and low computational demand, while Extra Trees is ideal when maximum accuracy is needed. Data characteristics also influence model outcomes. Our results, based on weather and pesticide inputs, suggest that ensemble tree models generalize well; however, performance may shift when using remote sensing or agrotechnological data. For instance, JHAJHARIA and MATHUR [20] found Random Forest outperformed Decision Tree, Lasso, and SVR using remote sensing data.

Finally, while deep learning is often regarded as highly accurate, its significant computational demands raise questions about its practical value in large-scale agriculture. Future work should benchmark deep learning models against tree-based ensembles across varied datasets to assess whether increased accuracy justifies the added complexity and resource use. Besides, it should be mentioned that our study still has several limitations, such as missing soil data (soil fertility is crucial for crop yields estimation), as well as absence of data on fertilization, irrigation and tillage (these factors are among the most influential agrotechnological factors that shape crop productivity). Also, deep learning algorithms were not investigated because of computation power limitations.

Conclusions. This study confirmed the effectiveness of advanced machine learning methods – particularly tree-based models – for crop yield prediction based on agroecological and management factors. Extra Trees achieved the highest accuracy ($R^2 = 0.991$), with ensemble methods also performing well, though not consistently better than top individual models. Random Forest is recommended when computational efficiency is a priority, while linear and kernel models proved insufficient for this task. Feature importance analysis showed crop type as the strongest predictor, followed by pesticide use, temperature, and rainfall, highlighting both biological and environmental influences. Overall, tree-based and

ensemble models offer reliable, interpretable tools for large-scale yield prediction. These findings support sustainable agriculture, food security planning, and precision farming. Future research should compare deep learning against strong ensemble models and incorporate broader economic and environmental data.

REFERENCES

- [1] TILMAN D., C. BALZER, J. HILL, B. L. BEFORT (2011) Global food demand and the sustainable intensification of agriculture, *Proc. Natl. Acad. Sci. USA*, **108**(50), 20260–20264, <https://doi.org/10.1073/pnas.1116437108>.
- [2] HALDER M., A. DATTA, M. K. H. SIAM et al. (2023) A systematic review on crop yield prediction using machine learning. In: *Proc. Int. Conf. on Intelligent Systems & Networks*, Singapore, Springer, 658–667, https://doi.org/10.1007/978-981-99-4725-6_77.
- [3] VOZHEHOVA R., M. MALIARCHUK, I. BILIAIEVA et al. (2020) Spring row crops productivity prediction using normalized difference vegetation index, *J. Ecol. Eng.*, **21**(6), 176–182, <https://doi.org/10.12911/22998993/123473>.
- [4] WILLCOCK S., J. MARTÍNEZ-LÓPEZ, D. A. HOOFTMAN et al. (2018) Machine learning for ecosystem services, *Ecosyst. Serv.*, **33**, 165–174, <https://doi.org/10.1016/j.ecoser.2018.04.004>.
- [5] JAMES G., D. WITTEN, T. HASTIE, R. TIBSHIRANI (2013) *An Introduction to Statistical Learning: with Applications in R*, Vol. **103**, New York, Springer, 426 pp, <https://doi.org/10.1080/24754269.2021.1980261>.
- [6] DHALIWAL D. S., M. M. WILLIAMS (2024) Sweet corn yield prediction using machine learning models and field-level data, *Precis. Agric.*, **25**(1), 51–64, <https://doi.org/10.1007/s11119-023-10057-1>.
- [7] JABED M. A., M. A. A. MURAD (2024) Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability, *Heliyon*, **10**(24), e40836, <https://doi.org/10.1016/j.heliyon.2024.e40836>.
- [8] JEONG J. H., J. P. RESOP, N. D. MUELLER et al. (2016) Random forests for global and regional crop yield predictions, *PLoS One*, **11**(6), e0156571, <https://doi.org/10.1371/journal.pone.0156571>.
- [9] DOUCET M. S., D. A. BRIDGE, R. A. GRIMLUND, A. E. SHAMOO (1994) An application of stratified sampling techniques for research data, *Account. Res.*, **3**(4), 237–247, <https://doi.org/10.1080/08989629408573856>.
- [10] PANDEY D., K. NIWARIA, B. CHOURASIA (2019) Machine learning algorithms: a review, *Int. Res. J. Eng. Technol.*, **6**(2), 916–922, <https://doi.org/10.21275/ART20203995>.
- [11] YANG J. M., J. Y. YANG, S. LIU, G. HOOGENBOOM (2014) An evaluation of the statistical methods for testing the performance of crop models with observed data, *Agric. Syst.*, **127**, 81–89, <https://doi.org/10.1016/j.agsy.2014.01.008>.
- [12] ARCHER K. J., R. V. KIMES (2008) Empirical characterization of random forest variable importance measures, *Comput. Stat. Data Anal.*, **52**(4), 2249–2260, <https://doi.org/10.1016/j.csda.2007.08.015>.

- [13] WOLPERT D. H. (1992) Stacked generalization, *Neural Netw.*, **5**(2), 241–259, [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1).
- [14] BREIMAN L. (2001) Random forests, *Mach. Learn.*, **45**(1), 5–32, <https://doi.org/10.1023/A:1010933404324>.
- [15] JHAJHARIA K., P. MATHUR, S. JAIN, S. NIJHAWAN (2023) Crop yield prediction using machine learning and deep learning techniques, *Procedia Comput. Sci.*, **218**(1), 406–417, <https://doi.org/10.1016/j.procs.2023.01.023>.
- [16] PANIGRAHI B., K. C. R. KATHALA RAO, M. SUJATHA (2023) A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models, *Procedia Comput. Sci.*, **218**(13), 2684–2693, <https://doi.org/10.1016/j.procs.2023.01.241>.
- [17] PATIL Y., H. RAMACHANDRAN, S. SUNDARARAJAN, P. SRIDEVIPONMALAR (2025) Comparative analysis of machine learning models for crop yield prediction across multiple crop types, *SN Comput. Sci.*, **6**(1), 64, <https://doi.org/10.1007/s42979-024-03602-w>.
- [18] MAHESWARI M. U., R. RAMANI (2023) A comparative study of agricultural crop yield prediction using machine learning techniques. In: *Proc. 9th Int. Conf. on Advanced Computing and Communication Systems (ICACCS)*, Vol. **1**, IEEE, 1428–1433, <https://doi.org/10.1109/ICACCS57279.2023.10112854>.
- [19] SATPATHI A., P. SETIYA, B. DAS et al. (2023) Comparative analysis of statistical and machine learning techniques for rice yield forecasting for Chhattisgarh, India, *Sustainability*, **15**(3), 2786, <https://doi.org/10.3390/su15032786>.
- [20] JHAJHARIA K., P. MATHUR (2023) Prediction of crop yield using satellite vegetation indices combined with machine learning approaches, *Adv. Space Res.*, **72**(9), 3998–4007, <https://doi.org/10.1016/j.asr.2023.07.006>.

¹*Department of Irrigated Agriculture and Decarbonization of Agroecosystems,
Institute of Climate-Smart Agriculture of NAAS,
vil. Khlibodarske, Odessa, 67667, Ukraine
e-mails: pavel.likhovid@gmail.com, G_Ludmila15@ukr.net,
maksimovsinger@gmail.com*

²*Department of Agriculture, Kherson State Agrarian and Economic University,
23 Stritenska St, Kherson, Ukraine
e-mails: averchev1966@gmail.com, docent6977@gmail.com*

³*Department of Manufacturing and Processing Agricultural Production
Technologies Named after Academician V. G. Pelykh,
Kherson State Agrarian and Economic University,
23 Stritenska St, Kherson, Ukraine
e-mail: karashchuk_g@ksaeu.kherson.ua*