Baltic J. Modern Computing, Vol. 12 (2024), No. 4, 387-414
https://doi.org/10.22364/bjmc.2024.12.4.03

An Improved Approach to Prediction of Maize Disease Occurrence Based on Weather Monitoring and Machine Learning: Case of the Forest-Steppe and Northern Steppe of Ukraine

Grygorii DIACHENKO[1], Ivan LAKTIONOV[1], Artem VIZNIUK[2], Vyacheslav GOREV[1], Vita KASHTAN[1], Kostiantyn KHABARLAK[1], Yana SHEDLOVSKA[1]

[1]Dnipro University of Technology, av. Dmytra Yavornytskoho, 19, UA49005 Dnipro, Ukraine
[2]CLOUD FLOW LLC, str. Khotkevycha Hnata, 12, UA02094 Kyiv, Ukraine

Diachenko.G@nmu.one,Laktionov.I.S@nmu.one,artem.viznuk@gmail.com,
Gorev.V.M@nmu.one,Kashtan.V.Yu@nmu.one,Khabarlak.K.S@nmu.one,
Shedlovska.Y.I@nmu.one,

ORCID 0000-0001-9105-1951, ORCID 0000-0001-7857-6382, ORCID 0000-0002-0594-2633, ORCID 0000-0002-9528-9497, ORCID 0000-0002-0395-5895, ORCID 0000-0003-4263-0871, ORCID 0000-0003-4931-4070

Abstract. Agriculture is becoming an increasingly computerised and knowledge-intensive industry in the context of the current global digitalisation and intellectualisation of production processes. This plays a crucial role in ensuring food security at the national and global levels. The subject of the study is software components and computer models for increasing the efficiency of the transformation of weather data based on machine learning. Three types of machine learning algorithms were investigated: Linear regression, Random Forest and Feedforward neural network. The best results were obtained using Random Forest. The main scientific and applied effect of the study in this article is a substantiated approach to improving software and hardware solutions of information technologies for monitoring the soil and climatic conditions of agricultural open-field crop production. This involved the development of software components and computer models for predicting the probability of maize diseases. This effect has been achieved through the implementation of comprehensive studies that include: preliminary approximation of input datasets; identification of models for predictive analytics of the probability of occurrence of specific types of maize diseases in certain agroclimatic zones, taking into account the cumulative impact of climatic parameters and the probability of disease occurrence; formalised accounting of expert experience in the field of crop stress tolerance; the development of software components in the Python programming language that can be integrated into the low-level data processing link.

Keywords: prediction, disease occurrence, random forest, approximation, maize, hyperparameter optimisation.

1. Introduction

1.1. Relevance of the topic and research motivation

Nowadays, in the light of global trends in the development of science and technology in the context of global digitalisation and intellectualisation of production processes, it is worth noting that agriculture is becoming an increasingly computerised and knowledge-intensive industry, which in turn plays a crucial role in ensuring food security at the national and global levels. At the same time, the open-field crop production industry faces a significant number of challenges caused by social, environmental and economic factors, namely: global climate change; the probabilistic nature of pricing for fertilisers, seeds and material and technical resources required for the full cycle of agricultural production; dysfunctional and unstable agricultural markets and supply chains. Given the diversity and many destabilising factors that determine the quality and quantity of agricultural crop production, it is increasingly important to develop and implement scientifically substantiated solutions to preserve crops throughout the full cycle of cultivation. Therefore, practitioners in the field of crop production must have stable and prompt access to precise information obtained through computerised acquisition and intellectual transformation of large amounts of measurement data distributed in time and space on the soil and climatic conditions of agricultural areas (Ceccarelli et al., 2022; WEB, a).

Contemporary methods and tools for computerisation and intellectualisation of production and technological processes of agricultural enterprises allow for the practical implementation of an information-oriented approach to ensuring high levels of crop stress resistance. This, in turn, stimulates positive dynamics of environmental, economic and social aspects of sustainable development of the agricultural sector on a global (WEB, b) and national scale (WEB, c). To date, the use of intellectualised information and computer technologies has proven to be an effective tool for monitoring and diagnosing the current and predicted integrated soil and climatic conditions of agricultural production. The deployment and application of contemporary information and computer technologies can enhance the efficiency of planning and implementing agrotechnical procedures through online monitoring of informative and destabilising factors regarding crop cultivation conditions with automatic DSS based on the processing of large amounts of measurement data using software tools based on ML and AI algorithms.

The importance of solving scientific and applied problems in the development and research of intelligent information and communication technologies for agrotechnical purposes is justified by the scale of open-field crop production in global and national formats, as shown in Fig. 1. The statistical data used to construct the diagrams in Fig. 1a and Fig. 1b is obtained by analysing the analytical resource (WEB, d), which is summarised by the FAO from 2013 to 2022 (the date of the last data update).

In order to localise the focus of applied research on the development of intelligent tools for predicting the probability of occurrence of specific types of crop diseases, the statistical data of the analytical resource (WEB, d) are analysed and logically summarised in terms of the most profitable and popular crops for the agroclimatic conditions of the forest-steppe and northern steppe of Ukraine (Dnipro and Cherkasy regions), namely wheat, maize and sunflower. The averaged data in relative form

(percentage of the total share of relevant crop production indicators in Ukraine) for the period from 2013 to 2022 by the respective crop types is shown in Fig. 2.

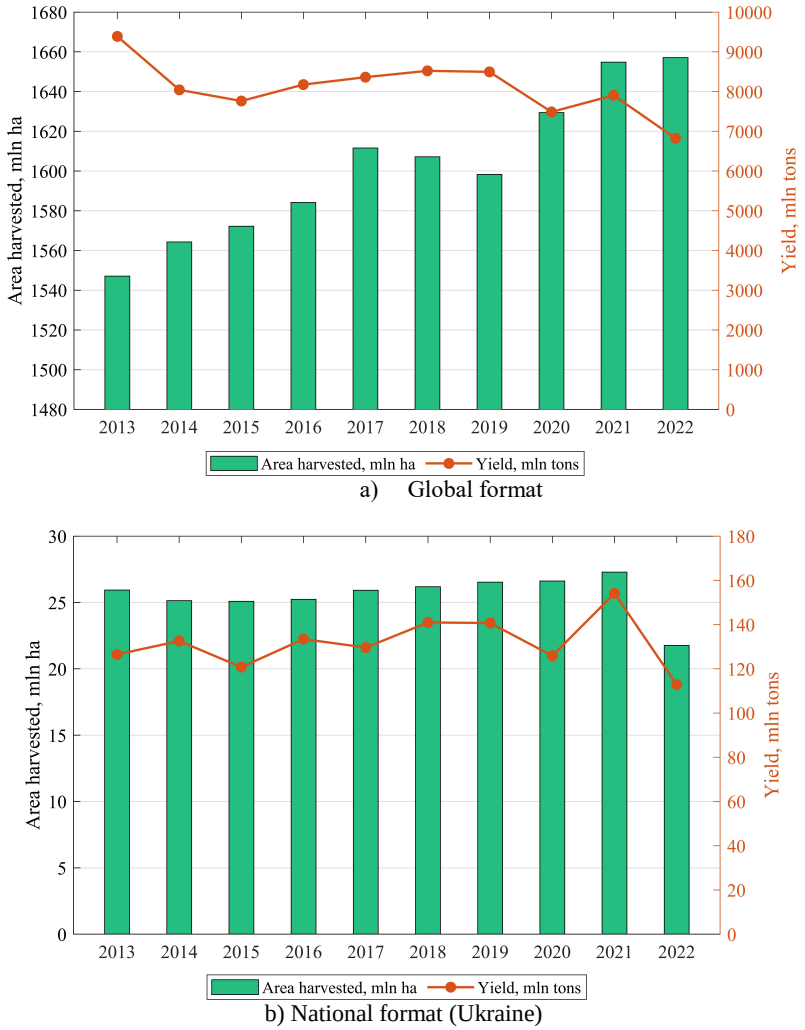
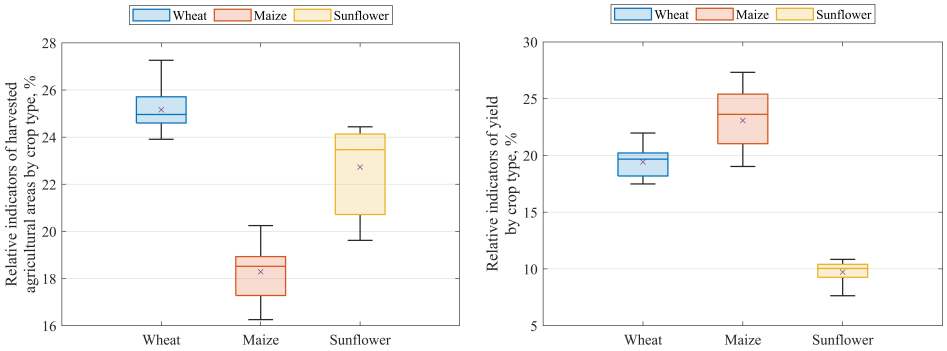


Figure 1. Statistical indicators of agricultural activities in crop production on a global and national scale

Based on the analysis of the statistical indicators shown in Fig. 1 and Fig. 2, the following has been established:

1. In recent years, the global dynamics of cultivated agricultural areas has been steadily increasing (an absolute increase of 110 million hectares, which is 7.1% in relative terms, over the period from 2013 to 2022). At the same time, there is no steady positive trend in the growth of crop yields: from 9388.6 million tonnes (6.07 t/ha) in 2013 to 6823.5 million tonnes (4.12 t/ha) in 2022. This evidence is an objective

confirmation that during the full cycle of cultivation, crops are significantly affected by a combination of destabilising factors (pests, climatic parameters, adequacy and timeliness of planning and implementation of agrotechnical procedures for crop preservation, and other), which in turn negatively affects the volume and quality of crop production.



a) Harvested agricultural areas b) Yield

Figure 2. Averaged relative indicators of agricultural open-field crop production by crop type for the period 2013-2022 in Ukraine

2. In Ukraine, taking into account the negative dynamics of cultivated agricultural areas due to the full-scale military aggression of Russia and taking into account the strategic importance of cultivation of cereal crops (maize: 18.3% of the total agricultural area and 23.1% of the total harvest; wheat: 25.2% of the total agricultural area and 19.4% of the total harvest; sunflower: 22.7% of the total agricultural area and 9.7% of the total harvest), ensuring national food security and competitive export potential of Ukraine in the global agricultural market, significantly increases the importance of rational use of agricultural land resources, including those affected by military actions, as well as preserving crops throughout the full cycle of agricultural crop production.

Therefore, taking into account the aforementioned, as well as the global trends of digitalisation and computerisation of the agricultural sector on a global and national scale, it can be concluded that it is relevant and important to develop and research with the subsequent implementation of software and hardware solutions of information and communication technologies aimed at enhancing the stress resistance of agricultural crops by detecting the probability of occurrence of specific types of diseases in specific types of crops in real-time.

The main scientific and applied problem that this article aims to solve is the development of the theory of engineering intelligent software and hardware solutions for the timely detection of the probability of specific types of grain crop diseases based on online measurements of a set of informative soil and climatic parameters of open-field crop production facilities. The main aspects that require additional research and detail this scientific and applied problem are as follows: identification of dynamic models for predicting the probability of crop diseases occurrence, taking into account the complex impact of a set of informative soil and climatic parameters at current and previous moments of time; justification of approaches to improving the efficiency of the

preprocessing stage of input datasets on measured soil and climatic parameters; consideration of agroclimatic conditions, types and periods of crop vegetation when developing computer models for predicting specific types of diseases; implementation of software components of the soil and climate monitoring system for agrotechnical purposes based on identified models, taking into account the possibility of their integration into microcomputer devices of the edge level of monitoring systems.

Thus, the main aim of this article is to substantiate scientific and applied approaches to improving software and hardware solutions for predicting the probability of occurrence of crop diseases in the example of maize by developing a method for increasing the efficiency of the preprocessing stages of input datasets and identifying dynamic models of regression analysis of the output function (probability of disease occurrence) based on ML algorithms, which will allow justifying the principles of improving the stress resistance of crops during the full cycle of their cultivation in open-field conditions.

The object of this study is the dynamic processes of intelligent processing of distributed measurement data on the soil and climatic conditions of agricultural enterprises of open-field crop production. The subject of the study is software methods and computer models for improving the efficiency of intelligent transformation of data from soil and climatic monitoring of agrotechnical facilities of open-field crop production.

1.2. Review, analysis and systematisation of relevant literature sources

Given the importance of justifying the ways of solving the scientific and applied problem addressed in this article, it should be noted that this determines the actualisation of the implementation of contemporary intelligent information technologies for integrated monitoring and management of agrotechnical processes in agricultural enterprises. On the one hand, this is due to the high level of scientific intensity of the problem under study, and on the other hand, it has a strong impact on the success of solving a significant set of environmental, economic and social issues. These include increasing the profitability and investment attractiveness of the agricultural sector, optimising the use of fertilisers and agrochemicals, ensuring a stable supply of quality agricultural products to the public, and so forth.

In analysing and summarising the state-of-the-art scientific achievements that correlate with the subject area of this article, the principle of hierarchical decomposition of the subject area was used according to the following logical chain: patterns of influence of climatic parameters on the efficiency of crop cultivation – approaches to predicting the probability of crop diseases based on the computerised acquisition and intellectual analysis of climatic factors – methods, approaches and models for preprocessing time series of observation results when applying ML and AI algorithms – methods and models for identifying dynamic models of computerised monitoring processes.

The results of the analysis of the relevant scientific sources correlated with the aim, object and subject of this article are presented in Table 1.

Table 1. Results of analysis and generalisation of relevant scientific and applied research

Object and subject of the study	Scientific and applied effect obtained	Scientific source
Scientific publications devoted to the identification of patterns of influence of climatic parameters on the efficiency of crop production		
Patterns of climate change impact on the occurrence and development of plant diseases	The potential negative aspects of the impact of climate change on the occurrence and development of plant diseases are analysed, in particular, according to the criteria of yield losses, effectiveness of agricultural management strategies, and geographical distribution. Promising areas of research to minimise the negative impact of climate change on the emergence and development of plant diseases are discussed	Chakraborty et al., 2000
Modelling the processes of climate change impact on the occurrence of agricultural crop diseases	Modern achievements in the field of research on the impact of climate change on the occurrence and development of agricultural crop diseases obtained by modelling methods are comprehensively analysed. The effectiveness of using approaches based on multimodel ensembles in the study of this problem is proved. Priority areas of research to ensure food and environmental security are discussed	Newbery et al., 2016
Patterns of influence of changes in climate parameters on the efficiency of the agricultural sector, as well as strategies to reduce the negative effects of climate dynamics on the agricultural system as a whole	The results of a detailed analysis of the patterns of influence of the dynamics of climatic parameters and factors on the agricultural sector are presented in terms of the following characteristics: efficiency of the agricultural system as a whole, qualitative and quantitative indicators of grain crops, the occurrence of pests and diseases, the use of agrochemicals, and other. Comprehensive justification of strategies to reduce the effects of climate change and adaptation of approaches to crop cultivation in agricultural practice	Yadav et al., 2021
Scientific publications devoted to the study of approaches to predicting the probability of crop diseases based on computerised acquisition and intellectual analysis of climatic factors		
Methods and tools for intelligent monitoring of the probability of agricultural crop diseases	Modern scientific and applied advances in the field of methods and tools for predicting the probability of crop diseases occurrence at the pre-symptomatic stage are comprehensively analysed and argued thoroughly	Fenu and Malloci, 2021
Structural and functional organisation of agricultural crop disease prediction systems	The structural and functional organisation of the system for predicting the occurrence of plant diseases based on fuzzy logic with a minimised set of input meteorological data has been proposed	Tilva, 2013
Methods of weather data analysis for rice yield prediction	The results of a comparative study of regression analysis methods based on climatic indices and ANNs of the multilayer perceptron type for predicting rice yields in West Bengal are presented. The input variables are measured data on minimum and maximum temperature, precipitation, and relative humidity, and their past dynamics are taken into account. The output variable is the rice yield.	Gupta, 2023

Information solutions for the aggregation of measurement data with integrated software for predictive analytics of the soil and climatic conditions	The effectiveness of wireless infocommunication systems for the acquisition and intelligent transformation of measurement data in predicting the dynamics of soil and climatic conditions of agricultural open-field crop production was experimentally proved	Laktionov et al., 2021; Laktionov et al., 2023a
Methods and models for predicting the occurrence of fungal diseases in wheat	The effectiveness of integrating ML algorithms, AI and weather models for seasonally adaptive forecasting of wheat fungal disease in the agroclimatic conditions of Morocco was studied	El Jarroudi et al., 2020
Methods and approaches to assessing the risk of potato diseases based on measurement monitoring of climatic parameters and AI algorithms	An AI-based approach to predicting potato late blight disease in the Sardinia region is substantiated on the basis of historical data (2016-2019) on temperature, humidity, precipitation, wind speed and solar radiation from different locations. It was proved that climatic factors (temperature, humidity and wind speed) play a key role in predicting potato diseases	Fenu and Malloci, 2020
Scientific publications devoted to the development of methods, approaches and models for preprocessing time series of observation results when applying ML and AI algorithms		
Methods for improving the efficiency of categorical data classification based on ML algorithms	A method of preprocessing input data when applying ML algorithms to reduce the dimensionality of an input dataset with missing values based on the Pairwise and Listwise methods is developed and described in detail	Ruiz-Chavez et al. 2018
Evaluation of the impact of data preprocessing algorithms on the performance of ML algorithms	The efficiency of using the SPQR-approximation method in the development of ML algorithms is studied. The results obtained are comparable to those obtained using the Principal Component Analysis (PCA) method, without fundamentally changing the initial dataset	Amato and Di Lecce, 2023
Data preprocessing algorithms using open-source ML libraries of Python	A methodology is developed and the effectiveness of using specialised Python libraries in solving problems of preprocessing input data arrays, which are subsequently used in ML algorithms, is evaluated	Pandey et al., 2020
Methodology for preprocessing time series of observation results	A new data preprocessing methodology is developed to extract information about the most informative variables when predicting a time series of observations in the oil refining industry. The proposed methodology is based on adding dynamic knowledge, reducing noise, reducing the dimensionality of the dataset and selecting informative features. Prediction metrics (MAE, MSE, SMAPE, and lag) are evaluated at each dynamic step, after which the final solution is generated by the system and sent to experts for further process optimisation	Cortes-Ibanez et al., 2020
Methods and tools for preprocessing numerical time series data	The results of a detailed analysis of methods for preprocessing numerical time series data according to various criteria, as well as an empirical analysis of the impact of data preprocessing methods on the performance of AI algorithms, are presented. In addition, the results of the analysis of the effectiveness of integrating preprocessing methods at the edge computing level, which reduces the server load and energy consumption, are presented	Tawakuli et al., 2024
Performance and effectiveness of automated ML tools in analysing time series of observation results	The performance of automated ML tools (AutoGluon, Auto-Sklearn, and PyCaret) when working with time series data is analysed and evaluated. It was found that the performance of each tool in solving the tasks of analysing and predicting time series data varies greatly depending on the specific datasets	Westergaard et al., 2024

Scientific publications devoted to the development of methods and models for identifying dynamic models of computerised monitoring processes

Methods for identifying parameters of mechatronic systems based on ML and AI	A comprehensive analysis of multiple modelling approaches to the identification of mechanical and electronic systems based on neural networks (feed-forward neural network (FNN), convolutional neural network (CNN), long short-term memory (LSTM), transformer), as well as machine learning methods (physically informed neural network (PINN) and sparse identification of nonlinear dynamics (SINDy)) is carried out. The possibilities of real application of AI methods and ML algorithms mentioned herein are presented on the example of the identification of parameters of a DC gear motor. Potential areas of application of the relevant models, as well as promising areas for further research, are indicated	Ayankoso, Olejnik, 2023
Approaches to analysing local patterns in time series data for anomaly identification	A new approach to the identification of anomalies in the time series of data based on the algorithm for maximising mathematical expectation by analysing the probabilistic behaviour of local patterns is proposed. The effectiveness of the proposed approach in decision-making for detecting anomalies in the time series of observation results is proved	Kotera et al., 2023
Methods for identifying models for predicting the probability of grain crop disease	A method for identifying informative parameters of the model for predicting the probability of maize diseases occurrence based on computerised measurements of informative climatic factors with consideration of values at previous moments of time and software processing of the results based on the adaptive neuro-fuzzy inference system (ANFIS) is developed. The possibility of integrating the developed methods and models into the edge level of Internet of Things (IoT) agrotechnical monitoring systems is proved	Laktionov et al., 2023*b*

Based on the analysis of the known relevant results of scientific and applied research on the development of methods, models and approaches to the advancement and improvement of software and hardware solutions for predictive analytics of time series of measurement monitoring results (see Table 1), has been found that AI and ML algorithms are highly effective and feasible in the implementation of software components for intelligent processing of time series of soil and climate monitoring results. In addition, it has been established that the efficiency and productivity of the use of intelligent algorithms for transforming measurement data significantly depends on the quality of the implementation of the stages of preprocessing the input datasets and identifying the parameters of predictive models for predicting the probability of agricultural crop diseases. It has also been established that the existing results of research on the development and application of algorithms for intelligent processing of climate monitoring data in detecting the occurrence of crop diseases are characterised by the following limitations: insufficient degree of elaboration of issues related to the impact of soil and climatic indicators dynamics on the effectiveness of identification and development of computer models for assessment and prediction of the state of agricultural facilities of open-field crop production; insufficient degree of substantiation of the formalised description of the impact of a set of soil and climatic parameters on predicting the probability of occurrence of specific types of crop diseases during the full cycle of their cultivation, taking into account the cumulative destabilising effect of

informative soil and climatic parameters and the probability of disease occurrence at previous points in time; insufficient degree of elaboration of the issues of integration of software components of intelligent data transformation with decision-making support for the integrated impact of soil and climatic parameters on the efficiency of agricultural crops cultivation to the low-level link (peripheral microcomputer devices), which in turn limits the applied principles of introducing information technologies based on edge architecture into the production processes of agricultural enterprises; lack of known scientific research on methods of preliminary processing of input datasets (identification and approximation of abnormal time intervals: a steep decrease or increase in the probability of disease occurrence), taking into account specific aspects of the applied field of application of the results (agricultural objects of open-field crop production in the forest-steppe and northern steppe of Ukraine).

It can therefore be noted that the scientific and applied problem of this article, which is to develop a theory of building intelligent software and hardware solutions for the timely detection of the probability of the occurrence of specific types of grain crop diseases based on online measurements of a set of informative soil and climatic parameters of open-field crop production facilities, is relevant. A potential approach to justifying ways to solve this problem is to develop methods to increase the efficiency of the preprocessing stage of input soil and climatic data, followed by the identification and software implementation of dynamic models for predicting the probability of crop diseases based on ML algorithms. The investigated approach to predicting the probability of occurrence of maize diseases, in contrast to the previously known ones, is as follows: approximation of abnormal parts of the time series of the training dataset (a steep decline in the probability function of the disease occurrence); identification of informative input features with consideration of their values at previous moments of time; comparative analysis of ML algorithms in terms of prediction accuracy by the metrics MAE (mean absolute error), RMSE (root mean square error) and R^2 (coefficient of determination); software implementation of the method for predicting the probability of maize diseases occurrence, taking into account the agroclimatic features of the forest-steppe and northern steppe of Ukraine, considering the possibility of integrating the relevant software components into the edge level of the information and communication system for agro-monitoring of the soil and climatic condition of open-field crop production facilities. The proposed approach is a further development of previously known software and hardware solutions in the field of agrotechnical monitoring and is a computer-oriented solution to the decision-making support for the optimisation of agrotechnical procedures to increase the stress resistance of crops to the dynamics of soil and climatic factors during the full cycle of their cultivation, based on contemporary scientific and practical achievements of the theory of identification of dynamic processes, methods of regression analysis, and ML algorithms.

1.3. Novelty and main contributions of the article

The main scientific and applied effect of the study of this article is a substantiated approach to improving the software and hardware solutions of infocommunication technologies for monitoring the soil and climatic conditions of agricultural open-field crop production through the development of software components and computer models for predicting the probability of maize diseases, which, unlike the known ones, implement online measurements of air temperature, relative humidity, precipitation, and leaf wetness duration, taking into account the types and periods of vegetation of crops

and specific diseases in specific agroclimatic conditions (forest-steppe and northern steppe of Ukraine), which are subject to software processing based on identified dynamic ML models with preliminary approximation of input datasets. This approach allows to increase the accuracy of predicting the probability of maize diseases during the full cycle of cultivation and can be implemented in the form of software components of microcomputer devices that form the edge computing link of information and communication monitoring systems for agrotechnical purposes.

A detailed description of the main scientific and applied results of this article is as follows: the algorithm for preprocessing (approximation) of abnormal areas of input datasets (a steep decrease in the probability of disease occurrence) with exponential dependencies is developed and investigated, which allows to reduce the destabilising effect of such abnormal areas in training datasets: depending on the type of ML algorithm (linear regression, random forest, feedforward neural network), the coefficient of determination for the system without data preprocessing varies from 0.901 to 0.987, and for the system with data preprocessing – from 0.986 to 0.997; the optimal value of the hyperparameter (*timesteps*=9) was identified, which is responsible for the number of previous time intervals for detecting physicochemical parameters (the dimension of the *timesteps* hyperparameter is hours), which should be taken into account when assessing the probability of maize diseases; the types of structural and algorithmic organisation of the system (presence or absence of data on the output parameter of the probability of occurrence of specific diseases at previous moments of time) according to the criterion of prediction accuracy, which allowed to justify the need to build software and hardware components of the system with the probability of occurrence of the disease at previous moments of time as an additional input parameter: depending on the type of ML algorithm (linear regression, random forest, feedforward neural network), the coefficient of determination for the system with the probability of occurrence of the disease occurrence at previous moments varies from 0.986 to 0.997, and for the system without the probability of the disease occurrence at previous moments – from 0.150 to 0.498; the effectiveness of different types of ML for the system with consideration of the probability of disease occurrence at previous moments of time under the optimal value of *timesteps*=9 and preprocessing of input datasets by metrics (MAE, RMSE and R^2) was investigated, which resulted in the establishment that the Random Forest algorithm is optimal for the agroclimatic conditions of the northern steppe (Dnipro region) of Ukraine and for the forest-steppe (Cherkasy region) of Ukraine. Quantitative indicators of prediction accuracy for the northern steppe: MAE=0.212, RMSE=0.88 and R^2=0.997, and for the forest-steppe conditions: MAE=0.206, RMSE=1.374 and R^2=0.995.

The generalised practical effect of the study of the article is a substantiated structural and algorithmic organisation of the computer-oriented technology of intelligent infocommunication monitoring with decision-making support to increase the stress resistance of agricultural crops. This is achieved by detecting the probability of occurrence of specific crop diseases during their cultivation in open-field conditions. This effect is a further advancement of well-known intelligent systems for predicting the probability of crop diseases based on computerised measurements of climatic parameters and their subsequent automated processing based on ML algorithms.

This is possible due to the software and hardware implementation of the following functionality: preliminary approximation of input datasets; identification of dynamic models of predictive analytics of the probability of occurrence of specific types of maize diseases in certain agroclimatic zones, taking into account the cumulative effect of soil

and climatic parameters and the output function (probability of disease occurrence) in previous periods of time; formalised accounting of many years of expert experience in the field of stress tolerance of agricultural field crops; software implementation of intelligent components that can be integrated into the low-level data processing link, which in turn allows for an approach to integrated agro-climatic monitoring based on edge architecture.

1.4. Structure and organisation of the article

The structural and logical organisation of the presentation of the study in this article is as follows: the first section presents the results of justification of the relevance, scientific and applied problem, aim, object and subject of the study, as well as systematic results of the analysis of relevant scientific sources of information in the subject area with an indication of research gaps that require further scientific and applied research; in the second section, a description of materials, methods and approaches to carrying out studies in this subject area is presented; in the third section, the main quantitative and qualitative results of studies on improving approaches to predicting the probability of occurrence of maize diseases based on ML algorithms for agroclimatic conditions of the forest-steppe and northern steppe of Ukraine are presented; the fourth section contains information on the results of a critical comparison of the results of this article with previously known ones, as well as a justification for further promising areas of research in this subject area; the fifth section provides general conclusions.

2. Materials and methods

In this section, questions on intellectual data analysis and model building based on ML algorithms are discussed. Computer experimental studies were carried out in the Google Collab cloud environment using the Python 3.10.12 programming language. The following Python libraries were used for data analysis, visualisation and machine learning: NumPy for performing linear algebra operations; Pandas for statistical data analysis; Matplotlib for creating charts and graphs; Keras for creating and evaluating the effectiveness of ML models; Sklearn for creating linear regression and Random Forest models; Fast-ML for dividing data into training, validation, and testing samples.

2.1. Data description

In this study, a dataset of professional weather stations from Metos by Pessl Instruments using the FieldClimate IoT platform is used, access to which is provided by Metos Ukraine LLC. The data is collected from September 2022 to September 2013 with a 1-hour sampling interval for two agroclimatic zones: the northern steppe (Dnipro region) and the forest-steppe (Cherkasy region) of Ukraine. The type of crop is maize. The diagnosed disease is Fusarium Head Blight. The dataset contains 17345 observations and 6 variables. The variables include datetime (date and time of the sensor polling), informative soil and climatic parameters (air temperature, ^{o}C; relative humidity, %; precipitation, mm; and leaf wetness duration, min), and the target variable for this dataset – disease probability, %. Thus, the Metos by Pessl Instruments weather station

dataset is a set of time-dependent data points. This means that each data point is associated with a specific timestamp.

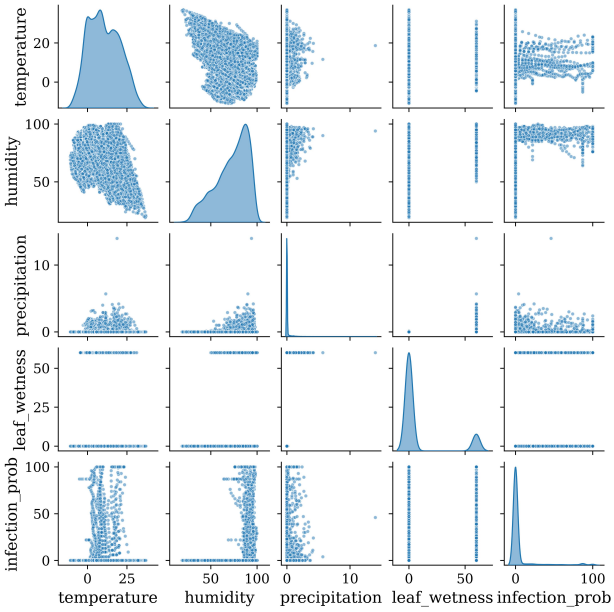
In Table 2, the statistical indicators are presented, such as the number of observations for each zone, the mean, standard deviation, median, minimum and maximum values for the collected climatic parameters and the probability of disease occurrence.

Table 2. Dataset descriptive statistics

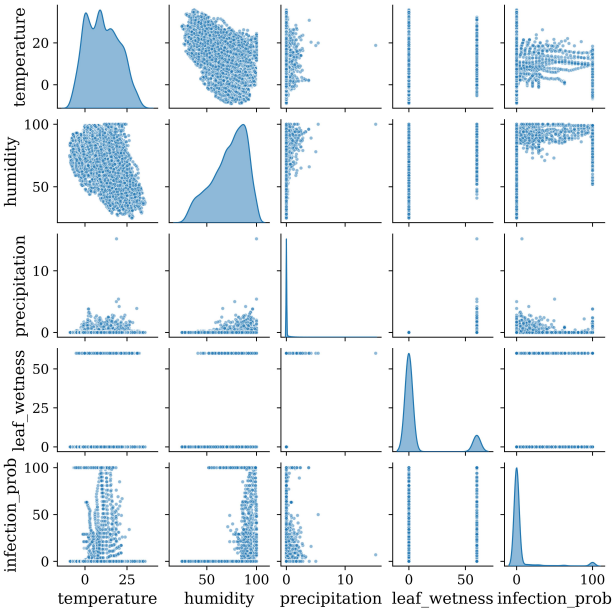
Statistical data	Air temperature,°C	Relative humidity, %	Precipitation, mm	Leaf wetness duration, min	Disease probability, %
The northern steppe of Ukraine					
Count	8658				
Mean	10.62	72.8	0.07	8.7	4.84
Standard deviation	9.74	17.32	0.34	21.13	17.61
Median	9.65	76	0	0	0
Min value	-10.9	19	0	0	0
Max value	37	100	14	60	100
The forest-steppe of Ukraine					
Count	8687				
Mean	10.64	72.4	0.06	8.5	5.7
Standard deviation	9.58	17.07	0.32	20.93	20.43
Median	9.90	76	0	0	0
Min value	-9.00	25	0	0	0
Max value	36	100	15.2	60	100

The scattering matrices with histograms of the distribution of climatic parameters and disease probabilities, as well as pair plots showing the mutual dependencies of changes in input (climatic parameters) and output (probability of maize disease occurrence) parameters, are shown in Fig. 3.

From the pair plots in Fig. 3 for the two agroclimatic zones it can be seen that most of the non-zero values of the probability of disease occurrence were observed in the temperature range from 0 °C to 20 °C. Also, almost all of the positive disease probabilities corresponded to relative humidity values above 85%. From the pair plot for leaf wetness time, it can be seen that for this climatic indicator, only two possible values were present in the data – 0 min or 60 min. This is interpreted as the following logical condition: whether the leaves were wet or not during the last hour.



a) The northern steppe



b) The forest-steppe

Figure 3. Pair plots

2.2. Data preprocessing

Since the datetime feature does not contain useful information for predicting the probability of maize disease, it was not considered as an input value in the ML models studied. However, datetime is used to arrange the data in chronological order, and the timestamps should be equidistant in the time series. The chronological order is achieved by sorting the data frame by timestamps. There are no missing values in the dataset under study, which eliminates the need for any explicit handling of such rows of data. This ensures that the entire dataset is complete and ready for further preprocessing and analysis without the need to apply any processing techniques or remove incomplete observations. A sample of 500 rows from the northern steppe data is shown in Fig. 4, which shows a typical local area of the dataset with anomalies. In Fig. 4, the probability of the disease occurring at some intervals gradually increases and then drops steeply to 0, which can negatively affect the predictive qualities of the model and, therefore, requires additional processing.

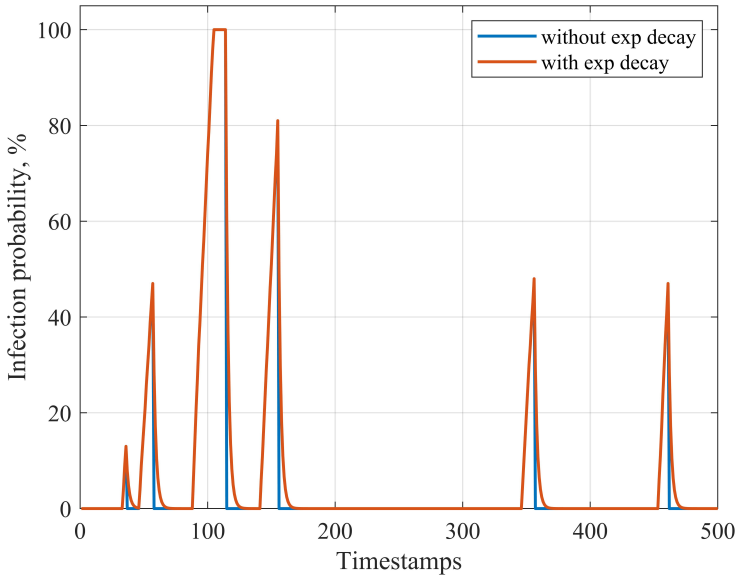


Figure 4. Change in disease probability for a sample of 500 values from the northern steppe dataset

To solve this problem, it is assumed that the use of an exponential decay function to smooth these abnormal intervals can improve the accuracy of model predictions. By implementing this approach, the expected effect is to mitigate abrupt transitions in the probability of disease occurrence, thereby contributing to a more continuous and consistent presentation of the underlying data models. This hypothesis suggests that exponential smoothing improves the predictive performance of the model by effectively reflecting gradual changes in the probability of disease occurrence and minimising the destabilising effect of steep drops to zero, which correlates with the actual physical processes of wetting and drying of the crop leaf cover.

To "smooth out" the intervals where the probability drops steeply to zero, the exponential decay function is used (1):

$$f(t) = Ae^{-kt}, \tag{1}$$

where A is the value of the probability of the disease before it drops to zero; t is the time (argument of the exponential decay function): at t=0, $f(0)=A$; k is the exponential decay constant (a larger value corresponds to a faster decrease in the function, a smaller value to a smoother decrease).

Thus, based on the hypothesis of using an exponential decay function to smooth the intervals of steep decrease to zero of the probability of disease occurrence, the research task of determining the optimal exponential decay coefficient arises (1). The choice of an appropriate decay factor involves finding a balance between preserving significant variations in the dataset and reducing the destabilising effect of such anomalous areas in the training datasets, depending on the type of ML algorithm. This task involves computational experiments with subsequent analysis to determine the impact of different decay coefficients on prediction accuracy.

Some ML models, such as neural networks, are trained using a numerical gradient descent optimisation method that achieves faster convergence on equally scaled data. The input data was scaled using normalisation and standardisation. Taking into account the distribution of leaf wetness duration and disease probability (Fig. 3), these indicators were normalised using the following equation (2) (Cao et al., 2016):

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{2}$$

where x' is the normalised value of the numerical characteristic x; max(x) and min(x) are the maximum and minimum values of x, respectively.

Precipitation, air temperature, and relative humidity were standardised using (3) (Wan, 2019):

$$x' = \frac{x - \bar{x}}{\sigma}, \tag{3}$$

where x' is the standardised value of the numerical characteristic x; $\bar{x}$ is the mean value of x in the training set; σ is the standard deviation of x in the training set.

2.3. Machine learning models

The machine learning process involves several important steps, including data acquisition and preprocessing, model selection, training the selected model on the dataset, evaluating its performance, and finally deploying the model. Throughout this iterative process, iterative experimentation and refinement are important to improve the performance of the model. The main objective is to develop a model that can effectively generalise to data that was not part of the training set and efficiently solve the prediction problem.

Effective hyperparameter tuning is a prerequisite for selecting the optimal model. The purpose of hyperparameter tuning is to use the knowledge gained from previous training iterations to improve the performance of the model. Fine-tuning the algorithm parameters helps to gradually improve the performance of the model. In this study, the search for optimal parameter configurations was performed using computer experimentation.

To take into account the history of changes in climate parameters and disease probabilities, an additional hyperparameter (*timesteps*) was introduced, which determines the number of previous timestamps of climate parameters and disease probabilities that should be used as input values to the machine learning model. For instance, for a value of *timesteps*=1, the current climate parameters together with the climate parameters and the probability of disease an hour ago are used as input values. For a *timesteps*=2, the current climate parameters together with the climate parameters and the disease probability for the last 2 hours are used as input values.

Given the introduction of a hyperparameter (*timesteps*) that regulates the depth of inclusion of historical data on climate parameters and disease probability in the studied ML models, the research task is to choose the optimal value of this parameter. The hyperparameter (*timesteps*) directly determines the amount of historical data fed into the model input and, as a result, affects its complexity. The objective of this study is twofold: first, to empirically investigate the impact of different values of timesteps on model performance and complexity, and second, to determine the optimal value that maximises prediction accuracy while minimising model complexity. ML pipeline used for the study is shown in Fig. 5.

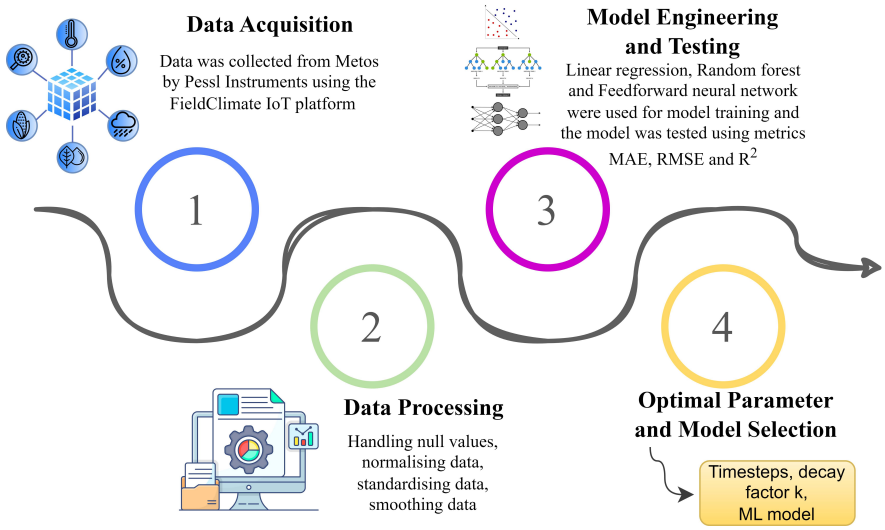


Figure 5. Machine learning pipeline

After preliminary data processing, the following ML models were trained:

1. Linear Regression: this model serves as the baseline, initially assuming a linear relationship between the target variable (occurrence of the disease) and the input variables. It provides a straightforward method to understand the correlation between predictors and disease occurrence.

2. Random Forest: an ensemble model consisting of multiple decision trees constructed from various subsets of the training data. By averaging the outcomes of these trees, random forest mitigates overfitting issues and enhances predictive accuracy. It is particularly adept at capturing complex relationships between predictors and disease occurrence due to its robustness against noise and non-linearity.

3. Feedforward Neural Network: unlike linear regression, this model can discern intricate patterns within the data owing to the inclusion of non-linear activation functions in its hidden layers. By iteratively adjusting weights and biases through backpropagation, the neural network can learn complex mappings between input features and the probability of Fusarium Head Blight disease occurrence.

In this study, an open-loop prediction approach is adopted, which involves the prediction of the next time step in a sequence using only the input data. This method relies on collecting true values from the data source and using them as input when making predictions for subsequent time steps. For instance, to predict the value for time step $t+1$, the data collected from time steps $t-timesteps$ through t is utilized. Subsequently, to predict the value for time step $t+2$, the true value for time step $t+1$ is recorded, and then the data from $(t+1)-timesteps$ through $t+1$ is used as input for the next prediction.

The training of ML models is done using two distinct approaches: firstly, the utilization of machine learning models solely fed with climatic data and their historical values; and secondly, the incorporation of machine learning models fed with both climatic data and disease probability, alongside their corresponding historical values.

Algorithm for training and selecting the most appropriate model for predicting the occurrence probability of Fusarium Head Blight disease in maize:

1. Data acquisition: climate data, including disease occurrence probabilities, is collected.

2. Feature engineering: preceding climate parameter values and disease probabilities for the last w hours are incorporated, where w is defined by the value of the hyperparameter "*timesteps*".

3. Data splitting: the dataset for the northern steppe (Dnipro region) is divided into training, validation, and testing sets using a 70:15:15 ratio, respectively. The dataset for the forest-steppe (Cherkasy region) is used only for testing purposes.

4. Data scaling: the training set is scaled to ensure uniformity in feature magnitudes.

5. Scaling consistency: the validation and testing sets are scaled using the statistical parameters derived from the training set, such as mean and standard deviation values of the climate parameters.

6. Model training: the coefficients of the model are adjusted using the training set. Regression metrics are computed on the training set.

7. Validation: regression metrics are calculated on the validation set to evaluate the performance of the model.

8. Overfitting assessment: the performance of the model on the validation set is compared to its performance on the training set to determine if overfitting has occurred. If overfitting is detected, the process proceeds to hyperparameter tuning (Step 8.1), otherwise, it advances to Step 9.

8.1. Hyperparameter tuning: model hyperparameters are adjusted to optimize performance.

8.2. Model refinement: the adjusted model is re-trained using the training set.

9. Testing: the performance of the considered models is assessed on the independent testing set to gauge their generalization capabilities.

10. Model selection: the most optimal model is selected based on predefined regression metrics, such as mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2), chosen to best represent the predictive accuracy and goodness of fit for the problem at hand.

2.4. Performance metrics

After the models were trained, they were evaluated by various performance indicators. These are quantitative indicators that allow assessing the ability of the applied ML models to make predictions. Table 3 shows the regression metrics used and the formulas for their calculation.

Table 3. Regression metrics used in this study

No.	Regression metric	Definition	Formula
1	MAE (mean absolute error)	Measure of the average size of the errors in a collection of observations expressing the same phenomenon, without taking their direction into account.	$MAE = \frac{1}{m}\sum_{i=1}^{m}\lvert y_i - y\_hat_i\rvert$
2	RMSE (root mean squared error)	Measure of the average difference between values predicted by a model and the actual values.	$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - y\_hat_i)^2}$
3	R^2 (coefficient of determination)	Measure of the goodness of fit of a model.	$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - y\_hat_i)^2}{\sum_{i=1}^{m}(y_i - \overline{y_i})^2}$
			where m – is the size of the dataset, y_i – is the true target value of the observation i, y_hat_i – is the predicted target value for the observation i, $\overline{y_i}$ – is the mean value of the target variable.

3. Results

3.1. Study by the criterion of the timesteps hyperparameter

The performance of the considered ML models was investigated for different values of the hyperparameter *timesteps*=(3, 5, 7, 9) in two modes: at the first stage – without taking into account the values of the disease probability (Table 4), and at the second stage – taking into account the values of the disease probability for the last *timesteps* hours (Table 5). Tables 4 and 5 show the results of the ML models.

Table 4. Results of ML models on the testing set of the northern steppe at different values of the timesteps hyperparameter, using only climate parameters as inputs

Timesteps	Linear Regression	Random Forest	FNN
3	R^2=0.138, RMSE=17.13, MAE=8.12	R^2=0.389, RMSE=14.42, MAE=5.463	R^2=0.246, RMSE=16.019, MAE=6.51
5	R^2=0.113, RMSE=14.47, MAE=6.92	R^2=0.333, RMSE=12.55, MAE=4.56	R^2=0.233, RMSE=13.457, MAE=5.17
7	R^2=0.139, RMSE=15.48, MAE=7.53	R^2=0.422, RMSE=12.69, MAE=4.495	R^2=0.306, RMSE=13.908, MAE=5.205
9	R^2=0.138, RMSE=16.307, MAE=7.902	R^2=0.457, RMSE=12.93, MAE=4.56	R^2=0.233, RMSE=15.379, MAE=5.434

Table 5. Results of ML models on the testing set of the northern steppe at different values of the timesteps hyperparameter, using the previous disease probabilities as additional inputs

Timesteps	Linear Regression	Random Forest	FNN
3	R^2=0.961, RMSE=3.61, MAE=0.85	R^2=0.964, RMSE=3.46, MAE=0.379	R^2=0.961, RMSE=3.609, MAE=0.75
5	R^2=0.901, RMSE=4.83, MAE=0.79	R^2=0.958, RMSE=3.11, MAE=0.357	R^2=0.922, RMSE=4.266, MAE=0.68
7	R^2=0.901, RMSE=5.01, MAE=0.93	R^2=0.972, RMSE=2.77, MAE=0.30	R^2=0.92, RMSE=4.71, MAE=0.85
9	R^2=0.973, RMSE=2.85, MAE=0.95	R^2=0.987, RMSE=1.96, MAE=0.27	R^2=0.965, RMSE=3.26, MAE=0.919

As can be seen from the results presented in Tables 4 and 5, the most optimal value for the *timesteps* hyperparameter according to the regression metrics is 9.

The use of additional previous disease probabilities as input values resulted in a significant improvement (Table 5) compared to models that took only historical climate indicators as inputs (Table 4).

Among the models that accepted disease probabilities as input values, Random Forest showed the highest accuracy on the test data: R^2=0.987, RMSE=1.96, MAE=0.27.

The hyperparameters of the Random Forest and FNN models are given below:

Random Forest: max_depth=5 (maximum tree depth), n_estimators=10 (number of trees).

Feedforward neural network: 12, 6, 4, and 4 neurons in the hidden layers, 1 neuron in the output layer, optimiser=AdamW (weight_decay=0.25), activation functions – ReLU for the hidden layers and linear at the output, batch_size=256, epochs=500.

In Fig. 6, the experimental and predicted by the Random Forest model values of the disease probability for *timesteps*=9 are shown, taking into account the previous disease probabilities.

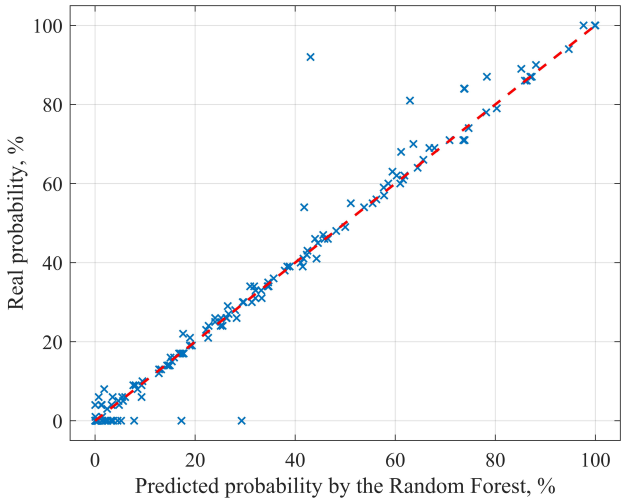


Figure 6. Probabilities of the disease at timesteps=9 - experimental and predicted by the Random Forest model

3.2. Study with exponential decay

In order to improve the results obtained with the hyperparameter *timesteps*=9, a study was conducted using exponential decay for values of the decay constant *k*=(0.25, 0.5, 0.75). The results of the comparison are presented in Tables 6 and 7 for the models that use only climate parameters and their prehistory and the models that additionally use the past probabilities of the disease, respectively.

Table 6. Results of ML models on the testing set of the northern steppe at timesteps=9 with climate parameters as inputs when using exponential decay

Value of k	Results of models		
	Linear regression	**Random forest**	**FNN**
0.25	R^2=0.15, RMSE=16.19, MAE=7.91	R^2=0.498, RMSE=12.447, MAE=4.47	R^2=0.33, RMSE=14.37, MAE=5.083
0.5	R^2=0.15, RMSE=16.19, MAE=7.91	R^2=0.498, RMSE=12.447, MAE=4.47	R^2=0.31, RMSE=14.58, MAE=5.29
0.75	R^2=0.15, RMSE=16.19, MAE=7.91	R^2=0.498, RMSE=12.447, MAE=4.47	R^2=0.318, RMSE=14.05, MAE=5.181

As can be seen from the values of the regression metrics in Tables 6 and 7, the results for all ML models improved when using exponential decay to approximate the intervals where the disease probability dropped steeply to zero. Changing the exponential decay coefficient *k* had almost no effect on the results, so for further studies, the value of *k*=0.75 was chosen, since at this value, the disease probability decreases most quickly and the approximated values are as close as possible to the original data.

Table 7. Results of ML models on the testing set of the northern steppe at timesteps=9 with the previous probabilities of the disease as additional inputs and with decay

Value of k	Results of models		
	Linear regression	**Random forest**	**FNN**
0.25	R^2=0.996, RMSE=1.006, MAE=0.331	R^2=0.997, RMSE=0.87, MAE=0.212	R^2=0.989, RMSE=1.83, MAE=0.639
0.5	R^2=0.996, RMSE=1.006, MAE=0.331	R^2=0.997, RMSE=0.88, MAE=0.212	R^2=0.986, RMSE=2.09, MAE=0.83
0.75	R^2=0.996, RMSE=1.006, MAE=0.331	R^2=0.997, RMSE=0.88, MAE=0.212	R^2=0.987, RMSE=2.03, MAE=0.703

In Fig. 7, the experimental and predicted by the Random Forest model disease probabilities on the test data are compared when applying exponential decay and when using the disease probabilities at previous time points as additional input values.

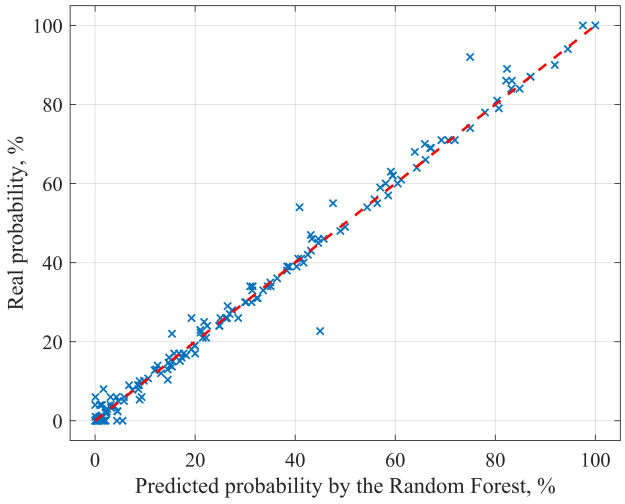


Figure 7. Experimental and predicted by the Random Forest model probabilities of the disease when applying exponential decay

3.3. Comparison of results for different agroclimatic zones

With the optimal preprocessing hyperparameters (*timesteps*=9 and *k*=0.75), the performance of ML and AI models with only climate parameters and their values at previous time points as inputs and models with the history of disease probabilities as additional inputs is compared in Tables 8 and 9, respectively.

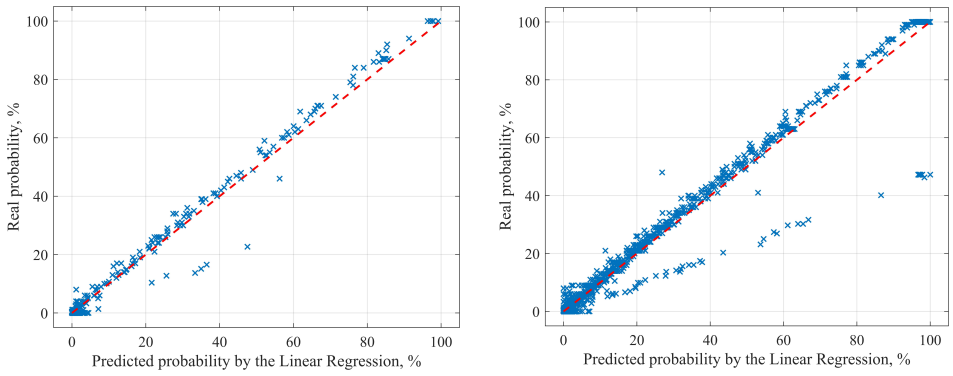
Table 8. Comparison of ML models at timesteps=9 and k=0.75 with climate parameters as inputs

Model type	Northern steppe sample	Forest steppe sample
Linear Regression	R^2=0.15, RMSE=16.19, MAE=7.91	R^2=0.134, RMSE=18.88, MAE=8.61
Random Forest	R^2=0.498, RMSE=12.447, MAE=4.47	R^2=0.362, RMSE=16.20, MAE=6.45
FNN	R^2=0.31, RMSE=14.58, MAE=5.29	R^2=0.260, RMSE=17.46, MAE=6.98

Table 9. Comparison of ML model results at timesteps=9 and k=0.75 with disease probabilities as additional inputs

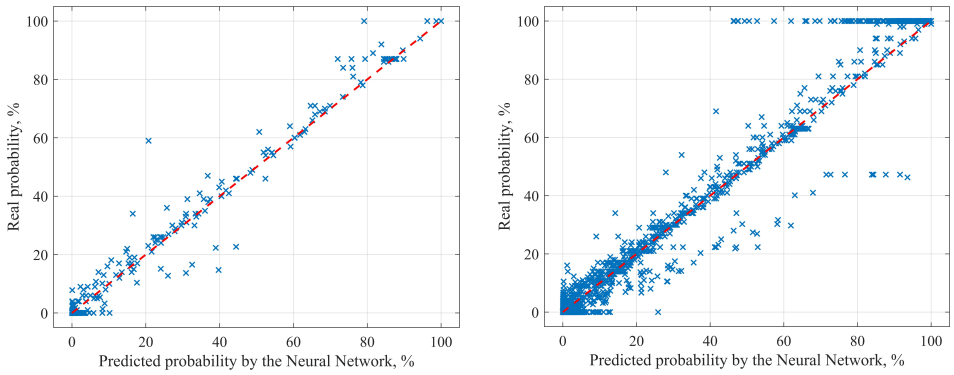
Model	Northern steppe sample	Forest steppe sample
Linear Regression	R^2=0.996, RMSE=1.006, MAE=0.331	R^2=0.986, RMSE=2.33, MAE=0.553
FNN	R^2=0.986, RMSE=2.09, MAE=0.83	R^2=0.976, RMSE=3.09, MAE=0.74
Random Forest	R^2=0.997, RMSE=0.88, MAE=0.212	R^2=0.995, RMSE=1.374, MAE=0.206

In Figs. 8 – 10, the experimental and predicted probabilities by Linear regression, Random Forest and Feedforward neural network from Table 9 are compared for the datasets of the agro-climatic conditions of the northern steppe and forest-steppe of Ukraine.



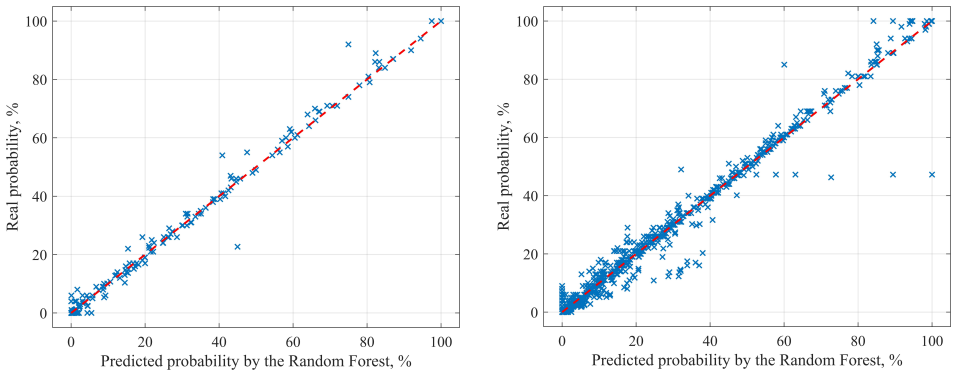
a) The northern steppe b) The forest-steppe

Figure 8. Results of model investigation based on the Linear regression algorithm



a) The northern steppe b) The forest-steppe

Figure 9. Results of model investigation based on the FNN algorithm



a) The northern steppe b) The forest-steppe

Figure 10. Results of model investigation based on the Random Forest algorithm

By comparing the effectiveness of different types of ML models for the system with consideration of the probability of disease occurrence at previous time points under the optimal value of *timestep*=9 and preprocessing of input datasets, it was found that for the agro-climatic conditions of the northern steppe (Dnipro region) and forest-steppe (Cherkasy region) of Ukraine, the Random forest algorithm has the best predictive characteristics by the metrics MAE, RMSE and R^2.

3.4. Synthesised structural and algorithmic organisation

Based on the research conducted to substantiate computer models and software tools for intelligent processing of climatic parameters in order to detect the predicted values of the probability of maize diseases for the agro-climatic conditions of the forest-steppe and

northern steppe of Ukraine, the structural and algorithmic organisation of the relevant software and hardware has been substantiated, as shown in Fig. 11.

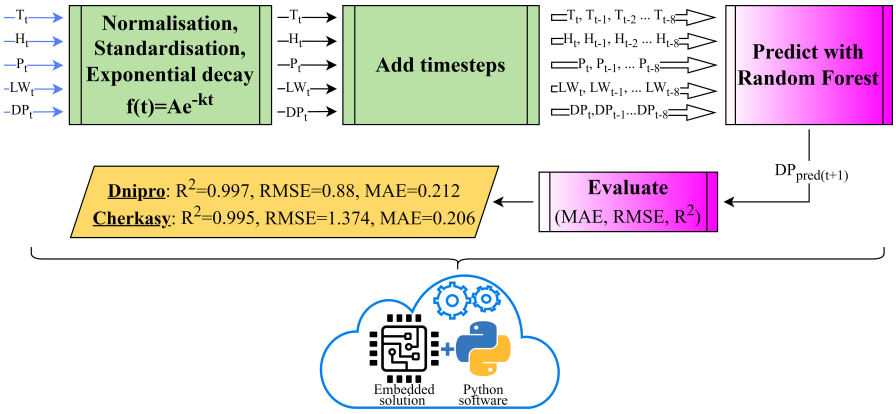


Figure 11. Structural and algorithmic organisation of the developed system

The symbols in Fig. 11 are as follows: *T* – air temperature, ºC; *H* – relative humidity, %; *P* – precipitation, mm; *LW* – leaf wetness duration, min; *DP* – probability of disease occurrence, %; *t* – sample time.

Thus, the developed structural and algorithmic organisation of the system (see Fig. 11) can be used as a basis for predicting the probability of maize diseases using edge computing technology according to the results of computerised online monitoring and automatic intelligent processing of influential climatic parameters.

4. Discussion and suggestions for future investigations

In this article, the effectiveness of scientific and applied provisions for solving the actual research problem of advancing the theory of building intelligent software and hardware solutions for detecting the probability of occurrence of specific types of grain crop diseases at the pre-symptomatic stage based on online measurements of a set of informative climatic parameters of open-field crop production facilities has been developed and proved. This effect has been achieved through the development of computer models and software components for the intelligent transformation of climatic parameters, which is a further advancement of the known results in the field of software and hardware solutions for information technologies for agrotechnical purposes, namely:

1. An algorithm for approximating abnormal regions of the time series of the training sample of climate data by an exponential relationship during maize cultivation in the agroclimatic conditions of the forest-steppe and northern steppe of Ukraine has been developed and proved to be effective, which allows increasing the efficiency of the data preprocessing stage when applying ML algorithms.

2. An approach to improving the efficiency of predicting the probability of Fusarium Head Blight disease occurrence is proposed by identifying the optimal value of the

hyperparameter (*timesteps*) responsible for the number of hours of preliminary accounting of climatic parameters and the probability of the disease occurrence.

3. Software components for the intelligent transformation of the cumulative effect of climatic parameters on the probability of maize disease occurrence based on ML algorithms in Python have been developed. As a result of these studies, computer experiments have shown that the software implementation based on the Random Forest algorithm is optimal for the agro-climatic conditions of the forest-steppe and northern steppe of Ukraine.

4. A detailed structural and algorithmic organisation of the process of computerised monitoring and intellectual transformation of climate data for detecting the probability of the occurrence of Fusarium Head Blight disease in maize is substantiated, taking into account current trends in edge computing.

Promising areas for further research of the proposed approach to predicting the probability of occurrence of crop diseases are: integration of the developed computer models and software components into the industrial solutions of computerised weather stations used, followed by long-term experimental testing in real operating conditions; expansion of the list of diagnosed diseases and types of crops; implementation and research of the proposed approaches based on an expanded list of ML algorithms; comprehensive assessment of technical and economic indicators of the studied software and hardware implementations of information technologies for predicting the probability of agricultural crop diseases.

5. Conclusion

In this article, the effectiveness of scientific and applied approaches to improving software and hardware solutions of information technologies for predicting the probability of occurrence of grain crop diseases on the example of Fusarium Head Blight of maize is developed and proved. This effect was achieved through the implementation of comprehensive research to develop a method for improving the efficiency of the stages of preprocessing the input arrays of training data (climatic parameters) and identifying dynamic models for regression analysis of the output function (probability of disease occurrence) based on machine learning and artificial intelligence algorithms, namely:

1. An analysis of the state-of-the-art in the field of development and implementation of intelligent information technologies for agrotechnical monitoring for diagnosing crop diseases based on computerised measurements and intelligent processing of climatic parameters has been carried out, which allowed localising the current research gaps in the subject area under study.

2. A computerised methodology for studying software tools for processing climatic parameters has been developed to improve the efficiency of predicting the Fusarium Head Blight disease of maize, which allowed the computer experiment to assess the feasibility of using ML in predictive analytics of the combined effect of air temperature, relative humidity, precipitation and leaf wetness duration on the probability of the above-mentioned disease of maize.

3. The model of accounting for climatic parameters and the probability of occurrence of Fusarium Head Blight at previous moments of time on the predicted value of the probability of occurrence of this disease has been identified and implemented: the optimal value of the *timesteps* parameter (responsible for the number of hours of data

prehistory) is 9 hours for the agroclimatic conditions of the forest-steppe and northern steppe of Ukraine.

4. An algorithm for preprocessing local patterns of the input data sets with a steep decrease in the probability of disease occurrence by exponential dependencies has been developed and implemented in Python. This allowed to improve the quality of data processing by reducing the negative impact of such abnormal areas in the training datasets: the coefficient of determination for the system without preliminary exponential data approximation varies from 0.901 to 0.987, and for the system with preliminary exponential data approximation – from 0.986 to 0.997, depending on the type of ML algorithm used (linear regression, feedforward neural network, random forest).

5. It has been established that the optimal for the agro-climatic conditions of the forest-steppe and northern steppe of Ukraine is the use of the random forest algorithm in the development of software and hardware solutions for detecting the probability of Fusarium Head Blight disease in maize: for the northern steppe (Dnipro region) – R^2=0.997, RMSE=0.88, MAE=0.212 and the forest-steppe (Cherkasy region) – R^2=0.995, RMSE=1.374, MAE=0.206.

6. The requirements to the structural and algorithmic organisation of the system of intellectual information technology for detecting the probability of occurrence of grain crop diseases in the agroclimatic conditions of Ukraine have been substantiated.

7. A list of promising directions for further research to improve the efficiency of computerised monitoring of agrotechnical facilities of open-field crop production has been formed.

Acknowledgement

This research was carried out as part of the scientific project ‘Development of software and hardware of intelligent technologies for sustainable crop production in wartime and post-war’ funded by the Ministry of Education and Science of Ukraine at the expense of the state budget (state registration number 0124U000289).

List of abbreviations:

ANFIS	adaptive neuro-fuzzy inference system
ANN	artificial neural network
CNN	convolutional neural network
DSS	decision support system
FAO	Food and Agriculture Committee of the United Nations
FNN	feedforward neural network
IoT	Internet of Things
LSTM	long short-term memory
MAE	mean absolute error
ML	machine learning
MSE	mean squared error
PCA	principal component analysis
PINN	physics-informed neural network
RMSE	root mean square error
SINDy	sparse identification of nonlinear dynamics
SMAPE	symmetric mean absolute percentage error

References

Amato, A., Di Lecce, V. (2023). Data preprocessing impact on machine learning algorithm performance. *Open Computer Science*, **13** (1): 1–16. https://doi.org/10.1515/comp-2022-0278.

Ayankoso, S., Olejnik, P. (2023). Time-Series Machine Learning Techniques for Modeling and Identification of Mechatronic Systems with Friction: A Review and Real Application. *Electronics*, **12** (17):3669: 1–27. https://doi.org/10.3390/electronics12173669.

Cao, X.H., Stojkovic, I., Obradovic, Z. (2016). A robust data scaling algorithm to improve classification accuracies in biomedical data. *BMC Bioinformatics*, **17** (359): 1–10. https://doi.org/10.1186/s12859-016-1236-x.

Ceccarelli, T., Chauhan, A., Rambaldi, G., Kumar, I., Cappello, C., Janssen, S., McCampbell, M. (2022). Leveraging automation and digitalization for precision agriculture: Evidence from the case studies. Background paper for The State of Food and Agriculture 2022. *FAO Agricultural Development Economics Technical Study No. 24*. Food and Agriculture Organization of the United Nations, Rome, 120 p. https://doi.org/10.4060/cc2912en.

Chakraborty, S., Tiedemann, A.V., Teng, P.S. (2000). Climate change: potential impact on plant diseases. *Environmental Pollution*, **108** (3): 317–326. https://doi.org/10.1016/S0269-7491(99)00210-9.

Cortes-Ibanez, J.A., Gonzalez, S., Valle-Alonso, J.J., Luengo, J., Garcia, S., Herrera, F. (2020). Preprocessing methodology for time series: An industrial world application case study. *Information Sciences*, **514**: 385–401. https://doi.org/10.1016/j.ins.2019.11.027.

El Jarroudi, M., Lahlali, R., El Jarroudi, H., Tychon, B., Belleflamme, A., Junk, J., Denis, A., El Jarroudi, M., Kouadio, L. (2020). Employing Weather-Based Disease and Machine Learning Techniques for Optimal Control of Septoria Leaf Blotch and Stripe Rust in Wheat. In: *Ezziyyani, M. (eds) Advanced Intelligent Systems for Sustainable Development (AI2SD'2019). AI2SD 2019. Advances in Intelligent Systems and Computing*, **1103**: 157–165. https://doi.org/10.1007/978-3-030-36664-3_18.

Fenu, G., Malloci, F.M. (2020). Artificial Intelligence Technique in Crop Disease Forecasting: A Case Study on Potato Late Blight Prediction. In: *Czarnowski, I., Howlett, R., Jain, L. (eds) Intelligent Decision Technologies. IDT 2020. Smart Innovation, Systems and Technologies*, **193**: 79–89. https://doi.org/10.1007/978-981-15-5925-9_7.

Fenu, G., Malloci, F.M., (2021). Forecasting Plant and Crop Disease: An Explorative Study on Current Algorithms. *Big Data and Cognitive Computing*, **5** (1):2: 1–24. https://doi.org/10.3390/bdcc5010002.

Gupta, A., Sarkar, K., Dhakre, D., Bhattacharya, D. (2023). Weather based crop yield prediction using artificial neural networks: A comparative study with other approaches. *MAUSAM*, **74** (3): 825–832. https://doi.org/10.54302/mausam.v74i3.174.

Kotera, K., Yamaguchi, A., Ueno, K. (2023). Learning Local Patterns of Time Series for Anomaly Detection. *Engineering Proceedings*, **39** (1):82: 1–10. https://doi.org/10.3390/engproc2023039082.

Laktionov, I., Diachenko, G., Koval, V., Yevstratiev, M., 2023*a*. Computer-Oriented Model for Network Aggregation of Measurement Data in IoT Monitoring of Soil and Climatic Parameters of Agricultural Crop Production Enterprises. *Baltic Journal of Modern Computing*, **11** (3): 500–522. https://doi.org/10.22364/bjmc.2023.11.3.09.

Laktionov, I., Diachenko, G., Rutkowska, D., Kisiel-Dorohinicki, M. (2023*b*). An Explainable AI Approach to Agrotechnical Monitoring and Crop Diseases Prediction in Dnipro Region of Ukraine. *Journal of Artificial Intelligence and Soft Computing Research*, **13** (4): 247–272. https://doi.org/10.2478/jaiscr-2023-0018.

Laktionov, I.S., Vovna, O.V., Kabanets, M.M., Sheina, H.O., Getman, I.A. (2021). Information model of the computer-integrated technology for wireless monitoring of the state of microclimate of industrial agricultural greenhouses. *Instrumentation Mesure Metrologie*, **20** (6): 289 – 300. https://doi.org/10.18280/i2m.200601.

Newbery, F., Qi, A., Fitt, B.D.L. (2016). Modelling impacts of climate change on arable crop diseases: progress, challenges and applications. *Current Opinion in Plant Biology*, **32**: 101–109. https://doi.org/10.1016/j.pbi.2016.07.002.

Pandey, N., Patnaik, P.K., Gupta, S. (2020). Data Pre-Processing for Machine Learning Models using Python Libraries. *International Journal of Engineering and Advanced Technology*, **9** (4): 1995–1999. https://doi.org/10.35940/ijeat.D9057.049420.

Ruiz-Chavez, Z., Salvador-Meneses, J., Garcia-Rodriguez, J. (2018). Machine Learning Methods Based Preprocessing to Improve Categorical Data Classification. In: *Yin, H., Camacho, D., Novais, P., Tallon-Ballesteros, A. (eds) Intelligent Data Engineering and Automated Learning – IDEAL 2018. IDEAL 2018. Lecture Notes in Computer Science*, **11314**: 297–304. https://doi.org/10.1007/978-3-030-03493-1_32.

Tawakuli, A., Havers, B., Gulisano, V., Kaiser, D., Engel, T. (2024). Survey: Time-series data preprocessing: A survey and an empirical analysis. *Journal of Engineering Research*: 1–38. https://doi.org/10.1016/j.jer.2024.02.018.

Tilva, V., Patel, J., Bhatt, C. (2013). Weather based plant diseases forecasting using fuzzy logic. In: *2013 Nirma University International Conference on Engineering (NUiCONE)*: 1–5. https://doi.org/10.1109/NUiCONE.2013.6780173.

Wan, X. (2019). Influence of feature scaling on convergence of gradient iterative algorithm. *Journal of Physics: Conference Series*: 1–6. https://doi.org/10.1088/1742-6596/1213/3/032021

Web (a): *ITU and FAO, 2020. Status of Digital Agriculture in 18 countries of Europe and Central Asia. International Telecommunication Union*. Available at: https://fao.org/3/ca9578en/CA9578EN.pdf [30 August 2024].

Web (b): *EU Science Hub: Agricultural monitoring*. Available at: https://joint-research-centre.ec.europa.eu/scientific-activities-z/agricultural-monitoring_en [28 August 2024].

Web (c): *National Economic Strategy 2030*. (in Ukrainian). Available at: https://nes2030.org.ua/ [28 August 2024].

Web (d): *FAOSTAT: Food and agriculture organization of the united nations*. Available at: https://www.fao.org/faostat/en/#data/QCL [26 August 2024].

Westergaard, G., Erden, U., Mateo, O.A., Lampo, S.M., Akinci, T.C., Topsakal, O. (2024). Time Series Forecasting Utilizing Automated Machine Learning (AutoML): A Comparative Analysis Study on Diverse Datasets. *Information*, **15** (1):39: 1–20. https://doi.org/10.3390/info15010039.

Yadav, P., Jaiswal, D.K., Sinha, R.K. (2021). 7 – Climate change: Impact on agricultural production and sustainable mitigation. In: *Singh, S., Singh, P., Rangabhashiyam, S., Srivastava, K.K. (eds) Global Climate Change*: 151–174. https://doi.org/10.1016/B978-0-12-822928-6.00010-1.

Received September 9, 2024, accepted November 5, 2024