ДОПОВІДІ
НАЦІОНАЛЬНОЇ
АКАДЕМІЇ НАУК
УКРАЇНИ

https://doi.org/10.15407/dopovidi2025.04.017
UDC 519.7+004.8

V.H. Hnatiienko[1], https://orcid.org/0009-0000-2678-5158
H.M. Hnatiienko[1], https://orcid.org/0000-0002-0465-5018
O.L. Zozulya[2], https://orcid.org/0000-0003-3500-3423
V.Ye. Snytyuk[1], https://orcid.org/0000-0002-9954-8767
V.V. Schwartau[3], https://orcid.org/0000-0001-7402-5559

[1] Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
[2] Syngenta LLC, Kyiv, Ukraine
[3] Institute of Plant Physiology and Genetics of the NAS of Ukraine, Kyiv, Ukraine
E-mail: g.gna5@ukr.net

Site-specific sunflower yield forecasting based on spatial analysis and machine learning

Presented by Academician of the NAS of Ukraine V.V. Morgun

The study focuses on the development of an intelligent yield forecasting system using satellite data, geospatial data and climate indicators. The introduction of modern information technologies, in particular machine learning and big data analysis methods, provides agricultural professionals with strategic advantages, reducing the risks of excessive pesticide use and promoting sustainable agricultural development. This study aims to optimize desiccant application in sunflower cultivation by modeling potential yield losses based on data obtained during the growing season. The use of digital solutions is relevant for crop production, as it increases the accuracy of forecasts and the efficiency of management decisions, while reducing costs and increasing the productivity of agrophytocenoses.

Keywords: *satellite data, climate indicators, machine learning, big data analysis, vegetation indices, FAO, loss forecasting, desiccation.*

Introduction. The active development of digital agronomy opens up broad prospects for intensifying the development of the agricultural sector, while at the same time given a rise to a number of complex tasks and challenges. Amid climate change, market price fluctuations and growing demands on the efficiency of natural resources use, the need to harmonize Ukrainian legislation in the field of plant protection with European standards with accurate budget planning and opti-

Citation: Hnatiienko V.H., Hnatiienko H.M., Zozulya O.L., Snytyuk V.Ye., Schwartau V.V. Site-specific sunflower yield forecasting based on spatial analysis and machine learning. *Dopov. Nac. akad. nauk Ukr.* 2025. No. 4. P. 17—26. https://doi.org/10.15407/dopovidi2025.04.017

mization of plant care methods is becoming increasingly important. The modern development of digital technologies, artificial neural networks, artificial intelligence and new approaches to statistical data processing allows farmers to move to a new level of agricultural production. Therefore, the search for the application of these new methods of obtaining and processing information is an urgent problem for the agricultural sector. After all, the successful application of such modern methods is an essential factor in ensuring food security [1].

The shortcomings and limitations of traditional statistical approaches, which provide only a rough estimate of yields, become particularly apparent in the face of the demands of modern agricultural production. While these methods can model potential performance, they often fail to meet the need for accurate and detailed planning. At the same time, artificial intelligence, with its capabilities of deep analysis of large amounts of data and machine learning, opens up new horizons for the development of the agricultural sector of the economy. Digitalization of processes in agrophytocenoses has great potential for the development of crop production, facilitating the creation of innovative solutions to optimize agrotechnical measures and improve production efficiency. The transition to the use of these advanced technologies requires not only the development of new tools and methods, but also a profound rethinking of approaches to the management of agricultural processes.

In [2], the authors developed and presented a mathematical framework for minimizing sunflower yield losses. The study was based on spatial analysis of satellite images. The scientific results were obtained by applying machine learning methods.

Analysis of the latest research and publications. Modern yield forecasting systems using artificial intelligence cover a wide range of technologies. Here are some important areas of research in this area.

1. The use of deep neural networks (DNNs) can accurately predict yield by analyzing data on genotype, weather and soil characteristics, as well as the historically identified productivity zones in each field, demonstrating an average accuracy of 85—89 %. However, the main disadvantage of this approach is its limitation to field-level forecasts, which does not allow taking into account microclimatic and soil variations within a field that are important for detailed forecasting [3].

2. Machine learning using traditional algorithms provides high accuracy in predicting the overall field yield. However, despite the theoretical possibilities of detailed analysis, as a rule, detailed forecasting for individual plots is not realized [4].

3. Some studies have applied recurrent neural networks using reinforcement learning to predict yield [5], achieving an average accuracy of 93.7 %. This study also did not address distributed (spatially resolved) forecasting.

4. In study [6] stratified sampling for potato yield forecasting using empirical equations based on NDVI and SAVI indices was considered. The authors point out that the forecasting is performed with errors of 3.8—7.5 %, but the test and training samples are formed on data of the same fields, so the possibilities of generalization and practical application of the method were not analyzed.

Modern approaches to yield forecasting, including the use of artificial intelligence technologies, have made it possible to achieve significant results in processing large amounts of data and providing accurate forecasts at the whole-field level. Despite their effectiveness in identifying general yield trends, the methods discussed above have significant limitations, especially when it comes to achieving resolution at the scale of individual field plots. Forecasting is complicated by the fact that many factors are completely unpredictable, such as rainfall, number of days of potential vegetation, natural disasters, etc.

The main problem is that most existing methods are designed to predict total field yields and do not take into account internal variations that can be critical for effective management of agronomic measures. This limitation does not allow for a detailed productivity map, which, in turn, limits the potential of such systems in a number of key tasks. In particular, optimization of differentiated application of fertilizers and plant-protection products, maintenance of water regime, and analysis of the impact of different combinations of parameter values on the productivity of individual plots remain beyond the capabilities of these technologies. It is quite difficult to compare certain technologies due to the high variability of individual plots. Thus, the development of detailed analysis and forecasting methods at the individual plot level is becoming a priority for improving yield forecasting systems, opening up new prospects for precision agriculture.

Objective of the study. The main purpose of this study is to improve the accuracy of yield forecasting

This will make it possible to predict variations in crop productivity formation due to uneven maturation of sunflower and avoid yield losses. These losses can be significantly reduced by desiccation. However, in each specific case the question arises whether these losses are significant enough to invest in an additional agrotechnical measure — desiccation. The question also arises whether it makes sense to use differentiated application of the product (pesticide) to specific plots as an alternative to continuous spraying of the field. This is achieved by introducing the ability to accurately predict yields in individual plots of the field, to predict losses in each area, to optimize sowing and plant care conditions, including sowing dates, sowing density, timing and intensity of herbicide and fungicide applications.

Materials and methods. To build the model, we identified predictors that can affect the nonuniform ripening, namely: sowing date, weather conditions, soil moisture, FAO (or hybrid maturity group), predecessor crops, etc. Such a forecast allows the farmer to assess the economic feasibility of the planned agrotechnical protection measures, apply selective treatment of individual plots, reducing the pesticide burden on the environment.

The task of forecasting yields is extremely complex, comparable to weather forecasting. It requires not only taking into account a large number of parameters, but also identifying the key factors that have the greatest impact on the result.

Yield depends on many parameters: chemical composition and structure of the soil, its moisture, pH, types of fertilizers and methods of their application, etc. Other important parameters include information on weather conditions: air temperature, precipitation, soil moisture and solar radiation intensity. The presence and activity of pests and diseases also have a significant impact on yields. Agrotechnical measures are equally important: tillage, crop rotation, sowing and harvesting methods. In addition, genetic characteristics of seeds, their resistance to diseases and adaptability to weather conditions should be taken into account.

It is necessary to develop methods for predicting the yield of each field area based on the analysis of detailed data on plant condition. Such data include maps of reflected solar radiation intensity obtained from satellite images in different spectra, which are converted into vegetation indices NDVI, NDWI, CLg, CLr, GLI [7]. Meteorological data are also important: temperature, precipitation, wind speed and direction, cloud cover, solar radiation, and atmospheric pressure. These data are supplemented by information on agrotechnical measures, including herbicide and fungicide treatments, the ripeness group of the sunflower hybrid (five groups from early to late), and sowing density. All these data affect the maturity rate of sunflower. A dataset containing yield data in tons per hectare for each field plot is used to train and validate the model.

Results and discussion. Let us introduce some notations to further describe the data structure and methods.

Let X_i be the matrix of field i, $x_{ijkl} \in X_i$ be the elements of the matrix;

$i = 1, 2, \ldots, g$ — field numbers; g — number of fields in the training set;

$j \in J = \{j_1, j_2, \ldots, j_n\}$ — days of observations; n — number of days of observations conducted for the field;

$k = 1, 2, \ldots, m$ — are the row numbers of the matrix X_i; m is the number of field plots; each plot corresponds to a row of the matrix;

$l \in L = \{NDVI, NDWI, GLI, CLr, CLg, \textit{wind speed}, \textit{seeding density}, \ldots\}$ — parameters of input information.

The input data for yield forecasting are extremely voluminous due to a wide range of parameters and the length of the observation period. Their structure is shown in Fig. 1. The data is a multidimensional array of dimension $m \times |L| \times n$.

In order to reduce the dimensionality of the input data vector and increase the forecasting efficiency, preliminary analysis and selection of the most informative features is performed. One of the tools of this process is correlation analysis, which allows to identify statistical relationships between parameters. Features that are highly correlated with each other are identified, and among them, and the most informative ones are determined by expert judgment, while others are discarded to reduce the dimensionality of the data set. This increases the model efficiency by reducing the computational burden.

The following steps are performed at the preprocessing stage.

Step 1. Removal of outliers for each day separately for each field using the z-score method [8, 9].

The z-score for each item is determined by the formula:

$$z_{ijkl} = \frac{x_{ijkl} - \mu_{ijl}}{\sigma_{ijl}},$$

where $x_{ijkl} \in X_i$; $\mu_{ijl} = \frac{\sum_{k=1}^{m} x_{ijkl}}{m}$, is the average value ofthe parameter l for day j in the matrix X_i; $\sigma_{ijl} = \sqrt{\frac{\sum_{k=1}^{m} (x_{ijkl} - \mu_{ijl})^2}{m}}$ is the standard deviation ofthe parameter l for day j in the matrix X_i; z_{ijkl} is the z-score of the element x_{ijkl}.

Heuristic E1. We consider the element x_{ijkl} to be an outlier if the value is $|z_{ijkl}| > 2.9$. In this case, the entire row k for the day j of the matrix X_i is removed from the training set.

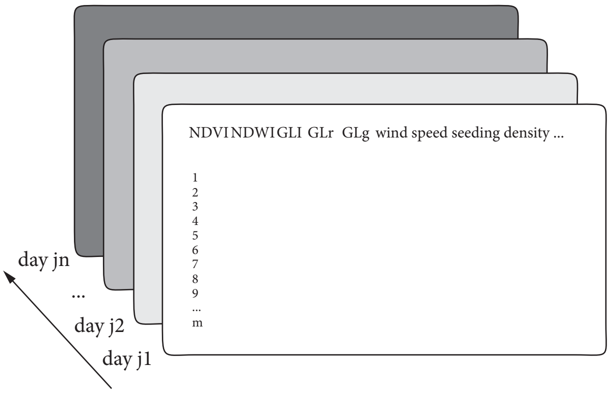

Fig. 1. Schematic representation of the structure of the training dataset before preprocessing

Step 2. Data aggregation:

For each matrix X_i, $i = \overline{1, g}$, for each sequence of elements $x_{ijkl} \in X_i$, $j \in J$, the aggregates are calculated using the formulas:

$$x_{ikl}^{\min} = \min_{j \in J}(x_{ijkl}),\ x_{ikl}^{\text{mean}} = \frac{1}{|J|}\sum_{j \in J} x_{ijkl},\ x_{ikl}^{\max} = \max_{j \in J}(x_{ijkl}),$$

forming a matrix of aggregated values X_i'.

Step 3. Combining data into a common dataset X: $X = \bigcup_{i = \overline{1, g}} X_i'$.

Step 4. Repeated removal of outliers on the merged dataset X:

$$z_{ql}' = \frac{x_{ql}' - \mu_l'}{\sigma_l'},$$

where $x_{ql}' \in X$, is the matrix element corresponding to row q and the parameter l;

$\mu_l' = \frac{\sum_{q=1}^{s} x_{ql}'}{s}$ — is the average value of the parameter l in the matrix X;

$\sigma_l' = \frac{\sum_{q=1}^{s} (x_{ql}' - \mu_l')^2}{s}$ — is the standard deviation of the parameter l in the matrix X;

z_{ql}' — z-score of the element x_{ql}'.

Heuristic E2. We consider the element x_{ql}' to be an outlier if $|z_{ql}'| > 2.9$. In this case, the entire row q of the matrix X is removed from the training set.

Repeated removal of outliers is an important step because removing outliers separately for each field does not guarantee the absence of erroneous observations in the aggregated dataset. When combining different fields, it is possible that data that were considered normal for one field become abnormal in the context of the overall set due to differences in scale, distributions, or other characteristics. Therefore, it is necessary to remove outliers again to ensure the consistency and homogeneity of all data.

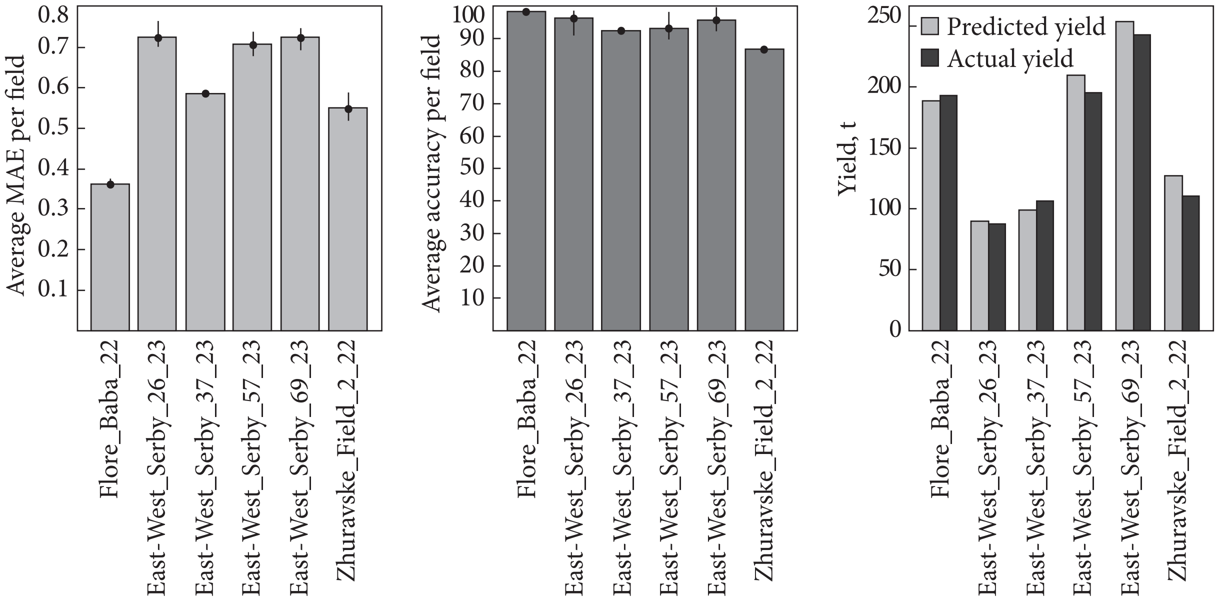

Fig. 2. Visualization of the obtained accuracy indicators [2]

To record the data on the dates of herbicide and fungicide treatments that are part of the use of crop protection products, an algorithm similar to one-hot encoding was implemented. Instead of using the number of days from the date of sowing to the moment of treatment, the dates were encoded as categorical variables representing predefined day-range buckets. The experts established typical time ranges for fungicide (34—82 days from sowing date) and herbicide (24—62 days) applications. These observation day ranges were categorized separately. For fungicides, these are {34, 46, 58, 70, 82}, and for herbicides, {24, 33, 43, 52, 62}. Each date of chemical application refers to the closest category. For example, if fungicides were applied on the 48th day after sowing, the vector for this case would consist of the elements {0, 1, 0, 0, 0}, which corresponds to category 46.

As a result of applying the described data processing methods, a vector of traits describing its development during the ripening period was constructed for each field plot. One of the models was built with the Light Gradient Boosting Machine (LightGBM) algorithm [10]. The model was used to predict yield for each field area in isolation.

Results of studying the efficiency of the yield forecasting system [2]

Field	MAE	Accuracy	Predicted harvest, tons	Actual harvest, tons	Area, ha
Flora__Baba__22	0.360007	98.299489	189.184446	192.400052	101
East-West__Serby_26__23	0.722826	96.284523	89.576487	87.459699	26.6
East-West __Serby_37__23	0.585210	92.651682	98.978797	106.251481	37
East-West __Serby_57__23	0.705239	93.220749	209.482780	195.217092	57.4
East-West __Serby_69__23	0.724573	95.723041	253.647770	242.73062	69
Zhuravske__Field_2__22	0.548212	86.796210	126.477451	109.757490	29.9

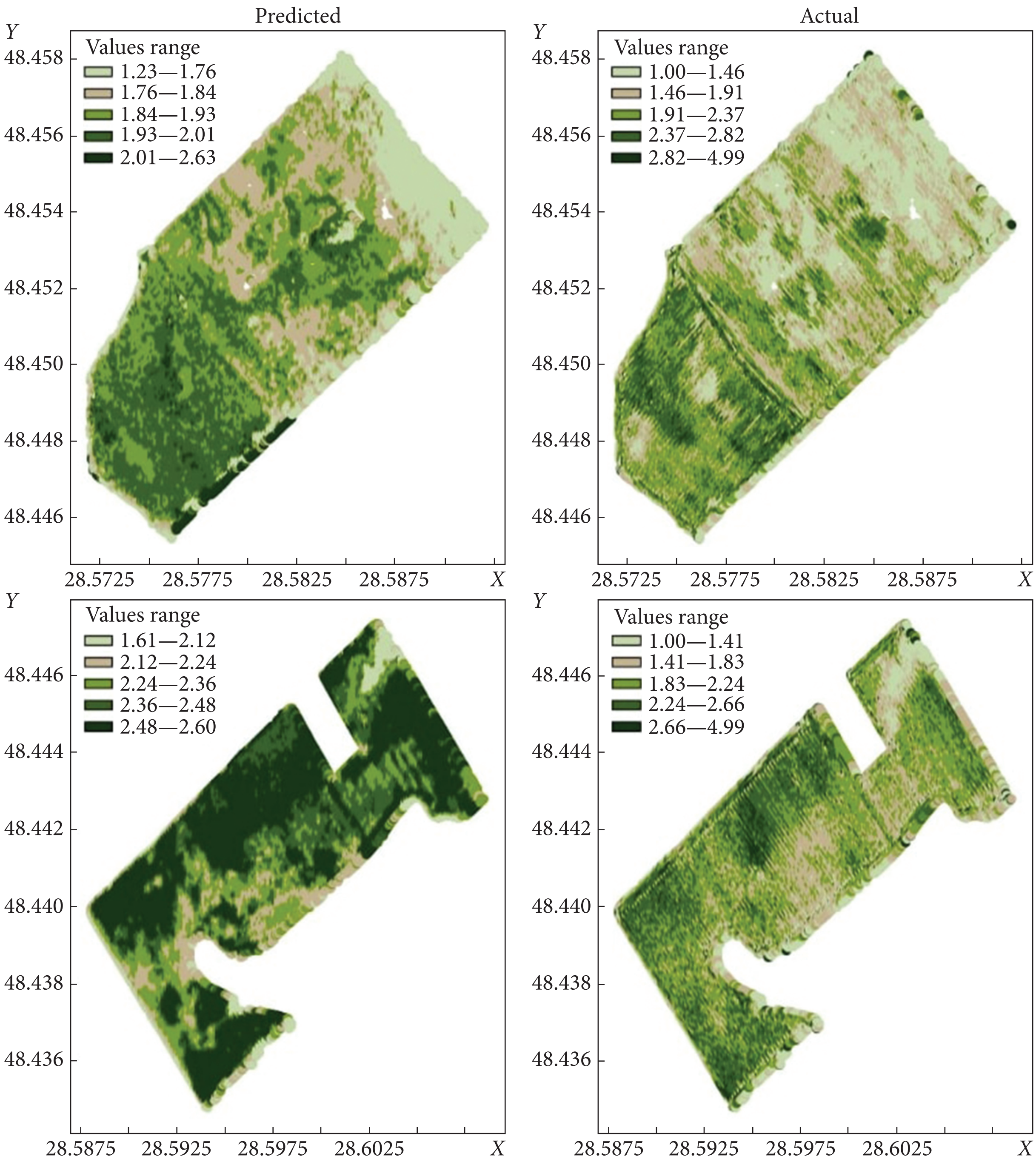

Fig. 3. On the left is a map of projected yields, on the right is the actual yield

In addition to the traditional approach based on the vector representation of monitoring parameters for a single site, it is proposed to consider the spatial context by analyzing an additional dimension that covers a set of neighboring sites. This allows taking into account spatial dependencies and obtaining comprehensive information on the state of agricultural areas [11, 12].

The combination of these approaches provides a synergistic effect [13], increases the accuracy of forecasting [1, 14], and expands the possibilities of analyzing the studied objects. To implement spatial context analysis, a computer vision model based on U-Net architecture was developed to effectively identify high-productivity zones and zones with potential yield reduction by analyzing spatial relationships between plots.

Since agricultural fields have a variety of sizes, this study used a method of splitting the images into smaller parts, known as patches, to process the data efficiently. This approach allows detailed segmentation of each part of the field separately, after which the resulting patches are matched together to form a complete segmented image of the field. In the process of overlaying the patches, a weighting method was used, allowing for smoother merging of the image parts. Each pixel in the overlapping areas receives a weight depending on its distance to the center of the patch. This procedure contributes to a softer and more natural transition between segments.

The U-net model developed for land plot segmentation allows for large-scale analysis of plant development conditions, taking into account the general features of the field and the individual characteristics of each plot. After the segmentation is completed, an additional stage of analysis is performed for each field area using the LightGBM model: forecasting is performed taking into account the defined field segment.

As part of Syngenta's experimental research, a dataset has been created for a number of fields to determine the accuracy of yield prediction. Each field is divided into separate plots for which the model is used to predict yields. The forecast for each plot is compared with the actual harvest data, and the root mean square error (RMSE) is calculated.

Next, the yield of each plot (tons per hectare) is converted to tons by multiplying by the plot area. This way, the total yield of the entire field is calculated from the predicted data and compared to the actual total yield. Total-field forecast accuracy is assessed as the percentage error between predicted and actual yields.

Although the key indicator is the accuracy of the total yield, the accuracy of the prediction for individual plots is also important for further research. This will allow the system to be scaled up and generalized, ensuring high accuracy in both general forecasting and individual plot application.

The results of the yield forecasting system efficiency analysis are shown in Table and Fig. 2.

The use of the developed models allowed us to obtain estimates of potential yields with high accuracy. At the same time, the forecasting accuracy is not stable enough: the minimum value is 87.62 %, the maximum is 97.88 %. The mean absolute error (MAE) across all field plots is 0.608. The average accuracy of total yield forecasting is 92.78 %.

During the study we faced a constraint of a very limited training and testing dataset. The training sample contains only 8 fields, which significantly limits the generalizability of the model. This contrasts with modern studies that use hundreds and sometimes thousands of fields for training. Expanding the training dataset will allow additional patterns to emerge, thereby improving model accuracy and enabling more reliable yield forecasts in different agroclimatic conditions and regions. This is important to improve the accuracy of forecasts and make the model more versatile and suitable for a wide range of applications.

Fig. 3 shows examples of forecast visualization. Each figure shows a map of actual yields on the left and a map of predicted yields on the right.

Conclusions. The introduction of a model capable of generating high-resolution forecasts localized to individual plots opens up new prospects for the application of digital approaches in agriculture. This approach allows for an in-depth analysis of the impact of local factors on plant development and yield, which helps to identify optimal conditions or negative factors for their growth. In addition, it opens up opportunities to optimize the variable-rate application of fertilizers and crop-protection chemicals, significantly increasing the efficiency of agricultural practices and minimizing the negative impact of human activity on the environment.

The high accuracy of forecasts provided by the model enables highly reliable budget planning for farming enterprises. This, in turn, allows agricultural producers to effectively plan and optimize costs, ensuring more efficient resource management and increased overall profitability. Thus, the implementation of the described model opens up significant opportunities for increasing productivity and ensuring sustainable development of the agricultural sector.

In further research, enriching the training set by adding data from a variety of geographical locations and growing conditions will be key to maximizing the model's versatility, allowing it to work effectively with a wider range of agroecosystems.

Optimization of the training set by balancing its structure by removing overrepresented data and augmenting under-represented categories plays an important role in improving the accuracy of forecasts. This will allow the model to better adapt to the variability of conditions and characteristics of different types of crops. The presented digital solutions are promising for further development and integration with nutrient management and crop protection systems as part of agrophytocenosis management, as well as a way to ensure the country's food security.

REFERENCES

1. Schwartau, V. V. (2024). Biological factors affecting food security in Ukraine: According to the materials of scientific report at the meeting of the Presidium of NAS of Ukraine, February 7, 2024. Visn. Nac. Acad. Nauk Ukr., No. 4, pp. 15-24 (in Ukrainian). https://doi.org/10.15407/visn2024.04.015
2. Hnatiienko, V. H., Hnatiienko, H. M., Zozulya, O. L. & Snytyuk, V. Ye. (2024). Method of forecasting yield of agricultural crops using multifactor analysis and neural networks. Scientific Bulletin of Uzhhorod University. Series of Mathematics and Informatics, 44, No. 1, pp. 93-105 (in Ukrainian). https://doi.org/10.24144/2616-7700.2024.44(1).93-105
3. Khaki, S. & Wang, L. (2019). Crop yield prediction using deep neural networks. Front. Plant Sci., 10, 621. https://doi.org/10.3389/fpls.2019.00621
4. Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C. & Athanasiadis, I. N. (2021). Machine learning for large-scale crop yield forecasting. Agric. Syst., 187, 103016. https://doi.org/10.1016/j.agsy.2020.103016
5. Elavarasan, D. & Vincent, P. M. D. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. IEEE Access, 8, pp. 86886-86901. https://doi.org/10.1109/ACCESS.2020.2992480
6. 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. https://doi.org/10.1371/journal.pone.0162219
7. Zozulia, O. L., Schwartau, V. V., Mykhalska, L. M., Kovel, O. L., Hnatienko, H. M., Snitiuk, V. E., Domrachev, V. M. & Tmenova, N. P. (2023). Modern methods of digital monitoring in crop production. Kyiv, Vid A do Ya (in Ukrainian).
8. Anusha, P. V., Anuradha, Ch., Murty, P. S. R. C. & Kiran, Ch. S. (2019). Detecting outliers in high dimensional data sets using z-score methodology. Int. J. Innov. Technol. Explor. Eng., 9, No. 1, pp. 48-53. https://doi.org/10.35940/ijitee.A3910.119119
9. Jiao, L., Huo, L., Hu, C. & Tang, P. (2001). Refined UNet: UNet-based refinement network for cloud and shadow precise segmentation. Remote Sens., 12, No. 12. https://doi.org/10.3390/rs12122001

10. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T.-Y. (2017, December). LightGBM: A highly efficient gradient boosting decision tree. Proceeding of the 31st Conference on Neural information processing systems 30 (NIPS 2017), (pp. 3146-3154), Long Beach, CA, USA.
11. Bilan, S., Hnatiienko, V., Ilarionov, O. & Krasovska, H. (2023, September). The technology of selection and recognition of information objects on images of the earth's surface based on multi-projection analysis. Proceedings of the 3th International Scientific Symposium "Intelligent Solutions" (IntSol-2023), (pp. 23-32), Kyiv — Uzhhorod, Ukraine.
12. Hnatiienko, H., Domrachev, V. & Saiko, V. (2021). Monitoring the condition of agricultural crops based on the use of clustering methods. Proceeding of the 15th International Conference Monitoring of geological processes and ecological condition of the environment, Vol. 2021 (pp. 1-5), Kyiv, Ukraine. https://doi.org/10.3997/2214-4609.20215K2049
13. Hnatiienko, V. & Snytyuk, V. (2024, October). Site-specific forecasting of agricultural crop yield as a technology and service. Proceedings of the 8th International Scientific and Practical Conference Applied information systems and technologies in the digital society (AISTDS 2024), (pp. 44-55). Kyiv, Ukraine.
14. Hnatiienko, V. & Hnatiienko, H. (2024). Integration of machine learning and deep learning methods for sunflower yield prediction. Management of Development of Complex Systems, 59, pp. 225-234. https://doi.org/10.32347/2412-9933.2024.59.225-234

Received 05.05.2025

В.Г. Гнатієнко[1], https://orcid.org/0009-0000-2678-5158
Г.М. Гнатієнко[1], https://orcid.org/0000-0002-0465-5018
О.Л. Зозуля[2], https://orcid.org/0000-0003-3500-3423
В.Є. Снитюк[1], https://orcid.org/0000-0002-9954-8767
В.В. Швартау[3], https://orcid.org/0000-0001-7402-5559

[1] Київський національний університет ім. Тараса Шевченка, Київ, Україна
[2] ТОВ "Сингента", Київ, Україна
[3] Інститут фізіології рослин і генетики НАН України, Київ, Україна
E-mail: g.gna5@ukr.net

РОЗПОДІЛЕНЕ ПРОГНОЗУВАННЯ ВРОЖАЙНОСТІ СОНЯШНИКА НА ОСНОВІ ПРОСТОРОВОГО АНАЛІЗУ ТА МАШИННОГО НАВЧАННЯ

Дослідження присвячено розробленню інтелектуальної системи прогнозування врожайності з використанням супутникових та геоінформаційних даних і кліматичних показників. Впровадження сучасних інформаційних технологій, зокрема методів машинного навчання та аналізу великих даних, надає фахівцям аграрного сектору стратегічні переваги, що дає можливість знижувати ризики надмірного використання пестицидів і сприяти сталому розвитку сільського господарства. Це дослідження спрямоване на оптимізацію використання десикантів на соняшнику шляхом моделювання обсягів можливих втрат врожаю на основі одержаних у період вегетації культури даних. Використання цифрових рішень є актуальним для рослинництва, оскільки забезпечує підвищення точності прогнозів та ефективності управлінських рішень, сприяючи зменшенню витрат та збільшенню продуктивності агрофітоценозів.

Ключові слова: *супутникові дані, кліматичні показники, машинне навчання, аналіз великих даних, вегетаційні індекси, ФАО, прогнозування втрат, десикація.*