DOI: 10.32347/2412-9933.2024.59.225-234
UDC 004.9

Hnatiienko Vladyslav
Researcher trainee, *https://orcid.org/0009-0000-2678-5158*
Syngenta LLC*, Kyiv*
Hnatiienko Hryhorii
PhD, associate professor department of intellectual technologies,
https://orcid.org/0000-0002-0465-5018
Taras Shevchenko National University of Kyiv, Kyiv

INTEGRATION OF MACHINE LEARNING AND DEEP LEARNING METHODS FOR SUNFLOWER YIELD PREDICTION

Abstract. *Practical experience in yield forecasting demonstrates that it is a complex multifactorial task requiring precise and reliable methods for resolution. The use of machine learning and deep learning is crucial for achieving better results in digital agronomy. This paper is dedicated to the development of an intelligent system for sunflower yield forecasting. Based on the analysis of scientific publications and practical experience, the main issues of data processing are summarized, and schemes for their resolution are proposed. The main stages of the work included the study of the current state of digital agronomy, the selection of an approach, the development of a data processing method, the software implementation of the model, and testing. The key challenges were the limited data sets and the complexity of choosing the optimal approach to avoid overfitting. Combining different analysis methods allowed the creation of a powerful system that surpasses traditional approaches, though it requires more data for training. The study's conclusions show that machine learning and deep learning methods, such as LightGBM and U-Net, along with the proposed data processing methods, achieve high accuracy in forecasting. The model demonstrated the ability to generalize knowledge to new fields and to build detailed yield maps. Further research includes the development of a method for generating combinations of plant care options, the adaptation of computer vision methods with optimized algorithms to reduce computational complexity, and the expansion of the system's functionality to include cybersecurity aspects. The proposed system significantly enhances the efficiency of sunflower yield forecasting, contributing to the development of digital agronomy.*

Keywords: Agricultural sector; yield forecasting; satellite data; machine learning; computer vision

Introduction

Today, digital agronomy is in a stage of development and faces challenges, particularly in yield forecasting using modern technologies. The capabilities of forecasting methods in the agricultural sector are quite limited. Traditional statistical methods provide only approximate values of potential yield, while artificial intelligence-based methods currently cannot guarantee high reliability and informativeness, as these technologies are aimed at forecasting the total yield of the entire field, which significantly limits their application. Therefore, considerable attention is paid to the implementation of artificial intelligence in the agricultural sphere, as this direction has great potential, although it requires further research and improvement.

The key task of this study is to develop a system that includes an optimization approach to data processing and artificial intelligence methods for forecasting, which in combination should effectively perform both local forecasts and global analysis of agricultural lands. This approach will achieve high forecasting accuracy by recognizing and accounting for complex vegetation development patterns, going beyond traditional analysis.

It is necessary to create a model capable of forecasting yield with high accuracy and reliability, based on comprehensive data analysis. To achieve this, a large number of information sources need to be considered, including satellite images in various spectra, meteorological data, as well as specific field and crop characteristics such as plant hybrid, planting density, planting date, and chemical treatment history.

Based on these data, an approach will be developed that takes into account key factors of agricultural production, from the genetic characteristics of crops to environmental conditions, ensuring high forecasting accuracy.

Review of Existing Research

The most studied are traditional yield forecasting methods. These methods are based on the application of empirically established relationships determined through

correlation analysis. They rely on historical data and statistical studies of the interconnections between various agronomic factors and the final field productivity. Classical machine learning methods, such as decision trees, support vector machines (SVM), linear and multivariate regression, are often used to analyze these dependencies and build forecasts. These methods allow modeling complex dependencies between data, although they have limitations in accounting for nonlinear dependencies.

Previous studies have used vegetation indices to establish relationships between them and the final field productivity. By analyzing the correlation between vegetation indices and actual yield, empirical equations were developed that formed the basis of forecasting models. This process involved directly deriving dependencies from the collected data on the relationship between indices and yield without resorting to complex machine learning methods [1].

Researchers have also actively applied machine learning methods to improve yield forecasting accuracy, particularly through the efficient selection of key parameters influencing the result. The use of regression models allowed for a detailed analysis of the dependencies between a set of input data and the final yield. This approach contributed to the development of more accurate forecasting models and opened up possibilities for a deeper understanding of the impact of various agronomic and environmental factors on crop productivity [2]. However, a drawback of these methods is their limited ability to generalize to new data, particularly information about fields outside the training sample.

To identify more complex relationships in the data, such as weather conditions and plant genotype, deep learning methods were used. The development of neural networks, as well as their combination in complex systems with integrated networks [3], allowed for deeper and more specific information processing. This approach ensured the detection of hidden dependencies between various factors and the final field productivity, opening new opportunities for accurate yield forecasting.

However, despite the high efficiency of deep learning methods, researchers have not considered the possibility of detailing forecasts to individual local areas of the field. Developed mathematical models focus on determining the overall total yield at the entire field level, without taking into account the potential heterogeneity of conditions or productivity of different areas. This limitation reduces the potential use of these models for precise agricultural management.

In their studies, specialists have resorted not only to supervised learning methods but also applied reinforcement learning approaches, expanding the capabilities of deep neural networks. In particular, the use of recurrent neural networks in combination with the q-learning method achieved an average forecasting accuracy of 93.7% [4], which is a significant achievement for solving such a complex problem as yield forecasting. However, this study focused on forecasting the total yield, not extending its focus to detailed forecasting for individual areas. Although theoretically, the model could be adapted for such detailed forecasting, researchers did not provide information on conducting such experiments.

Traditional methods also include time series analysis using the ARIMA method [5] and statistical methods, in particular, kriging [6], which is often applied in cases where the data is presented in a specific format that requires the adjustment of the corresponding mathematical model.

It is also worth noting that there are systems that provide paid access for forecasting [7, 8] and for detailed analysis of plant health status [9]. However, specific accuracy indicators or examples of forecasts are not provided in open access.

Definition of Research Direction

In this study, a comprehensive approach combining decision tree-based methods with computer vision techniques was chosen for forecasting. This approach aims to leverage the advantages of both methods to create more accurate and reliable predictions. The systems are intended to complement each other, enhancing capabilities and increasing the stability of forecasts. The decision tree method ensures high prediction accuracy based on numerical and categorical data, while computer vision techniques provide additional information through image analysis, allowing for consideration of visual aspects of field conditions.

The models will analyze different aspects of information using their unique interpretations. Decision trees can effectively handle large volumes of data, considering the impact of numerous factors such as weather conditions, soil type, and agronomic practices. Simultaneously, computer vision methods analyze satellite images of fields, identifying visual patterns that may indicate the state of vegetation and potential issues. Their combination will enable a comprehensive examination of available data and a more holistic analysis of field vegetation, thereby creating the prerequisites for effective and accurate forecasting.

Main part of the research

Considering the mentioned limitations imposed by the unresolved issue of detailed forecasting and the imperfection of current approaches, solving it will be a significant step in the development of digital agronomy. Implementing accurate and detailed forecasting will improve the efficiency of agronomic production processes, optimize resource use, and reduce costs. This will enable more informed decisions regarding planting,

fertilization, irrigation, and harvesting, ultimately leading to increased yield and product quality.

From a scientific progress perspective, solving this problem will make a significant contribution to the development of agronomic sciences. A system capable of providing detailed forecasts with high accuracy will exceed the capabilities of current solutions. This will enhance current agronomic practices and create a foundation for further research in this direction. Such a breakthrough in forecasting could open new horizons for research in optimizing agrotechnical measures, managing natural resources, and adapting to climate change.

The aim of this study is to develop highly accurate yield forecasting methods, which includes creating detailed forecast maps. The main objective is to ensure the reliability and accuracy of the obtained forecasts by developing an efficient model capable of accounting for various factors affecting yield. The study aims to achieve high forecasting accuracy and create tools that will allow these forecasts to be visualized as maps, reflecting the local characteristics of fields.

The main tasks of the study are:

– To develop a data processing algorithm to maximize the accuracy that models can achieve when trained on this data.

– To develop an efficient architecture for the forecasting system that can fully exploit the potential of the models used.

– To implement a yield forecasting system that provides highly accurate detailed forecasts.

– To develop a user interface to simplify information presentation and ease of experimentation.

– To analyze the obtained results and propose possible further improvements.

For yield forecasting, it is important to consider factors that directly affect plant conditions and, consequently, their development: information about weather conditions and the application of chemical treatments in the field. Additionally, monitoring indicators that provide expanded information about the field and the plants within it are crucial: plant hybrid, sowing date, sowing density, and satellite images.

Research on yield forecasting demonstrates that the development and health status of plants can be effectively tracked by analyzing the amount of absorbed solar radiation [1; 2], which is supported by high correlations and the ability of artificial intelligence models to make accurate predictions based on this information [3; 4]. This is explained by the fact that plants use light for photosynthesis, and thus the amount of absorbed light directly correlates with their health and productivity. Photosynthesis is a key process in the life of a plant, during which sunlight is used to convert carbon dioxide and water into oxygen and glucose, which serves as an energy resource for growth and development. Hence, the amount of absorbed solar radiation directly affects the efficiency of photosynthesis: the more light is absorbed, the more energy can be synthesized by the plant. This, in turn, promotes better growth, development, and health of the plant, providing it with the necessary conditions to achieve maximum productivity. Therefore, by tracking the amount of absorbed solar radiation, valuable information about the condition of the plants and their yield potential can be obtained.

By capturing reflected solar radiation, one can get an idea of the extent to which it has been absorbed by the plant. Reflected and absorbed solar radiation together form the overall intensity of sunlight, which is relatively stable. Even under conditions of limited resolution – in this study, images with a resolution of 10 by 10 meters are used – the obtained data provide valuable insights into plant development, as confirmed by high correlations achieved even when analyzing only a single vegetation index [1].

In this study, vegetation indices are actively used for the analysis of plant conditions, particularly NDVI, NDWI, GLI, CLg, and CLr. These indices are crucial for accurate yield forecasting as they reflect various aspects of the physiological state of plants. These indices focus on measuring the chlorophyll content in leaves, which is an indicator of photosynthetic activity and the overall health of plants [10].

Combining these indices allows for a comprehensive analysis of the impact of different weather conditions and management interventions on yield, enabling an accurate determination of plant conditions at any stage of their development. Such an approach is critical for developing effective yield forecasting models.

Ideally, it would be appropriate to analyze satellite data in its pure, unmodified form, which would provide the most complete picture of vegetation status and its interaction with the environment.

However, in practice, access to pure, unprocessed satellite data is often limited. Practical aspects of obtaining and using satellite data often require compromises, particularly due to limitations imposed by data providers. The satellite data for this study is obtained from SkyGlyph, which provides it exclusively in the form of vegetation indices among the available options.

The aim of this study is to develop a system to enhance the accuracy of yield forecasting. The main focus is on minimizing deviations between the predicted and actual yield figures. The task can be formulated as follows.

To construct a detailed yield map, the field Y is divided into individual plots $y_i \in Y$, $i = \overline{1,n}$, where n is the number of plots in the field Y, and a separate forecast is made for each plot. Let the predicted yield value be denoted as

$$\hat{y}_i = f_\theta(x_i),$$

where $x_i \in X$ – the vector of input information describing the state of plot i during the period from sowing to the hundredth day after the sowing date inclusive;

X – data matrix of all field plots Y;

θ – forecast model parameters;

f – the functional relationship between the input data of the field state and the yield, which is established by machine learning methods.

Then the forecasting task can be described as follows:

$$L(f_\theta(x_i), y_i) \to min, i = \overline{1,n},$$

where L – loss function that reflects the deviation of the predicted values from the actual ones.

The input data for yield forecasting is extremely voluminous and multidimensional. It encompasses a wide variety of data, from weather conditions to specifics of agronomic practices. To build an effective forecasting model based on such data, a significant number of observations need to be conducted. The complexity of the information requires large volumes of data because the more complex the information, the harder it is to identify hidden dependencies and patterns for quality training of the artificial intelligence model.

The need for a large number of observations for analyzing multidimensional data in agronomic research can be explained by the significant complexity of interactions that exist between different agronomic factors. Each of these factors-whether it be temperature, humidity, soil type, precipitation levels, or solar radiation – can uniquely affect yield. For the model to adequately reflect these complex interactions and variability in conditions, it must be trained on a sufficiently large amount of data that provides a representative coverage of all possible scenarios.

Additionally, each dimension in a multidimensional dataset can contribute additional information that needs to be integrated and analyzed in the context of other dimensions. For example, the dependency of yield on temperature may vary depending on soil moisture levels or how these conditions have changed throughout the growing season. Such multi-level dependency makes it impossible to adequately generalize based on a small amount of data, as this may lead to underestimating important interactions or overestimating less significant relationships.

Furthermore, to identify statistically significant patterns in multidimensional data, it is necessary to have enough data to reduce noise and detect true patterns instead of random fluctuations. Without a large number of observations, models can become overfitted, meaning they become too specialized to the dataset they were trained on and lose the ability to generalize to new data or conditions, significantly reducing the quality of predictions.

In this study, the sample size is limited due to difficulties associated with obtaining data on harvested crops, which requires individual negotiations with each farm, and the lack of automation in the process of uploading satellite images, leading to significant time costs for replenishing the training sample. Consequently, to effectively realize the potential of the available information in the context of accurate yield forecasting, a number of simplifications have been proposed. These simplifications aim to optimize the model training process, allowing for more accurate predictions based on a limited amount of data.

Thus, the main task of data preprocessing is to minimize their dimensionality while preserving the maximum amount of information contained in them. The process includes correlation analysis to select important features, dimensionality reduction by finding minimum, mean, and maximum values, and a two-stage outlier removal based on z-score [14]. The algorithm is presented in the study [11].

For forecasting the yield of individual plots in this study, the LightGBM model [12] is applied. Given that the number of available observations is limited, the use of overly complex models, such as deep neural networks, is found to be ineffective. This is because complex models, under conditions of limited data, often lead to overfitting, where the model adapts too closely to the training sample and loses the ability to generalize to new data.

Choosing an ensemble model of decision trees, such as LightGBM, is a justified decision under these conditions. This model allows for effective use of limited data without overfitting due to its structure and optimization algorithms. While LightGBM may not detect very complex hidden patterns in the data as some other artificial intelligence models can, it is effective in finding fundamental dependencies, enabling sufficiently high forecasting accuracy. Thanks to its properties, LightGBM ensures high efficiency in forecasting and generalizing knowledge in unknown conditions. Therefore, LightGBM becomes the optimal choice for solving the task of predicting the yield of individual plots.

To compensate for the limitations associated with the local nature of LightGBM's forecasting, which does not take into account contextual information about surrounding land plots, the study employs a computer vision model with a U-Net architecture [15]. This model can segment the field into productivity zones using satellite imagery data, thereby providing the ability to assess the overall condition of the zone in which a specific plot is located.

Through field segmentation, the U-Net model allows for the identification of different productivity zones, taking into account direct information about the 10 by 10 meter plot and details about the condition of neighboring plants, possible soil density fluctuations, the

presence of underground water flows, the overall adaptation of plants to specific growing conditions, and other indicators that can be inferred from visual information. This information enables a "broader view" and a better understanding of the general conditions for plant development in the studied area.

This approach significantly expands the potential of the LightGBM model, as information about which performance segment each plot belongs to is added to the input parameters. In other words, significant progress in yield forecasting accuracy is achieved through a combined approach that includes extended visual analysis along with deep vector analysis. The extended visual analysis allows the model to identify spatial relationships and detect complex patterns that often go unnoticed but may arise due to phenomena such as terrain fluctuations or water flows. This analysis is complemented by deep vector analysis, which focuses on studying individual plots of the field. This enables a detailed assessment of their condition, which is critical for the accuracy of overall forecasts. The synergy created by the combined use of these two methods significantly enhances the accuracy and reliability of predictions.

As a result of applying the described data processing methods, an information vector was created for each field plot, describing its development during maturation. The Light Gradient Boosting Machine (LightGBM) was used for training the predictive model, which is an ensemble model that sequentially builds and complicates decision trees using gradient descent to optimize loss. The use of LightGBM allowed for effective yield forecasting, providing the ability to analyze each individual plot of the field in isolation.

Mathematically, LightGBM optimizes the following objective function:

$$L(\theta) = \sum_{i=1}^{n} l\left(y_i, f(x_i;\theta)\right) + \sum_{j=1}^{m} \Omega\left(g_j\right) \quad (1)$$

where $l\left(y_i, f(x_i;\theta)\right)$ — the loss function that evaluates the discrepancy between the actual value y_i and the predicted value f based on the model parameters θ;

g – decision trees, which together form an ensemble model;

m – a set of trees in the model;

$\Omega(g_i)$ — the regularization term, which controls the model's complexity, particularly the penalty for the number and depth of decision trees [12].

A key feature of LightGBM (1) is its ability to process large volumes of data at high speed, thanks to the optimized distribution of tree branching points based on histograms and the use of gradient boosting with Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) [13]. These technologies significantly reduce computational costs and improve the accuracy of analyzing each individual field plot, allowing detailed yield forecasting considering local agronomic conditions.

To provide contextual information, a specialized computer vision model based on the U-net architecture was developed. This model is designed to analyze dependencies in data over large areas without being limited to local subspaces, enabling a global assessment of the state of agricultural lands. Due to its deep segmentation capability, the model effectively identifies areas with potentially high yield and delineates zones where lower productivity is expected.

The U-net model has an encoder-decoder structure, where the encoder sequentially reduces the image size while extracting key features, and the decoder expands the reduced image back to its original size while retaining important details. Mathematically, each layer in U-net can be described as follows:

$$a^{l+1} = \sigma(W^l * a^l + b^l) \quad (2)$$

where $*$ denotes the convolution operation;

σ – non-linear activation function (e.g., ReLU);

W^l, b^l – weights and biases at the l-th layer;

a^l – activation at the l-th layer;

Convolutional layers effectively process spatial information, while up-convolution (transposed convolution) layers restore the details and dimensions of the image [15]. Due to this structure, U-net can assess dependencies in data over large areas, not limited to analyzing only local subspaces, and shows high efficiency in performing satellite image segmentation tasks [17; 18]. This allows the model (2) to globally evaluate the condition of agricultural fields, identifying areas with potentially high yield and highlighting zones where lower productivity is expected.

Since agricultural fields have diverse sizes, a method of dividing images into smaller parts, known as patches, was applied for efficient data processing, as illustrated in fig. 1 and fig. 2. This approach ensures detailed segmentation of each part of the field separately, after which the obtained patches are combined to form a cohesive segmented image of the field. This process simplifies the processing of large land areas and enhances segmentation accuracy through detailed analysis of each fragment.

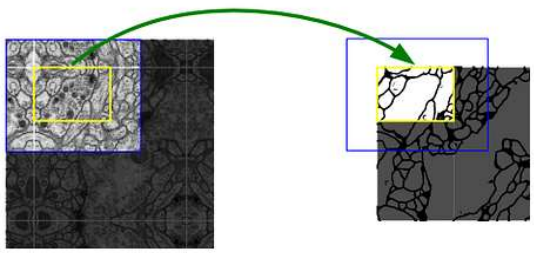

Figure 1 – An example of extracting an image patch for subsequent transformation by a neural network into part of the original image [15]

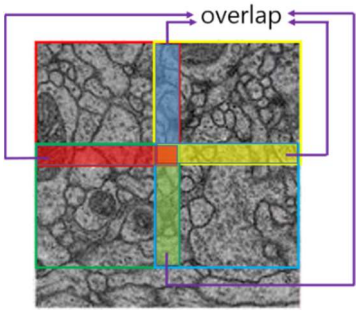

Figure 2 – An example of patch overlaying to form a continuous image

In the process of patch overlaying, a method with weighting coefficients was used, allowing for a smoother integration of image parts. Each pixel in the overlapping areas receives a weight depending on its distance from the center of the patch, contributing to a smoother transition between segments. This approach minimizes the risk of sharp differences at the patch boundaries, ensuring high quality of the final segmented image and accurate representation of yield variability in the fields. The formula for determining the weighting coefficient $w(x, y)$ for a pixel located at a distance (x, y) from the center of the patch can be presented as follows:

$$w(x, y) = e^{\frac{-x^2+y^2}{2\sigma^2}} \quad (3)$$

where x and y are the distances of the pixel from the center of the patch along the respective axes;

σ – the parameter that determines the width of the Gaussian curve's "bell," i.e., the level of smoothing at the edges of the patches [16].

Thus, the coefficient value varies based on the pixel's distance from the center of the patch (3), allowing for the effective overlaying of overlapping parts, smoothing discrepancies in predicted values.

Since the initial sample was limited, several additional steps were introduced to optimize the training.

Firstly, the input data vector was limited to satellite data only. The complete input information contains more than 80 channels, which leads to overfitting in the case of a small number of observations. Using only satellite data allowed us to reduce the size of the input vector and decrease the risk of overfitting while ensuring sufficient informativeness for the model.

Secondly, patches from the image were not selected sequentially without overlap but using a stride step of 1. This means that the next image covers most of the previous one, except for the first column of pixels in the case of a horizontal step and the top row in the case of a vertical step. This approach allows for an increase in the number of training examples and improves model generalization by utilizing the maximum amount of data from the available images.

Thirdly, the following augmentations were used: horizontal flipping, vertical flipping, rotations of 30, 60, and 90 degrees, and various combinations of these transformations. These augmentations are permissible since field data are equivalent in any direction of capture. Deformations were not applied, only rotations and flips, which were accordingly applied to the output images as well. This ensures that each pixel of the input data corresponds to each pixel of the output regardless of the chosen augmentation. These measures help reduce the risk of overfitting and improve the model's robustness.

The results obtained using the U-Net model reveal the significant potential of this approach in accurately forecasting yield. The model demonstrates high prediction accuracy. However, it is necessary to address one of the substantial drawbacks— the tendency of the proposed model to overfit. This aspect can significantly limit its effectiveness in real-world conditions, especially when access to a large amount of high-quality data is restricted. Due to its complexity, the U-Net model requires substantial data volumes for training, necessitating a meticulous approach to data collection and analysis. Therefore, the results obtained in limited conditions may not reflect the full potential of the approach, which is crucial to consider when testing and using it in practical tasks.

Table 1 presents the accuracy metrics of linear regression, Table 2 shows the accuracy metrics of the proposed approach, and fig. 4 visualizes the obtained data.

The results obtained using linear regression clearly demonstrate that the task of yield forecasting is very complex. The analyzed data have complex nonlinear dependencies that linear regression is not always able to adequately capture. This is evidenced by cases where the model produces negative yield predictions, which result from incorrect data interpretation. Such results indicate the necessity of using more sophisticated models capable of accounting for nonlinearity and greater data variability to achieve more accurate and reliable forecasts.

*Table 1 – **Accuracy metrics of linear regression***

Field	RMSE	Forecasted Yield (t)	Actual Yield (t)	Accuracy
Flora__Dahtaliya__22	0.6613	154.26	135.85	86.44%
Flora__Teklivka__22	0.2644	88.71	77.77	85.93%
East-West__Serby_26__23	0.2716	27.98	41.08	68.12%
East-West__Serby_57__23	0.3355	82.60	115.99	71.21%
East-West__Serby_69__23	0.3948	91.00	141.65	64.24%
East-West__Serby_56__23	0.2396	-0.21	27.11	-0.76%
Zhuravske__Field_3__22	0.2438	29.95	48.54	61.69%

Table 2 – ***Accuracy metrics of the proposed approach***

Field	RMSE	Forecasted Yield (t)	Actual Yield (t)	Accuracy
Flora__Dahtaliya__22	0.6701	197.47	135.85	54.64%
Flora__Teklivka__22	0.4496	115.60	77.77	51.36%
East-West__Serby_26__23	0.2144	42.51	41.08	91.30%
East-West__Serby_57__23	0.2565	115.83	115.99	94.78%
East-West__Serby_69__23	0.2764	131.30	141.65	92.12%
East-West__Serby_56__23	0.2122	32.27	27.11	70.24%
Zhuravske__Field_3__22	0.1194	48.83	48.54	92.46%

The proposed approach to yield forecasting significantly enhances the model's ability to detect dependencies in the data, as evidenced by a high accuracy of up to 95% in fields such as "East-West__Serby_57__23" and "East-West__Serby_69__23". This underscores the potential of more complex models for handling the nonlinear characteristics of data frequently encountered in agronomic research.

On the other hand, the model's tendency to overfit is an obvious drawback, especially noticeable in the fields "Flora__Dahtaliya__22" and "Flora__Teklivka__22", where the accuracy is the lowest, approximately 55% and 51%, respectively. This indicates that the model may be overly optimized for the training dataset and unable to adequately generalize knowledge to new data, particularly when the available data does not fully represent all possible conditions. This is also partially evident in the field "East-West__Serby_56__23", where the accuracy dropped to 70.24%, which may indicate issues with accounting for certain agronomic conditions in the model.

Examples of forecast visualizations are presented in fig. 5 and fig. 6. On the left part of the images is the index snapshot of the field, in the middle is the predicted yield, and on the right part is the actual yield.

As previously mentioned, the system has great potential for further development and can become a powerful foundation for new research aimed at advancing the agricultural industry.

The development of a method for generating and analyzing combinations of possible plant care options opens up new prospects for creating comprehensive decision-making systems [19]. This will enhance the accuracy of individualized forecasts for each plot and help determine the most effective strategies for agricultural practices based on a deep analysis of potential yield.

The adaptation of computer vision methods using more optimized algorithms to reduce the computational complexity of the system will allow for the optimization of training and application processes of the model, thereby expanding its practical value in digital agronomy tasks [20]

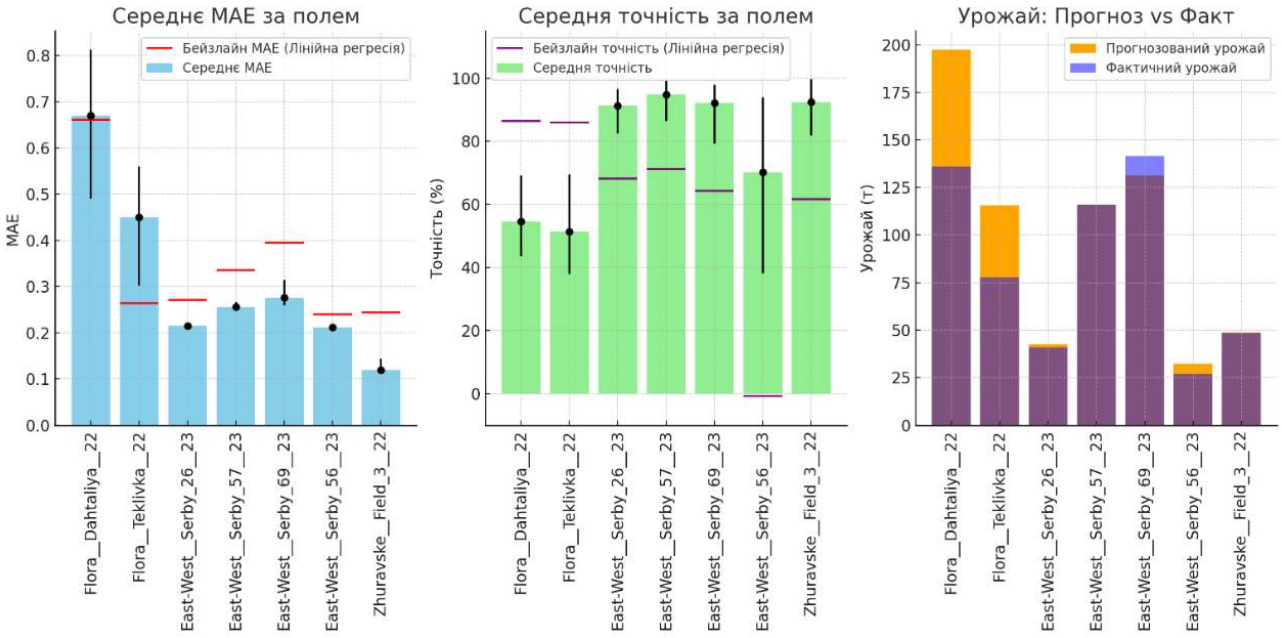

Figure 4 – Visualization of the obtained accuracy metrics

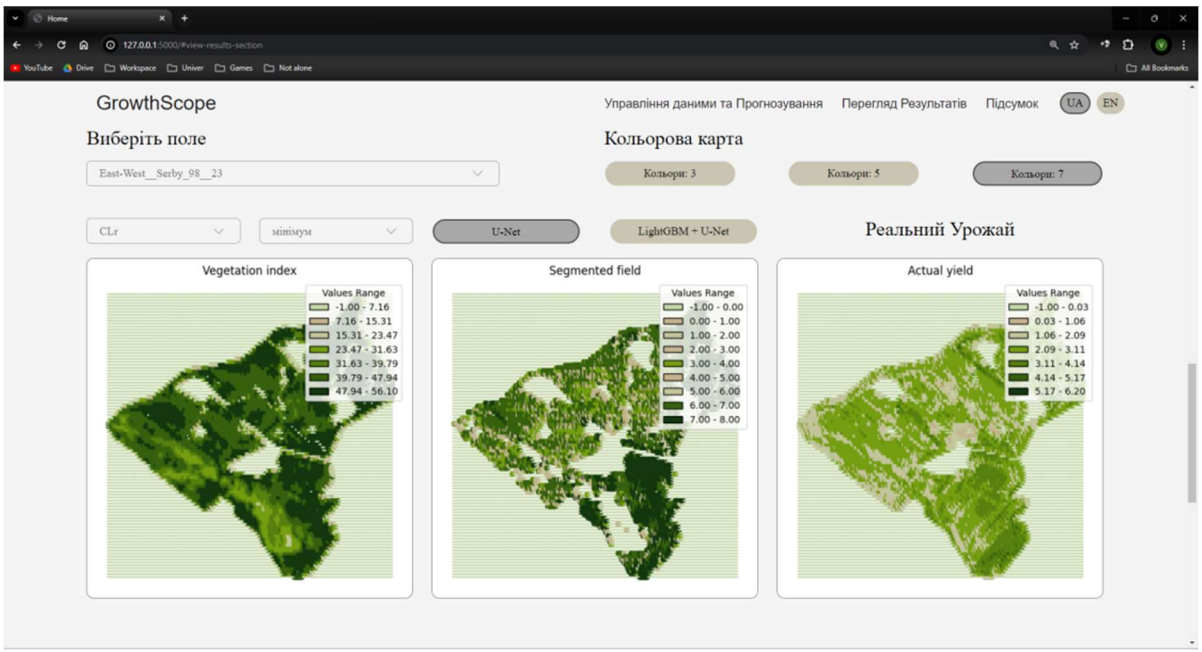

Figure 5 – ***Forecasting results, example 1***

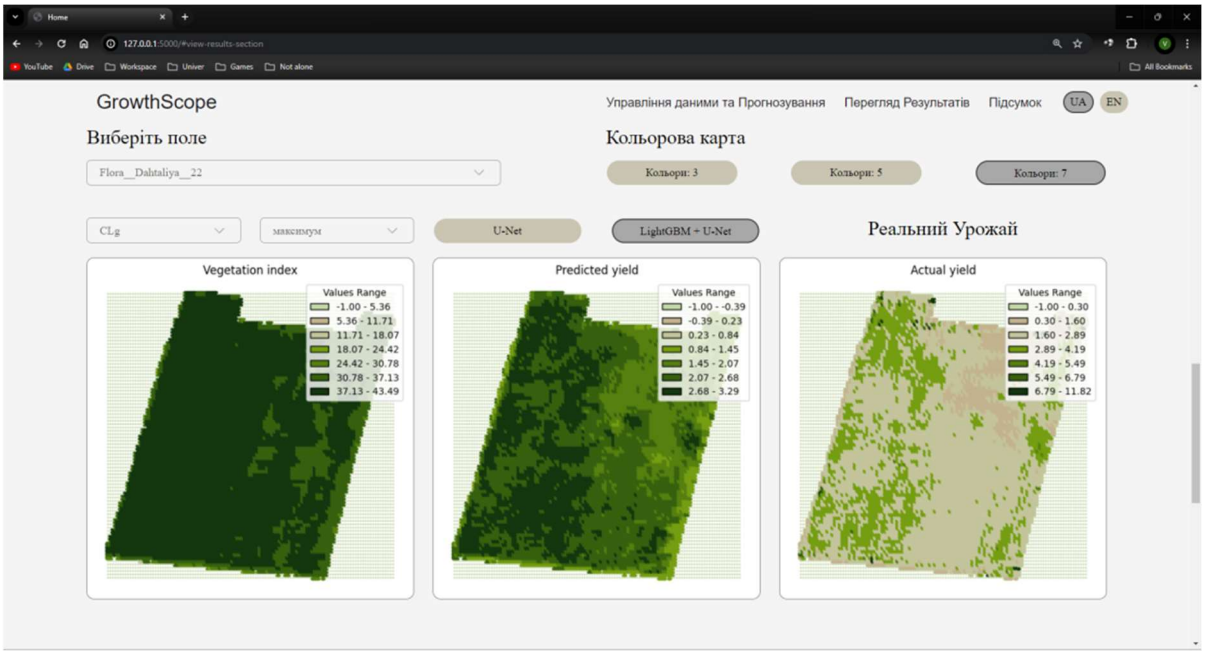

Figure 6 – ***Forecasting results, example 2***

Conclusions from the research

Accurate yield forecasting is a complex task with many uncertainties and unknown factors, significantly complicating problem-solving. An important aspect is detailed forecasting, which requires a specific data format and properly selected and tuned models.

In this study, an analysis of the issues was conducted, existing solutions were reviewed, and an approach to solving the posed problems was proposed. The combination of classical machine learning methods with deep learning has great potential and shows high efficiency even with a limited dataset.

To achieve high accuracy and stability, the system requires further enhancement of informational support, extended testing, and tuning. The limited dataset does not allow for a comprehensive analysis and obtaining reliable results.

The proposed model demonstrates results indicating the prospects of this research direction and can be applied for the further development of approaches to solving agronomic problems, such as precise budget planning and optimization of agronomic measures to improve yield and prevent significant losses in critical situations.

References

1. Al-Gaadi, K. A., Hassaballa, A. A., Tola, E., Kayad, A. G., Madugundu, R., Alblewi, B., Assiri, F. (2016). Prediction of Potato Crop Yield Using Precision Agriculture Techniques. *PLoS One*, 11(9):e0162219. doi: 10.1371/journal.pone.0162219. PMID: 27611577; PMCID: PMC5017787.

2. Paudel, D., Boogaard, H., Wit, A. de, Janssen, S., Osinga, S., Pylianidis, C., Athanasiadis, I. N. (2021). Machine learning for large-scale crop yield forecasting. *Agric. Syst.,* 187, 103016, 10.1016/j.agsy.2020.103016.

3. Khaki, S. & Wang, L. (2019). Crop yield prediction using deep neural networks. *Front. Plant Sci.*, 10, 621.

4. Elavarasan, D. & Vincent, P. M. D. (2020). Crop Yield Prediction Using Deep Reinforcement Learning Model for Sustainable Agrarian Applications. *IEEE Access*, 8, 86886-86901, doi: 10.1109/ACCESS.2020.2992480

5. Box, George, Jenkins, Gwilym. Time Series Analysis: Forecasting and Control

6. Oliver, M. A., Webster, R. (2014). A tutorial guide to geostatistics: Computing and modelling variograms and kriging. https://doi.org/10.1016/j.catena.2013.09.006

7. Cropwise: офіційний веб-сайт. URL: https://www.cropwise.com/

8. Climate FieldView: офіційний веб-сайт. URL: https://www.climatefieldview.com.ua/

9. Xarvio: офіційний веб-сайт. URL: https://www.xarvio.com/ua/uk.html

10. Zozulya, O. L., Shvartau, V. V., Mikhalska, L. M., Kovel, O. L., Hnatiienko, H. M., Snytyuk, V. Y., Domrachev, V. M., Tmienova, N. P. (2024). Kyiv : From A to Z. Modern methods of digital monitoring in crop production: Monograph, 254.

11. Hnatiienko, Vladyslav, Snytyuk, Vitaliy. (2024). Intellectual analysis and prediction of desiccation efficiency based on satellite images. Materials of the 1st International Scientific and Practical Conference "Information Systems and Technologies: Results and Prospects", Kyiv, Ukraine. K.: FIT KNUTSH, P. 340–343.

12. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *NeurIPS*, 30:3146–3154.

13. Xi, X. (2023). The role of LightGBM model in management efficiency enhancement of listed agricultural companies. *Applied Mathematics and Nonlinear Sciences*.

14. Anusha, P. V., Anuradha, C., Murty, P. S. R. C. & Kiran, C. S. (2019). Detecting Outliers in High Dimensional Data Sets using Z-Score Methodology. *International Journal of Innovative Technology and Exploring Engineering*, 9(1), 48-53. https://doi.org/10.35940/ijitee.A3910.119119.

15. Ronneberger, Olaf, Fischer, Philipp, Brox, Thomas. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation.

16. Rifat, Kurban. (2023). Gaussian of Differences: A Simple and Efficient General Image Fusion Method

17. Jiao, L., Huo, L., Hu, C. & Tang, P. (2020). Refined Unet: Unet-Based Refinement Network for Cloud and Shadow Precise Segmentation. *Remote Sensing*, 12 (12). doi:10.3390/rs12122001.

18. Waleed, Alsabhan, Turky, Alotaiby, Basil, Dudin. (2022). Detecting Buildings and Nonbuildings from Satellite ImagesUsing U-Net. *Computational Intelligence and Neuroscience*, 4831223. https://doi.org/10.1155/2022/4831223

19. Hnatiienko, H. M., Snytyuk, V. Y., Hnatiienko, V. H., Zozulya, O. L. (2022). Application of models and methods of artificial intelligence in determining the yield of agricultural crops. Applied systems and technologies in the information society: coll. theses of reports and sciences. reported participants of the VI International Scientific and Practical Conference/ by general ed. V. Pleskach, V. Zosimov, M. Pyrog // Kyiv: Kyiv national. University named after Taras Shevchenko, P. 90–98.

20. Bilan, Stepan, Hnatiienko, Vladyslav, Ilarionov, Oleh & Krasovska, Hanna. (2023). The Technology of Selection and Recognition of Information Objects on Images of the Earth's Surface Based on Multi-Projection Analysis. CEUR Workshop Proceedings, 3538, 23. Selected Papers of the III International Scientific Symposium “Intelligent Solutions”. Symposium Proceedings Kyiv – Uzhhorod, Ukraine, September 27-28.

Стаття надійшла до редколегії 29.07.2024

Гнатієнко Владислав Григорович
Дослідник-стажист,
orcid.org/0009-0000-2678-5158
ТОВ «Сингента», Київ
Гнатієнко Григорій Миколайович
Кандидат технічних наук, доцент кафедри інтелектуальних технологій,
orcid.org/0000-0002-0465-5018
Київський національний університет імені Тараса Шевченка, Київ

ІНТЕГРАЦІЯ МЕТОДІВ МАШИННОГО ТА ГЛИБИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ ВРОЖАЙНОСТІ СОНЯШНИКА

Анотація. *Практичний досвід прогнозування врожайності показує, що це складна багатофакторна задача, яка вимагає точних та надійних методів вирішення. Використання машинного та глибинного навчання є ключовим у досягненні кращих результатів у цифровій агрономії. Стаття присвячена побудові інтелектуальної системи прогнозування врожайності соняшника. На основі аналізу наукових публікацій та практичного досвіду, узагальнено основні проблеми обробки даних і запропоновано схеми їх вирішення. Основні етапи роботи включали дослідження поточного стану цифрової агрономії, вибір підходу, розробку методу обробки інформації, програмну реалізацію моделі та тестування. Ключовими викликами стали обмеженість наборів даних та складність вибору оптимального підходу для уникнення перенавчання. Поєднання різних методів аналізу дозволило створити потужну систему, яка переважає традиційні підходи, хоча потребує більше даних для навчання. Висновки дослідження показують, що методи машинного та глибинного навчання, такі як LightGBM і U-Net, разом із запропонованими методами обробки даних, досягають високої точності у прогнозуванні. Модель продемонструвала здатність до узагальнення знань на нові поля та до побудови детальних карт врожайності. Подальші дослідження включають розробку методу для генерації комбінацій варіантів догляду за рослинами, адаптацію методів комп'ютерного зору з оптимізованими алгоритмами для зменшення обчислювальної складності та розширення функціоналу системи з включенням аспектів кібербезпеки. Запропонована система значно підвищить ефективність прогнозування врожайності соняшника, сприяючи розвитку цифрової агрономії.*

Ключові слова. Сільськогосподарський сектор; прогнозування врожайності; супутникові дані; машинне навчання; комп'ютерний зір

Link to publication

APA Hnatiienko, V. & Hnatiienko, H. (2024). Integration of machine learning and deep learning methods for sunflower yield prediction. *Management of Development of Complex Systems*, 59, 225–234, dx.doi.org\10.32347/2412-9933.2024.59.225-234.

ДСТУ Гнатієнко В. Г., Гнатієнко Г. М. Інтеграція методів машинного та глибинного навчання для прогнозування врожайності соняшника. *Управління розвитком складних систем.* Київ, 2024. № 59. С. 225 – 234, dx.doi.org\10.32347/2412-9933.2024.59.225-234.