UDC 005.631.11:004
DOI: http://doi.org/10.17721/1728-2713.111.13

Vitaliy ZATSERKOVNYI[1], DSc (Engin.), Prof.
ORCID ID: 0009-0003-5187-6125
e-mail: vitalii.zatserkovnyi@gmail.com

Victor VOROKH[1], PhD Student
ORCID ID: 0009-0005-0112-8422
e-mail: fainkucha@gmail.com

Olga HLOBA[1], Student
ORCID ID: 0009-0003-4923-3374
e-mail: olgagloba73@knu.ua

Olesia LIASHCHENKO[1], PhD (Philol.), Assoc. Prof.,
ORCID ID: 0000-0003-4649-3667
e-mail: Lyashchenko1981@ukr.net

Iryna SIUIVA[1], PhD (Law), Assoc. Prof.,
ORCID ID: 0000-0002-5001-2750
e-mail: isiuiva.knu@gmail.com

[1]Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

APPLICATION OF MACHINE LEARNING METHODS AND REMOTE SENSING DATA FOR CROP YIELD FORECASTING

(Представлено членом редакційної колегії д-ром геол. наук, ст. дослідником Олександром МЕНЬШОВИМ)

Background. ***Forecasting agricultural crop yields has always been a complex task, particularly in the context of climate instability and increasing pressure on resources. Given the limitations of classical mathematical models in such a complex field as agricultural analytics, data-driven approaches and machine learning-based methods are becoming increasingly important. The combination of satellite imagery, agrochemical soil analysis, and artificial intelligence algorithms is particularly promising for building flexible and accurate forecasts.***

Methods. ***This study analyzes two agricultural fields located in different regions of Ukraine with varying natural conditions. A comprehensive dataset was collected, including topographic features (elevation, slope, topographic wetness index), spectral indices from Sentinel-2A and Landsat 8 satellites (specifically, NDMI and GNDVI), and soil chemical composition. Correlation analysis was used to identify which indicators are most closely associated with yield levels. Yield prediction models were developed using Random Forest and Gradient Boosting algorithms, adapted to field subplots of 5 ha and 1 ha.***

Results. ***The analysis revealed that vegetation condition and crop water balance (NDMI, GNDVI) are the most effective indicators in explaining yield variability. Meanwhile, surface temperature showed a clearly negative impact, suggesting potential heat stress during the grain filling periods. Gradient Boosting demonstrated particularly high sensitivity to spatial detail, reaching a prediction accuracy of R^2=0.801 at the 1 ha grid level. In contrast, Random Forest proved to be a robust method with lower sensitivity to data scale.***

Conclusions. ***The study demonstrates that combining satellite imagery, soil analysis results, and machine learning methods can significantly improve the accuracy of crop yield prediction. The developed models incorporate vegetation indices along with factors describing crop growing conditions. A comparison of various algorithms was also conducted under different levels of spatial detail. The results indicate that the proposed approach can be effectively applied in precision agriculture, particularly for agronomic planning and crop monitoring.***

Keywords: ***artificial intelligence (AI), machine learning (ML), remote sensing (RS), random forest (RF), gradient boosting (GB), normalized difference moisture index (NDMI), green normalized difference vegetation index (GNDVI), precision agriculture (PA), Correlation Analysis (CA).***

Background

Predicting potential crop yields, selecting suitable crops, assessing their profitability, and minimizing associated risks have been fundamental challenges in agriculture since its inception.

Forecasting is a crucial component of modern information technologies and decision support systems, applied both in the design of complex systems (such as energy, agrotechnical, and information and communication systems) and in their management under uncertainty. Subsequent agricultural planning involves developing crop rotation systems, pasture management plans, soil cultivation methods, fertilizer application strategies, and packages of agrotechnologies adapted to different intensification levels (Semeniaka et al., 2024).

Currently, preference is given to advanced technologies, particularly precision agriculture systems for crop management, which are based on satellite and computer technologies (Makedonska, Zatserkovnyi, & Tustanovska, 2018). Therefore, yield assessment and prediction can be approached in various ways. The application of traditional classical mathematical modeling approaches (such as systems of econometric equations, adaptive models, methods of nonlinear dynamics, etc.) does not always allow for obtaining adequate results, as they may lead to problems that cannot be solved by known methods or algorithms. Consequently, researching new classes of mathematical models is a relevant and promising task. In this context, artificial intelligence and machine learning algorithms present one of the most effective approaches. It is an approach in which historical data or examples are utilized to initially develop and subsequently improve the predictive model.

In Ukrainian studies, the effectiveness of indices like NDVI and NDMI as key yield predictors has been demonstrated (Kravchenko, & Danylenko, 2022; Ivanenko, & Sakhno, 2021). Their use enables the timely identification

ISSN 1728-3817

of drought-affected areas that impact crop yields. Similar conclusions are supported by international research as well. For instance, a model based on deep Gaussian processes achieved high accuracy in maize yield prediction (You et al., 2017), while the integration of NDVI, surface temperature, and field parameters have been justified as a foundational approach to predictive analytics (Lobell et al., 2015).

Machine learning (ML) approaches are applied across diverse domains-from retail (for analyzing customer behavior) (Ayodele, 2010) to agriculture (McQueen et al., 1995) and even telecommunications usage prediction (Witten et al., 2016). ML is a branch of artificial intelligence focused on developing algorithms that allow computers to learn from data without explicit programming. Within ML, computers "learn" to make predictions, recognize patterns, and support decision-making based on input data. ML is a valuable tool for determining which crops to cultivate and what agricultural operations to perform during the growing season. Under the pressures of global climate change and economic constraints, highly accurate crop yield forecasting becomes critical to effective farming. However, natural, biological, and technological factors exert complex and sometimes opposing influences on prediction outcomes, at times leading to forecast errors exceeding 15 %.

Depending on the research objective, ML models can be either descriptive or predictive. Descriptive models are used to gain insight from data and explain past events, while predictive models are intended to forecast future outcomes (Alpaydin, 2010). It is important to recognize that forecasting in precision agriculture is not a trivial task, as it depends on multiple datasets, including climatic conditions, weather, soil properties, fertilizer use, and seed variety (Xu et al., 2019).

ML, as a branch of AI, emphasizes learning and offers a practical approach for improving predictions, discovering patterns and correlations, and extracting knowledge from existing datasets. The application of ML algorithms to the analysis of these high-dimensional agricultural data opens new possibilities for enhancing forecast accuracy and optimizing agrotechnical operations.

Among spatial modeling approaches, Gradient Boosting and Random Forest methods have proven to be the most effective (Gnatienko, Sorochynskyi, & Derkach, 2024; Shin, Kim, & Lee, 2021), showing high resilience across varying spatial scales and adaptability to specific field conditions. Other researchers have employed multimodal approaches that integrate drone and satellite imagery (Maimaitijiang et al., 2020), or deep neural networks, which improved forecast accuracy by 10–15 % (Kuwata, & Shibasaki, 2015; Melnyk, 2023). The use of LSTM architectures has also shown promise in regional yield assessments (Melnyk, 2023), while a satellite image-based approach to crop classification in Ukraine achieved over 88 % accuracy (Kussul et al., 2017).

Precision agriculture includes Variable Rate Seeding, which allows for the site-specific adjustment of seeding and fertilization rates based on soil and field properties (Samko et al., 2025). Indeed, the correlation analysis of collected data in yield forecasting is a complex task requiring processing numerous parameters and precisely identifying the key factors that determine final outcomes. Crop yield is also significantly influenced by the topographical features of the site – such as elevation and slope, which determine the distribution of surface and groundwater, as well as local water balance and erosion conditions.

Methods

This study employs two machine learning methods: Random Forest and Gradient Boosting.

Random Forest (RF) is an ensemble method that constructs numerous individual decision trees, each trained on random subsets of features and data samples. The method operates as follows: for each tree, a random subset of the data (with replacement) and a subset of features are selected; the predictions of all trees are averaged (for regression) or determined by majority vote (for classification). Key advantages of this method include resistance to overfitting due to the result averaging, robustness when dealing with noisy or non-representative data, automatic handling of missing values, insensitivity to feature scaling, and the ability to assess the importance of individual features, thereby enhancing model interpretability (Breiman, 2001).

Gradient Boosting is a sequential ensemble technique where each new tree is trained to minimize the errors of the previous one by optimizing a loss function using gradient descent. Initially, a simple tree (e.g., with a depth of 1–3) is built, and at each step, a new tree is added that is trained on the residual errors of the model. This method achieves high predictive accuracy even on complex datasets and offers flexibility through hyperparameter tuning (learning rate, tree depth, number of iterations), making it effective for yield prediction tasks (Friedman, 2001). Crop condition, assessed using spectral indices (calculated from Sentinel-2A and Landsat 8 satellite imagery) and band combinations, reflects the growth stage and physiological status of crops in various zones of the field. Agronomic soil characteristics – such as the chemical composition of macro- and micronutrients, pH, and organic matter create the foundational conditions for plant nutrition. The integration of these three data groups provides a comprehensive overview of productivity factors and can serve as a basis for accurate yield forecasting and the optimization of differentiated agronomic measures (Table 1).

Field 1, with an area exceeding 309.6 hectares, is located in Varva settlement of the Pryluky district, Chernihiv region, near the village of Berizka (Fig. 1). The soils in the study area are predominantly typical low humus chernozems and degraded light loam chernozems, which require particular attention to the preservation of their fertility and structure. The terrain within the study area is heterogeneous, with elevation differences of up to 20 meters and gentle slopes with inclinations up to 4° (Zatserkovnyi et al., 2025a). The levels of potassium and phosphorus are moderately high. In terms of quality, the soils are average, with an average organic matter content of 3.7 %. The field is managed by the agricultural company "Kernel". The average corn grain yield on this field ranges from 7 to 9 t/ha.

Field 2, covering 119.2 hectares, is located near the village of Lysivtsi in the Tovste community of the Chortkiv district, Ternopil region (Fig. 2), and is cultivated by the scientific-production agri-enterprise "El Gaucho". The field mainly consists of podzolized chernozems, primarily formed on loess soils. The slope gradient ranges from 1° to 5°. The soils are of high quality (scoring between 70 and 80 points), with an estimated organic matter content of approximately 5.5 %. The potassium and phosphorus levels are high, and the soil pH is moderately alkaline.

To reduce the dimensionality of the input dataset and improve prediction efficiency, a preliminary analysis of the collected data is conducted to select the most informative parameters. One of the main methods used for this purpose is correlation analysis, which helps establish statistical relationships between different indicators. The correlation matrix reflects the degree of linear association between pairs of variables and is based on the Pearson correlation coefficient, which is calculated using formula (1).

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

where x_i and y_i – are individual observations (values) for the two variables X and Y, respectively. $\bar{x}$ and $\bar{y}$ – are the mean

values of all observations for the variables X and Y (Pearson correlation coefficient, n.d.).

Table 1

List of attributes to be used for calculations and modeling

Parameter	Description	Acronym
Relief		
Elevation	Absolute height above sea level, m.	El
Slope gradient	Surface slope angle (°)	Sl
Topographic Position Index	Difference between the elevation of a point and the mean elevation of its surrounding area	TPI
Topographic Wetness Index	The ratio of the specific catchment area to the tangent of the slope angle, used to assess the potential for moisture accumulation	TWI
Soil		
Exchangeable calcium.	Calcium concentration in soil exchangeable cations, mg/kg	Ca
Exchangeable potassium	Concentration of potassium in soil exchangeable cations, mg/kg.	K
Mobile phosphorus	Concentration of available phosphorus in soil, mg/kg	P
Exchangeable sodium	Concentration of sodium in soil exchangeable cations, mg/kg	Na
Exchangeable magnesium	Concentration of magnesium in soil exchangeable cations, mg/kg	Mg
Organic matter	Content of organic matter in soil, %	Org_M
pH (KCl)	Soil acidity extracted with KCl solution, pH units	pH_KCl
Crop condition		
Green vegetation index	Vegetation activity indicator, (NIR – GREEN) / (NIR + GREEN)	GNDVI
Surface temperature	Temperature of the land surface or vegetation derived from the satellite thermal channel (°C)	LST
Moisture index	Indicator of leaf moisture content in plants, (NIR – SWIR) / (NIR + SWIR)	NDMI

Fig. 1. Location of the study field No. 1 (authors' own elaboration based on OpenStreetMap data)

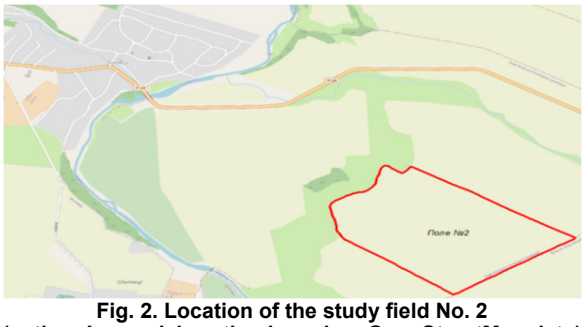


Fig. 2. Location of the study field No. 2 (authors' own elaboration based on OpenStreetMap data)

Results

In the case of the study field No. 1, the correlation matrix (Fig. 3) shows the strongest positive relationship between yield and GNDVI (r = 0.78), as well as between yield and NDMI (r = 0.77), highlighting the significant role of water balance and vegetation status in yield formation. During the active grain filling phase, these indices are directly related to the crop condition, which explains the high correlation coefficients. The TWI index shows a moderate positive correlation (r = 0.36). Soil water availability is a key factor; however, TWI is not a direct indicator of moisture but rather reflects the potential for its accumulation (depending on topography, drainage, etc.). Therefore, even if a location has the potential to retain water, it does not necessarily mean that moisture will actually be available, as this depends on rainfall, evaporation, and other factors.

Potassium (r = –0.49) shows a moderate negative correlation with yield. The surface temperature of the corn crops grown for grain (r = –0.59) indicates a strong negative correlation. Higher surface temperatures of crops often signal water stress. When a plant experiences water deficiency, its leaf and surface temperature rise. Thus, this strong negative relationship is entirely expected. Exchangeable magnesium, organic matter, and slope steepness indicators show the weakest correlations with yield. High potassium values may result from uneven fertilizer distribution or poor availability due to other factors, meaning that potassium is present but not effectively utilized. As a result, an inverse relationship may be observed: a high amount of potassium in certain areas does not indicate better yields and may, in fact, indicate the opposite.

Correlation matrix of the study field No. 2 (Fig. 4): The strongest positive correlation with yield is observed for NDMI (r = 0.50) and GNDVI (r = 0.52).

The terrain, specifically slope steepness (r = –0.45), shows a noticeable correlation with yield. This likely indicates that field No. 2 has more pronounced topographic contrasts, which have a stronger impact on yield compared to field No. 1. The surface temperature of the crops during the full grain filling period is negatively correlated with corn yield (r = –0.41), which is a typical result for corn under high-temperature conditions during this critical growth phase. Unlike in the previous field, magnesium shows a strong positive correlation here (r = 0.52). The weakest correlations with grain corn yield are observed for exchangeable sodium and mobile phosphorus.

The analysis of the correlation results revealed that mobile phosphorus, as well as exchangeable magnesium and sodium, exhibit weak correlations with corn yield in the study fields No. 1 or No. 2. Specifically, the correlation coefficients for these indicators fall within low ranges, indicating their limited explanatory power in the context of the selected dataset. According to standard statistical analysis procedures, variables showing weak correlation would typically be excluded from further modeling or forecasting schemes to improve accuracy. However, considering that these parameters are fundamental in agrochemical soil analysis and are traditionally used to assess yield potential, the decision was made to retain them in the study. It is likely that the weak correlation is due to the limited data volume or data quality. This suggests that with

a more representative dataset, collected over broader spatial or temporal scales, these indicators may demonstrate a more significant influence on corn yield.

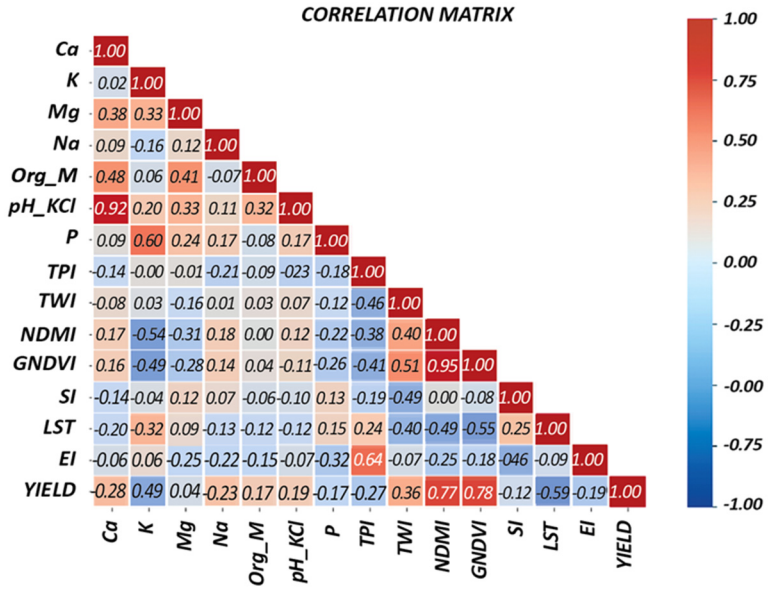


Fig. 3. Matrix of linear correlation coefficients for the study field No. 1 (authors' own elaboration)

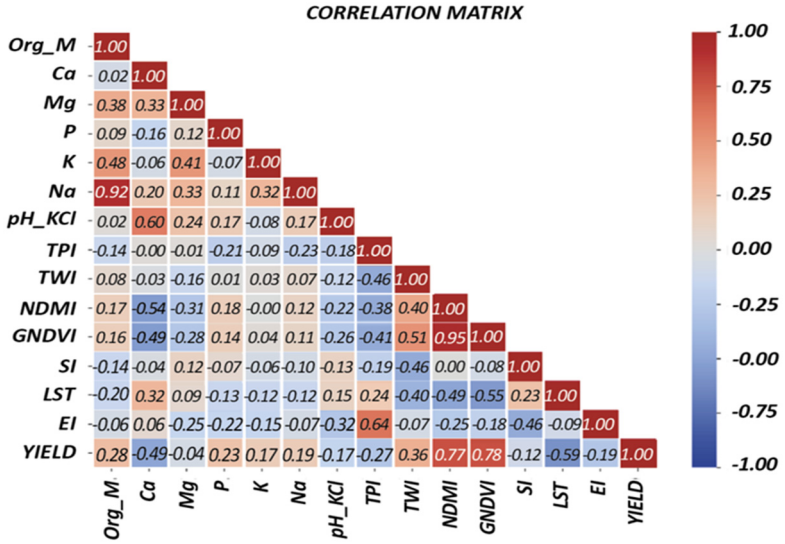


Fig. 4. Matrix of linear correlation coefficients for the study field No. 2 (authors' own elaboration)

Construction of yield prediction models for grain corn based on collected data from the fields studied.

To build a high-resolution yield prediction model, the study fields are divided into a certain number of subplots, and a separate forecast is generated for each. Let the actual yield value of the *i*-th subplot be denoted as y_i, the predicted value as – $\overline{y_i}$ (formula (2)).

$$\overline{y_i} = f_0(x_i), \quad (2)$$

where $x_i \in X$ – is the input data vector describing the condition of the *i*-th subplot, θ – represents the parameters of the yield prediction model, f – is the functional relationship between the input field condition data and the crop yield of the agricultural crop (Hnatiienko et al., 2024)

Training and validation of the model will be performed using training datasets, namely, data collected from fields No. 1 and No. 2, where each field plot includes yield data in tons per hectare. Since the study fields No. 1 and No. 2 are in different geographic zones and are managed by different agricultural producers, separate yield prediction models will be developed for each field, considering the parameter values (Table 1) for subplots with areas of 5 ha and 1 ha.

The code listings for the presented models are provided in the supplementary materials.

Validation of grain corn yield prediction models.

To ensure an objective assessment of the accuracy of the yield prediction models, validation of the results must be performed using statistical metrics. The application of quantitative accuracy indicators is a key stage in the study, as it enables the comparison of different models and the identification of the most effective prediction approaches.

In this study, the following commonly used metrics are applied to evaluate model performance: Root Mean Square

Error ($RMSE$) (formula (3)), Mean Absolute Error (MAE) (formula (4)), Mean Absolute Percentage Error ($MAPE$) (formula (5)), and Coefficient of Determination ($R2$) (formula (6)) (Kumar, 2024).

$$RMSE = \sqrt{\frac{\sum_i^n (y_i - \bar{y}_i)^2}{n}} \quad (3)$$

$$MAE = \frac{\sum_i^n |y_i - \bar{y}_i|}{n} \quad (4)$$

$$MAPE = \frac{1}{n}\sum_i^n \left|\frac{y_i - \bar{y}_i}{y_i}\right| \quad (5)$$

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)}{\sum_i^n (y_i - \bar{y}_i)} \quad (6)$$

As part of the study on the effectiveness of the grain corn yield prediction system, the results obtained are presented in Tables 2 and 3.

Analyzing the results, it can be noted that the Random Forest model demonstrates more stable accuracy when the scale of input data changes, compared to Gradient Boosting. Specifically, when transitioning from 5-hectare to 1-hectare subplots, the coefficient of determination R^2 for Random Forest increased from 0.614 to 0.776 (+0.162), whereas for Gradient Boosting, the increase was more significant – from 0.562 to 0.801 (+0.239).

A similar trend in performance improvement is observed across other accuracy metrics for the study field No. 1. The Mean Squared Error (MSE) for the Random Forest model decreased from 0.101 to 0.079 (approximately a 22% improvement), while for Gradient Boosting it decreased from 0.115 to 0.071 (approximately a 38 % improvement). This indicates that Gradient Boosting responds more strongly to an increase in data volume, showing a more substantial reduction in error when using a finer grid (1 ha).

It is also worth noting that the Mean Absolute Error (MAE) for Random Forest decreased from 0.276 to 0.234 (around a 15 % improvement), whereas for Gradient Boosting it dropped from 0.294 to 0.235 (around a 20 % improvement). Similarly, the Mean Absolute Percentage Error ($MAPE$) for both models declined to approximately the same level (0.029), though Gradient Boosting showed a greater improvement.

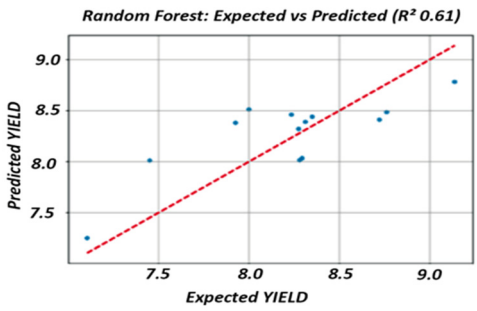
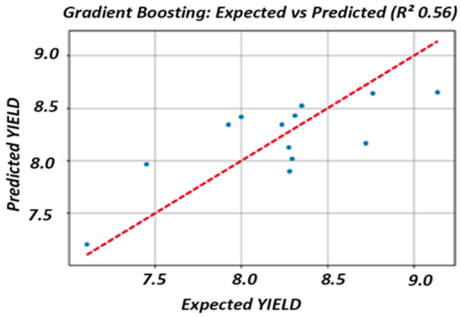
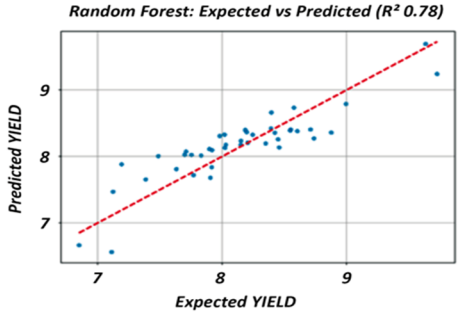
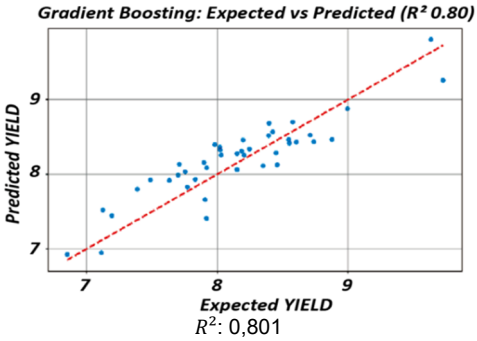
Therefore, the grid resolution (subplot size) matters when predicting yield, but its effect depends on the chosen machine learning algorithm.

The analysis of results for the study, field No. 2 confirms the patterns observed for field No. 1 (Table 3). Specifically, the Random Forest model (Fig. 5) demonstrates high accuracy stability regardless of subplot size: the increase in the coefficient of determination when reducing the grid scale is minimal (from 0.571 to 0.575). In contrast, Gradient Boosting remains more sensitive to spatial granularity: when moving from 5 ha to 1 ha subplots, the model's accuracy increases more significantly (from 0.480 to 0.647).

In conclusion, a finer grid generally improves prediction results, but the degree of this effect depends on the selected algorithm. This implies that the optimal subplot size should be chosen considering both the model's performance and the practical capabilities for data collection.

Table 2

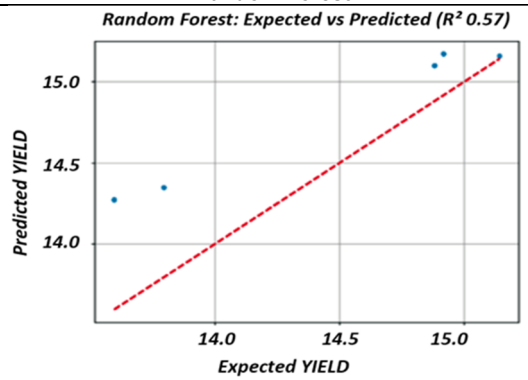
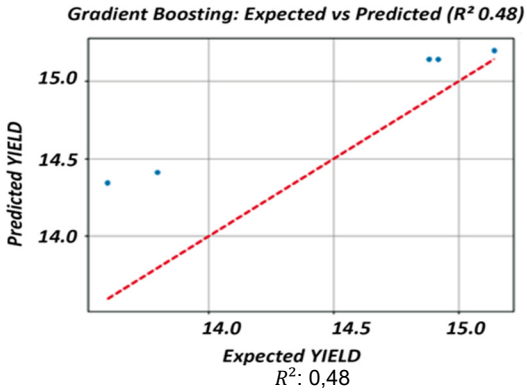
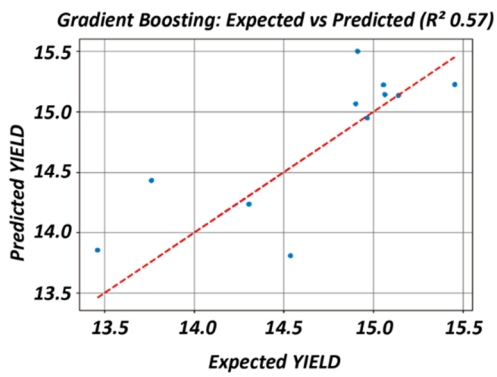
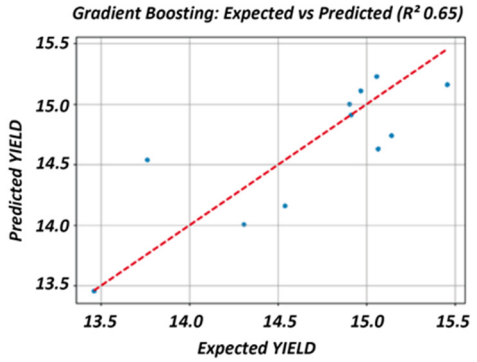
Validation results of grain corn yield prediction models for the study field No. 1 depending on the subplot size and the machine learning method

Subplot size	Machine learning method	
	Random Forest	Gradient Boosting
5 ha	 R^2: 0,614 MSE: 0,101 MAE: 0,276 $MAPE$: 0,034	 R^2: 0,562 MSE: 0,115 MAE: 0,294 $MAPE$: 0,036
1 ha	 R^2: 0,776 MSE: 0,079 MAE: 0,234 $MAPE$: 0,029	 R^2: 0,801 MSE: 0,071 MAE: 0,235 $MAPE$: 0,029

Source: authors' own elaboration.

Table 3

Validation results of grain corn yield prediction models for the study field No. 2 depending on the subplot size and the machine learning method

Subplot size	Machine learning method	
	Random Forest	Gradient Boosting
5 ha	 R^2: 0,571 MSE: 0,176 MAE: 0,344 $MAPE$: 0,025	 R^2: 0,48 MSE: 0,213 MAE: 0,381 $MAPE$: 0,027
1 ha	 R^2: 0,575 MSE: 0,146 MAE: 0,283 $MAPE$: 0,02	 R^2: 0,647 MSE: 0,121 MAE: 0,273 $MAPE$: 0,019

Source: authors' own elaboration.

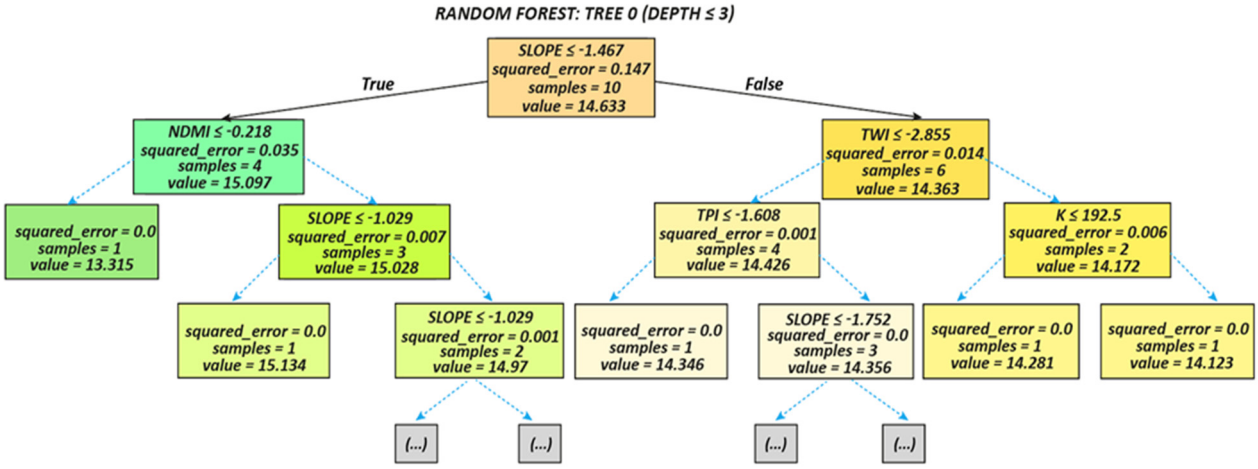


Fig. 5. Example of the first decision tree for the study field No. 2 with 5 ha subplots (authors' own elaboration)

Based on the developed machine learning models, grain corn yield prediction was carried out for two agricultural fields with different levels of spatial detail (grids of 1 ha, 2 ha, and 5 ha). The predictors included the vegetation indices GNDVI and NDMI, as well as concomitant variables: soil moisture indicators, soil types, temperature regime, and crop rotation information. Model accuracy was evaluated by comparing the predicted and actual yield values.

For field No. 1, the most accurate results were achieved by the Gradient Boosting model using a 1 ha grid: the

predicted yield was 5.67 t/ha, which was only 0.16 t/ha lower than the actual value of 5.83 t/ha. As the level of spatial aggregation increased, accuracy slightly decreased: 5.60 t/ha at 2 ha and 5.63 t/ha at 5 ha. The Random Forest model, in turn, produced less accurate predictions ranging from 5.55 to 5.61 t/ha for the different grids, with a maximum deviation of 0.28 t/ha.

For field No. 2, the Gradient Boosting model again proved to be the most effective at the 1 ha resolution: the predicted yield was 5.39 t/ha versus an actual yield of 5.52 t/ha (an error of 0.13 t/ha). Other spatial resolutions produced similar results: 5.34 t/ha (2 ha) and 5.37 t/ha (5 ha). The Random Forest model's predictions ranged between 5.29 and 5.35 t/ha.

The obtained results confirm the suitability of Gradient Boosting for accurate yield prediction, especially under conditions of high spatial resolution. In contrast, Random Forest provides stable yet slightly less precise estimates, which can be beneficial in situations with limited access to high-resolution data.

Discussion and conclusions

A comprehensive analysis of the factors influencing grain corn yield was conducted using remote sensing data, agrochemical soil analysis results, and topographic characteristics. Correlation analysis made it possible to identify key indicators with the strongest impact on yield, particularly the spectral indices (GNDVI, NDMI).

Yield prediction models were constructed using machine learning methods – Random Forest and Gradient Boosting – for two study fields, with subplot sizes of 5 ha and 1 ha. Their performance was evaluated using the Mean Absolute Error (MAE) and the Coefficient of Determination (R^2), enabling the identification of each method's advantages and limitations under conditions of spatially heterogeneous agrolandscapes. Validation results demonstrated that Gradient Boosting is more sensitive to spatial data granularity, significantly improving prediction accuracy with finer grids (1 ha). On the other hand, Random Forest provides consistent results even at coarser spatial resolution, highlighting its robustness when data availability is limited or data volume is smaller.

Overall, the study confirms the importance of integrating multicomponent datasets and considering the field spatial structure to improve yield prediction accuracy. The developed methodology has strong potential for implementation in precision agriculture systems to optimize agronomic practices and enhance resource management efficiency.

Authors' contribution: Vitalii Zatserkovnyi – conceptualization, formal analysis, methodology, review and editing; Viktor Vorokh – conceptualization, methodology; Olga Hloba – formal analysis, data processing; Iryna Siuiva – editing, publication analysis, Olesia Liashchenko – review and editing.

References

Alpaydin, E. (2010). *Introduction to machine learning* (2nd ed.). The MIT Press.

Ayodele, T. O. (2010). *Introduction to machine learning*. IntechOpen. https://www.intechopen.com/chapters/10703

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Friedman, H. J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gnatienko, I., Sorochynskyi, I., & Derkach, M. (2024). Comparative analysis of ensemble learning methods for yield prediction based on remote sensing. *Journal of Precision Agriculture and Data Science*, 12(1), 45–58 [in Ukrainian]. [Гнатієнко, І., Сорочинський, І., & Деркач, М. (2024). Порівняльний аналіз методів ансамблевого навчання для прогнозування врожайності на основі дистанційного зондування. *Журнал прецизійного землеробства та наук про дані*, *12*(1), 45–58].

Ivanenko, O., & Sakhno, S. (2021). Use of NDVI and NDMI indices for crop yield estimation in Ukrainian agricultural systems. *Agroecology and Land Management*, *24*(3), 71–78. [In Ukrainian]. [Іваненко, О., & Сахно, С. (2021). Використання індексів NDVI та NDMI для оцінки врожайності в агросистемах України. *Агроекологія і землекористування*, 24(3), 71–78].

Kravchenko, P., & Danylenko, V. (2022). Assessment of vegetation stress indicators for yield forecasting in arid zones of Ukraine. *Ukrainian Journal of Remote Sensing*, *10*(2), 33–41. [In Ukrainian]. [Кравченко, П., & Даниленко, В. (2022). Оцінка індикаторів вегетаційного стресу для прогнозування врожайності в посушливих регіонах України. *Український журнал дистанційного зондування*, 10(2), 33–41].

Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14(5), 778–782. https://doi.org/10.1109/LGRS.2017.2681128

Kuwata, K., & Shibasaki, R. (2015). Estimating crop yields with deep learning and remotely sensed data. In *Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 858–861). IEEE. https://doi.org/10.1109/IGARSS.2015.7325924

Lobell, D. B., Thau, D., Seifert, C., Engle, E., & Little, B. (2015). A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*, 164, 324–333. https://doi.org/10.1016/j.rse.2015.04.021

Maimaitijiang, M., Sagan, V., Sidike, P., Maimaitiyiming, M., Hartling, S., Peterson, K. T., & Fritschi, F. B. (2020). Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote Sensing of Environment*, 237, 111599. https://doi.org/10.1016/j.rse.2019.111599

Makedonska, I. O., Zatserkovnyi, V. I., & Tustanovska, L. V. (2018). Application of GIS technologies and remote sensing in precision farming. In *17th International Conference on Geoinformatics – Theoretical and Applied Aspects*. EAGE Publications BV. https://doi.org/10.3997/2214-4609.201801835

McQueen, R. J., Garner, S. R., Nevill-Manning, C. G., & Witten, I. H. (1995). Applying machine learning to agricultural data. *Computers and Electronics in Agriculture*, 12(4), 275–293. https://doi.org/10.1016/0168-1699(95)98601-9

Melnyk, O. (2023). Application of LSTM neural networks for regional crop yield forecasting based on EO data. *Environmental Modeling and Assessment*, *28*(1), 109–123 [in Ukrainian]. [Мельник, О. (2023). Застосування нейронних мереж LSTM для регіонального прогнозування врожайності на основі даних дистанційного зондування. *Екологічне моделювання та оцінювання*, 28(1), 109–123].

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58, 240–242.

Samko, M., Zatserkovnyi, V. I., Vorokh, V., Tsyguliov, I., & Ilchenko, A. (2025, April). *Monitoring using UAVs in precision farming technologies* [Conference presentation]. XVIII International Scientific Conference "Monitoring of Geological Processes and Ecological Condition of the Environment", Kyiv, Ukraine.

Semeniaka, V., Zatserkovnyi, V. I., Vorokh, V. V., Ilyin, L., & Myronchuk, T. (2024, October 7). *Differential technologies for precision agriculture* [Conference paper]. International Conference of Young Professionals "GeoTerrace-2024", European Association of Geoscientists & Engineers. https://doi.org/10.3997/2214-4609.2024510098

Shin, J., Kim, N., & Lee, H. (2021). Evaluation of machine learning models for predicting crop yield at field level. *Computers and Electronics in Agriculture*, 182, 106037. https://doi.org/10.1016/j.compag.2021.106037

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann. https://doi.org/10.1016/c2009-0-19715-5

Xu, X., Gao, P., Zhu, X., Guo, W., Ding, J., Li, C., & Wu, X. (2019). Design of an integrated climatic assessment indicator (ICAI) for wheat production: A case study in Jiangsu Province, China. *Ecological Indicators*, 101, 943–953. https://doi.org/10.1016/j.ecolind.2019.01.059

You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep Gaussian processes for crop yield prediction based on remote sensing data. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence* (pp. 4559–4566). AAAI Press. https://doi.org/10.1609/aaai.v31i1.11273

Zatserkovnyi, V., Vorokh, V., Hloba, O., Mironchuk, T., & Siuiva, I. (2025). Agrochemical analysis of soils in precision farming technologies: a case study of the Chernihiv region. *Visnyk of Taras Shevchenko National University of Kyiv. Geology*, *1*(108), 85-93. https://doi.org/10.17721/1728-2713.108.12

Отримано редакцією журналу / Received: 12.09.25
Прорецензовано / Revised: 28.10.25
Схвалено до друку / Accepted: 16.12.25

Віталій ЗАЦЕРКОВНИЙ[1], д-р техн. наук, проф.
ORCID ID: 0009-0003-5187-6125
e-mail: vitalii.zatserkovnyi@gmail.com

Віктор ВОРОХ[1], асп.
ORCID ID: 0009-0005-0112-8422
e-mail: fainkucha@gmail.com

Ольга ГЛОБА[1], студ.
ORCID ID: 0009-0003-4923-3374
e-mail: olgagloba73@knu.ua

Олеся ЛЯЩЕНКО[1], канд. філол. наук, доц.
ORCID ID: 0000-0003-4649-3667
e-mail: Lyashchenko1981@ukr.net

Ірина СЮЙВА[1], канд. юр. наук, доц.
ORCID ID: 0000-0002-5001-2750
e-mail: isiuiva.knu@gmail.com

[1]Київський національний університет імені Тараса Шевченка, Київ, Україна

ЗАСТОСУВАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ТА ДАНИХ ДИСТАНЦІЙНОГО ЗОНДУВАННЯ ЗЕМЛІ В ПРОГНОЗУВАННІ ВРОЖАЙНОСТІ

В с т у п . ***Прогнозування врожайності сільськогосподарських культур завжди було непростим завданням, особливо в умовах кліматичної нестабільності та зростаючого тиску на ресурси. Зважаючи на обмеження класичних математичних моделей у такій складній галузі, як аграрна аналітика, нині все більшої ваги набувають підходи, основані на даних і машинному навчанні. Особливо перспективним виглядає поєднання супутникових знімків, агрохімічного аналізу ґрунтів та алгоритмів штучного інтелекту для побудови гнучких і точних прогнозів.***

М е т о д и . ***Проаналізовано два сільськогосподарські поля, розташовані в різних регіонах України, із різними природними умовами. Було зібрано масив даних: топографічні параметри (висота, ухил, водозбірний потенціал), спектральні індекси із супутників Sentinel-2A та Landsat 8 (зокрема, NDMI, GNDVI), а також хімічний склад ґрунту. За допомогою кореляційного аналізу визначено, які саме показники тісніше пов'язані з рівнем урожайності. Побудовано моделі прогнозу врожайності на основі Random Forest та Gradient Boosting, з адаптацією під розділення полів на підділянки розміром 5 та 1 га.***

Р е з у л ь т а т и . ***Аналіз показав, що стан вегетації та водний баланс культури (NDMI, GNDVI) найкраще пояснюють варіації врожайності. Водночас такі показники, як температура поверхні, мають чіткий негативний вплив, що може вказувати на тепловий стрес у періоди наливу зерна. Gradient Boosting продемонстрував особливо добру чутливість до просторової деталізації – на сітці 1 га точність прогнозу досягала $R^2 = 0,801$. Натомість Random Forest показав себе як стійкий і менш чутливий до масштабу даних метод.***

В и с н о в к и . ***У досліджені доведено, що поєднання супутникових знімків, результатів аналізу ґрунтів та методів машинного навчання дає змогу поліпшити точність прогнозування врожайності. У моделі включено показники вегетації та характеристики умов вирощування культур. Також проведено порівняння різних алгоритмів при різній деталізації просторових даних. Отримані результати свідчать про те, що запропонований підхід може бути корисним у практиці точного землеробства, особливо для агрономічного планування та моніторингу посівів.***

К л ю ч о в і с л о в а : ***штучний інтелект (ШІ), машинне навчання (ML), дистанційне зондування Землі (ДЗЗ), random forest (RF), gradient boosting (GB), normalized difference moisture index (NDMI), green normalized difference vegetation index (GNDVI), прецизійне землеробство (PZ), кореляційний аналіз (КА).***

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.