CHAPTER 1

Crop yields prediction using machine learning and remote sensing data in digital agriculture

Pavlo Lykhovyd

Abstract

An analytical review of contemporary scientific literature highlights the high potential of remote sensing data for developing mathematical models to predict agricultural crop productivity at local, regional, and national scales. By integrating multispectral and agrometeorological observations derived from Earth observation platforms, remote sensing enables comprehensive quantification of crop status and environmental dynamics across varying spatial and temporal resolutions. The integration of remote sensing technologies, particularly vegetation and agrometeorological indices derived from Earth observation data, has opened new avenues for yield prediction within the framework of digital agriculture.

As a result of the present study, a suite of machine learning models was developed to assess and predict the productivity of major crops cultivated in Ukraine. These models utilized remote sensing-derived parameters to capture the biophysical conditions of crops throughout the growing season. Vegetation indices, such as NDVI, EVI, NDMI, and VHI, as well as satellite-derived values of PET and LST, were used to build the models for major crop yields. Different machine learning approaches, including regression, ensemble, and decision tree-based algorithms, were applied to the input datasets. A transparent methodological framework was implemented, encompassing data preprocessing, temporal aggregation of indices, model calibration using cross-validation, and independent testing to ensure reproducibility and robustness. The results of statistical evaluation confirmed the reliability, high accuracy, and quality of approximation of most of the developed models, thereby supporting their applicability for both scientific analysis and practical decision-making in the agricultural sector.

In addition to model development, a comprehensive mathematical assessment of the relationship between vegetation indices and crop yields was conducted. This analysis contributed to the theoretical advancement of remote crop monitoring

 DOI: 10.21303/000-0000-0000-0-0.ch1

by deepening the understanding of how spectral and meteorological data correlate with plant growth and productivity, particularly when sourced from aerospace imagery. Such relationships were further used to optimize feature selection and enhance model interpretability within a data-driven digital agriculture framework.

At the local scale, the highest predictive performance was observed in regression models for soybean, grain corn, and common bean yields ($R^2 \geq 0.9$, MAPE $< 10\%$). On the regional level, the most accurate models were obtained for sunflower, soybean, rapeseed, and cereals ($R^2 \approx 0.5–0.7$, MAPE $< 20\%$). On the national scale, the best results were recorded for sugar beet, sunflower, and barley ($R^2 \approx 0.3–0.7$). The developed models not only demonstrate strong potential for improving yield forecasting but also represent a valuable contribution to the evolving field of digital and space-based agriculture. Their operational integration into digital agriculture systems can provide real-time decision support for optimizing resource allocation, irrigation scheduling, and risk management. Future research will focus on multisource data fusion, spatiotemporal harmonization, and adaptive learning architectures to improve model transferability across heterogeneous agroecological zones.

Keywords

Data analysis, forecasting, gradient boosting, mathematical analysis, neural networks, random forest, regression analysis, statistics, sustainability, vegetation indices.

1.1 Introduction

Accurate crop yield prediction is essential to ensure food security, rational resource management, and agricultural planning. The integration of machine learning (ML) algorithms, both conventional and innovative, with remote sensing data is a rapidly developing and methodologically distinct branch of agricultural science, offering scalable, precise, and timely yield forecasts across diverse crops and regions. Rapid development and implementation of the above-mentioned technologies have led to the transformation of conventional agriculture into digital agriculture, which is a new paradigm of applying data analytics, artificial intelligence (AI) and ML, sensors, drones, and satellites to optimize farming practices, improve resource efficiency, and boost crop productivity. Digital agriculture embraces climate-smart, environmentally friendly, and precision farming, enabling data-driven decisions to enhance both quantitative and qualitative performance of crop yields, optimize management of natural, financial, labor, and technological resources, and ensure the sustainability of the entire food system, from production to supply chains [1]. As digital agriculture becomes a mainstream trend in global agricultural development, Ukraine also

actively implements these innovations to remain competitive among other powerful agrarian countries of the world [2].

Machine learning-driven yield predictions based on satellite data are the cornerstone of digital agriculture. A scientifically grounded and transparent methodological framework – encompassing data acquisition, preprocessing, feature engineering, model calibration, and validation – provides the foundation for establishing crop yield prediction as an independent discipline within agricultural data science [3].

The earliest documented attempts to apply predictive analysis for crop yields date back to 1977, when researchers first demonstrated the potential of satellite data for forecasting yields and optimizing irrigation scheduling in hard wheat cultivation [4]. Linear regression in various modifications is one of the most popular traditional ML algorithms for yield prediction. When combined with data derived from vegetation indices it provides strong predictive power for crop yields. For example, models using MODIS-derived enhanced vegetation index (EVI) and triangular vegetation index (TVI) achieved R^2 values up to 0.70 for cotton and provided reliable forecasts during key crop stages [5]. In sugarcane, multivariate linear regression (MLR) models predicted biomass with up to 89% accuracy four months before harvest, outperforming more complex machine learning models in some cases [6]. For soybean and rice, linear regression using normalized difference vegetation index (NDVI) and climate variables yielded mean absolute percentage errors below 10% [7].

While linear regression is valued for its simplicity and interpretability, studies consistently show that advanced ML models – such as support vector machines (SVM), random forests (RF), and deep learning (DL) – often outperform linear regression, especially with large, complex datasets or when capturing non-linear relationships in agricultural ecosystems [8, 9]. However, the selection of appropriate ML algorithms requires careful methodological justification, considering data dimensionality, noise, and computational constraints. For example, long short-term memory (LSTM) and convolutional neural networks (CNN) are valued for their ability to model temporal dependencies, manage data gaps, and provide highly nonlinear representations, resulting in strong predictive reliability for crop yield estimation. However, these models require large, well-structured datasets, significant computational resources, and advanced interpretability techniques to overcome their "black-box" nature, which limits their operational deployment in some agricultural contexts [10]. Random forests, gradient boosting ensembles, and SVM models are often referred to as balanced solutions for crop yields prediction because of their scalability, high accuracy, relatively low computational requirements, and simplicity of implementation in practice [11]. However, SVMs are prone to underfitting complex interactions among explanatory variables and yield components, which in certain cases results in lower

prediction accuracy compared to linear regression models. Usually, the performance depends strongly on dataset characteristics and preprocessing strategies.

Apart from ML algorithms the accuracy and reliability of yield prediction depend critically on the quality, representativeness, and temporal consistency of input data. Remote sensing-derived vegetation indices (NDVI, EVI, etc.), climate data (precipitation, air and/or land surface temperature), and soil parameters are the most influential predictors in crop forecasting [12]. High-resolution satellite imagery (e.g., retrieved from Sentinel-2, Landsat-8, CubeSat) enables field-level and regional yield estimation, with cumulative or peak values of the vegetation indices often providing the best results [13]. Satellite imagery of low resolution is mainly used in large-scale regional or state-level studies, as it is not suitable for field-scale implementation. Using lower resolution images also results in reduced predictive power of ML models and requires additional refinement [14].

The primary methodological challenges in ML-based crop yield prediction include limited availability of ground truth datasets, the interpretability of complex models, and the adaptation of trained models for deployment within real-world agricultural systems and embedded platforms. Additional barriers involve selecting optimal model architectures for specific agroecological conditions, harmonizing multisource data streams, and ensuring standardized preprocessing workflows [15]. Therefore, the question of reliable crop yields prediction using ML models is still not fully resolved within the international scientific community and requires further careful investigation, considering not only global trends in digital agriculture, but also regional and national peculiarities of agro-industrial development.

The main goal of this chapter is to present the latest methodological advances and empirical findings in ML-based crop yield prediction for Ukraine, emphasizing the integration of remote sensing-derived vegetation indices within digital agriculture frameworks. The chapter also outlines key achievements, current limitations, and future research directions focused on multisource data fusion, spatiotemporal harmonization, and adaptive ML systems capable of operating across heterogeneous agroecological zones.

1.2 Local models

The results of crop yield prediction at the local-field level are summarized in **Table 1.1**. According to the results of mathematical analysis, quantitative statistical analysis, different yield prediction models based on satellite vegetation indices show varying degrees of effectiveness depending on the crop and its

phenological phase. To ensure methodological transparency and comparability across models, the quality of prediction was primarily assessed using the correlation coefficient (*R*) to capture the strength of the linear relationship and the mean absolute percentage error (MAPE, %) to quantify the average prediction deviation.

Table 1.1 Local models of crop yields based on remote sensing NDVI values

Crop	Phenological phase	Model	*R*	MAPE, %
Common beans	Flowering – pod formation	$Y = 4.4278 - 6.7325 \times NDVI + 22.8703 \times NDVI^2$	0.90	7.12
Grain corn	Tasseling – ear formation	$Y = -8.0534 + 22.7549 \times NDVI + 8.5707 \times NDVI^2$	0.99	8.75
Sweet corn	Tasseling	$Y = -63.8839 + 335.391 \times NDVI - 355.5484 \times NDVI^2$	0.65	28.13
Grain sorghum	Flowering	$Y = 10.0315 - 42.1255 \times NDVI + 52.1927 \times NDVI^2$	0.93	17.62
Soy-beans	Second internode	$Y = -2.4740 + 9.6725 \times NDVI - 0.5885 \times NDVI^2$	0.99	3.75
Winter wheat	Earing	$Y = -16.7294 + 51.0578 \times NDVI - 24.2807 \times NDVI^2$	0.85	13.42

Source: author's research

Models classified as highly efficient ($R \geq 0.90$; MAPE $\leq 10\%$) demonstrated superior predictive power and operational reliability. These values of statistical parameters indicate models with high robustness, suitable not only for academic validation but also for integration into decision-support systems within digital agriculture platforms. According to the results, such models have been obtained for:

1. Soybeans ($R = 0.99$; MAPE = 3.75%). The exceptionally high correlation coefficient indicates an almost perfect relationship between the vegetation index at the second internode stage and the final crop yield. The extremely low MAPE value confirms that the model predicts soybean productivity almost flawlessly. This performance likely reflects the crop's strong physiological linkage between early canopy development and final reproductive output, enabling the NDVI signal at early stages to serve as a reliable proxy for yield potential.

2. Grain corn ($R = 0.99$; MAPE = 8.75%). The high correlation coefficient indicates a strong direct relationship between the indices obtained at the critical phase of crop tasseling and final grain yield. Although the error is slightly higher than for soybeans, it is still very low. for operational-level yield forecasting and can be directly integrated into farm management systems for early-season productivity estimation.

3. Common beans (R = 0.90; MAPE = 7.12%). A correlation coefficient of 0.90 is considered high, indicating a strong direct relationship between the vegetation index value and crop yield, and an error of less than 10% makes the forecast reasonable for management decisions. The stability of this model across multiple growing seasons highlights its potential for incorporation into regional-scale precision agriculture frameworks.

Models of moderate efficiency (R = 0.85–0.90; MAPE = 10–20%) retained high analytical value but required contextual calibration for atypical conditions. These models include the following:

1. Winter wheat (R = 0.85; MAPE = 13.42%). The correlation coefficient of 0.85 indicates a close relationship between NDVI and crop yield. Although moderately accurate, this model benefits from ensemble calibration or hybridization with agrometeorological predictors (e.g., cumulative precipitation, thermal time) to improve stability under climate anomalies.

2. Grain sorghum (R = 0.93; MAPE = 17.62%). The high correlation coefficient (0.93) indicates a strong relationship between the studied parameters (viz., NDVI and crop yield), but the MAPE is close to 20%, indicating that while the model successfully captures general yield trends, the magnitude of predicted values remains sensitive to environmental variability. This situation may arise if yield is influenced by factors that are not sufficiently reflected by the NDVI values. Incorporating soil moisture or evapotranspiration features could enhance its representativeness in future adaptive model designs.

Such models have limited practical application but remain valuable for exploratory analysis and methodological refinement. These include the sweet corn yield model. The yield data for sweet corn (R = 0.65; MAPE = 28.13%). The model for the tasseling phenological phase has an average correlation coefficient (0.65) and a high error, which makes it unsuitable for accurate forecasting. The reduced model performance can be attributed to several methodological and biological factors:

1. Yield dependency on non-spectral attributes such as kernel quality, pollination efficiency, and microclimatic stressors (e.g., drought, pests), which are not adequately captured by NDVI-based models.

2. Potential model underfitting caused by limited feature diversity and linear algorithm constraints, emphasizing the need for nonlinear or ensemble learning strategies.

Overall, the results indicate that yield modeling using NDVI is a powerful tool, but its effectiveness is not universal for all crops and agroecological conditions. It depends on:

1. Biological and morphological traits of each crop, where yield is strongly coupled with canopy development and photosynthetic efficiency.

2. Accurate identification of phenological phases corresponding to maximum spectral-yield correlation, as even minor timing deviations can degrade model accuracy.

3. Comprehensive data acquisition and preprocessing – including normalization, outlier removal, and synthetic augmentation where applicable – to improve sample representativeness.

4. Selection of the most appropriate ML architecture and analytical framework tailored to data structure, ensuring compliance with statistical assumptions and minimizing overfitting risks.

For convenience, based on the results of a yield scale for the crops studied was created linking NDVI intervals to expected yield classes (**Table 1.2**). The developed scale can be used for a rapid preliminary assessment of yield potential across fields using spatial NDVI mosaics, supporting early-stage management interventions within digital agriculture platforms.

Table 1.2 Yield scale of the studied crops (t/ha) depending on the NDVI value in the corresponding phenological phase

NDVI value	Grain corn	Sweet corn	Sorghum	Soybeans	Winter wheat	Common beans
0.3	–	3.4–6.0	–	0.3–0.4	–	–
0.4	2.2–2.7	9.5–17.0	1.3–1.8	1.2–1.4	–	1.3–1.5
0.5	5.0–6.0	10.5–19.0	1.6–2.4	2.1–2.3	2.3–3.1	1.6–1.9
0.6	7.9–9.4	6.7–12.0 †	2.9–4.1	3.0–3.3	4.4–5.9	2.4–2.8
0.7	11.0–13.1	–	5.0–7.2	3.8–4.2	6.1–8.1	3.6–4.1
0.8	14.0–17.0	–	8.0–11.4	4.7–5.1	7.4–9.8	–

Source: author's research
Note: "–" indicates vegetation index values that are atypical for the model; † – the decrease in sugar corn yield at higher vegetation index values is because the NDVI ≥ 0.6 variant is a dead variant in the input data set

In future research, multisource data fusion – combining NDVI with additional indices such as NDMI, VHI, and LST, along with meteorological and soil data – will enhance predictive robustness and adaptability. Furthermore, implementing adaptive ML architectures capable of dynamic retraining based on updated satellite observations will improve scalability and reliability under heterogeneous agroecological conditions.

1.3 Regional and national models

Regional (zonal) models are crucial for understanding the agroecological gradients of crop cultivation, as well as for planning agricultural policy and ensuring food security. Based on the results of a pilot short-term (2012–2021) study of MODIS-based NDVI and EVI dynamics, combined with yield data for major crops in the Kherson region of Ukraine, it was demonstrated that it is possible to predict crop productivity based on regional vegetation index values, provided that agricultural land is pre-delineated using the region's vegetation mask provided by the NextGIS service.

It was shown that it is possible to accurately predict the yield of winter crops based on the regional NDVI or EVI values obtained in May using regression or artificial neural network (ANN) modeling. The mean absolute percentage error (MAPE) of the predictions varied: 5.7 and 6.5% for winter wheat when forecasting using NDVI and EVI values, respectively; 8.9 and 10.6% for winter barley when forecasting using NDVI and EVI values, respectively. The high quality of the models was confirmed by the coefficients of determination (0.90 and 0.89 for winter wheat when forecasting using NDVI and EVI values, respectively; 0.82 and 0.77 for winter barley when forecasting using NDVI and EVI values, respectively). Based on the results of ML modeling using a linear regression algorithm, a productivity scale for winter grain crops in the Kherson region was developed (**Tables 1.3** and **1.4**).

Table 1.5 contains the results of crop yield modeling at two levels:

– regional (Kherson region);

– national (Ukraine).

Analysis of correlation coefficients (R) and mean absolute percentage error (MAPE) values allows assessing the performance of these models.

Table 1.3 Grain yield scale of winter grain crops in the Kherson region of Ukraine depending on the MODIS NDVI value in May

MODIS NDVI	Winter wheat, t/ha	Winter barley, t/ha
0.3–0.4	< 0.40	< 0.1
0.4–0.5	0.4–2.5	0.1–2.2
0.5–0.6	2.5–4.6	2.3–4.5
0.6–0.7	4.7–6.7	4.5–6.7
0.7–0.8	6.8–8.8	6.8–9.0
> 0.8	> 8.8	> 9.0

Source: author's research

Table 1.4 Grain yield scale of winter grain crops in the Kherson region of Ukraine depending on the MODIS EVI value in May

MODIS EVI	Winter wheat, t/ha	Winter barley, t/ha
0.2–0.3	0.4–2.8	0.2–2.6
0.3–0.4	2.8–5.2	2.6–5.1
0.4–0.5	5.2–7.6	5.1–7.5
0.5–0.6	7.6–9.9	7.6–10.0
> 0.6	> 10.00	> 10.0

Source: author's research

Table 1.5 Regional and national models of crop yields depending on remote sensing data

Crop	Time span	Region	Model	*R*	MAPE, %
Winter wheat	June	Kherson	$Y = -2.2349 + 8.9551 \times NDVI$	0.67	18.47
Winter barley	May	Kherson	$Y = -3.7256 + 12.5616 \times NDVI$	0.71	20.68
Winter rye	June	Kherson	$Y = -1.8686 + 6.2818 \times NDVI$	0.59	19.80
Oats	June	Kherson	$Y = -1.1838 + 4.5031 \times NDVI$	0.54	27.72
Millet	June	Kherson	$Y = -1.1751 + 4.7315 \times NDVI$	0.76	24.34
Winter wheat	Season	Kherson	$Y = 0.2296 + 5.0346 \times VHI$	0.86	10.40
Spring wheat	Season	Kherson	$Y = 0.3548 + 4.5446 \times VHI$	0.86	9.67
Spring barley	Season	Kherson	$Y = -0.8136 + 6.3248 \times VHI$	0.78	20.96
Grain corn	Season	Kherson	$Y = 2.7183 + 5.65358 \times VHI$	0.54	13.90
Sunflower	Season	Kherson	$Y = -0.1453 + 2.5882 \times VHI$	0.73	21.89
Potato	June	Kherson	$Y = -3.5970 + 27.7080 \times NDVI$	0.57	10.04
Vegetables	May	Kherson	$Y = -18.2710 + 86.6660 \times NDVI$	0.63	21.07
Fruits	August	Kherson	$Y = 0.3748 + 14.4510 \times NDVI$	0.33	34.53
Winter rapeseed	April	Kherson	$Y = 4.0723 \times NDVI$	0.95	25.33
Soybeans	August	Kherson	$Y = 2.9677 \times NDVI$	0.98	18.28
Sunflower	July	Kherson	$Y = 6.2328 \times NDVI$	0.99	13.24
Spring wheat	Season	Ukraine	$Y = 0.3814 + 7.3537 \times NDVI$	0.54	MSE = 0.33
Winter rapeseed	Season	Ukraine	$Y = 3.9245 + 0.0003 \times PET - 0.0986 \times LST$	0.57	MSE = 0.19
Peas	Season	Ukraine	$Y = 7.0695 - 0.0062 \times PET$	0.62	MSE = 0.12

Source: author's research

Modeling at the regional level, as a rule, gives more accurate results, since it captures local soil, climatic, and technological factors. In this case, the models are divided into two groups:

1. Models based on average monthly indices.

These models use data for a single month corresponding to a critical phenological stage:

– *high-performance models*. Models for sunflower (July, $R = 0.99$, MAPE = 13.24%), soybeans (August, $R = 0.98$, MAPE = 18.28%), and winter rapeseed (April, $R = 0.95$, MAPE = 25.33%) demonstrate very strong correlations between vegetation index values and yield outcomes. This indicates that the yield of these crops strongly depends on their condition during a specific critical growth phase. For example, for sunflower this corresponds to the flowering and seed formation stage. The relatively higher errors for rapeseed and soybeans may occur because the models accurately predict the general trend but do not account for localized agroecological or spectral factors. For rapeseed, this may also relate to the reflectance properties during its bright-yellow flowering phase, which may distort spatial imagery and vegetation index calculations;

– *moderate-performance models*. The models for winter barley ($R = 0.71$) and winter wheat ($R = 0.67$) show moderate correlations which makes them useful but not suitable for operational forecasting. The relatively high prediction errors (~ 20%) confirm that additional factors should be incorporated – such as climatic and soil parameters – or that more robust ML algorithms may improve accuracy;

– *low-performance models*. The models for oats ($R = 0.54$) and winter rye ($R = 0.59$) show weak relationships and large prediction errors, which makes them unsuitable for reliable forecasting. The particularly low accuracy of the fruit yield model ($R = 0.33$) indicates its limited theoretical and practical value. This result is expected, since fruit yield depends on many factors – flowering weather, pollination success, pest pressure, and harvest conditions. In addition, this group of crops is highly heterogeneous, which further reduces forecasting accuracy.

2. Models based on annual average indices.

These models use vegetation indices averaged over the entire growing season. This can smooth out temporal variations but also reduce model sensitivity:

– *high-performance models*. The models for winter wheat ($R = 0.86$) and spring wheat ($R = 0.86$) show high correlation and low error (MAPE < 11%). This suggests that for these crops, the overall canopy biomass condition throughout the season is an excellent indicator of final yield;

– *moderate-performance models*. The model for spring barley ($R = 0.78$) is also reasonably effective but has a slightly higher MAPE (20.96%). This indicates the influence of unmodeled factors, or an insufficient dataset size;

– *low-performance models*. The models for grain corn (R = 0.54) and sunflower (R = 0.73) have relatively weak to moderate correlations with high MAPE. This confirms that for these crops it is more important to assess their condition during key phenological periods rather than based on seasonal averages.

National-scale models are considerably less accurate than regional models due to large-scale heterogeneity in agroclimatic conditions and cultivation technologies. The correlation for spring wheat (R = 0.54), winter rapeseed (R = 0.57), and peas (R = 0.62) is moderate, confirming that simple models based only on average annual indices cannot adequately reflect the spatial complexity of national yield patterns. Such models are more suitable for assessing broad productivity trends, rather than for precise forecasting. Of course, more sophisticated ML algorithms can partially improve forecast accuracy, but in this case the interpretability of the predictive function is often lost, and therefore its integration into digital decision-support or precision farming systems becomes limited.

Temporal resolution is a decisive factor: for most of the studied crops, especially sunflower, soybean, and rapeseed, models based on critical-phase vegetation data are significantly more accurate than those using seasonal averages. This confirms that the forecasting process should account for phenological timing, and not just average data. Regional-level models consistently outperform national ones due to the lower variability of factors affecting yield within one region. Such models are especially effective for cereals and oilseeds (e.g., soybean, wheat, sunflower) In contrast, for crops with complex yield-forming mechanisms index-based modeling alone remains insufficient.

To improve the efficiency of regional and national models, it is necessary to integrate additional predictors such as precipitation, air temperature, and soil type. Their inclusion helps reduce relative errors and improve model reliability, even in cases of high correlation coefficients. In addition, expanding the input datasets will enable the use of more advanced ML algorithms capable of capturing nonlinear and complex dependencies between vegetation indices and crop yields.

An analysis of crop productivity in Ukraine, based on vegetation indices (NDVI, VHI, NDMI) and remote sensing-derived parameters (LST – land surface temperature, and PET – potential evapotranspiration), revealed the following patterns **(Table 1.6)**:

– NDMI: this index shows moderate correlations with yield for most crops, particularly spring barley (R = 0.50), peas (R = 0.49), spring wheat (R = 0.48), and winter wheat (R = 0.47). These relationships emphasize the crucial role of moisture availability in determining yield under Ukraine's frequent water deficit conditions:

– NDVI: this index demonstrates moderate correlations, notably for cereals such as winter barley (R = 0.51) and spring wheat (R = 0.48), indicating that overall biomass condition effectively reflects crop productivity potential;

– VHI: VHI, which integrates thermal and vegetation parameters, exhibits lower correlations with yield than NDVI and NDMI across most crops. This suggests that the direct influence of moisture availability captured by NDMI is more critical to yield formation than the composite stress captured by VHI;

– LST: LST demonstrates the strongest (inverse) relationships with yield among the studied indices – sunflower ($R = -0.62$), maize ($R = -0.57$), and winter wheat ($R = -0.58$). This underscores that elevated surface temperatures during critical phenological stages are a dominant yield-limiting factor;

– PET: PET also shows significant negative correlations with yield – maize ($R = -0.55$), spring barley ($R = -0.57$), sunflower ($R = -0.57$), and winter wheat ($R = -0.55$). This further confirms that water stress resulting from elevated temperature regimes critically constrains yield formation. Conversely, for sugar beet and soybean, correlations with LST and PET remain relatively weak, likely due to crop-specific adaptive mechanisms or the influence of additional factors not fully captured by thermal and moisture indices.

Table 1.6 Pairwise correlation between NDVI, VHI, NDMI, LST, PET and yield of major crops in Ukraine

Crop	NDVI	NDMI	VHI	LST	PET
Spring wheat	0.48	0.48	0.44	–0.50	–0.53
Rapeseed	0.24	0.38	0.19	–0.47	–0.45
Spring barley	0.45	0.50	0.42	–0.55	–0.57
Winter barley	0.51	0.43	0.48	–0.43	–0.45
Grain corn	0.43	0.43	0.35	–0.57	–0.55
Sunflower	0.30	0.43	0.22	–0.62	–0.57
Soybeans	0.33	0.25	0.29	–0.24	–0.28
Sugar beets	0.24	0.27	0.24	–0.09	–0.16
Winter wheat	0.35	0.47	0.26	–0.58	–0.55
Oats	0.21	0.39	0.19	–0.36	–0.32
Rye	0.23	0.35	0.19	–0.32	–0.32
Peas	0.30	0.49	0.23	–0.54	–0.54

The results of machine learning (ML) modeling demonstrated that the yield of most major crops in Ukraine primarily depends on two key environmental factors – moisture availability (NDMI) and temperature regime (LST and PET).

Across most crops, negative correlations with LST and PET are stronger than positive correlations with NDVI and NDMI. This indicates that crop yields in Ukraine

are more frequently constrained by environmental stress factors (heat and drought) than determined by biomass-driven productive potential.

Effective yield prediction models should incorporate combinations of these indices. In particular, integrating air and land surface temperature parameters with hydrological indicators substantially enhances forecasting precision.

Therefore, effective yield forecasting for Ukrainian agroecosystems requires explicit consideration of climatic influences, particularly under the increasing frequency of droughts and temperature extremes.

Furthermore, the study evaluated the performance of several ML algorithms for yield prediction in Ukraine using remote sensing datasets (**Table 1.7**). The efficiency of the models varied across crop types and algorithmic approaches.

Table 1.7 Comparison of linear regression, random forest and gradient boosting algorithms in forecasting the yield of major agricultural crops in Ukraine using aerospace monitoring data

Crop	ML Model								
	Linear regression			Random forest			Gradient boosting		
	R^2	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE
Spring wheat	0.29	0.45	0.33	0.18	0.47	0.38	0.04	0.53	0.45
Rapeseed	0.33	0.34	0.19	–0.56	0.35	0.20	–1.37	0.38	0.30
Spring barley	0.25	0.49	0.43	0.49	0.36	0.26	0.31	0.43	0.35
Winter barley	0.35	0.53	0.42	0.44	0.47	0.36	0.35	0.50	0.41
Grain corn	0.04	1.26	2.67	0.13	1.32	2.42	0.08	1.36	2.56
Sunflower	0.09	0.40	0.25	0.54	0.31	0.13	0.47	0.29	0.15
Soybeans	0.08	0.36	0.25	0.16	0.36	0.23	0.21	0.33	0.22
Sugar beets	–0.12	7.00	110.77	0.57	5.21	33.85	0.63	4.77	28.56
Winter wheat	0.03	0.72	0.64	0.16	0.56	0.61	0.19	0.53	0.57
Oats	0.11	0.40	0.26	0.18	0.35	0.24	0.09	0.37	0.27
Rye	0.09	0.60	0.47	–0.04	0.59	0.60	0.11	0.55	0.51
Peas	0.39	0.28	0.12	–0.14	0.37	0.23	–0.64	0.41	0.33

The highest model performance (high R^2 and low MAE, MSE) was observed for the following crops:

– sugar beet: all ML algorithms performed well, with gradient boosting ($R^2 = 0.63$) and random forest ($R^2 = 0.57$) significantly outperforming linear regression. This suggests that sugar beet yield depends on nonlinear and complex relationships effectively captured by tree-based ensemble methods;

– spring barley: random forest achieved the best results ($R^2 = 0.49$), outperforming both linear regression ($R^2 = 0.25$) and gradient boosting. This confirms its suitability for spring barley yield prediction, likely due to its ability to model complex variable interactions;

– sunflower: random forest also provided the most accurate predictions for sunflower ($R^2 = 0.54$), surpassing both linear regression and gradient boosting, confirming its effectiveness for this crop;

– winter barley: all models produced comparable results, with random forest ($R^2 = 0.44$) and linear regression ($R^2 = 0.35$) performing best. This implies that yield relationships for winter barley are less complex, allowing linear models to capture a substantial portion (35%) of variability.

Moderate ML model performance (R^2 values ranging from 0.10 to 0.30) was observed for the following crops:

– winter wheat and oats: for these crops, all models yielded low performance ($R^2 < 0.20$), suggesting either insufficient dataset representativeness or that key yield determinants lie beyond the scope of remote sensing–derived predictors;

– spring wheat, soybean, rye, corn: for these crops, all models produced low R^2 values, indicating limited predictive capability of the available data. In particular, models for corn and rye demonstrated very low performance, confirming their limited practical applicability.

Negative R^2 values were observed for certain crops, including:

– rapeseed, rye, peas: for rapeseed, rye, and peas, random forest and gradient boosting produced negative R^2 values, suggesting that these complex models underperformed relative to simple linear regression. This likely reflects overfitting or an imbalance between model complexity and data volume or quality.

Overall, the results of the study suggest that, although there are limitations, the approach using machine learning and remote sensing data for crop yield prediction on a large scale is promising. However, to achieve higher accuracy, it is necessary to carefully select not only the ML algorithm but also collect proper dataset in size and quality and use appropriate data augmentation and normalization techniques prior to passing the data into models.

Future research should focus on integrating multisource data, including high-resolution satellite imagery, UAV-based observations, and in-situ sensor measurements, to capture diverse environmental and management factors affecting crop growth. Moreover, developing spatiotemporal harmonization techniques will allow models to effectively combine data across different scales and time frames, improving predictive reliability under variable climatic and phenological conditions. Another promising direction is the design of adaptive machine learning

frameworks that can adjust to heterogeneous agroecological zones, thereby enhancing model generalizability, resilience to noise and missing data, and scalability across diverse farming contexts. Finally, exploration of hybrid modeling approaches that combine process-based crop models with data-driven ML techniques could further strengthen decision-support capabilities in operational digital agriculture systems, providing actionable insights for precision resource management and yield optimization.

1.4 Conclusions

Based on the results of this study, a suite of machine learning (ML) models was developed to assess and predict the productivity of key crops cultivated in Ukraine, utilizing remote sensing-derived vegetation indices and agrometeorological parameters. Statistical evaluation demonstrated the robustness, high accuracy, and overall reliability of the majority of the developed models, confirming their suitability for both scientific research and practical applications. The results indicate that these models can be effectively integrated into operational digital agriculture systems, providing decision-support tools for resource management, yield optimization, and precision farming.

Furthermore, the mathematical analysis of vegetation indices and their relationships with crop productivity has contributed to advancing theoretical understanding in the field of remote monitoring of crop growth and development. In particular, the integration of aerospace imagery has proven to be a valuable tool for assessing the condition and productivity potential of agricultural crops. Moreover, the study highlights the importance of incorporating key climatic factors, phenological stages, and soil characteristics into predictive models to enhance practical applicability.

At the local scale, the highest prediction accuracy was achieved for regression models applied to soybean, grain corn, and common bean yields. At the regional level, the best-performing models were those for sunflower, soybeans, rapeseed, and cereals. Nationally, the most accurate predictions were obtained for sugar beet, sunflower, and barley. However, to further improve model generalizability, resilience to data noise, and scalability across diverse agroecological zones, future research should explore multisource data fusion, spatiotemporal harmonization, and adaptive machine learning frameworks. Such approaches would allow models to integrate high-resolution satellite, UAV, and ground-based sensor data, capture non-linear interactions, and support actionable recommendations for precision resource management in heterogeneous farming contexts.

References

1. Fuentes-Peñailillo, F., Gutter, K., Vega, R., Silva, G. C. (2024). Transformative Technologies in Digital Agriculture: Leveraging Internet of Things, Remote Sensing, and Artificial Intelligence for Smart Crop Management. Journal of Sensor and Actuator Networks, 13 (4), 39. https://doi.org/10.3390/jsan13040039
2. Derkach, O. D., Mykhaylichenko, Y. M. (2021). Digital agriculture: The experience of Ukraine. Mechanization in agriculture & Conserving of the resources, 67 (2), 52–56. Available at: https://stumejournals.com/journals/am/2021/2/52.full.pdf
3. Chlingaryan, A., Sukkarieh, S., Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and Electronics in Agriculture, 151, 61–69. https://doi.org/10.1016/j.compag.2018.05.012
4. Idso, S. B., Jackson, R. D., Reginato, R. J. (1977). Remote-Sensing of Crop Yields. Science, 196 (4285), 19–25. https://doi.org/10.1126/science.196.4285.19
5. de Siqueira, D. A. B., Vaz, C. M. P., da Silva, F. S., Ferreira, E. J., Speranza, E. A., Franchini, J. C. et al. (2024). Estimating Cotton Yield in the Brazilian Cerrado Using Linear Regression Models from MODIS Vegetation Index Time Series. AgriEngineering, 6 (2), 947–961. https://doi.org/10.3390/agriengineering6020054
6. Servia, H., Pareeth, S., Michailovsky, C. I., de Fraiture, C., Karimi, P. (2022). Operational framework to predict field level crop biomass using remote sensing and data driven models. International Journal of Applied Earth Observation and Geoinformation, 108, 102725. https://doi.org/10.1016/j.jag.2022.102725
7. Mateo-Sanchis, A., Piles, M., Muñoz-Marí, J., Adsuara, J. E., Pérez-Suay, A., Camps-Valls, G. (2019). Synergistic integration of optical and microwave satellite data for crop yield estimation. Remote Sensing of Environment, 234, 111460. https://doi.org/10.1016/j.rse.2019.111460
8. Jhajharia, K., Sharma, N. V., Mathur, P. (2025). A Machine Learning Model for Crop Yield Prediction Using Remote Sensing Data. International Research Journal of Multidisciplinary Scope, 6 (2), 577–590. https://doi.org/10.47857/irjms.2025.v06i02.03182
9. Lykhovyd, P. V. (2018). Prediction of sweet corn yield depending on cultivation technology parameters by using linear regression and artificial neural network methods. Biosystems Diversity, 26 (1), 11–15. https://doi.org/10.15421/011802
10. Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N. H., Islam, N. (2022). A Systematic Literature Review on Crop Yield Prediction with Deep Learning and Remote Sensing. Remote Sensing, 14 (9), 1990. https://doi.org/10.3390/rs14091990

11. Yang, S., Li, L., Fei, S., Yang, M., Tao, Z., Meng, Y. et al. (2024). Wheat Yield Prediction Using Machine Learning Method Based on UAV Remote Sensing Data. Drones, 8 (7), 284. https://doi.org/10.3390/drones8070284
12. Hara, P., Piekutowska, M., Niedbała, G. (2021). Selection of Independent Variables for Crop Yield Prediction Using Artificial Neural Network Models with Remote Sensing Data. Land, 10 (6), 609. https://doi.org/10.3390/land10060609
13. Ziliani, M. G., Altaf, M. U., Aragon, B., Houborg, R., Franz, T. E., Lu, Y. et al. (2022). Early season prediction of within-field crop yield variability by assimilating CubeSat data into a crop model. Agricultural and Forest Meteorology, 313, 108736. https://doi.org/10.1016/j.agrformet.2021.108736
14. Rembold, F., Atzberger, C., Savin, I., Rojas, O. (2013). Using Low Resolution Satellite Imagery for Yield Prediction and Yield Anomaly Detection. Remote Sensing, 5 (4), 1704–1733. https://doi.org/10.3390/rs5041704
15. Parashar, N., Johri, P., Khan, A. A., Gaur, N., Kadry, S. (2024). An Integrated Analysis of Yield Prediction Models: A Comprehensive Review of Advancements and Challenges. Computers, Materials & Continua, 80 (1), 389–425. https://doi.org/10.32604/cmc.2024.050240