PLANT SCIENCE TODAY
Vol x(x): xx–xx
https://doi.org/10.14719/pst.11454

eISSN 2348-1900

RESEARCH ARTICLE

Prediction of agricultural crop yields based on spatial vegetation indices and machine learning

P Lykhovyd[1*], L Hranovska[1], O Averchev[2], A Tomnytskyi[1], O Rudik[3] & D Maksymov[1]

[1]Department of Irrigated Agriculture and Decarbonization of Agroecosystems, Institute of Climate-Smart Agriculture, Odessa 67667, Ukraine
[2]Department of Agriculture, Kherson State Agrarian and Economic University, Kherson 73020, Ukraine
[3]Department of Climate-Oriented Agricultural Technologies, Institute of Climate-Smart Agriculture, Odessa 67667, Ukraine

*Correspondence email - pavel.likhovid@gmail.com

Received: 25 August 2025; Accepted: 01 November 2025; Available online: Version 1.0: xx xxxxxx 2025

Cite this article: Lykhovyd P, Hranovska L, Averchev O, Tomnytskyi A, Rudik O, Maksymov D. Prediction of agricultural crop yields based on spatial vegetation indices and machine learning. Plant Science Today (Early Access). https://doi.org/10.14719/pst.11454

Abstract

Accurate prediction of crop yields is not only a scientific challenge but also an economic necessity, as it directly influences food security, market stability and efficient resource allocation in agriculture. This study is driven by the hypothesis that the integration of satellite-based vegetation data with machine learning (ML) can substantially improve yield forecasting accuracy under semiarid climatic conditions, thereby reducing financial risks for farmers and agribusinesses. To test this, we developed and compared multiple data-driven prediction models for three key crops – peas, rapeseed and wheat – representing major contributors to regional agricultural income. We used freely available satellite imagery from the Sentinel-2 mission to calculate several vegetation indices that describe crop greenness, canopy structure and water content. These indices were analyzed to determine which combination best captures the relationship between crop condition and final yield. To ensure reliability, we expanded the dataset with controlled random noise and assessed model stability. Nine ML approaches were compared and the gradient boosting algorithm consistently delivered the most accurate results, achieving up to 99 % agreement with observed yields and fewer than 5 % average errors. The most informative vegetation indices differed among crops, revealing new interdisciplinary insights into how crop physiology and environmental stress interact with spectral indicators. The breakthrough of this research lies in demonstrating a crop-specific optimization strategy that connects remote sensing, agronomy and data science in a single predictive framework. This approach can be immediately applied to improve yield estimation systems at regional and national scales, potentially reducing forecasting uncertainty by 20–30 % and saving agricultural producers millions of euros annually through optimized input management and market planning. Future research should focus on integrating weather forecasts, soil moisture data and economic models to transform yield prediction into a comprehensive decision-support system for precision agriculture. These findings, therefore, provide a practical pathway toward data-driven, climate-resilient and economically sustainable crop production worldwide.

Keywords: modified chlorophyll absorption in reflectance index; modified soil-adjusted vegetation index; normalized difference vegetation index; peas; rapeseed; red-edge chlorophyll index

Introduction

Developing innovative methods for the accurate and timely prediction of agricultural crop yields is a crucial task in modern agricultural science. Yield predictions are used not only for scientific purposes but also to make rational operational decisions regarding agrotechnological adjustments, to consider changes in agrarian policy to enhance food security and to revise agricultural zones suitable for producing certain crops (1). Therefore, the task of yield modeling and prediction is of great importance and value, especially today, when military activities in Ukraine and the Middle East, climate change and the depletion of natural resources, particularly freshwater, have significantly aggravated the food crisis (2, 3).

There are different approaches to yield prediction in agricultural science. Based on the data sources that form their foundation, it is possible to classify all the approaches into the following groups:

Meteorologically and climatically based models

These models rely predominantly on weather and climate data as their main predictors of crop yield. They operate on the fundamental premise that weather factors (e.g., temperature, precipitation, solar radiation) are the most significant determinants of crop growth and final yield. For example, models predicting sweet corn yield based on the availability of life factors (heat, light and water) or soybean yield models based on the selected climatic variables (4, 5).

Models based on the results of empirical field trials

These models are built upon the outcomes of field trials and experimental data, focusing on the influence of human-controlled management practices and agrotechnological interventions. They are particularly useful for making management recommendations and optimizing inputs, but their usability is usually limited to the ecological conditions of the trial zones. It is not recommended to extrapolate the outcomes of agricultural experiments conducted in arid climates to humid climate zones. These models are among the most widespread in the scientific literature of the 20th century. Some recent examples include the model of sweet corn yields depending on fertilization, plant density and tillage and the model of grain sorghum yield depending on plant density and hybrids (6, 7).

Soil and site-specific models

This approach is comparatively rare in its pure form. Such models focus on the influence of soil properties and other static site-specific factors that affect crop productivity. They are crucial for precision agriculture applications and for understanding the spatial variability of yield within a field. For example, soil-based models of wheat, corn and cotton yields in Spain, which determine 56 to 84 % of the crops' productivity features, could be proposed (8).

Remote sensing-based models

These models use data acquired from satellites, drones or other aerial platforms to monitor crop health and development throughout the growing season. The data, often in the form of specific vegetation indices (VIs) – quantitative indicators of vegetation vigor and biomass – calculated from reflectance values in different spectral bands (such as red, near-infrared or short-wave infrared), provide a non-invasive way to assess crop conditions over large areas. Nowadays, this approach to yield prediction is one of the most interesting and promising ones due to its great scalability, low costs and operational and dynamic nature. There is a great variety of remote sensing-based crop model. Based on a recent academic review, such models are mainly based on data from the moderate resolution imaging spectroradiometer (MODIS), Sentinel-1 and Sentinel-2 and Landsat satellites, employ various configurations of deep learning (DL) and other ML algorithms and are mostly developed to predict the yields of major staple crops such as wheat, corn, soybean, barley, rice and rapeseed (9). Additionally, remote sensing data are most frequently used in hybrid (combined or integrated) crop models.

Hybrid models

This group represents the most advanced approach, combining data from multiple sources to create a more robust and accurate prediction system. They overcome the limitations of single-source models by leveraging the complementary nature of different data types. Such models usually use a combination of experimental data and remote sensing data, or meteorological data and remote sensing data, or meteorological and soil data. These models are generally more complicated than single-source models and if designed properly, they are the most accurate ones. However, in some cases, the pitfalls of overfitting (when a model learns noise rather than true relationships) and using excessive data with inappropriate modeling techniques have led to inferior performance of these models. For example, a hybrid model utilizing remote sensing (normalized difference vegetation index, NDVI) and meteorological data (air temperature and rainfall amounts) was used to predict grain corn yields in China (10).

In addition to the data sources, crop models can be classified by the algorithms used to create the model and generate predictions. From this perspective, all models are divided into four major groups:

Process-based (mechanistic) models

These models are based on the underlying physiological and biophysical processes of crop growth and development. They simulate the interactions between the plant and its environment by using a set of differential equations and parameters that describe a wide range of processes, from photosynthesis and respiration to nutrient uptake and water balance. Such models are usually used in decision support systems for scientists and agricultural practitioners like decision support system for agrotechnology transfer (DSSAT), agricultural production systems simulator (APSIM) and world food studies model (WOFOST) (11).

Statistical (empirical) models

These models use statistical relationships between historical crop yield data and various predictor variables (e.g., weather, soil or management data) to build a predictive model. They do not necessarily account for the biophysical processes of crop growth. This type of model is more popular in scientific community, embracing such approaches as regression analysis and time series forecasting algorithms (12).

Machine learning models

This group of models uses computational algorithms capable of learning complex, non-linear relationships between a vast number of input variables and crop yield. They are particularly well-suited for handling large datasets, including remote sensing data and gridded climate data. However, it must be noted that this group could be merged with empirical models, as regression and time series analysis are frequently considered to be part of ML in a broader sense. The most popular approaches in modern agricultural science include support vector machines (SVM), random forests (RF), gradient boosting (XGBoost) and artificial neural networks (ANN). Usually, these complex models perform better compared to simple regression, but they require significantly larger datasets, more computational power and expertise. Additionally, although the outcomes are great, it is almost impossible to know how the machine achieved the prediction result – a phenomenon known as the “black box” problem (13).

Hybrid models

These models combine elements from two or more of the above groups to leverage the strengths of each approach. The goal is to improve predictive accuracy, enhance interpretability and extend the applicability of the model. As an example, there are models combining regression analysis and artificial neural networks for predicting bean yields based on remote sensing data or models for predicting wheat yields using a combination of biophysical and ML models (14, 15).

In this study, all data processing and model training were implemented using widely available open-source software tools and libraries, including Python (version 3.13) and packages such as scikit-learn for ML and XGBoost for gradient boosting. Sentinel-2 satellite imagery was accessed through the OneSoil online platform, which provides preprocessed vegetation indices at 10 m spatial resolution. These clarifications are provided to ensure reproducibility and accessibility of the research for readers from various disciplines.

Integrating remote sensing with ML for crop yield prediction directly contributes to economic sustainability by linking technological innovation to profitability and risk reduction. Accurate forecasts enable farmers, cooperatives and policymakers to optimize inputs such as water and fertilizers, reduce unnecessary expenditures and improve timing of field operations, which collectively lower production costs and increase returns per hectare. Moreover, reliable yield prediction enhances market planning and price stability, reducing financial uncertainty for both producers and supply chains. In this way, economic viability becomes an integral outcome of precision agriculture innovations, ensuring that environmental improvements translate into measurable profitability and long-term sustainability (16).

In addition, improved crop forecasting has far-reaching benefits across multiple dimensions of the agricultural sector. At the farm level, accurate yield predictions allow producers to plan input use, labor and irrigation schedules more efficiently, thereby minimizing waste and reducing production costs as mentioned above. For agribusinesses and cooperatives, early and reliable forecasts support logistics planning, storage management and contract pricing, improving overall supply chain stability. At the policy level, accurate yield information enables governments to anticipate food supply fluctuations, manage trade balances and design timely support or intervention measures to ensure food security. The economic value of better forecasting is substantial: even a modest 5–10 % improvement in yield prediction accuracy can translate into millions of dollars in annual savings through optimized input management and reduced post-harvest losses. Furthermore, by reducing uncertainty, farmers can make more informed market decisions and secure better insurance and credit conditions, directly enhancing the profitability and financial resilience of agricultural enterprises. Thus, accurate crop forecasting is not merely a technical advancement – it is a cornerstone of economically viable and risk-resilient agriculture (17).

Considering the importance and relevance of developing robust, accurate and efficient yield prediction models, the main goal of our study was to establish the most suitable remotely sensed vegetation indices for predicting the yields of wheat, rapeseed and peas in the semi-arid climate of southern Ukraine (Bsk zone according to the Köppen climate classification, where "Bsk" indicates a cold semi-arid steppe climate) (18). The most accurate ML technique to be recommended for yield prediction based on the vegetation indices.

To address the needs of modern agribusiness, this study responds to the growing industrial demand for decision-support tools that can convert satellite-derived vegetation data into actionable forecasts of crop productivity. The agricultural sector increasingly relies on digital and data-driven solutions to maintain profitability while reducing environmental impact. Therefore, understanding how different vegetation indices can be used in combination with ML methods to predict crop yields with high accuracy is not only of academic value but also of direct operational importance for farmers, cooperatives and agrotechnology developers.

Research hypothesis

It is hypothesized that integrating remotely sensed vegetation indices with ML algorithms enables accurate, cost-effective and scalable prediction of crop yields, thereby simultaneously supporting environmental sustainability through optimized input use and enhancing economic efficiency through improved decision-making. The urgency of this investigation lies in bridging the gap between environmental and economic objectives of agriculture – proving that advanced digital monitoring and predictive analytics can deliver measurable profitability while promoting sustainable land and water management.

Materials and Methods

The study was conducted on representative agricultural fields cultivated with three major crops – wheat, rapeseed and peas – under typical temperate and semi-arid farming conditions. To ensure the methodology is universally reproducible, the workflow was designed to be independent of any specific location, season or climatic zone. All procedures described below can be repeated for any region or year using equivalent data sources and software tools.

Crop yield data were obtained from production fields or experimental plots, with each plot representing a homogeneous management unit. After harvesting using a self-propelled combine or plot harvester, grain samples were weighed and adjusted to a standardized reference moisture content (14 % for wheat and peas; 8 % for rapeseed) to ensure comparability across studies. Each yield observation was georeferenced and linked to corresponding coordinates on satellite imagery to enable spatial analysis.

Vegetation indices were extracted from Sentinel-2 satellite imagery (10 m spatial resolution, level-2A surface reflectance data) using the OneSoil or any equivalent precision agriculture platform. The peak (maximum seasonal) values of vegetation indices for each crop were identified and used as predictors for yield modeling. The following indices were analyzed due to their proven relevance to crop health and productivity:

Normalized difference vegetation index (NDVI)

Quantifies vegetation vigor and biomass by contrasting near-infrared (NIR) and red light reflectance; widely used for growth monitoring and yield estimation (19).

Normalized difference moisture index (NDMI)

Evaluates canopy water status based on NIR and shortwave infrared (SWIR) reflectance; useful for drought and irrigation management (20).

Normalized difference red-edge index (NDRE)

Sensitive to chlorophyll concentration and early plant stress; supports nutrient deficiency detection (21).

Modified soil-adjusted vegetation index (MSAVI)

Reduces soil background influence in areas with partial canopy cover, enhancing vegetation signal accuracy (22).

Red-edge chlorophyll index (RECI)

Estimates chlorophyll content and plant photosynthetic capacity; useful for fertilizer optimization (23).

Photochemical reflectance index (PRI)

Reflects light-use efficiency and photosynthetic activity; applied to monitor stress and water management efficiency (24).

Modified chlorophyll absorption ratio index (MCARI)

Estimates chlorophyll concentration and detects plant stress or nutrient imbalance (25).

After linking each yield record to corresponding vegetation indices, the dataset for each crop was augmented using a Gaussian noise algorithm to increase sample diversity and prevent model overfitting. The augmentation added random noise (mean = 0, standard deviation = 0.05 of each feature range) to each index value while maintaining realistic biophysical ranges.

To determine which vegetation indices most strongly influence crop yield, a Pearson correlation matrix was computed for each crop. Variables with high pairwise correlation coefficients ($r > 0.85$) were excluded to minimize multicollinearity and improve model interpretability. The remaining indices were used as input variables for ML analysis.

Each dataset was then randomly divided into training (80 %) and testing (20 %) subsets to allow unbiased model validation. All data preprocessing steps - normalization, random shuffling and partitioning - were performed using standardized routines from the Scikit-learn library (Python). Nine ML algorithms were implemented to model the relationship between vegetation indices and crop yields:

Multiple linear regression (MLR)

Fits a linear equation between dependent and independent variables; assumes linearity and is useful for baseline comparison (26).

Random forest regression (RFR)

Ensemble of decision trees providing robust performance and resistance to overfitting (27).

Support vector regression (SVR)

Uses kernel-based optimization to capture nonlinear dependencies (28).

Extreme gradient boosting (XGBoost)

Gradient boosting algorithm constructing sequential tree ensembles with high predictive accuracy (29).

K-nearest neighbors regression (KNR)

Predicts yield based on the mean of the *k* most similar samples in the feature space (30).

Lasso regression (LR)

Applies L1 regularization for variable selection and sparsity enhancement (31).

Multi layer perceptron (MLP)

Artificial neural network trained via backpropagation; captures complex nonlinear patterns in vegetation-yield relationships (32).

Ridge regression (RR)

Applies L2 regularization to reduce overfitting and stabilize coefficients (33).

Elastic net (EN)

Combines L1 and L2 regularization to handle multicollinearity and improve generalization (34).

Model performance was evaluated using four statistical metrics widely used in regression analysis:

- R^2 (coefficient of determination) – Measures model explanatory power.
- Root mean square error (RMSE) – Quantifies overall prediction error magnitude.
- Mean absolute error (MAE) – Indicates average prediction deviation.
- Mean absolute percentage error (MAPE) – Expresses model error as a relative percentage of observed values (35).

All analyses were conducted in Python version 3.13 using the libraries Pandas, Numpy, Scikit-learn and Matplotlib. Code was executed in Visual Studio Code IDE. All functions, parameters and random seeds were explicitly defined to ensure complete reproducibility. The full pipeline – from data preparation to model validation - can be replicated on any machine with the same software environment and access to equivalent satellite imagery and yield dataset.

Results

Correlation matrix and multicollinearity analysis

Correlation matrices for the yields of the studied crops and corresponding peak values of the vegetation indices are provided in Fig. 1. High positive or negative correlations point out which indices are most strongly associated with yield. These relationships were used to identify the most influential vegetation indices for each crop, forming the basis for subsequent ML modeling and supporting the research hypothesis that specific remote-sensing indicators can accurately predict crop yield. It was established that the strength of the relationship between yield and vegetation indices varies significantly depending on the crop type. For example, pea yield showed a strong correlation with NDVI ($R = 0.95$) and MSAVI ($R = 0.95$), while the weakest connection was recorded for PRI ($R = 0.54$). For rapeseed and wheat, the correlations were generally weaker. The strongest connections were observed for rapeseed with RECI ($R = 0.75$) and NDRE ($R = 0.72$) and for wheat with MCARI ($R = 0.60$) and NDVI ($R = 0.58$).

Apart from detecting the strongest pairwise correlations, multicollinearity was assessed to ensure that the models utilized independent and informative indices. This is crucial for industrial applications, because using redundant indices could complicate operational decision-making without improving predictive accuracy. Based on this analysis, the indices selected for modeling were NDVI and MCARI for peas, MSAVI and RECI for rapeseed and NDVI and MCARI for wheat.

From an industrial and managerial perspective, this analysis indicates which remotely sensed indices provide the most actionable information for yield forecasting, enabling optimized resource allocation and operational planning.

Crop models evaluation

Crop modeling revealed the following regularities in yield prediction (Table 1):

- Lasso and ElasticNet regression models consistently performed poorly for peas and wheat, indicating that these simpler models are not suitable for practical yield forecasting. For rapeseed, Lasso regression produced the least accurate results.
- XGBoost consistently achieved the highest accuracy for all crops. Although for rapeseed the advantage over RFR was smaller, XGBoost shows substantially lower MAE and MAPE values, making it the most reliable model for operational use.
- Prediction accuracy was highest for peas and lowest for wheat, but all XGBoost models had R^2 values within the good to very good range and MAPE < 5 %, demonstrating industrially relevant precision for yield planning and financial forecasting.
- Visual evaluations are presented for peas, rapeseed and wheat

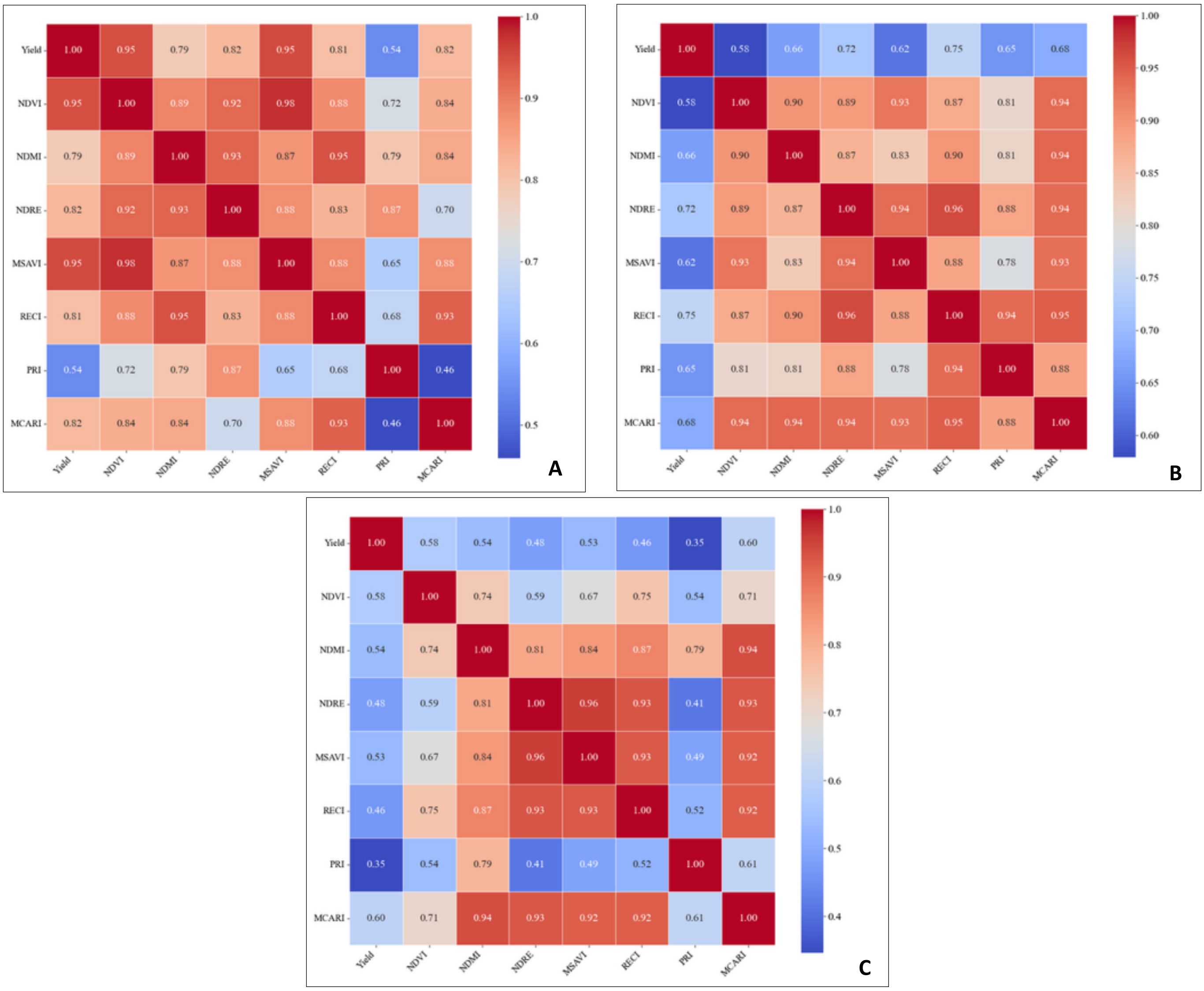

Fig. 1. Correlation matrices between crop yields and vegetation indices for peas (**A**), rapeseed (**B**) and wheat (**C**). Each matrix displays Pearson correlation coefficients between yield and seven vegetation indices NDVI, NDMI, NDRE, MSAVI, RECI, PRI and MCARI.

Table 1. Evaluation metrics for different ML approaches to the prediction of peas, rapeseed and wheat yields based on remote sensing data, explaining the strength of the relationship between vegetation indices and crop yields (R^2 values) and the accuracy of yield prediction for each crop, expressed in absolute (RMSE and MAE) and relative (MAPE) errors

Crop	Model	R^2	RMSE, t ha^{-1}	MAE, t ha^{-1}	MAPE, %
Peas	MLR	0.9075	0.1034	0.0842	4.5453
	RFR	0.9832	0.0441	0.0313	1.6117
	SVR	0.9290	0.0906	0.0729	3.7372
	XGB	0.9867	0.0393	0.0221	1.0906
	KNR	0.9741	0.0547	0.0421	2.2256
	LR	-0.0007	0.3401	0.2793	16.1916
	MLP	0.6184	0.2100	0.1703	9.8905
	RR	0.8670	0.1240	0.0977	5.3352
	EN	-0.0007	0.3401	0.2793	16.1916
Rapeseed	MLR	0.7425	0.4579	0.3731	29.2778
	RFR	0.9682	0.1608	0.0921	6.5136
	SVR	0.8547	0.3440	0.2250	17.2404
	XGB	0.9676	0.1624	0.0405	2.0178
	KNR	0.9198	0.2556	0.1880	13.6223
	LR	-0.0178	0.9104	0.8214	66.0205
	MLP	0.6711	0.5175	0.4744	36.8630
	RR	0.7311	0.4679	0.3982	29.9851
	EN	0.2036	0.8053	0.7405	59.1609
Wheat	MLR	0.3041	0.2973	0.2623	10.5797
	RFR	0.7607	0.1743	0.1040	4.0779
	SVR	0.4602	0.2618	0.2188	8.7408
	XGB	0.7797	0.1673	0.0840	3.2692
	KNR	0.6846	0.2001	0.1245	4.8555
	LR	-0.0046	0.3572	0.2714	11.3399
	MLP	0.2922	0.2998	0.2644	10.7675
	RR	0.3003	0.2981	0.2602	10.5553
	EN	-0.0046	0.3572	0.2714	11.3399

(Fig. 2-4). These plots illustrate predicted vs. actual yields and residuals, allowing managers to quickly assess the reliability of each model. Models with points tightly clustered along the ideal fit line (e.g., XGBoost) are the most trustworthy for decision-making. This visualization highlights which models provide the most precise yield predictions, directly supporting the research hypothesis that integrating specific vegetation indices with optimal ML techniques enables accurate and economically meaningful crop yield forecasting.

From a managerial and economic standpoint, the results highlight XGBoost as the most effective and actionable model for crop yield prediction. Using the identified vegetation indices, it enables reliable production estimation, optimized input use, reduced operational costs and improved financial planning for farms and agribusinesses.

Discussion

Machine learning models leveraging remote sensing vegetation indices are widely used for crop yield prediction. Ensemble methods (like random forest), advanced DL models and hybrid approaches consistently outperform simpler models, especially when integrating multiple indices.

The most frequently used vegetation indices include NDVI, soil-adjusted vegetation index (SAVI), enhanced vegetation index (EVI), vegetation condition index (VCI), normalized difference water index (NDWI) and some other modifications of these common indicators like green NDVI (GNDVI). Our study focused both on commonly used indices like NDVI and SAVI, but also employed some less studied indices like PRI, NDRE and MCARI to fill the gap in theoretical knowledge about the relationship of major crop yields with these parameters (36–38).

As for ML models, the top five are represented by RFR with typical R^2 values of 0.71-0.87 depending on vegetation indices and crops; SVR with slightly lower typical R^2 values of 0.66-0.77, mainly performing best for small and medium-sized datasets; and MLR, LR and RR models, which usually have R^2 values within 0.73-0.77 and are very popular because of their simplicity and low demands for computational power and dataset preparation (39–41). The highlighted results correspond to the outcomes of our study, where XGBoost and RFR were the best ML algorithms for yield predictions.

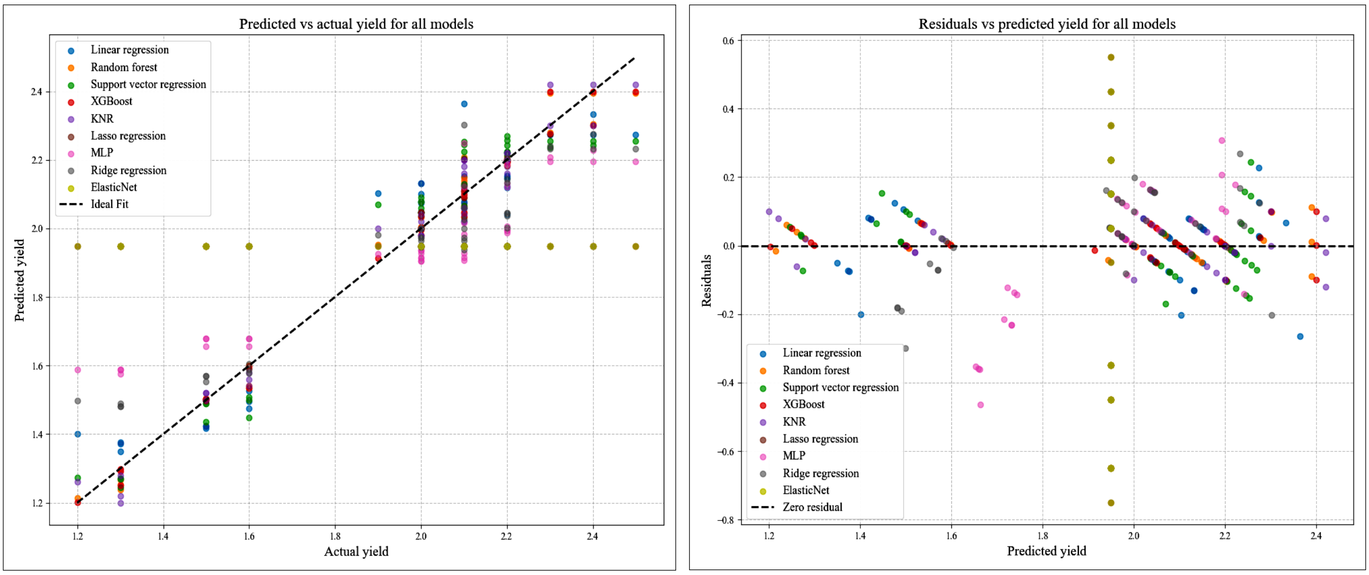

Fig. 2. Performance comparison of different ML models in predicting pea yields based on vegetation indices. The bar chart displays the discrepancy between the actual and predicted yields of the crop for nine ML algorithms, illustrating their relative accuracy and reliability.

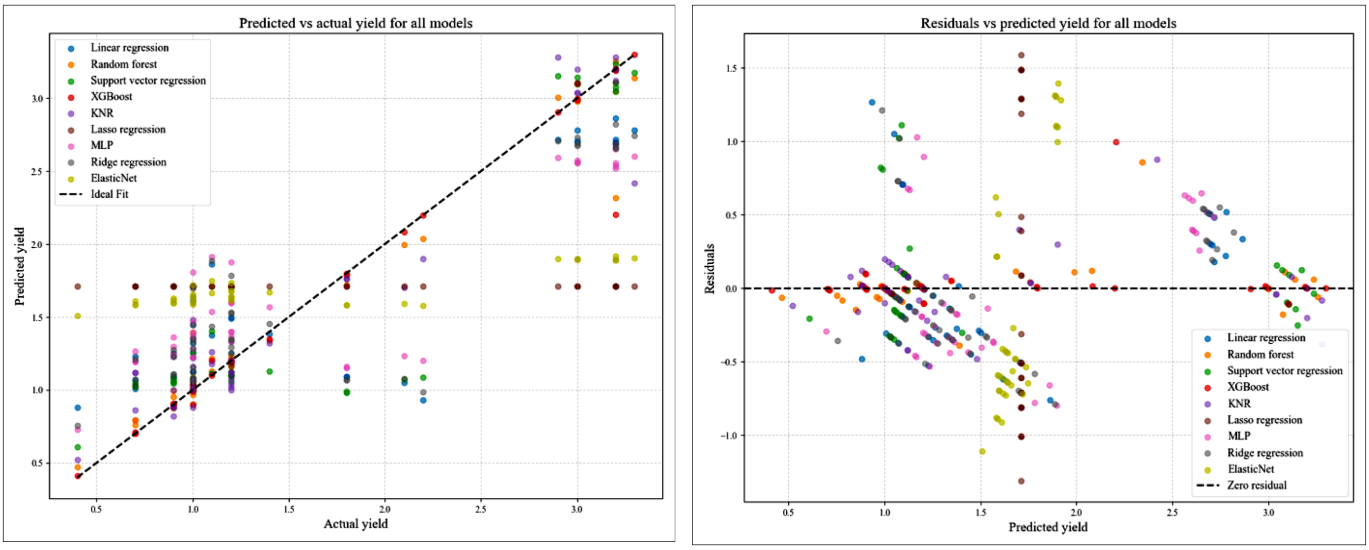

Fig. 3. Performance comparison of different ML models in predicting rapeseed yields based on vegetation indices. The bar chart displays the discrepancy between the actual and predicted yields of the crop for nine ML algorithms, illustrating their relative accuracy and reliability.

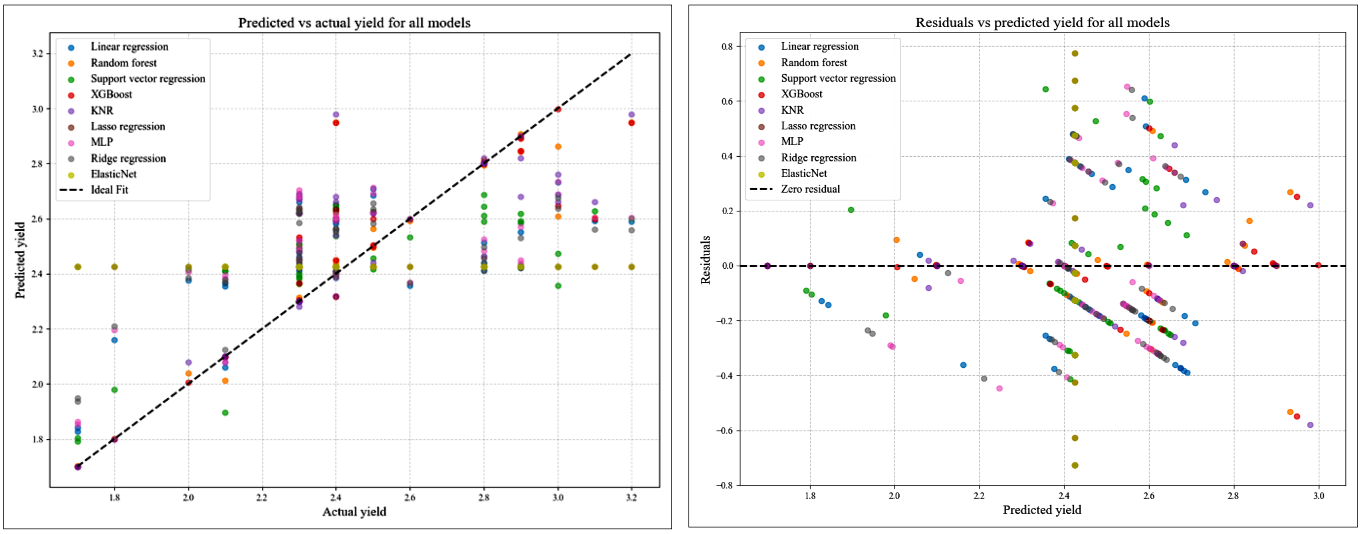

Fig. 4. Performance comparison of different ML models in predicting wheat yields based on vegetation indices. The bar chart displays the discrepancy between the actual and predicted yields of the crop for nine ML algorithms, illustrating their relative accuracy and reliability.

The best performance is recorded for DL models, such as deep neural networks (DNN) and convolutional neural networks (CNN). Such models have R^2 values greater than 0.85, are good at handling large datasets with missing data and provide the best performance if accuracy and reliability are the highest priorities. However, their “black box” nature and high computational demands limit their use and implementation, especially in embedded systems, where advanced techniques for making CNNs more lightweight are required prior to their implementation (42–44). In some cases, DL models are not superior to simpler regression algorithms, for example, when there are no non-linear relationships between the studied parameters (45).

In general, ensemble methods (especially random forest) and deep learning models provide the most accurate and robust crop yield predictions when using remote sensing vegetation indices, particularly when multiple indices and phenological features are combined. Simpler models are less effective for complex, nonlinear relationships. Integrating diverse vegetation indices and advanced feature selection further enhances prediction accuracy.

Comparing the outcomes of our study with recent research, we found strong agreement regarding the superior performance of the XGBoost ML algorithm in processing remote sensing-derived vegetation indices to predict crop yields. For wheat, XGBoost consistently ranks among the top-performing models, often outperforming or matching random forest and other algorithms in both accuracy and computational efficiency. In previous studies, R^2 values for wheat yield prediction using XGBoost reached 0.86, with NDVI and EVI identified as the most informative indices (46). While XGBoost has been extensively studied for wheat, there is a notable lack of research applying it to peas and rapeseed using satellite-derived vegetation indices. For peas, some studies have successfully implemented NDVI and MSAVI within an artificial neural network framework, achieving R^2 values up to 0.88 – slightly lower than the accuracy obtained in our study with XGBoost (47). For rapeseed, research on ML-driven yield prediction using vegetation indices is particularly limited. Random forest has been reported as the best-performing model in previous studies, with R^2 values up to 0.65 (48). In our work, random forest regression also performed strongly, but the coefficient of determination reached 0.97. This substantial difference in accuracy can be attributed to the scale of prediction: our model provides local, field-level forecasts, whereas the referenced study targets regional-scale yield predictions.

As our study has certain limitations, such as a short time span, a limited spectrum of the studied crops and absence of deep learning models, further research should be conducted in this direction to fill existing gaps and provide a scientifically sound and robust substantiation for combining ML models with remote sensing data for precise prediction of yields in major crops.

Recent studies confirm that XGBoost, which is the best-performing model in our research, often surpasses or matches the predictive accuracy of other leading ML approaches, including random forest, support vector regression, artificial neural networks and even advanced deep learning architectures, when applied to crop yield prediction from remote sensing data. The algorithm’s major advantages lie in its ability to efficiently process high-dimensional, non-linear and multi-source datasets, which are typical in remote sensing applications. XGBoost combines several key benefits, including high predictive accuracy, fast computation, moderate hardware requirements, robustness against overfitting and relatively good interpretability. Reported coefficients of determination (R^2) for XGBoost-based yield prediction models generally range between 0.82 and 0.96 (49, 50). In comparison, R^2 values for other commonly used algorithms are typically 0.80–0.93 for random forest (50, 51), 0.80–0.99 for categorical boosting and up to 0.80 for various linear regression algorithms (52, 53). Therefore, XGBoost remains one of the most reliable and efficient tools for crop yield prediction using remote sensing data, combining strong performance with computational efficiency and practical interpretability - qualities that make it highly suitable for both scientific research and operational agricultural applications.

In a broader industrial context, modern computing methods such as digital twin and immersive 3D simulation technologies offer the potential to extend ML-based crop yield prediction into fully virtualized agro-industrial systems. By integrating real-time remote sensing data with virtual farm models, digital twins can simulate multiple scenarios, optimize resource allocation and predict both agronomic and economic outcomes before actual implementation. This approach aligns with recent advances in industrial metaverse technologies, where interconnected cyber-physical systems, autonomous data-driven algorithms and extended reality tools enable techno-economic forecasting, operational optimization and value co-creation across complex production environments (54, 55).

Beyond the scope of direct crop yield prediction, recent advances in artificial intelligence (AI) and digital twin technologies demonstrate how ML models can be integrated into broader techno-economic simulations. Generative artificial intelligence of things (AIoT) systems, multisensory extended reality environments and big data simulation tools now enable the creation of virtual digital twins capable of estimating the economic impacts of technological and agronomic innovations within specific regional contexts (56). These developments highlight a growing convergence between predictive modeling, industrial metaverse frameworks and sustainability analytics. Furthermore, as illustrated by recent studies on corporate social responsibility performance across economic sectors, the integration of AI-driven predictive models with economic indicators provides valuable insights for enhancing the environmental and financial sustainability of agricultural production (57). In this regard, the presented ML-based approach to crop yield prediction may serve as a foundational element of future digital twin systems designed to simulate, optimize and evaluate both the environmental and economic dimensions of agricultural decision-making.

The main driving mechanisms behind the superior performance of the XGBoost and random forest algorithms in our study lie in their ability to capture complex, non-linear interactions between vegetation indices and crop yield, while maintaining stability under multicollinearity and data heterogeneity. In particular, indices such as NDRE and MCARI, which are less frequently applied in yield modeling, showed strong explanatory power for leguminous crops like peas, suggesting a higher sensitivity of these indices to canopy nitrogen content and photosynthetic efficiency. The field-level focus of this research also contributed to the higher R^2 values obtained, as it allowed the models to detect subtle spatial variations in vegetation structure and moisture that are often smoothed out in regional-scale analyses. This implies that the effectiveness of ML algorithms in yield prediction depends not only on model architecture but also on the spatial resolution and spectral diversity of the input data. These findings extend current knowledge by emphasizing that the integration of multiple, physiologically grounded vegetation indices can reveal crop-specific sensitivities and improve model generalizability, ultimately supporting the development of adaptive, data-driven tools for precision agriculture and resource-efficient management.

Conclusion

This study confirms that remote sensing-derived vegetation indices can be effectively used to predict yields of peas, rapeseed and wheat. The research hypothesis – that ML models based on vegetation indices can accurately predict crop yields, with XGBoost outperforming other algorithms – was strongly supported by the results. Among all tested models, XGBoost demonstrated the best performance, achieving the lowest mean absolute percentage error (MAPE < 5 %) and the highest coefficients of determination ($R^2 = 0.78$ –0.99) across all studied crops. The combination of NDVI and MCARI proved optimal for peas and wheat yield prediction, while MSAVI and RECI were most suitable for rapeseed. Conversely, Lasso regression and ElasticNet showed the weakest performance, indicating their limited suitability for yield forecasting applications. These findings underscore the high potential of ensemble-based ML models for integrating satellite-derived vegetation indices into precise, scalable and cost-efficient yield forecasting systems, supporting both scientific and industrial applications in data-driven, climate-resilient agriculture.

Authors' contributions

PL performed statistical analysis and graphic work. LH performed literature review and analysis. OA performed manuscript writing. AT performed critical revision and final approval of the manuscript. OR performed manuscript formatting. DM performed data collection. All authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest: Authors do not have any conflict of interest to declare.

Ethical issues: None

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Gemini to format the reference list. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Sharifi A. Yield prediction with machine learning algorithms and satellite images. J Sci Food Agric. 2021;101(3):891-96. https://doi.org/10.1002/jsfa.10696
2. Ashlyn SA. A food insecurity labyrinth: Unveiling the causes and responses to Pakistan's food crisis (2022-present). J Contemp Politics. 2024;3(1):32. https://doi.org/10.53989/jcp.v3i1.10
3. El Bilali H, Ben Hassen T. Disrupted harvests: How Ukraine-Russia war influences global food systems-A systematic review. Policy Stud. 2024;45(3-4):310-35. https://doi.org/10.1080/01442872.2024.2329587
4. Lykhovyd P. A life factor approach to the yield prediction: A comparison with a technological approach in reliability and accuracy. J Ecol Eng. 2019;20(6):177-83. https://doi.org/10.12911/22998993/108630
5. Torsoni GB, de Oliveira Aparecido LE, Dos Santos GM, Chiquitto AG, da Silva Cabral Moraes JR, de Souza Rolim G. Soybean yield prediction by machine learning and climate. Theor Appl Climatol. 2023;151:1709. https://doi.org/10.1007/s00704-022-04341-9
6. Lykhovyd PV. Prediction of sweet corn yield depending on cultivation technology parameters by using linear regression and artificial neural network methods. Biosyst Divers. 2018;26(1):11-15. https://doi.org/10.15421/011802
7. Boiko MO. Implementation of non-linear neural networks for grain sorghum yields modelling in the conditions of Southern Steppe of Ukraine. Bull Dnipro State Agrar Econ Univ. 2016;(2):118-23.
8. De la Rosa D, Cardona F, Almorza J. Crop yield predictions based on properties of soils in Sevilla, Spain. Geoderma. 1981;25(3-4):267-74. https://doi.org/10.1016/0016-7061(81)90040-9
9. Joshi A, Pradhan B, Gite S, Chakraborty S. Remote-sensing data and deep-learning techniques in crop mapping and yield prediction: A systematic review. Remote Sens. 2023;15(8):2014. https://doi.org/10.3390/rs15082014
10. Zhu X, Guo R, Liu T, Xu K. Crop yield prediction based on agrometeorological indexes and remote sensing data. Remote Sens. 2021;13(10):2016. https://doi.org/10.3390/rs13102016

11. Gavasso-Rita YL, Papalexiou SM, Li Y, Elshorbagy A, Li Z, Schuster-Wallace C. Crop models and their use in assessing crop production and food security: A review. Food Energy Secur. 2024;13(1):e503. https://doi.org/10.1002/fes3.503

12. Hoshmand R. Statistical methods for environmental and agricultural sciences. CRC Press; 2017.

13. Araujo SO, Peres RS, Ramalho JC, Lidon F, Barata J. Machine learning applications in agriculture: Current trends, challenges, and future perspectives. Agronomy. 2023;13(12):2976. https://doi.org/10.3390/agronomy13122976

14. Lavrenko S, Lykhovyd P, Lavrenko N, Ushkarenko V, Maksymov M. Beans (*Phaseolus vulgaris* L.) yields forecast using normalized difference vegetation index. Int J Agric Technol. 2022;18(3):1033–44.

15. Feng P, Wang B, Li Liu D, Waters C, Xiao D, Shi L, et al. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. Agric For Meteorol. 2020;285:107922. https://doi.org/10.1016/j.agrformet.2020.107922

16. Shafi U, Mumtaz R, Anwar Z, Ajmal M, Khan M, Mahmood Z, et al. Tackling food insecurity using remote sensing and machine learning-based crop yield prediction. IEEE Access. 2023;11:108640-108657. https://doi.org/10.1109/access.2023.3321020

17. Pant J, Pant RP, Singh MK, Singh DP, Pant H. Analysis of agricultural crop yield prediction using statistical techniques of machine learning. Mater Today: Proc. 2021;46:10922-10926. https://doi.org/10.1016/j.matpr.2021.01.948

18. Peel MC, Finlayson BL, McMahon TA. Updated world map of the Köppen-Geiger climate classification. Hydrol Earth Syst Sci. 2007;11 (5):1633–44. https://doi.org/10.5194/hess-11-1633-2007

19. Perry E, Sheffield K, Crawford D, Akpa S, Clancy A, Clark R. Spatial and temporal biomass and growth for grain crops using NDVI time series. Remote Sens. 2022;14(13):3071. https://doi.org/10.3390/rs14133071

20. Cahyono BE, Putri PO, Subekti A, Nugroho AT, Nishi K. Analysis of soil moisture as an indicator of land quality using vegetation index (SAVI and NDMI) retrieved from remote sensing data in Jember-Indonesia. In: AIP Conference Proceedings. 2022. p. 020006. https://doi.org/10.1063/5.0078761

21. Davidson C, Jaganathan V, Sivakumar AN, Czarnecki JMP, Chowdhary G. NDVI/NDRE prediction from standard RGB aerial imagery using deep learning. Comput Electron Agric. 2022;203:107396. https://doi.org/10.1016/j.compag.2022.107396

22. Voitik A, Kravchenko V, Pushka O, Kutkovetska T, Shchur T, Kocira S. Comparison of NDVI, NDRE, MSAVI and NDSI indices for early diagnosis of crop problems. Agric Eng. 2023;27. https://doi.org/10.2478/agriceng-2023-0004

23. Nadjla B, Assia S, Ahmed Z. Contribution of spectral indices of chlorophyll (RECl and GCI) in the analysis of multi-temporal mutations of cultivated land in the Mostaganem plateau. In: 2022 7th International conference on image and signal processing and their applications (ISPA). 2022. p. 1-6. https://doi.org/10.1109/ISPA54004.2022.9786326

24. Garbulsky MF, Peñuelas J, Gamon J, Inoue Y, Filella I. The photochemical reflectance index (PRI) and the remote sensing of leaf, canopy and ecosystem radiation use efficiencies: A review and meta-analysis. Remote Sens Environ. 2011;115(2):281–97. https://doi.org/10.1016/j.rse.2010.08.023

25. Wu C, Niu Z, Tang Q, Huang W. Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation. Agric For Meteorol. 2008;148(8-9):1230–41. https://doi.org/10.1016/j.agrformet.2008.03.005

26. Marill KA. Advanced statistics: Linear regression, part II: Multiple linear regression. Acad Emerg Med. 2004;11(1):94-102. https://doi.org/10.1197/j.aem.2003.09.006

27. Liu Y, Wang Y, Zhang J. New machine learning algorithm: Random forest. In: Liu B, Ma M, Chang J, editors. International conference on information computing and applications. 2012. p. 246–52. https://doi.org/10.1007/978-3-642-34062-8_32

28. Basak D, Pal S, Patranabis DC. Support vector regression. Neural Inf Process-Letters Rev. 2007;11(10):203–24.

29. Chen J, Zhao F, Sun Y, Yin Y. Improved XGBoost model based on genetic algorithm. Int J Comput Appl Technol. 2020;62(3):240–45. https://doi.org/10.1504/IJCAT.2020.106571

30. Aravind T, Reddy BS, Avinash S. A comparative study on machine learning algorithms for predicting the placement information of under graduate students. In: 2019 third International conference on I-SMAC (IoT in social, mobile, analytics and cloud) (I-SMAC). 2019. p. 542–46. https://doi.org/10.1109/I-SMAC47947.2019.9032654

31. Reid S, Tibshirani R, Friedman J. A study of error variance estimation in lasso regression. Stat Sin. 2016;35–67.

32. Taud H, Mas JF. Multilayer perceptron (MLP). In: Olmedo MTC, Paegelow M, Mas JS, Escobar F, editors. Geomatic approaches for modeling land change scenarios. Cham: Springer International Publishing; 2017. p. 451–55. https://doi.org/10.1007/978-3-319-60801-3_27

33. Saleh AME, Arashi M, Kibria BG. Theory of ridge regression estimation with applications. John Wiley & Sons; 2019.

34. Rauschenberger A, Glaab E, van de Wiel MA. Predictive and interpretable models via the stacked elastic net. Bioinformatics. 2021;37(14):2012–16. https://doi.org/10.1093/bioinformatics/btaa535

35. Tatachar AV. Comparative assessment of regression models based on model evaluation metrics. Int Res J Eng Technol (IRJET). 2021;8 (09):2395-3056.

36. Helland IS. On the interpretation and use of R^2 in regression analysis. Biometrics. 1987;61-69.

37. Moreno JJM, Pol AP, Abad AS, Blasco BC. Using the R-MAPE index as a resistant measure of forecast accuracy. Psicothema. 2013;25 (4):500-506. https://doi.org/10.7334/psicothema2013.23

38. Khan S, Iqbal J, Khan M, Malik N, Khan F, Khan K, et al. Using remotely sensed vegetation indices and multi-stream deep learning improves county-level corn yield predictions. Eur J Agron. 2025;164:127496. https://doi.org/10.1016/j.eja.2024.127496

39. Jhajharia K, Mathur P. Prediction of crop yield using satellite vegetation indices combined with machine learning approaches. Adv Space Res. 2023;72(9):3998-4007. https://doi.org/10.1016/j.asr.2023.07.006

40. Arshad S, Kazmi S, Javed M, Mohammed S. Applicability of machine learning techniques in predicting wheat yield based on remote sensing and climate data in Pakistan, South Asia. Eur J Agron. 2023;147:126837. https://doi.org/10.1016/j.eja.2023.126837

41. Aghighi H, Azadbakht M, Ashourloo D, Shahrabi H, Radiom S. Machine learning regression techniques for the silage maize yield prediction using time-series images of landsat 8 OLI. IEEE J Sel Top Appl Earth Obs Remote Sens. 2018;11:4563-77. https://doi.org/10.1109/JSTARS.2018.2823361

42. Muruganantham P, Wibowo S, Grandhi S, Samrat N, Islam N. A systematic literature review on crop yield prediction with deep learning and remote sensing. Remote Sens. 2022;14:1990. https://doi.org/10.3390/rs14091990

43. Tripathi A, Tiwari R, Tiwari S. A deep learning multi-layer perceptron and remote sensing approach for soil health based crop yield estimation. Int J Appl Earth Obs Geoinf. 2022;113:102959. https://doi.org/10.1016/j.jag.2022.102959

44. Chen Z, Chen J, Ding G, Huang H. A lightweight CNN-based algorithm and implementation on embedded system for real-time face recognition. Multimed Syst. 2023;29(1):129-38. https://doi.org/10.1007/s00530-022-00973-z

45. Jiao S, Gao Y, Feng J, Lei T, Yuan X. Does deep learning always outperform simple linear regression in optical imaging? Opt Express. 2020;28(3):3717-31. https://doi.org/10.1364/OE.382319

46. Han Y, Tang R, Liao Z, Zhai B, Fan J. A novel hybrid GOA-XGB model for estimating wheat aboveground biomass using UAV-based multispectral vegetation indices. Remote Sens. 2022;14(14):3506. https://doi.org/10.3390/rs14143506

47. Hara P, Piekutowska M, Niedbała G. Prediction of pea (*Pisum sativum* L.) seeds yield using artificial neural networks. Agric. 2023;13(3):661. https://doi.org/10.3390/agriculture13030661

48. Okupska E, Gozdowski D, Pudełko R, Wójcik-Gront E. Cereal and rapeseed yield forecast in Poland at regional level using machine learning and classical statistical models. Agric. 2025;15(9):984. https://doi.org/10.3390/agriculture15090984

49. Li Y, Zeng H, Zhang M, Wu B, Zhao Y, Yao X, et al. A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering. Int J Appl Earth Obs Geoinformation. 2023;118:103269. https://doi.org/10.1016/j.jag.2023.103269

50. Mouafik M, Fouad M, El Aboudi A. Machine learning methods for predicting *Argania spinosa* crop yield and leaf area index: A combined drought index approach from multisource remote sensing data. AgriEng. 2024;6(3):2283. https://doi.org/10.3390/agriengineering6030134

51. Yang S, Li L, Fei S, Yang M, Tao Z, Meng Y, Xiao Y. Wheat yield prediction using machine learning method based on UAV remote sensing data. Drones. 2024;8(7):284. https://doi.org/10.3390/drones8070284

52. Razavi M, Nejadhashemi A, Majidi B, Razavi H, Kpodo J, Eeswaran R, et al. Enhancing crop yield prediction in Senegal using advanced machine learning techniques and synthetic data. Artific Intel Agric. 2024;14:99-114. https://doi.org/10.1016/j.aiia.2024.11.005

53. Manjunath M, Palayyan B. An efficient crop yield prediction framework using hybrid machine learning model. Revue d'Intelligence Artificielle. 2023;37(4):1157-67. https://doi.org/10.18280/ria.370428

54. Chatterjee S, Kliestik T, Rowland Z, Bugaj M. Immersive collaborative business process and extended reality-driven industrial metaverse technologies for economic value co-creation in 3D digital twin factories. Oecon Copernic. 2025;16(1):125. https://doi.org/10.24136/oc.3596

55. Stefko R, Michalikova KF, Strakova J, Novak A. Digital twin-based virtual factory and cyber-physical production systems, collaborative autonomous robotic and networked manufacturing technologies, and enterprise and business intelligence algorithms for industrial metaverse. Equilibrium. 2025;20(1):389-425.

56. Zvarikova K, Gajanova L, Horak J. Exploring CSR performance as a proxy for competitive advantage across sectors in the Central European countries. Oecon Copernic. 2024;15(3):991-1020.

57. Kliestik T, Kral P, Bugaj M, Durana P. Generative artificial intelligence of things systems, multisensory immersive extended reality technologies, and algorithmic big data simulation and modelling tools in digital twin industrial metaverse. Equilibrium. 2024;19(2):429-61.

Additional information

Peer review: Publisher thanks Sectional Editor and the other anonymous reviewers for their contribution to the peer review of this work.

Reprints & permissions information is available at https://horizonepublishing.com/journals/index.php/PST/open_access_policy

Publisher's Note: Horizon e-Publishing Group remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Indexing: Plant Science Today, published by Horizon e-Publishing Group, is covered by Scopus, Web of Science, BIOSIS Previews, Clarivate Analytics, NAAS, UGC Care, etc
See https://horizonepublishing.com/journals/index.php/PST/indexing_abstracting

Publisher information: Plant Science Today is published by HORIZON e-Publishing Group with support from Empirion Publishers Private Limited, Thiruvananthapuram, India.