

Article Type: Research Article

J Name: Modern Phytomorphology

Short name: MP

ISSN: ISSN 2226-3063/eISSN 2227-9555

Year: 2025

Volume: 19

Page numbers: 425-430

DOI: 10.5281/zenodo.17994141
(10.5281/zenodo.2025-19-425-430)



Short Title: Remote sensing data and machine learning to predict yields of major crops on regional scale

RESEARCH ARTICLE

Remote sensing data and machine learning to predict yields of major crops on regional scale

Pavlo Lykhovyd^{1*}, Raisa Vozhehova¹, Liudmyla Hranovska¹, Oleksandr Averchev², Anatolii Tomnytskyi¹, Nataliia Avercheva³, Mariia Nikitenko², Serhii Haievskiy⁴, Oleg Shalar⁵

¹Department of Irrigated Agriculture and Decarbonization of Agroecosystems, Institute of Climate-Smart Agriculture of NAAS, Odesa, Ukraine

²Department of Agriculture, Kherson State Agrarian and Economic University, Kropyvnytskyi, Ukraine

³Department of Social and Behavioral Sciences, Kherson State Agrarian and Economic University, Kropyvnytskyi, Ukraine

⁴Department of Plant Sciences and Agricultural Engineering, Kherson State Agrarian and Economic University, Kropyvnytskyi, Ukraine

⁵Department of Management, Marketing and Information Technologies, Kherson State Agrarian and Economic University, Kropyvnytskyi, Ukraine

*Corresponding author: Pavlo Lykhovyd, Department of Irrigated Agriculture and Decarbonization of Agroecosystems, Institute of Climate-Smart Agriculture of NAAS, Odesa, Ukraine, E-mail: pavel.likhovid@gmail.com

Received: 02.02.2025, Manuscript No: mp-25-160951 | Editor Assigned: 05.02.2025, Pre-QC No. mp-25-160951 (PQ) | Reviewed: 26.11.2025, QC No. mp-25-160951 (Q) | Revised: 10.12.2025, Manuscript No. mp-25-160951 (R) | Accepted: 17.12.2025 | Published: 24.12.2025

Abstract

Remote sensing and machine learning tandem is a powerful tool for crop monitoring and yield prediction. Current study is devoted to the evaluation of the relationship between five remote sensing indicators, affecting crop yields, such as NDVI, NDMI, VHI, LST and PET, as well as establishing the connectivity of these indices with the yields of twelve major crops cultivated in Ukraine during the period 2015-2023. Yielding data were retrieved from official statistical bodies of Ukraine. Remote sensing data were calculated and generalized through the Google earth engine platform through the requests to the API in JavaScript. Correlation and regression analysis were performed using common methodologies, as well as more robust machine learning techniques like random forest and gradient boosting regression were also applied for yield prediction. It was determined that the strongest correlation (0.93-0.96) is between such indices as NDVI and VHI, LST and PET, VHI and PET. As for crop models, for most crops Random Forest algorithms provided the best overall quality of fitting and prediction accuracy, followed by Gradient Boosting Regression. However, in case of crops with smaller datasets, such as spring wheat, rapeseed and peas, linear regression provided better accuracy than more robust machine learning methods. The study emphasizes the importance of using the tandem of machine learning and remote sensing in modern agriculture to improve crop monitoring and yield prediction, contributing to sustainable agricultural practices and food security.

Keywords: Land surface temperature, Modeling, Normalized difference moisture index, Normalized difference vegetation index, Potential evapotranspiration, Vegetation health index

Introduction

Remote sensing is a rapidly developing branch of modern science and technology with a wide range of practical and scientific

applications in life sciences, especially geography, ecology, and agriculture. In agriculture, remote sensing data are usually derived from different satellite sensors and aerospace photography through specific calculations to provide an indirect description of certain features of the on-land objects, e.g., water bodies, plant groups, soil properties, etc. Different indices calculated based on the remote sensing data are crucial for monitoring crop health and predicting yields, and their combined use is important to ensure that comprehensive embrace of all the major factors affecting crop yields is provided. In our opinion, the most important remotely retrieved indicators, which could be used to predict crops productivity in Ukraine, are as follows:

Normalized Difference Vegetation Index (NDVI), which measures vegetation greenness by comparing the difference between Near-Infrared NIR (which vegetation strongly reflects) and red R spectral bands (which vegetation absorbs). High NDVI values indicate healthy, dense vegetation, while lower values suggest sparse or stressed crops. The values falling below 0.2 are usually associated with bare soil cover (Lykhovyd, 2021). Monitoring NDVI over time is frequently used for assessing crop development and estimating potential yields (Khan, et al. 2025).

Normalized Difference Moisture Index (NDMI) assesses vegetation water content by analyzing the difference between Near-Infrared NIR and Shortwave Infrared Reflectance SWIR. This index helps in detecting plant water stress, which can adversely affect crop yields. By monitoring NDMI, agricultural producers can make informed irrigation decisions to mitigate water stress and optimize yield outcomes. Besides, NDMI is strongly related to real moisture storage in the upper layers of soil (Lykhovyd and Sharii, 2024).

Vegetation Health Index (VHI), which combines NDVI and temperature data to assess vegetation health and monitor drought conditions. It's particularly useful for early warning of crop stress due to unfavorable weather conditions and is a basis for drought effects estimation in other models and indices, e.g., Agricultural Stress Index by FAO (Rojas, 2020).

Land Surface Temperature (LST) provides valuable insights on soil moisture storages and indirectly characterizes evapotranspiration rates. High LST values can indicate water stress or drought conditions, which negatively impact crop growth and yields in the vulnerable crops cultivated in the rain fed conditions (Lee, et al. 2021).

Potential Evapotranspiration (PET) estimates the amount of water that would be evaporated and transpired by plants under optimal conditions. It's a key parameter for understanding crop water requirements and scheduling irrigation. Accurate PET assessments help in managing water resources efficiently, ensuring crops receive adequate moisture to achieve optimal yields. The best practice is on-land measurement and calculations based on the direct meteorological observations conducted in the field. However, standard methodology can be implemented the same way to derive PET values from remotely sensed meteorological data, providing the possibility to cut the expenses for the indicator estimation. For example, (Stisen, et al. 2008) demonstrated an approach of successful computation of PET through remote sensing data from Advanced Very High-Resolution Radiometer (AVHRR) sensors.

Remote sensing data combined with modern machine learning algorithms is a powerful tool for yield modeling and prediction. Machine learning is a subfield of artificial intelligence that focuses on developing algorithms capable of learning from data and making decisions or predictions without explicit programming. These algorithms identify patterns within datasets of different sizes, format, type and architecture, enabling systems to improve their performance over time as they are exposed to more data (Janiesch, et al. 2021). In recent years, machine learning has become integral to various industries, including agricultural sciences. Combined with remote sensing data, it becomes a powerful tool for robust crop modeling and productivity prediction on the areas of different scale (Jeong, et al. 2024). Therefore, the aim of this study was to evaluate the possibility of implementation remote sensing data, namely NDVI, NDMI, VHI, LST and PET, in crop yields prediction on regional scale for the territory of Ukraine and estimate the overall accuracy of different machine learning approaches, namely, linear regression, Random Forest and Gradient Boosting, to yield prediction by the commonly used statistical parameters as coefficient of determination R^2 , and prediction errors MAE and MSE.

Materials and Methods

The study was performed for the croplands of Ukraine for the period 2015-2023. The yields of twelve major crops, such as spring and winter wheat, spring and winter barley, rapeseed, rye, oats, peas, sunflower, maize, soybeans and sugar beets were retrieved from the official statistics provided by every region of Ukraine in the corresponding chapters of statistical yearbooks. The yields were converted into t/ha and generalized. The remote sensing data were retrieved from Google earth engine through the corresponding pull requests into the system, written in JavaScript. The data were obtained in monthly format for every year of the study period by every separate region of the country. Further, these data were generalized and recalculated for average annual values for every region to simplify mathematical analysis. The average annual data on the indicators, namely, NDVI, NDMI, VHI, LST and PET, was subject to correlation analysis, and corresponding correlation matrix of the inter-relationship between every index was built. The json files, storing the data on the indicators, were further converted into .csv format and fed into Python 3 ecosystem to build corresponding models for crop yield predictions through the common unmodified algorithms of linear regression (Maulud and Abdulazeez, 2020), Random Forest regression (Schonlau and Zou, 2020) and Gradient Boosting regression (Friedman, 2001) algorithms. All the models were evaluated using statistical parameters of p-value, as well as R^2 , MAE and MSE (Willmott, et al. 1985). Determination coefficient value was interpreted as follows (Saura and Andante, 2018): <0.20-very weak relationship; 0.20-0.30-weak relationship; 0.30-0.40-moderate relationship (reasonable model); 0.40-0.70-strong relationship (good model); >0.70-very strong relationship (very good model). If reasonable, mathematical equations of linear regression models were provided. Modeling and computations were performed in Python

3 environment. Average annual values of each indicator were applied to develop the models. In case of lacking data for statistical processing for some years, these years were skipped for the corresponding region of Ukraine. General methodological workflow of the study is presented in Fig. 1.

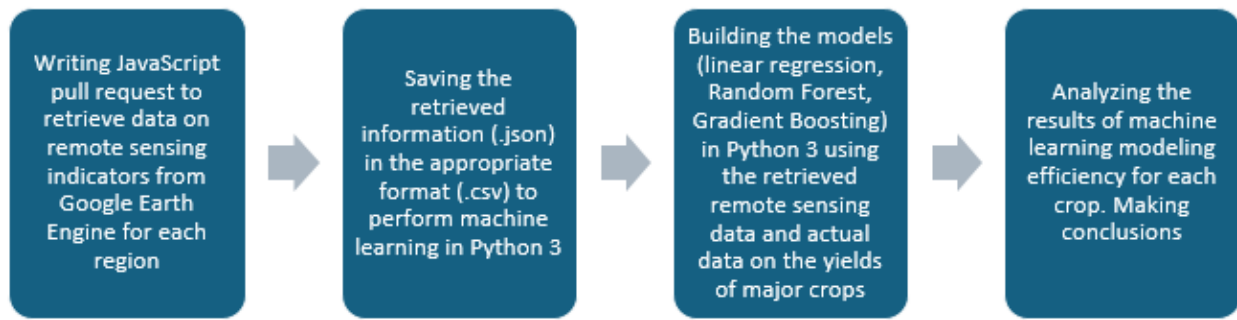


Figure 1. The code segment responsible for controlling the operation of the computer for the pre-sowing inoculation system.

Results and Discussion

First, it is necessary to analyse the correlation matrix for the studied remote sensing indicators (Tab. 1). The strongest correlation was established between NDVI and VHI, NDMI and LST and PET, LST and PET, and VHI and PET. This could be explained as follows. Strong direct relationship between NDVI and VHI is not surprising, as NDVI is a constituent for VHI calculation (Masitoh and Rusydi, 2019) and both indices are used to characterize general conditions of vegetation cover. The reverse strong correlation between NDMI and LST and PET could be put upon the fact that NDMI is an indicator of water stress in crops, and water stress is stronger with higher rates of PET and LST. PET and LST are strongly related because the higher the surface temperature is the greater evaporation rates are. Besides, strong correlation between VHI and PET is also explained by the affection of evaporation rates on plants health and vigor, because the higher PET is, the greater water stress is and as a result, plant vegetation conditions are altered (Gamal, et al. 2022).

Table 1. Correlation matrix for NDVI, VHI, NDMI, LST, and PET.

	NDVI	NDMI	VHI	LST	PET
NDVI	1	0.59	0.93	-0.5	-0.38
NDMI	0.59	1	0.43	-0.85	-0.85
VHI	0.93	0.43	1	-0.32	0.96
LST	-0.5	-0.85	-0.32	1	0.96
PET	-0.38	-0.85	-0.19	0.96	1

Most crops show positive correlations with NDVI, NDMI, and VHI. This is expected, as healthier, more vigorous vegetation (indicated by higher values of these indices) generally leads to higher crop yields. Spring wheat, spring barley, and winter barley show particularly strong positive correlations (Tab. 2). The greatest correlation with NDMI (meaning the strongest dependence on water supply) was established for spring barley, spring wheat and peas. The strongest correlation with NDVI was recorded for winter barley and spring wheat. Most crops show negative correlations with LST and PET. This suggests that higher land surface temperatures and higher potential evapotranspiration (leading to greater water stress) tend to be associated with lower crop yields. Sunflower and maize show particularly strong negative correlations with LST. The strongest negative correlation with PET was established for spring barley, sunflower and winter wheat, indicating high susceptibility of these crops to water stress events. The general trend is that higher NDVI, NDMI, and VHI, combined with lower LST and PET, are associated with higher yields for most crops.

Table 2. Pairwise correlation between NDVI, VHI, NDMI, LST, PET, and the yields of major crops.

	NDVI	NDMI	VHI	LST	PET
Spring wheat	0.48	0.48	0.44	-0.5	-0.53
Rapeseed	0.24	0.38	0.19	-0.47	-0.45
Spring barley	0.45	0.5	0.42	-0.55	-0.57
Winter barley	0.51	0.43	0.48	-0.43	-0.45

Maize	0.43	0.43	0.35	-0.57	-0.55
Sunflower	0.3	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	0.3	0.49	0.23	-0.54	-0.54

As for the comparison of three different approaches to crop yield prediction, it was observed that across most crops, the Random Forest model tends to have the highest R^2 and the lowest MAE and MSE, indicating superior predictive performance compared to the other two models (Tab. 3). Linear Regression generally exhibits the lowest R^2 and highest MAE/MSE, suggesting it's the least effective for this prediction task. This is likely due to the complex, non-linear relationships between remote sensing data and crop yields. However, it must be noted that linear regression outperformed more robust models in case of spring wheat, rapeseed and peas, which were represented by smaller datasets in comparison to other crops. It proves the hypothesis of no benefit of robust machine learning algorithms for small datasets (Jiao, et al. 2020). Gradient Boosting performs better than linear regression, but is often slightly outperformed by random forest, although in some cases, the results are comparable, as for example in the case of winter barley. For some crop/model combinations (especially Rapeseed and Peas with Random Forest and Gradient Boosting), the R^2 is negative. A negative R^2 indicates that the model performs worse than simply predicting the mean of the observed yields. This suggests potential issues with the models for these specific crops, such as overfitting.

Table 3. Comparison of linear regression, random forest and gradient boosting models for major crops yield prediction using remote sensing data.

Crop	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	0.2	-1.37	0.38	0.3
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	0.5	0.41
Maize	0.04	1.26	2.67	0.13	1.32	2.42	0.08	1.36	2.56
Sunflower	0.09	0.4	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	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	0.4	0.26	0.18	0.35	0.24	0.09	0.37	0.27
Rye	0.09	0.6	0.47	-0.04	0.59	0.6	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

Tab. 4 presents the best-performing multiple linear regression equations for predicting the Yield (Y) of three crops: spring wheat, rapeseed, and peas. These equations were selected based on statistical significance, meaning only variables with a p-value less than 0.05 were included in the final models. It is interesting that spring wheat's yield is better described by NDVI, and yields of rapeseed and peas are more dependent on indirect indicators affecting crop conditions, such as LST and PET. For every one-point increase in NDVI, the spring wheat yield is predicted to increase by 7.3537 t/ha. As for rapeseed, PET has a small, but still statistically significant, positive

coefficient (0.0003 t/ha), suggesting a slight increase in yield with increasing PET, although the effect is minimal and quite surprising, as most crops decrease yields with PET increment. However, as some studies suggest, it might be different for different crops cultivated in different soil and hydrologic conditions (Suliman, et al. 2024). LST, however, has a negative coefficient (-0.0986 t/ha), indicating that higher land surface temperatures are associated with lower rapeseed yields, which is commonly observed in practice. The yield of peas is predicted solely by PET. As PET increases by 1 mm, the predicted pea yield decreases by 0.0062 t/ha.

Table 4. Multiple linear regression models' equations for each crop where this method performed the best (the model equation was built including only regression coefficients with p-value<0.05)

Crop	Model equation
Spring wheat	$Y=0.3814+7.3537NDVI$
Rapeseed	$Y=3.9245+0.0003PET-0.0986LST$
Peas	$Y=7.0695-0.0062PET$

The presented study provides unique insights on using a combination of different remote sensing indicators to predict the yields of major crops, cultivated in Ukraine. Three popular machine learning approaches with a different level of robustness were implemented to derive crop yields depending on remote sensing data. The dataset itself is a unique one, providing the aggregated and generalized data, retrieved from Google earth engine, for Ukraine, including popular vegetation health indices (NDVI and VHI), crop water stress index (NDMI) as well as climatological indicators such as LST and PET. There are numerous studies of a comparative or even higher complexity level, but ours is different in terms of modeling approach, data set preparation, and remote sensing data used to predict crop yields. The most frequently used modeling approaches for yield prediction using aerospace images are artificial neural networks (Bayesian, long short-term memory, convolutional, deep, etc.). Different scientists used different inputs, e.g., NDVI, NDWI, EVI, LST, SAVI, and other remote sensing derived indicators, but never in such a combination as in the presented study (Muruganatham, et al. 2022). Thus, this research provides novel insights related to the questions of mutual relationship of the remotely sensed NDVI, NDMI, VHI, LST and PET, contributing to previously performed study on the inter-relationship between the vegetation indices (Lykhovyd, et al. 2023), as well as their relationship with the yields of twelve different crops, cultivated in Ukraine.

Conclusions

Remote sensing data is a crucial tool for monitoring vegetation cover and predicting crop yields. The combination of different remote sensing indicators, such as NDVI, NDMI, VHI, LST, and PET, provides comprehensive insights into crop health and yield predictions. Most crops are highly sensitive to Land Surface Temperature (LST) with negative correlations fluctuating between -0.36 and -0.62. The strongest positive correlation was established between crop yields and NDVI values, while the least relationship was attributed to VHI. Machine learning approaches, particularly Random Forest models, show promise in accurately predicting crop yields based on these indicators. However, for the crops with small initial datasets, e.g., spring wheat, rapeseed, and peas, multiple linear regression performed even better than more complicated and robust machine learning approaches as Random Forest and Gradient Boosting. The highest R^2 value achieved in the study was 0.63 in the Gradient Boosting model for sugar beets. The worst accuracy was established for rye. Therefore, the study underscores the importance of using advanced technologies and data analytics in modern agriculture to enhance crop monitoring and yield prediction, contributing to sustainable agricultural practices and food security.

References

- Friedman JH. (2001). Greedy function approximation: a gradient boosting machine. *Ann Stat.* **29**:1189-1232.
- Gamal R, El-Shirbeny M, Abou-Hadid A, Swelam A, El-Gindy AG, Arafa Y. (2022). Identification and quantification of actual evapotranspiration using integrated satellite data for sustainable water management in dry areas. *Agronomy.* **12**:2143.
- Janiesch C, Zscheck P, Heinrich K. (2021). Machine learning and deep learning. *Electron Mark.* **31**:685-695.
- Jeong S, Ko J, Ban JO, Shin T, Yeom JM. (2024). Deep learning-enhanced remote sensing-integrated crop modeling for rice yield prediction. *Ecol Inform.* **84**:102886.
- Jiao S, Gao Y, Feng J, Lei T, Yuan X. (2020). Does deep learning always outperform simple linear regression in optical imaging? *Opt Express.* **28**:3717-3731.
- Khan SN, Iqbal J, Khan MR, Malik NA, Khan FA, Khan K. (2025). Using remotely sensed vegetation indices and multi-stream deep learning improves county-level corn yield predictions. *Eur J Agron.* **164**:127496.
- Lee SJ, Kim N, Lee Y. (2021). Development of integrated crop drought index by combining rainfall, land surface temperature, evapotranspiration, soil moisture, and vegetation index for agricultural drought monitoring. *Remote Sens.* **13**:1778.
- Lykhovyd P, Averchev O, Fedorchuk M, Fedorchuk V. (2023). The relationship between spatial vegetation indices: A case study for the south of Ukraine. *Environ Ecol Res.* **11**:740-746.
- Lykhovyd PV, Sharii VO. (2024). Normalised difference moisture index in water stress assessment of maize crops. *Agronomy.* **7**:21-26.
- Lykhovyd PV. (2021). Seasonal dynamics of normalized difference vegetation index in some winter and spring crops in the South of Ukraine. *Agronomy.*

4:187-193.

Masitoh F, Rusydi AN. (2019). Vegetation Health Index (VHI) analysis during drought season in Brantas Watershed. *IOP Conf Ser Earth Environ Sci.* **389**:012033.

Maulud D, Abdulazeez AM. (2020). A review on linear regression comprehensive in machine learning. *J Appl Sci Technol Trends.* **1**:140-147.

Muruganatham P, Wibowo S, Grandhi S, Samrat NH, Islam N. (2022). A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sens.* **14**:1990.

Rojas O. (2020). Agricultural extreme drought assessment at global level using the FAO-Agricultural Stress Index System (ASIS). *Weather Clim Extrem.* **27**:100184.

Schonlau M, Zou RY. (2020). The random forest algorithm for statistical learning. *Stata J.* **20**:3-29.

Stisen S, Jensen KH, Sandholt I, Grimes DI. (2008). A remote sensing driven distributed hydrological model of the Senegal River basin. *J Hydrol.* **354**:131-148.

Suliman M, Scaini A, Manzoni S, Vico G. (2024). Soil properties modulate actual evapotranspiration and precipitation impacts on crop yields in the USA. *Sci Total Environ.* **949**:175172.

Willmott CJ, Ackleson SG, Davis RE, Feddema JJ, Klink KM, Legates DR. (1985). Statistics for the evaluation and comparison of models. *J Geophys Res Oceans.* **90**:8995-9005.