Assessing of Climate Impact on Wheat Yield using Machine Learning Techniques

Petro Hrytsiuk[a], Tetiana Babych[a], Serhii Baranovsky[a], Maksym Havryliuk[a]

[a] *The National University of Water and Environmental Engineering, Soborna str., 11, Rivne, 33000, Ukraine*

Abstract

The comprehensive informatization of society allows to receive an increasing amount on the agriculture yield data and climate data. Thus, it becomes possible to use the available data to forecast grain yields and grain prices, to analyze crop losses and more. Climatic factors play a decisive role in wheat yield fluctuations. The use of climate data bases makes it possible to build the yield prognostic models, which allow in advance (in 3 months) to estimate the future yield. Pre-harvest yield forecasting can assist grain producers in making the necessary arrangements for storage and marketing of the crop. Such forecasts will also help farmers in planning the logistics of their business.

In this work, the average decadal temperature values of April, May, and June and the monthly amounts of precipitation for six regions of the chernozem zone of Ukraine were chosen to study the impact of climate on the wheat yield. The task of this research is to assess the impact of climatic factors on detrended wheat yield values using machine learning techniques. The work uses an innovative approach, according to which detrended yield values are divided into two groups, labeled as "low yield" and "high yield". Five machine learning models were used as classifiers, which were adapted to the available data and demonstrated a classification accuracy about 80% on test samples. The linear discriminant analysis and the logistic regression model are the most effective classifiers and provide 87% classification accuracy.

Keywords

Wheat yield, climatic factors, machine learning, classification, cross-validation

1. Introduction

In Ukraine grain production is one of the main branches of the economy, which ensures the food needs of the population and a stable inflow of currency into the state budget. The average annual production of grain in Ukraine for 2019-2021 reached the level of 75 million tons, and the average annual export during this time amounted to 50 million tons [1] (Figure 1).

At the same time, the grain production volume experiences significant fluctuations, which is associated with the impact of changing climatic factors. Climate changes over the past 30 years have led to a change in the assortment of cultivated grain crops and the geography of their location [2, 3, 4]. In the chernozem zone of Ukraine and in the Polissia zone, there is an increase in the production of heat-loving crops, such as corn, soybeans, and sunflowers. At the same time, in recent years, the share of wheat in the total grain harvest has decreased from 50% to 40%, and the share of corn has increased from 15% to 42% [3, 5]. The steppe region of Ukraine is particularly sensitive to changes in climatic factors, where frequent droughts lead to a significant drop in grain yields. Therefore, this region is losing its leading position in grain production, instead, the share of the central and western regions of Ukraine is increasing.

In recent years, domestic consumption of grain in Ukraine did not exceed 20 million tons. This represents about 30% of total grain production, while the remaining 70% of grain is exported. In the

ICST-2023: Information Control Systems & Technologies, September 21-23, 2023, Odesa, Ukraine.
EMAIL: p.m.hrytsiuk@nuwm.edu.ua (P. Hrytsiuk); t.iu.babych@nuwm.edu.ua (T. Babych); s.v.baranovskyi@nuwm.edu.ua (S. Baranovsky); m.s.havryliuk@nuwm.edu.ua (M. Havryliuk)
ORCID: 0000-0002-3683-4766 (P. Hrytsiuk); 0000-0001-6927-7313 (T. Babych); 0000-0002-8056-2980 (S. Baranovsky); 0000-0003-1149-6251 (M. Havryliuk)

last three years alone, revenues from grain exports amounted to about 30 billion US dollars. Thus, grain production has turned into the largest source of foreign exchange earnings for the state and a guarantee of its successful economic development.

The basis for planning a long-term grain export strategy is grain yield forecasting. Forecasting yield is a difficult task, given the random nature of many impact factors. This determines the intelligent data analysis methods application with modern computer technologies using. Today, the latest methods of mathematical modeling are used to build predictive models, among which machine learning techniques and the technology of artificial neural networks take the leading place.

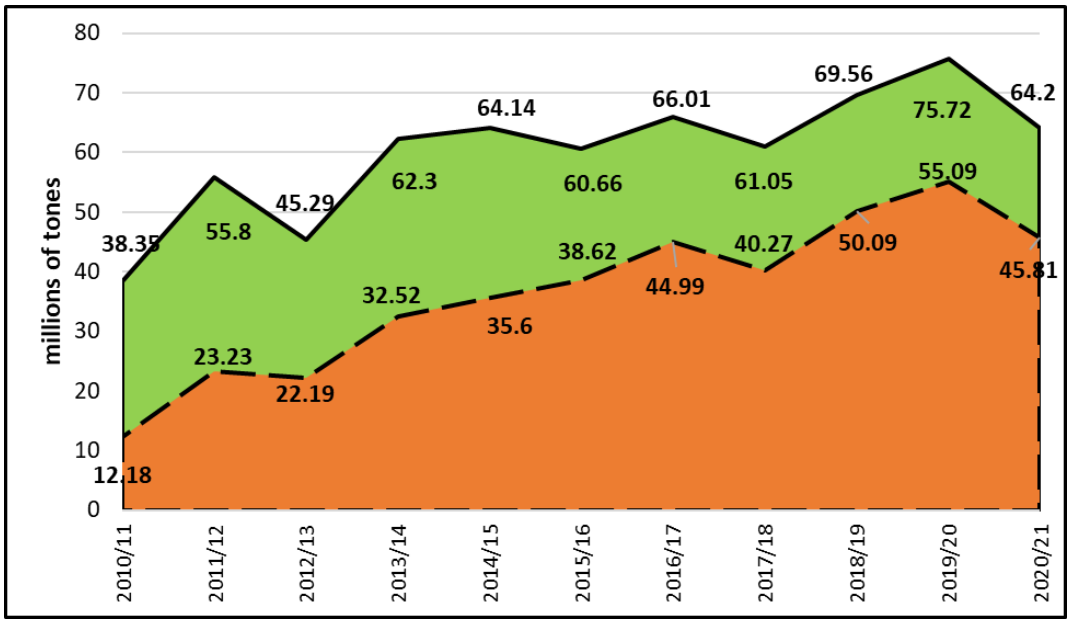

Figure 1: Dynamics of grain production and export from Ukraine [6]. Solid line – grain production, dashed line – grain export

2. Literature Review

The yield of agricultural crops largely depends on the weather and climate. Climate changes in recent years pose threats to sustainable agriculture. Understanding of climate impact on crop yields is important for the agricultural production development in the face of future climate change. Identification of climatic factors, which are the dominant causes of the agricultural yield fluctuations, is important for forecasting regional crop production. Many researches in recent years are devoted to the impact of climatic factors on the agricultural crops yield.

Wheat production is the basis of Ukrainian agriculture, but climate change threatens it at risk in some regions of Ukraine. In a large analytical review carried out as part of the German-Ukrainian Agricultural Policy Dialogue project [7], the impact of climate change on the winter wheat yield for Ukraine and apart for three separate ecological zones of Ukraine was assessed. According to the authors' conclusion, the main concern is the fertile steppe zone, where the climate is hotter and drier, and frequent droughts are also observed. Future climate changes may intensify these negative effects and cause desertification of this region.

Scientific and technical progress contributed to the arrival of large volumes of statistical data from various branches of agriculture. This greatly expanded the possibilities of using computer technologies for the analysis and modeling of climatic effects on the yield of agricultural crops. The advent of big data technology has led to the emergence of powerful new analytical tools, such as machine learning techniques, which have proven themselves in medicine, finance, and biology. In recent years, there have been publications describing the machine learning methods application to forecasting the agricultural crops yield.

The paper [8] aims to identify the best yield prediction model that can help farmers decide which crop to grow based on climate conditions and nutrients present in the soil. In an analysis of yield prediction by three different supervised machine learning models, the authors concluded that the best accuracy was achieved with the Random Forest Classifier in both Entropy and Gini Criterion.

Early and reliable seasonal yield forecasts are critical both at the state level and for farmers in particular. In the study [9], the authors developed an improved wheat yield forecasting system based on statistical regression. Data sources such as yield model output, extreme climate indices, and remote sensing data were combined into two models. More accurate forecasts of wheat yield are obtained in a model based on machine learning compared to a traditional multiple linear regression model. The optimal forecast events that produced reasonably accurate yield forecasts were those that provided one to two months of lead time to harvest. Droughts during the growing season have been identified as the most significant extreme climatic phenomenon causing crop losses in the wheat belt. In the paper [10] proposed predictive models for state-level wheat yield variations in Australia. The models combine large-scale climate factors with more localized environmental predictions using machine learning techniques. The obtained results demonstrate reliable forecasting of wheat yield fluctuations at the state level three months before the start of harvest.

A study by [11] proposes a late fusion convolutional neural network (CNN-LF) architecture. The advantages of this model are that after training on sufficient data from one field, it can be effectively used in another field with less data available. The authors also see the prospect of using the CNN-LF model to solve geospatial problems.

Machine learning allows you to effectively analyze various aspects that affect the management of agriculture. A review [12] presents a description and comparison of different machine learning techniques for building predictive models of crop yields using different error measures. The involvement of these methods in the research will make it possible to analyze the soil, climate and water regimes, which in turn significantly affects the growth of crops and precision agriculture.

A demonstration of the more effective use of machine learning and regression methods compared to process-based models can also be seen in [13]. The authors analyzed the expediency of using various methods for forecasting long-term trends, the average level and extreme values of corn yield on the example of the USA. More powerful possibilities for predicting the yield of the machine learning algorithm in the case of the presence of only two climatic variables have been determined. It is emphasized that supplementing process-based models with machine learning techniques to reduce the scale of temporal and spatial models. Since the role of climate change in abnormal yield may be dominant, it is recommended to use machine learning skills to estimate it.

Climate change is having a negative impact on crop yields in India. Early forecasting of yield will be useful for making decisions regarding marketing, storage and agricultural business logistics planning. Intelligent methods of data analysis of weather characteristics are used in the work [14]. A website has been developed to predict the effect of climate parameters on crop yields with proven high forecast accuracy for all crops and regions of India considered in the study.

A study [15] focuses on the impact of climate change on rainfed agricultural production in Rwanda. Early sharing of information on expected crop production can help reduce the risk of food insecurity. In this regard, data mining techniques are used to predict future crops (ie Irish potatoes and maize) using the weather. Among other models, Random Forest is highlighted as recommended for early yield forecasting. The optimal values of the amount of precipitation and temperature at each stage were determined to achieve the optimal yield of agricultural crops.

The authors of the research [16] concentrated on the problem that is an obstacle to the use of machine learning in seasonal forecasting applications. This is a limited sample size of observational data for training models. To avoid this problem, the expediency of training various machine learning approaches on a large ensemble of climate models is considered. A new approach to seasonal precipitation forecasting is shown using the example of the western part of the USA. Several separate machine learning approaches were trained on simulations of large climate models, then their predictions were combined into an ensemble to further predict large-scale patterns of precipitation anomalies. The perspective of this approach to seasonal forecasting is determined. A large sample size is achieved, thus overcoming sampling problems and representing non-linear interactions.

In the paper [17] evaluated various aspects of modeling yield response to climate change. The research methods were linear regression and the ML technique "boosted regression trees" (BRT).

Prediction accuracy was significantly higher in BRT. However, the conclusions obtained when comparing the effects of climate change on individual grain crops in India make it necessary to be cautious in interpreting the results of a single model. The peculiarities of the climate and agricultural methods in different regions require special attention and careful analysis. The authors recommend comparing new approaches with traditional ones. It is emphasized that combining models, provided that the strengths of each of them are used, leads to better results.

Yield forecasting is an important aspect of food security. The work [18] is devoted to the investigation of issues the impact of weather on crop yield. It was established that the dependence of crop yield on weather conditions is non-linear. The significant impact of extreme weather indicators on the yield of agricultural crops has been confirmed, which justifies the need to include them to the model as predictors.

The artificial neural network model for yield dynamics modeling was also used in the works [19, 20].

3. Methodology

3.1. Data Collection

In terms of importance, the wheat occupies a central place among grain crops of Ukraine. The average annual wheat production in Ukraine for 2019-2021 was 26.5 million tons. During this time, the weight share of wheat in grain exports reached the level of 38%. The share of winter wheat in the total wheat harvest is 97%. In this study, we used statistical climate data and wheat yield data for the period 2000-2021 for the Sumy, Kharkiv, Poltava, Kyiv, Cherkassy, and Vinnytsia regions, which are located in the chernozem zone of the forest-steppe region of Ukraine. Climatic characteristics are taken by us from [21], yield data are taken from [1]. Successful vegetation of wheat in the period from April to June has a decisive impact on the yield [3]. To assess the climate impact on wheat yield, the average decadal temperature values for April, May, June and monthly amounts of precipitation for this period we used (please, check Table 1). Decadal temperature values make it possible to more accurately take into account the impact of air temperature at different stages of plant vegetation. We used monthly precipitation totals, taking into account the long-term aftereffect of precipitation on plant growth. Since machine learning techniques require a large amount of data, we combined the data of the six areas mentioned above into one dataset. The basis for this was the similar nature of climatic and soil characteristics of these areas. Statistical parameters of climatic factors and wheat yield are given in the Table 2.

Table 1
Definition variables

Variable	Definition	Period
t1	Average temperature	First decade of April
t2	Average temperature	Second decade of April
t3	Average temperature	Third decade of April
t4	Average temperature	First decade of May
t5	Average temperature	Second decade of May
t6	Average temperature	Third decade of May
t7	Average temperature	First decade of June
t8	Average temperature	Second decade of June
t9	Average temperature	Third decade of June
R10	Amount of precipitation	April
R20	Amount of precipitation	May
R30	Amount of precipitation	June
eps	Detrended wheat yield	Year

Table 2
Summary statistics of numerical features

	t1	t2	t3	t4	t5	t6	t7	t8	t9	R10	R20	R30	eps
Minimum	2.00	7.70	7.50	9.70	11.90	11.77	13.50	17.30	16.90	0.00	7.00	7.00	-17.98
Median	9.53	11.55	14.00	15.73	17.40	19.82	20.73	21.55	22.00	33.90	51.65	67.30	1.39
Average	9.54	11.91	14.36	16.39	17.60	19.77	20.29	21.74	22.20	35.55	60.27	75.70	0.00
Maximum	16.00	17.75	21.80	24.85	24.15	27.91	26.55	27.70	27.90	132.5	172.4	215.0	13.32
Standard Deviation	3.02	2.42	2.72	3.24	2.95	3.39	2.83	2.72	3.00	22.72	35.09	44.09	6.50

3.2. Analysis of wheat yield dynamics

An analysis of wheat yield dynamics in the regions of Ukraine over the past 22 years shows that the yield is increasing [1]. This was the result of significant investments that have flowed into the grain industry in recent years. As a result, the seed base improved, agrotechnical culture improved, and the logistics network of grain production developed. In 2021, Ukraine received a record harvest of grain and leguminous crops - 84 million tons. However, the tendency to increase grain yield is accompanied by significant fluctuations in yield, the cause of which is mostly the impact of weather and climate factors. The wheat yield dynamic in Cherkassy region can serve as an illustration (Figure 2). The largest negative yield deviations from the trend were observed in 2003, 2007, 2017 and 2020. The weather conditions of these years (cold April and hot May and June) were unfavorable for wheat vegetation.

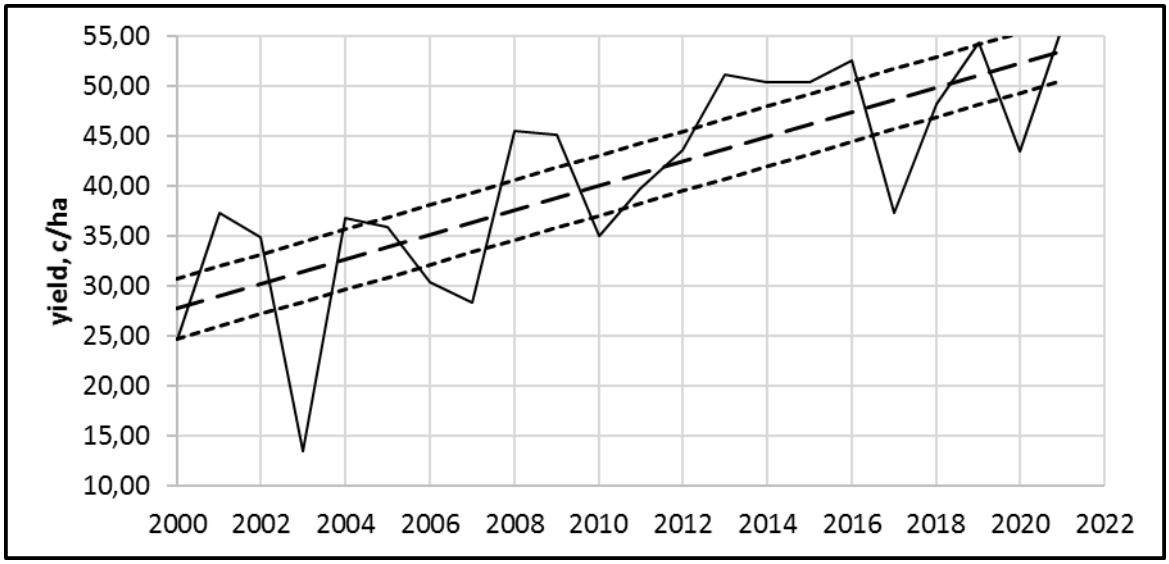

Figure 2: Wheat yield dynamics in the Cherkassy region. The dashed line is a linear trend. Dotted lines are high and low yield boundaries. Author's calculations according to [1]

To modeling of yield dynamics a linear trend model we used

$$tr_t = a_0 + ta_1. \quad (1)$$

Here a_0, a_1 - the trend coefficients, determined by statistical data using the least squares method [22]. An interval forecast is built on the linear trend basis, and for him the forecasting reliability level can be established. To construct an interval forecast of yield, it is necessary to check the hypothesis about a normal distribution of detrended yield *eps*

$$eps_t = y_t - tr_t. \quad (2)$$

To identify the distribution law of detrended yield, we used a combined sample of detrended yield for six regions of the Forest Steppe of Ukraine (132 values). We tested the fit of the detrended yield

distribution to three distribution laws: the normal distribution, the Laplace distribution, and the Gumbel distribution. The closest to the actual distribution of detrended yields is the normal distribution. But the reliability level of the conclusion regarding the correspondence of the detrended yields to the normal distribution is slightly lower than 95% (Figure 3). For further yield classification we accept the hypothesis about a normal distribution of detrended yields (Figure 4).

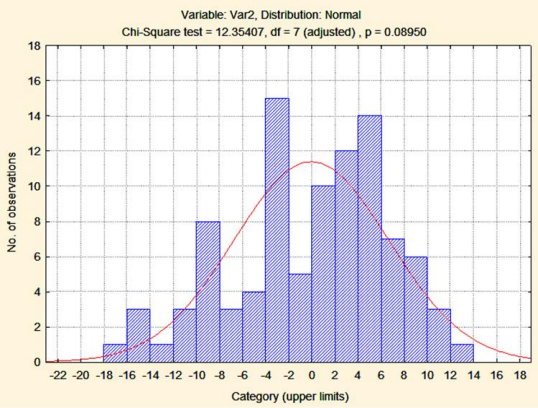

Figure 3: Verification of the normal distribution hypothesis for detrended yields. Combined sample for Sumy, Kharkiv, Poltava, Kyiv, Cherkassy, and Vinnytsia regions. Data for 2000-2021 [1]

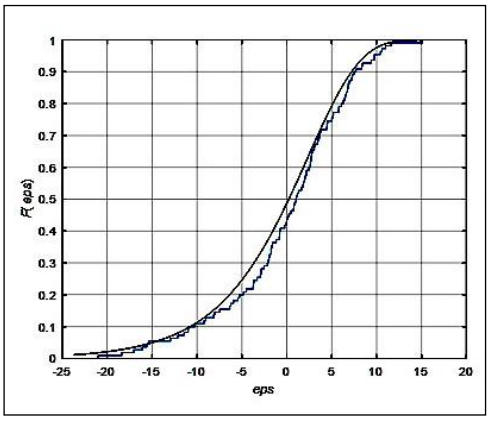

Figure 4: Integral curve of the normal distribution of detrended yields for 6 regions of the forest-steppe zone of Ukraine (black – theoretical distibution, blue – actual distribution)

3.3. Binarization of detrended yield

When making investment decisions and planning the logistics of agrarian business, it is not necessary to have an accurate yield forecast. Usually, in this case, it is enough to have an estimate of the future yield in terms of "high yield" – "low yield". At the same time, the term "high yield" means such a value of yield, which significantly exceeds the average level of yield; the term "low yield" means a yield value that is significantly lower than the average yield value. This approach enables the use of classification methods in yield forecasting.

We will use the hypothesis of a normal distribution of detrended yields for the binary classification of yield values using the categories "high yield" and "low yield". In this research our main task is to forecast low wheat yield values.

To the "low yield" group, we include those yield values that with a probability of $p < 0.33$ are located on the integral curve of the normal distribution of detrended yields (Figure 4), that is, those for which the condition is fulfilled

$$F\left(eps\right) < 0.33. \tag{3}$$

Yield values for which condition (3) is not fulfilled will be assigned to the "high yield" group. To implement a classification approach to yield prediction, we introduce a binary variable eps1 that takes only two values: 1 ("low yield") and 0 ("high yield"). At the same time, the value of the *eps1* variable is determined by the rule

$$eps1 = \begin{cases} 1, if\ F\left(eps\right) < 0.33; \\ 0, if\ F\left(eps\right) \geq 0.33. \end{cases} \tag{4}$$

According to the classification results, we get that the share of "low yield" values is 28.8%, the share of "high yield" values is 71.2%. We emphasize that the normal distribution of detrended yields is not a necessary condition for their classification. This hypothesis only simplifies the classification procedure.

3.4. Analysis of climatic factors impact on the wheat yield

The most important impacting factors that determines grain yield fluctuation are climatic factors. Therefore, assessing the impact of climatic factors on yield is an important prerequisite for successful wheat yield forecasting.

The linear correlation coefficients between climatic factors and wheat yield for Cherkassy region is presented in the Table 3. As you can see from the Table 3, the average temperature in April is positively correlated with the yield, the average temperature in May and early June is negatively correlated with the yield. Warm April, cool May and sufficient rainfall in May and June contribute to the successful vegetation of wheat in the region we studied.

Table 3

The linear correlation coefficients between climatic factors and wheat yield for Cherkassy region (authors' calculations (according to [1, 21]))

	t1	t2	t3	t4	t5	t6	t7	t8	t9	R10	R20	R30
eps	0.20	0.35	0.15	-0.13	-0.18	-0.25	-0.26	-0.16	0.09	-0.02	0.23	0.15

3.5. The multiple linear regression model. Features selection

We will build a model according to which the wheat yield is formed under the impact of 12 climatic factors (9 temperature and 3 related to precipitation). To build a model we will use the methods of multivariate correlation-regression analysis [22]. At the same time, the response *eps* is connected through the multiple regression equation with the factor features *t1, t2, t3, t4, t5, t6, t7, t8, t9, R10, R20, R30*. We believe that climatic factors affect not the average yield, but the deviation of the yield from trend value (detrended yield). Thus, we will consider the detrended yield eps_t as a response to the action of these factors accoeding to (2).

To model dependence, we will use the linear multiple regression equation, which has the following form:

$$eps = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_3 + \beta_4 t_4 + \beta_5 t_5 + \beta_6 t_6 + \beta_7 t_7 + \\ + \beta_8 t_8 + \beta_9 t_9 + \beta_{10} R_{10} + \beta_{20} R_{20} + \beta_{30} R_{30} + \varepsilon. \quad (5)$$

Here *β0,β1,β2,β3,β4,β5,β6,β7,β8,β9,β10,β20,β30* are model parameters, *t1, t2, t3, t4, t5, t6, t7, t8, t9, R10, R20, R30* are model factors, *eps* is response, ε is model residual. We used the least squares method [22] to determine of model parameters. Note that the use of this method is justified only when the residuals ε are normally distributed, the average value of the residuals is zero, and the residuals variance is constant.

The chernozem zone of the forest-steppe region of Ukraine is of interest for our research in view of its contribution to wheat production in Ukraine. To use of machine learning techniques, you need to have large data samples. Taking into account the similarity of soil and climatic conditions, we united several neighboring regions of Ukraine into a single region. This made it possible to obtain a sample that contains enough data. For our research, we used the united region, which includes Sumy, Kharkiv, Poltava, Kyiv, Cherkasy, and Vinnytsia regions of Ukraine. The corresponding data sample contains 132 samples, each containing 12 climate factors and detrended yield.

To process such large data sets, it is advisable to use specialized software. One of the most widely used software products designed for statistical data processing is the R programming environment with the RStudio extension [23, 24]. We used this software product for further calculations and evaluations.

When developing a statistical model of a phenomenon, the problem is to choose an algorithm that is optimal for a specific case. In recent decades, the implementation of machine learning methods to solve the problems of classification and regression (quantitative response prediction) has begun. These are such methods as: multiple regression method, logistic regression method, linear discriminant analysis, "random forest" method, support vector machines. We used these methods to predict and classify of yields. By comparing the obtained results, we can understand which methods are optimal for our task.

3.6. Machine Learning Algorithms

Recently, the machine learning technique is becoming more and more popular in agricultural production research. Supervised classification problems are widely used for data-driven decision making. Classification is an important form of data mining that helps formulate models describing different classes of data [25].

For categorical forecasting of wheat yield, we used several classification algorithms. A set of algorithms that have performed well in numerical experiments with real data are explained in more detail below.

Evaluation of classifiers. The following indicators are used for evaluating the performance of the classifier we built: matrix of errors (*Confusion matrix*), overall accuracy of classification (*Accuracy*), sensitivity of classification (*Sensitivity*), specificity of classification (*Specificity*) and the area under the ROC curve [26]. The Confusion matrix is built based on the results of classification by the model and the actual belonging of observations to classes [23]. Four cases are distinguished in the matrix:

- TP (*True Positives*) – the model correctly detected a low yield value;
- FP (*False Positives*) – the model wrongly recognized a high yield as a low yield;
- FN (*False Negatives*) – the model wrongly recognized a low yield as a high yield;
- TN (*True Negatives*) – the model correctly identified a case of high yield.

Using the values of the elements of the error matrix, the following performance indicators of the binary classifier can be determined:

- *Sensitivity SE = TP / (TP + FN)*, determines the share of correctly identified low-yield cases among all low-yield cases;
- *Specificity SP = TN / (TN + FP)*, determines the share of correctly identified high-yield cases among all high-yield cases;
- *Accuracy AC = (TP + TN) / (TP + FP + FN + TN)*, determines the overall probability of the test giving correct results.

Logistic regression model. As noted above, to solve many problems when planning an agrarian business, it is enough to have an estimate of the future yield in terms of "high yield" – "low yield". This approach enables the use of classification methods in yield forecasting. We binarized the detrended yield according to the rule (4). As a result, we got a new data set, which differs from the one described in paragraph 3.1 by replacing the numerical factor eps with the categorical factor *eps1*. Each of the 132 objects of the new data set is characterized by a 12-dimensional feature vector.

The logistic regression model looks like this

$$P = F\left(X\beta'\right), \tag{6}$$

where F is a function, the range of values of which belongs to the interval [0;1] and determines the low yield appearance probability P. To implement the function F, a logistic distribution function (logit model) is usually used:

$$F\left(z\right) = \frac{e^z}{1+e^z}. \tag{7}$$

Here, the parameter z is calculated from the ratio

$$z = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_3 + \beta_4 t_4 + \beta_5 t_5 + \beta_6 t_6 + \beta_7 t_7 + \\ + \beta_8 t_8 + \beta_9 t_9 + \beta_{10} R_{10} + \beta_{20} R_{20} + \beta_{30} R_{30}. \tag{8}$$

To choose the best model, it is necessary to estimate the value of the coefficients β_i of the logistic regression model. Usually, the maximum likelihood method [22] is used for this.

The logistic regression model allows you to classify the samples according to the rule

$$p = \begin{cases} 1, \; if \;\; p < 0.5; \\ 0, \; if \;\; p \geq 0.5. \end{cases} \tag{9}$$

The value $p = 1$ corresponds to the case of "low yield", the value $p = 0$ corresponds to the case of "high yield".

Linear discriminant analysis (LDA). LDA is a method of multivariate analysis that allows to evaluate the differences between two or more groups of objects according to several variables at the same time [27]. Discriminant analysis is based on the assumption that descriptions of objects of each *k*-th class are realizations of a multidimensional random variable distributed according to a normal law with mean μ_k and covariance matrix C_k. The task of discriminant analysis is to draw an additional axis *z(x)*

$$z(x) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m, \quad (10)$$

which passes through the point cloud in such a way that projections onto it provide the best resolution into two classes. Its position is given by a linear discriminant function (linear discriminant, LD) with weighting coefficients *β1, β2,…, βm* that determine the contribution of each output variable.

Decision tree model. Decision trees used in data mining are of two main types: a classification tree (the predicted result is the class identifier) and a regression tree (the predicted result is a real number) [28]. Decision trees split the space of objects according to some set of splitting rules. These rules make it possible to implement sequential dichotomous data segmentation. At each partitioning step, the amount of information about the variable under study (response) increases. When constructing a decision tree, it is important to set the optimal branching level.

The disadvantage of the decision tree method is instability: two trees built on the same training sample can give completely different resulting classes. This shortcoming can be eliminated by constructing ensembles of decision trees - a "Random Forest". The Random Forest classifier is based on bagging [29]. At the same time, several decision trees are built, repeatedly interpolating the data with replacement (bootstrap), and as a consensus answer, it gives the result of the voting of the trees (their average forecast). Boosting is another method for constructing a Random Forest [30].

Support vectors machine. The support vectors machine was developed by V. N. Vapnyk and A. Ya. Chervonenkis [31]. The basic idea of a support vectors classifier is to build a separating surface using only a small subset of points that lie in the zone critical for separation, while other correctly classified points of the training sample outside this zone are ignored by the algorithm. Since there can be many separating hyperplanes, the hyperplane that is the most distant from the training points, i.e., has the maximum *gap*, is selected from among them.

Method of cross-validation. Even when we have a large data set and random sampling has been applied to the constructing of training sample, the resulting model may be statistically unreliable. After all, another set of samples can lead to another model, which is significantly different from the first one. This shortcoming can be eliminated by cross-validation method. Let's briefly outline its essence [32].

1. First, you need to arbitrarily divide the initial data set into *k* groups ("folds") of approximately the same size.
2. One of the folds is selected as a data set for testing of model (testing set). The model is built based on the data of the remaining *k -1* folds that form the training set. We calculate the mean squared error MSE test error based on the testing set observations
3. The process described above is repeated k times, each time using a different set as a testing set.
4. We calculate the total test MSE as the average of *k* test MSEs. Similarly, we average other parameters of the model.

4. Results

4.1. Linear regression model

We will describe the method of building and verifying a linear regression model in the RStudio environment [23].

Model LM1. Guided by the principles of statistical modeling [33], we will divide all the data into two parts: the training sample (most of the initial data used to build the model) and the control sample (the rest of the data that did not fall into the training sample). The control sample data are new (unknown) to the constructed model and, therefore, they are used to assess the quality of the constructed model.

Let's break down the available data as follows: training sample - 92 rows, randomly selected (70% of the data); the control sample is the remaining 40 rows (30% of the data). The values of the estimates of the linear regression model are given in the Table 4.

Table 4
Values of estimates of the linear regression model LM1

	Estimate	Std. Error	t value	Pr(>|t|)	Signif. codes*
(Intercept)	2.417	8.879	0.272	0.786	
t1	0.116	0.240	0.483	0.631	
t2	0.009	0.291	0.030	0.977	
t3	1.129	0.315	3.584	0.001	***
t4	-0.719	0.245	-2.933	0.004	**
t5	-0.184	0.322	-0.572	0.569	
t6	0.402	0.280	1.437	0.155	
t7	-1.152	0.317	-3.632	0.001	***
t8	0.009	0.313	0.029	0.977	
t9	0.240	0.274	0.875	0.384	
R10	0.060	0.030	2.000	0.049	*
R20	0.037	0.018	2.087	0.040	*
R30	0.014	0.015	0.950	0.345	

*Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The LM1 model is generally adequate (F-statistic=4.964 on 12 and 79 DF; p-value = 4.611e-0.6), but most of the factors of this model are insignificant (all except *t3, t4, t7, R10, R20*). The quality of the built linear regression model can be improved by removing insignificant factors. Let's remove the factors *t1, t2, t5, t6, t8, t9, R10, R20, R30* from the model. We will get the following **LM2 model (**please, check Table 5**)**.

Table 5
Values of estimates of the linear regression model LM2

	Estimate	Std. Error	t-value	Pr(>|t|)	Signif. codes
(Intercept)	22.407	5.107	4.388	3.17e-05	***
t3	1.269	0.277	4.587	1.48e-05	***
t4	-0.985	0.210	-4.682	1.03e-05	***
t7	-1.189	0.218	-5.467	4.22e-07	***

The built linear regression model LM2 is adequate (F-statistic =15.44 on 3 and 88 DF; p = 3.75e-08) and all its factors are significant. The error of the model on the training sample is 5.366, the error on the control sample is 6.490. The model shows that factors *t3, t4* and *t7* have the greatest impact on wheat yield fluctuations. Warm April and cool May and June contribute to high wheat yields.

4.2. Logistic regression model

Let's proceed to the implementation of classification methods of wheat yield. We built a logistic regression model GLM for the data set, which unites 6 regions of the forest-steppe region of Ukraine. The feedback variable *eps* (the yield deviation from the trend) was previously binarized according to rule (4). We randomly split the data set into two parts: a training sample (70% of the data) and a test sample (30% of the data). The same variables that we selected in the multiple regression model were selected as factors of the GLM model. Based on the GLM model, a forecast was built on the test sample.

The description of the logistic regression model using factors *t3, t4, t7* is given in Table 6.

Table 6
Values of estimates of the logistic regression model GLM

	Estimate	Std. Error	z-value	Pr(>|z|)	Signif. codes
(Intercept)	-8.065	2.507	-3.218	0.0013	**
t3	-0.482	0.154	-2.135	0.0017	**
t4	0.312	0.107	2.906	0.0037	**
t7	0.423	0.127	3.333	0.0009	**

The model we built is fully adequate (*Residual deviance: 88.140 on 89 degrees of freedom; AIC=96.14)*.

The binary classification indicators calculated by us are as follows: *Sensitivity* SE = 0.370; *Specificity* SP = 0.970; *Accuracy* AC = 0.796. As we can see, the accuracy of the classifier (0.796) is high, but the sensitivity of the classifier (0.370) is low. This is a significant drawback, since out of 27 cases of low yield, the classifier recognized only 10. But one implementation of the model cannot judge its quality. The average values of the model characteristics for the five implementations are as follows: *Sensitivity* SE = 0.378; *Specificity* SP = 0.970; *Accuracy* AC = 0.798. An improved approach to the assessment of classification models, which consists in cross-validation of models, is implemented in the next section of the work.

4.3. A random forest model

A fragment of the decision tree of our problem is presented in Figure 5. At the first step, the algorithm determines the most significant factor and builds a dichotomy rule for it. Such a rule is the logical expression "$t_2 \geq 11$". This means that the average temperature of the second decade of April is the main factor affecting wheat yield. If its value exceeds 11°C, the yield is likely to be high. The next most important factor is factor *t5*. In the next step, the obtained classes are again divided into subclasses according to another rule. This makes it possible to clarify the general rule of classification.

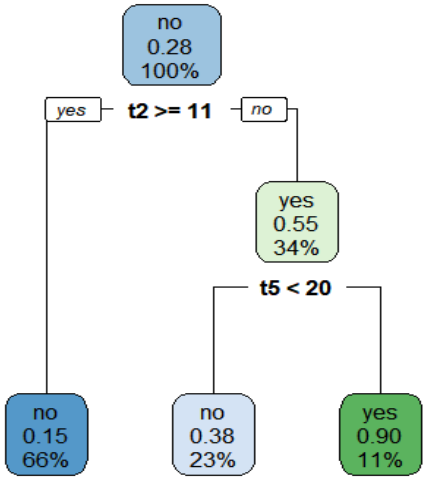

Figure 5: Classification of samples on one of the decision tree branches

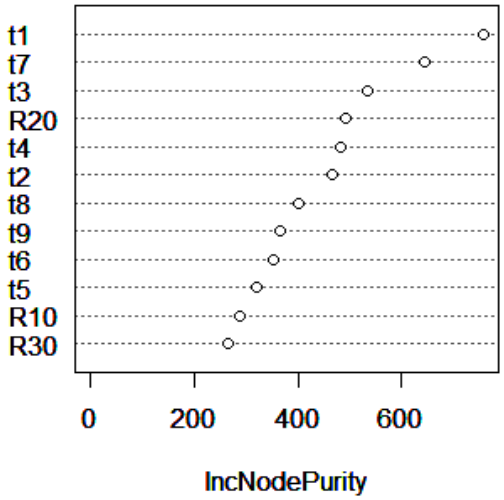

Figure 6: Relative importance of factors for binary samples classification

At the next stage, a group of trees is combined into a random forest. We built a regression-type random forest model based on the first data set using all impact factors. We obtained the following results. To achieve high classification accuracy of our data, it is necessary to use at least 100 trees. The total 500 trees were used in this case. The maximum number of branches that depart from the trunk is 4, the mean square of the prediction error is 22.779; percentage of data used 46.6%. The random forest method allows to evaluate the importance of each factor for a successful classification. The result of

such an assessment is presented in Figure 6. As you can see from the figure, the most important factors for successful classification in this method are factors *t1* and *t7*.

4.4. Comparison of the classification models effectiveness

We used five methods to build binary classification models: linear discriminant analysis (LDA), support vector machine with linear kernel function (SVML), support vector machine with radial kernel function (SVMR), random forest method (RF), logistic regression method (GLM). We compared them using the method described in [34]. The stages of this technique are as follows:

1. We divide the initial data into training and control samples. We allocate 70% of all data for training.
2. We train models.
3. The cross-validation procedure involves dividing the initial data into ten identical groups, one of which is a control group, and repeating the procedure three times. We carry out resampling of the received models and form the quality criteria of the models. The list of model quality criteria includes: accuracy *AC*, sensitivity *SE*, specificity *SP*, and *AUC* - the area under the ROC curve.

We build a table of forecasts for the test sample for all models (please, check Table 7).

Table 7
Prediction of low yield by different methods (test sample)

#	LDA	SVML	SVMR	RF	GLM	eps1	#	LDA	SVML	SVMR	RF	GLM	eps1
1	0	0	1	0	0	0	**21**	0	0	0	0	0	0
2	0	0	0	0	0	0	**22**	0	0	0	0	0	0
3	0	0	0	0	0	0	**23**	1	0	1	1	1	1
4	0	0	0	0	0	0	**24**	0	0	0	0	0	1
5	0	0	0	0	0	0	**25**	0	0	0	0	0	0
6	1	1	1	1	1	1	**26**	0	0	0	0	0	0
7	1	1	1	1	1	1	**27**	0	0	0	0	0	0
8	0	0	0	1	0	0	**28**	0	0	0	0	0	0
9	0	0	0	0	0	0	**29**	0	0	0	0	0	0
10	0	0	0	0	0	0	**30**	0	0	0	0	0	0
11	0	0	0	0	0	0	**31**	1	1	1	1	1	1
12	0	0	1	0	0	1	**32**	0	0	0	0	0	0
13	0	0	0	0	0	0	**33**	1	0	0	0	1	1
14	0	0	1	0	0	0	**34**	0	0	0	0	0	1
15	1	1	1	1	1	1	**35**	0	0	0	0	0	0
16	0	0	0	0	0	0	**36**	0	0	0	0	0	0
17	0	0	0	0	0	0	**37**	0	0	1	1	0	0
18	0	0	1	1	0	0	**38**	0	0	1	0	0	1
19	0	0	0	0	0	0	**39**	0	0	1	0	0	0
20	0	0	0	0	0	1							

Table 7 shows the results of predicting low wheat yield on the samples included in the test sample (39 samples). The last column of Table 8 contains the actual observed yield values in binary format ("1" – low yield, "0" – high yield). Analysis of Table 7 shows that in three cases out of 39 controls, none of the classification methods was able to predict the correct result. In two cases, only the SVMR method made the correct prediction. This can be explained by the impact of climatic factors outside the time interval of our research, which we did not take into account, or by the impact of other (non-climatic) factors. Among the tested machine learning techniques, the LDA and GLM showed the best

accuracy (Table 8). The SVMR method demonstrated the highest sensitivity, which is especially important for predicting low yields.

Table 8
Quality criteria of machine learning techniques

	Accuracy	Sensitivity	Specificity	AUC
LDA	**0.872**	0.545	1.000	0.684
SVML	0.821	0.364	1.000	0.727
SVMR	**0.769**	0.636	0.821	0.983
RF	0.769	0.455	0.893	0.973
GLM	**0.872**	0.545	1.000	0.688

A universal method for comparing the classifiers accuracy is ROC analysis [26]. We constructed ROC curves for the three recognition methods we used (LDA, SVMR, and RF) based on the complete table of initial data, which includes training and test samples (Figure 7). The area under the ROC curve AUC is a criterion for evaluating the classifier. For an ideal classifier, the ROC curve has the shape of a right angle. The best classifiers according to the AUC criterion are the support vector method (radial kernel function) and the random forest method.

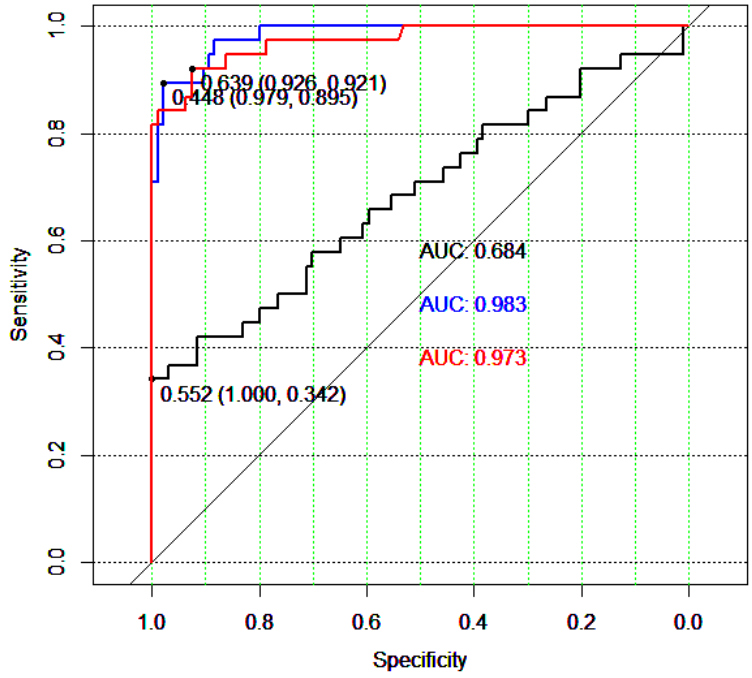

Figure 7: ROC curves for LDA (black), SVMR (blue) and RF (red) methods

5. Discussion

The purpose of this work was to use machine learning techniques to evaluate the impact of climatic factors on wheat yield fluctuation. This approach requires a large amount of data. To solve this problem, we combined the data from five regions of the chernozem zone of the forest-steppe region of Ukraine, which have similar soil and climatic characteristics. An analysis of wheat yields over the past 22 years showed that yields are increasing in all regions of Ukraine. This can be explained by the growth of investments in this industry, which has allowed the use of new high-performance technologies for grain cultivation, harvesting, storage and logistics. However, the growth in yield is uneven and is accompanied by yield deviations from the trend, mostly caused by the impact of climatic factors.

The weather and climate conditions of April, May and June play a decisive role in the formation of future wheat harvest, since it is during time that plants grow intensively. In previous studies different authors used time periods of different durations to integrate weather factors. We have shown that to estimate the impact of weather factors on yield, it is sufficient to use the average decadal values of temperature and monthly amounts of precipitation for the period from April to June. We have also shown that factors *t3, t4* and *t7* have the greatest impact on wheat yield fluctuations. A warm April and cool May and June contribute to high wheat yields.

We used a new approach in which trend deviations were classified into two groups: "low yield" and "high yield". Thus, the problem of yield forecasting was reduced to the classification problem. We employed five machine learning models as classifiers, including linear discriminant analysis, support vector machines with linear and radial kernel functions, random forest models, and logistic regression models. To increase statistical significance, we employed the cross-validation procedure with subsequent averaging of model parameters. After fitting the models to the available data, the optimized models demonstrated a classification accuracy of approximately 80% on test samples. However, some samples were not recognized correctly by any of the classifiers, possibly due to yield deviation caused by factors other then weather and climate.

The built models enable us to estimate future yield in advance (3 months before), which can support optimal investment and marketing decisions. Additionally, this technique can by applied to predict high yield cases and to classify yields into three categories: "high yield", "medium yield" and "low yield".

6. Conclusions

This research evaluated of climatic factors impact on wheat yield fluctuation using machine learning techniques. The traditional technique uses methods such as correlation analysis and linear regression and is limited by the requirement of a factor values normal distribution. Machine learning techniques are free from this limitation and allow yield prediction in categorical binary form.

When making investment decisions and planning the logistics of agrarian business it is enough to have an estimate of the future yield in terms of "high yield" – "low yield". We compared the various machine learning algorithms' accuracy by performing detailed experimental analysis of real data.

The linear discriminant analysis and logistic regression model have shown the best accuracy of detrended wheat yield classification and have provided prediction accuracy of 87% (on test sample). This predicting accuracy is very good for complex natural processes which we consider.

Moreover, the algorithms we used in our research are quite affordable in terms of implementation such that grain producers can use them for short-term yield forecasting.

The method proposed by us can be used to study the climate impact on other agricultural yield in different regions and countries. Each year we receive new sets of real climate data. Their use will increase the effectiveness of the developed models and provide more accurate estimates of the climate impact on grain yield.

Investigating the contribution of weather risk to total grain production risk is another promising area for future research. To solve this problem, it is necessary to investigate the issue of detrended yield distribution in more detail.

In our previous study, we explored the relationship between climate and wheat yield in Ukraine's steppe region [35]. This work was focused on the study of grain production processes in the chernozem zone of the forest-steppe region of Ukraine. In the next study, we plan to carry out a similar experimental analysis of the impact of climatic factors on wheat yield for the western region of Ukraine. This research will be final part of the series of studies devoted to the research of climate impact on wheat yield fluctuations in the regions of Ukraine.

Our research is a contribution to solving the problem of ensuring the sustainability of grain production in Ukraine. The obtained results can be used to support the economic development of Ukraine and solve the food problem in the world.

7. References

[1] State Statistics Service of Ukraine, 2022. URL: http://www.ukrstat.gov.ua.

[2] T. Adamenko, Climate change and agriculture in Ukraine: what farmers should know. German-Ukrainian agropolitical dialogue, Zapovit, Kyiv, 2019.
[3] P. Grytsyuk, L. Bachyshyna. Influence of change in climatic conditions on the dynamics of the crop yield of cereals in Ukraine, Economy of Ukraine 6 (2016) 68–75.
[4] IPCC, 2022: Climate Change 2022: Mitigation of Climate Change, in: P.R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, J. Malley (Eds.), Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK and New York, NY, USA. doi: 10.1017/9781009157926.
[5] P. M. Hrytsiuk, Analysis, modeling and forecasting of the dynamics of winter wheat yield in the regions of Ukraine, NUWM, Rivne, 2010.
[6] U. S. Department of Agriculture, 2022. URL: https://www.usda.gov/.
[7] D. Muller, A. Jungandreas, F. Koch, F. Shirhorn, The impact of climate change on wheat production in Ukraine. Report on agricultural policy (APD), 2016.
[8] M. Kalimuthu, P. Vaishnavi, M. Kishore, Crop prediction using machine learning, in: Proceedings of the Third International Conference on Smart Systems and Inventive Technology, ICSSIT '2020, IEEE, 2020, pp. 926–932. doi: 10.1109/ICSSIT48917.2020.9214190.
[9] P. Feng, B. Wang, D. L. Liu, C. Waters, D. Xiao, L. Shi, Q. Yu, Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique, Agricultural and Forest Meteorology, 285–286 (2020). doi.org/10.1016/j.agrformet.2020.107922.
[10] B. Wang, P. Feng, C. Waters, J. Cleverly, D. L. Liu, Q. Yu, Quantifying the impacts of pre-occurred ENSO signals on wheat yield variation using machine learning in Australia, Agricultural and Forest Meteorology, 291 (2020). doi.org/10.1016/j.agrformet.2020.108043.
[11] A. Barbosa, R. Trevisan, N. Hovakimyan, and N. F. Marti, Modeling yield response to crop management using convolutional neural networks, Computers and Electronics in Agriculture, 170 (2020). doi.org/10.1016/j.compag.2019.105197.
[12] D. Elavarasan, D. R. Vincent, V. Sharma, A. Y. Zomaya, and K. Srinivasan, Forecasting yield by integrating agrarian factors and machine learning models: a survey, Computers and Electronics in Agriculture 155 (2018) 257–282. doi: 10.1016/j.compag.2018.10.024.
[13] G. Leng, J. W. Hall, Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models, Environmental Research Letters, 15 4 (2020). doi: 10.1088/1748-9326/ab7b24.
[14] S. Veenadhari, B. Misra, C. Singh, Machine learning approach for forecasting crop yield based on climatic parameters, in: Proceedings of the International Conference on Computer Communication and Informatics, ICCCI '14, IEEE, 2014, pp. 1-5. doi: 10.1109/ICCCI.2014.6921718.
[15] M. Kuradusenge, E. Hitimana, D. Hanyurwimfura, P. Rukundo, K. Mtonga, A. Mukasine, C. Uwitonze, J. Ngabonziza, A. Uwamahoro, Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize, Agriculture, 13 1 (2023). doi.org/10.3390/agriculture13010225.
[16] P. B. Gibson, W. E. Chapman, A. Altinok, L. D. Monache, M. J. DeFlorio, D. E. Waliser, Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts, Communications Earth & Environment, 2 (2021). doi.org/10.1038/s43247-021-00225-4.
[17] B. S. Sidhu, Z. Mehrabi, N. Ramankutty, M. Kandlikar, How can machine learning help in understanding the impact of climate change on crop yields?, Environmental Research Letters, 18 2 (2023). doi.org/10.1088/1748-9326/acb164.
[18] V. S. Konduri, T. J. Vandal, S. Ganguly, A. R. Ganguly, Data science for weather impacts on crop yield, Frontiers in Sustainable Food Systems 4 (2020). doi.org/10.3389/fsufs.2020.00052.
[19] A. Crane-Droesch, Machine learning methods for crop yield prediction and climate change impact assessment in agriculture, Environmental Research Letters, 13 11 (2018), 1–12. doi: 10.1088/1748-9326/aae159.

[20] T. van Klompenburg, A. Kassahun, C. Catal, Crop yield prediction using machine learning: a systematic literature review, Computers and Electronics in Agriculture, 177 (2020). doi: 10.1016/j.compag.2020.105709.
[21] Central geophysical observatory the names of Boris Sreznevsky, 2022. URL: http://cgo-sreznevskyi.kyiv.ua/.
[22] N. R. Draper, H. Smith, Applied Regression Analysis. 3th ed., Wiley, New York, NY, 1998. doi: 10.1002/9781118625590.
[23] N. Zumel, J. Mount, Practical data science with R, Manning Publications Co., New York, NY, 2020.
[24] V. Babenko, A. Panchyshyn, L. Zomchak, M. Nehrey, Z. Artym-Drohomyretska, T. Lahotskyi, Classical machine learning methods in economics research: macro and micro level example, WSEAS Transactions on Business and Economics, 18 22 (2021) 209–217. doi.org/10.37394/23207.2021.18.22.
[25] S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, in: Proceedings of the International conference on Management of data, SIGMOD '00, Association for Computing Machinery, New York, NY, 2000, pp. 427–438. doi: 10.1145/342009.335437
[26] T. Fawcett, An Introduction to ROC Analysis, Pattern Recognition Letters, 27 8 (2006) 861–874. doi:10.1016/j.patrec.2005.10.010.
[27] A. Afifi, S. Azen, Statistical Analysis, Second Edition: A Computer Oriented Approach, Academic Press, New York, NY, 1979.
[28] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and regression trees, Brooks/Cole Publishing, Monterey, 1984. doi:10.2307/2530946.
[29] L. Breiman, Bagging Predictors, Machine Learning, 24 (1996) 123–140.
[30] J. Friedman, Stochastic Gradient Boosting, Computational Statistics and Data analysis, 38:4 (2002) 367-378. doi:10.1016/S0167-9473(01)00065-2.
[31] T. Hastie, R. Tibshirani, J. Friedman, Model Assessment and Selection, in: The Elements of Statistical Learning, Springer Series in Statistics, Springer, New York, NY, 2009, pp. 219-259. doi:10.1007/978-0-387-84858-7_7.
[32] D. Berrar, Cross-Validation, in: S. Ranganathan, M. Gribskov, K. Nakai, C. Schönbach (Eds.), The Encyclopedia of Bioinformatics and Computational Biology, Academic Press, Oxford, 2019, pp. 542–545. doi: 10.1016/B978-0-12-809633-8.20349-X.
[33] J. Gareth, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, Springer, New York, NY, 2013. doi:10.1007/978-1-4614-7138-7.
[34] M. Kuhn, K. Johnson, Applied Predictive Modeling, Springer, New York, NY, 2013. doi.org/10.1007/978-1-4614-6849-3.
[35] P. Hrytsiuk, T. Babych, B. Krasko, Classification methods of the yield forecasting, Herald of Khmelnytskyi national university. Technical sciences, 3 309 (2022) 209-216. doi: 10.31891/2307-5732-2022-309-3-209-216.