UDC 633.11:631.5:004.89
DOI: 10.15587/1729-4061.2025.340563

COMPARATIVE ASSESSMENT OF MACHINE LEARNING ALGORITHMS FOR FORECASTING WHEAT YIELDS USING CLIMATE INDICATORS AND SATELLITE VEGETATION INDICES

Nurlan Kurmanov
PhD Doctor, Professor*

Zhaxat Kenzhin
PhD Doctor, Associate Professor
Department of Management and Innovation in Sports
Kazakh National Sports University
Mangilik El ave., B 2.2, Astana, Republic of Kazakhstan, 010000

Darkhan Baxultanov
Corresponding author
PhD Doctor, Senior Lecturer*
E-mail: baksultanov@gmail.com

Bolat Zhagalbayev
PhD Student
Higher School of Business and Digital Technologies
Turan-Astana University
Ykylas Dukenuly str., 29, Astana, Republic of Kazakhstan, 010013

Dinara Mussabalina
PhD, Postdoctoral Researcher
Department of Economic Specialties
Abai Kazakh National Pedagogical University
Dostyk ave., 13, Almaty, Republic of Kazakhstan, 050010

Meruyert Zhagalbayeva
Researcher
College of Economics and Management
Northwest A&F University
Taicheng rd., 3, Yangling, Shaanxi, China, 712100

Galiya Amrenova
Senior Lecturer*
*Department of Economics
L. N. Gumilyov Eurasian National University
Satbaev str., 2, Astana, Republic of Kazakhstan, 010000

The object of the study is wheat yield forecasting based on the integration of climatic indicators, satellite vegetation indices, and machine learning algorithms. The problem to be solved is the limited accuracy of traditional crop yield forecasting methods, which fail to capture the complex nonlinear and multidimensional interactions among climatic, biophysical, and agronomic factors, thereby reducing their applicability for global food security tasks. The proposed approach is applied to a dataset comprising 345 observations from 2001–2023, combining vegetation indices (MODIS), climatic parameters (ERA5), and official statistics on yield and sown areas.

The methodology included descriptive statistics, correlation analysis and forecasting models based on random forest, support vector machine and convolutional neural network. Model performance was assessed using coefficient of determination, root mean square error and mean absolute error. Random forest and support vector machine showed the highest accuracy ($R^2 = 0.85$ with low errors), while convolutional neural network was less effective due to the limited dataset. The analysis confirmed the decisive role of vegetation indices, especially the normalized difference vegetation index, together with precipitation, temperature and sown area.

The results address the identified research gap by demonstrating that the integration of climatic indicators and satellite vegetation indices significantly enhances the performance of machine learning models for wheat yield forecasting. In particular, the findings highlight the advantages of ensemble and support vector methods, which proved to be more robust and accurate under conditions of high climatic variability.

The practical value lies in the potential use of these models in early warning and decision-support systems for farmers and state institutions, improving agrotechnical planning, resource allocation, and reducing food security risks, thereby contributing to global food security

Keywords: wheat yield forecasting, random forest, support vector machine, convolutional neural network, normalized difference vegetation index, enhanced vegetation index, MODIS, ERA5

Received 01.07.2025
Received in revised form 05.09.2025
Accepted 15.09.2025
Published 29.10.2025

How to Cite: *Kurmanov, N., Kenzhin, Z., Baxultanov, D., Zhagalbayev, B., Mussabalina, D., Zhagalbayeva, M., Amrenova, G. (2025). Comparative assessment of machine learning algorithms for forecasting wheat yields using climate indicators and satellite vegetation indices. Eastern-European Journal of Enterprise Technologies, 5 (13 (137)), 72–80. https://doi.org/10.15587/1729-4061.2025.340563*

1. Introduction

Food security is still one of the key tasks of the global community, especially in the context of population growth and climate change [1]. Wheat is one of the most important grain crops accounting for a significant part of the world's grain production and consumption. Its yield directly affects the level of food independence of countries and sustainability of agricultural systems [2].

Climate changes, such as rising average annual temperatures, more frequent droughts and extreme weather events, increase the agricultural risks and require more accurate methods for monitoring and crop forecasting [3, 4]. Traditional agrometeorological and statistical analysis methods are limited in accuracy as they only consider linear relationships between factors and do not reflect the complex interaction of climatic and biophysical parameters [5].

In Republic of Kazakhstan, these challenges are particularly acute. Agriculture accounts for around 5% of GDP and provides more than 20% of employment, while wheat remains the country's main strategic crop. Republic of Kazakhstan is among the world's top ten wheat exporters, annually supplying 6–8 million tons. The Kostanay region alone produces up to one third of the national wheat harvest, making it a critical area for national food security and international grain trade. However, yields in this region are extremely volatile: during the drought of 2021, wheat production in Republic of Kazakhstan fell by more than 20%, while in 2023 excessive rainfall caused serious harvesting delays. Such instability not only threatens domestic food supply but also undermines Kazakhstan's export reliability in global markets.

Geopolitical and economic conditions further aggravate these risks. The disruption of Black Sea grain exports in recent years has increased demand for Kazakh wheat, while global price volatility intensifies the pressure on national producers. At the same time, climate projections for Central Asia indicate a growing frequency of droughts, heatwaves and uneven precipitation. Under these circumstances, reliable yield forecasting becomes not only a scientific task but also an urgent economic and political priority.

The use of Earth remote sensing data opens up significant prospects. Satellite vegetation indices (NDVI, EVI) have proven to be reliable indicators of the state of crops and biomass [6, 7]. Their integration with climate data (temperature, precipitation, soil moisture) improves the accuracy of crop productivity assessment [8]. Meanwhile, machine learning advancements open up opportunities for building more complex models that can add nonlinear dependencies and multidimensionality of factors to consideration [9, 10].

The random forest and support vector machine algorithms have proven highly effective in addressing crop yield forecasting issues, including regional studies where data volumes are limited [11, 12]. Simultaneously, the use of neural networks, such as convolutional (CNN) and recurrent (LSTM) ones, demonstrates high potential in the presence of large data amounts [13, 14]. However, their effectiveness is reduced under conditions of limited time series typical for regional samples.

Therefore, studies that are devoted to developing advanced approaches for crop yield forecasting through the integration of climatic indicators, satellite-derived vegetation indices, and machine learning methods are of high scientific relevance, as they contribute to improving the resilience of agricultural systems and strengthening global food security under conditions of increasing climatic variability.

2. Literature review and problem statement

The paper [15] shows that crop forecasting is one of the central tasks for agricultural analytics and sustainable food security management. It is shown that traditional methods based on regression models and agrometeorological equations have been used for decades; however, their effectiveness is limited by the ability to consider only linear relationships and a small number of factors [16].

But there are still unresolved questions related to the integration of multidimensional and nonlinear factors, especially under conditions of ever-increasing climate variability and the growing complexity of natural and anthropogenic interactions. The reasons for this can be objective difficulties connected with limited data availability, costly monitoring systems, and methodological restrictions, which make the corresponding researches less effective.

An option to overcome the relevant difficulties can be the use of remote sensing data. Vegetation indices, primarily NDVI and EVI, have proven their effectiveness as condition indicators for crops and biomass [6, 7]. This is the approach used in [8, 17], where their integration with climatic parameters (precipitation, temperature, soil moisture, solar radiation) significantly improves the accuracy of crop forecasting. However, unresolved issues remain regarding the stability of such indices in regions with strong interannual variability.

The paper [18] analyzes the influence of precipitation and temperature variability on agricultural production in Central Asia and shows that rainfall distribution during the growing season is a decisive factor for wheat yields. It is shown that the authors emphasize climatic parameters as the main drivers of productivity change. But the unresolved questions are related to the limited geographical scope of the analysis and the absence of consideration of socio-economic and technological dimensions. The reasons of it can be the review-oriented nature of the study and the lack of integrated datasets covering both climate and production factors. An option to overcome these difficulties can be the application of modeling approaches that jointly account for climatic, agronomic and institutional variables.

The paper [19] presents empirical evidence that no-till practices, crop residue management and diversified crop rotations improve winter wheat productivity under rainfed conditions in Uzbekistan. It is shown that the study provides valuable long-term experimental data on sustainable land management strategies. But the unresolved questions are related to the limited representativeness of local case studies and the absence of cross-country comparisons, which reduces the generalizability of the results. The reasons of it can be the resource-intensive nature of large-scale experimental trials and the difficulty of ensuring comparability across heterogeneous agroecological zones. An option to overcome these limitations can be the integration of field experiments with broader regional datasets and modeling tools.

The paper [20] examines the impact of climate change on food security in the five Central Asian countries and shows a nonlinear relationship between average temperature, precipitation and food security, as well as significant negative effects of extreme climatic events. But the unresolved questions are related to the lack of crop-specific yield predictors and the predominantly macroeconomic perspective. The reasons of it can be the principal difficulty of combining detailed agroecological variables with macro-level statistical indicators in a single analytical framework. An option to overcome these challenges can be the use of integrated econometric and systems modeling approaches that link climatic, biophysical and socio-economic factors.

This is the approach used in [18–20], however the existing results, while valuable, remain somewhat fragmented and do not yet provide a comprehensive framework for explaining the complex interactions among climatic, biophysical and socio-economic factors. These limitations highlight the need for more integrated methodological approaches that can capture the multifactorial nature of agricultural productivity. All this allows to argue that it is appropriate to conduct a study devoted to developing and testing comprehensive models for assessing the combined influence of climatic and biophysical determinants on wheat yield in Central Asia.

This approach was further advanced in [21, 22], where ensemble methods such as random forest and support vector machine demonstrated robustness to noise and relatively good accuracy with small datasets. At the same time, their ability to capture highly nonlinear dependencies remained limited. More recent studies [23–25] show that neural networks, including CNN and LSTM, provide improved forecast accuracy when applied to large and diverse datasets. For instance, [23] developed a multi-branch deep learning framework (DeepAgroNet) for Pakistan that integrated satellite, meteorological, and soil data, demonstrating strong predictive capacity, although the authors stressed that long time series and detailed spatial inputs are essential for model stability. In [24], EfficientNet combined with attention mechanisms achieved very high accuracy in identifying wheat diseases, yet the approach was focused mainly on plant health classification rather than direct yield forecasting, which limits its transferability. Study [25] proposed a CNN-LSTM model with Copula theory to quantify drought thresholds for wheat yield in China, successfully capturing nonlinear interactions between climate variables and yields, but its application was constrained to one province and may not generalize across regions. These examples confirm the potential of deep learning for agriculture, but they also highlight methodological fragmentation and the dependence of model effectiveness on the quality and availability of regional data. Our study addresses this gap by testing machine learning models on integrated climate and biophysical predictors within the Central Asian context.

All this allows to argue that it is appropriate to conduct a study devoted to filling the identified research gap. Despite significant progress in international literature, the Republic of Kazakhstan still suffers from insufficient studies devoted to crop yield modeling, especially in the steppe regions of Central Asia, where agriculture is exposed to high levels of climate risks. This is especially relevant for the Kostanay region, the largest grain-producing region of the Republic, where wheat yields are characterized by high interannual and spatial variability. The insufficient study of this region in the context of integrating satellite and climate data for crop forecasting creates a scientific and practical niche filled with this study.

That being said, the analysis of previous studies shows that despite notable progress, existing research remains fragmented and faces several unresolved challenges: limited ability of traditional models to capture nonlinear and multidimensional relationships; insufficient integration of climatic, biophysical, and agronomic variables; methodological constraints linked to data availability and quality; and lack of comprehensive frameworks capable of ensuring reliable forecasting under conditions of high climatic variability. These limitations define the general unresolved problem, which is the need for developing more robust and integrated approaches to crop yield forecasting that can adequately reflect the complexity of interactions among multiple determining factors.

3. The aim and objectives of the study

The aim of this study is to conduct a comparative assessment of machine learning algorithms for forecasting wheat yields using integrated climatic indicators and satellite vegetation indices. By reaching this aim let's seek to improve the accuracy of forecasts and create scientifically sound prerequisites for the implementation of early warning systems and support for management decisions in the agricultural sector.

To achieve this aim, the following objectives were accomplished:

– to collect and systematize data on wheat yields, climatic indicators, and satellite vegetation indices (2001–2023) and to perform descriptive and correlation analysis aimed at identifying the key factors determining yield variations;

– to build and compare forecasting models (random forest, support vector machine, convolutional neural network), evaluate their accuracy (R^2, RMSE, MAE), analyze predictor significance, and derive practical conclusions for early warning and decision-support systems in agriculture.

4. Materials and methods of research

4. 1. Object and hypothesis of the study

The object of the study is wheat yield forecasting using the integration of climatic indicators, satellite-derived vegetation indices, and machine learning algorithms.

The main hypothesis of the study is that the combination of climatic parameters (precipitation, temperature) with satellite vegetation indices (NDVI, EVI) within machine learning models (random forest, support vector machine, convolutional neural network) significantly improves forecast accuracy compared to traditional statistical approaches.

Assumptions made in the study include the representativeness of the selected time frame (2001–2023) for reflecting long-term yield dynamics, the reliability of MODIS and ERA5 datasets for capturing vegetation and climatic variability, and the adequacy of district-level statistics as proxies for local agricultural conditions.

Simplifications adopted in the study are the use of aggregated climatic indicators for the growing season (April–July) instead of daily or decadal data, the absence of crop-specific masks for vegetation indices (calculations were based on district-level land cover), and the limitation to three machine learning models without testing additional algorithms such as recurrent neural networks.

4. 2. Data sources and preprocessing

The study used panel data for the period between 2001 and 2023 that included 345 observations (23 years across 15 districts of the Kostanay region). Such structure allowed to simultaneously consider both temporal and spatial differences in wheat yield and factors determining it. The Kostanay region has been chosen as an object of analysis due to its key role in the agricultural sector of the Republic of Kazakhstan as the region provides up to a third of national wheat production. Additionally, it is characterized by high climatic variability consisting in alternation of dry and relatively wet years, as well as sharp temperature changes rendering the region representative for the study of factors affecting crops.

To ensure comparability of information, all indicators have been aggregated by the growing season (April through July): average total values have been calculated for vegetation indices and temperature, and precipitation respectively. The data were checked for omissions and brought to a unified format. As a result, a coherent array of 345 observations has been formed used to build and evaluate machine learning models.

4. 3. Variables used in the study

Table 1 lists variables used in the study.

Table 1

Description of variables used in the study

Indicators	Designation	Units of measure	Sources
Normalized vegetation index	NDVI	Index (−1...+1)	MODIS (NASA LP DAAC, MOD13Q1, 250 m, 16-day composites)
Improved vegetation index	EVI	Index (−1...+1)	MODIS (NASA LP DAAC, MYD13A1, 500 m, 16-day composites)
Precipitation total (April–July)	PRECIP	Mm	ERA5 (ECMWF, 0.25° × 0.25°, daily)
Average temperature (April–July)	TEMP	°C	ERA5 (ECMWF, 0.25° × 0.25°, daily)
Wheat sown area	AREA	Hectare	National Statistics Bureau of the Republic of Kazakhstan
Wheat yield	YIELD	Dt/ha	National Statistics Bureau of the Republic of Kazakhstan

Satellite indices NDVI and EVI have been calculated based on MODIS (NASA LP DAAC) data. Climate characteristics (precipitation and average temperature for April through July) have been obtained from the ERA5 database of the European Centre for Medium-Range Weather Forecasts (ECMWF). Data on sown areas and wheat yields by Kostanay region districts have been provided by the National Statistics Bureau of the Republic of Kazakhstan.

Satellite vegetation indices were calculated using standard formulas:

$$NDVI = \frac{NIR - RED}{NIR + RED}, \tag{1}$$

$$EVI = C + \frac{NIR - RED}{NIR + C_1 \times RED - C_2 \times BLUE + L}, \tag{2}$$

where *NIR* is near-infrared reflectance;

RED is red reflectance;

BLUE is blue reflectance;

$C = 2.5$, $C_1 = 6$, $C_2 = 7.5$, $L = 1$.

Formulas (1) and (2) are given according to [26, 27].

4. 4. Machine learning models

The following three machine learning algorithms have been tested for crop forecasting:

1) random forest (RF) – an ensemble of 500 decision trees, constructed with bootstrapping and random feature selection. The maximum depth was restricted to prevent overfitting, with the Gini criterion applied for splitting;

2) support vector machine (SVM) – regression version (SVR) with radial basis function kernel. The regularization parameter was set to $C = 1.0$, and the kernel coefficient γ was optimized via cross-validation;

3) convolutional neural network (CNN) – architecture included two convolutional layers (32 and 64 filters, kernel size 3 × 3), max-pooling, one fully connected dense layer (128 neurons, ReLU activation), and a linear output layer for regression. The network was trained using the Adam optimizer, batch size 32, for 100 epochs.

RF and SVM were selected due to their proven robustness on small datasets and ability to capture nonlinear dependencies. CNN was included experimentally to evaluate deep learning potential for yield forecasting, even under data limitations. Recognizing that temporal dependencies are essential in agricultural time series, it is possible to note that recurrent neural networks (LSTM, GRU) represent a promising direction for future research.

4. 5. Model evaluation metrics

The following three metrics have been used to quantify model accuracy: the coefficient of determination (R^2), the root mean square error (RMSE), and the mean absolute error (MAE).

Coefficient of determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2}. \tag{3}$$

Root mean square error (*RMSE*)

$$RMSE = \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}. \tag{4}$$

Mean absolute error (*MAE*)

$$MAE = \frac{1}{n} \sum_{i-1}^{n} \left| \left(y_i - \hat{y}_i\right)^2 \right|, \tag{5}$$

where y_i is the actual yield;

$\hat{y}_i$ is the value predicted by the model;

$\bar{y}_i$ is the average yield;

$n = 345$ is the number of observations.

Using these three metrics allowed to comprehensively assess both the degree of correspondence between actual and predicted values, and the average values of forecast errors.

5. Results of assessing machine learning models for wheat yield prediction

5. 1. Descriptive statistics and correlation analysis

Assessing began with descriptive statistical calculations and correlation analysis of the main factors influencing wheat yield.

The descriptive statistics presented in Table 2 give a general idea of distribution of the main variables used to forecast wheat yield.

Table 2

Descriptive statistics

Variables	Mean	Std. dev.	Min	Max
NDVI	0.434	0.123	0.168	0.674
EVI	0.465	0.399	0.120	1.832
PRECIP	4.801	2.004	0.051	10.720
TEMP	15.330	2.356	6.034	21.323
AREA	226524.8	102229.4	20200.0	604200.0
YIELD	10.6	3.847	2.545	22.9

Table 2 shows that both normalized vegetation index (NDVI) and enhanced vegetation index (EVI) show significant variability with average values of 0.434 and 0.465, respectively. NDVI ranges between 0.168 and 0.674, pointing to significant variations in vegetation cover from year to year. The range of EVI values (0.120–1.832) is wider, further reflecting differences in crop structure and photosynthetic activity intensity.

Climate indicators also show significant variability. The average precipitation for April through July is 4.8 mm, with a maximum of 10.72 mm, indicating interannual differences in moisture availability, a key factor for growing wheat in dry conditions. The average temperature during the same period is 15.3°C, with values varying between 6.0°C and 21.3°C, which includes years with both favorable and stressful temperature conditions.

The agronomic factor (the area under wheat) varies especially strongly from 20.2 thousand hectares to over 604 thousand hectares, with an average value of 226.5 thousand hectares. Such variability is associated with both climatic conditions and management decisions that directly affect production potential.

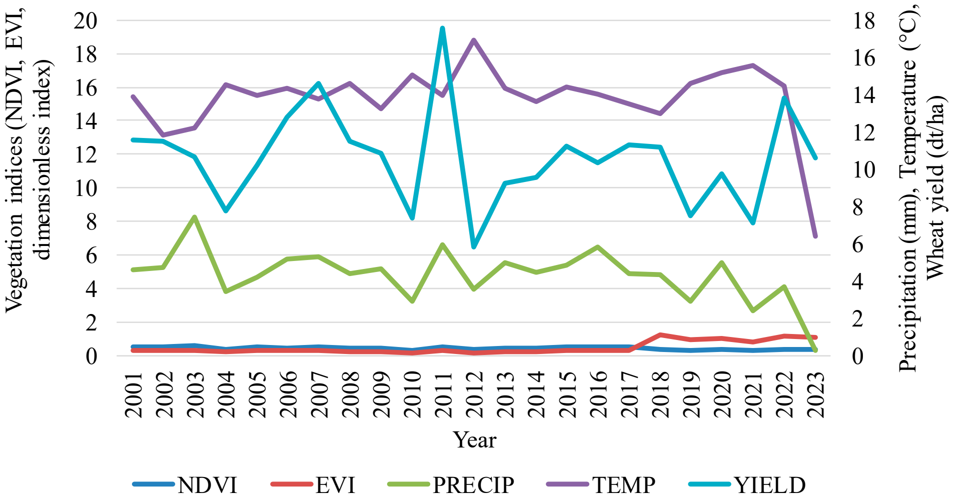


Fig. 1. Trends of wheat yield and main predictors (2001–2023): NDVI and EVI are shown on the left y-axis (dimensionless indices), whereas precipitation (mm), temperature (°C), and wheat yield (dt/ha) are presented on the right y-axis

Wheat yield demonstrates high interannual variability as well. Its average is 10.6 dt/ha, the minimum is 2.5 dt/ha, and the maximum is 22.9 dt/ha, which confirms productivity's sensitivity to a combination of natural and technological factors.

The correlation matrix in Table 3 reveals interconnections between predictors and their relationship with wheat yield.

Table 3

Correlation matrix

Variables	NDVI	EVI	PRECIP	TEMP	AREA	YIELD
NDVI	1.0	–0.021	0.749	–0.441	0.090	0.713
EVI	–0.021	1.0	–0.137	–0.349	–0.001	0.216
PRECIP	0.749	–0.137	1.0	–0.041	0.073	0.517
TEMP	–0.441	–0.349	–0.041	1.0	–0.080	–0.348
AREA	0.09	–0.001	0.073	–0.08	1.0	0.168
YIELD	0.713	0.216	0.517	–0.348	0.168	1.0

The strongest positive correlation is observed between Normalized Vegetation Index and crop yield ($r = 0.713$). This confirms the key role vegetation indices play in reflecting dynamics of crop growth and biomass accumulation. A positive correlation is also characteristic of precipitation ($r = 0.517$), which emphasizes the critical importance of moisture supply during the growing season. Concurrently, temperature has a negative relationship with crop yield ($r = -0.348$), which is consistent with the negative impact of high temperatures on wheat development, especially during sensitive growth phases.

The EVI shows only a weak positive correlation with crop yield ($r = 0.216$) indicating its less information content in the regional context compared to NDVI. There is virtually no correlation between yield and sown area ($r = 0.168$) attesting to the fact that expansion or reduction of sown areas do not lead to proportional changes in yield per hectare, and that climatic and biophysical factors play a decisive role.

The long-term dynamics of wheat yield and key predictors presented in Fig. 1 vividly confirm the findings.

Periods of low yield generally coincide with a decrease in the Normalized Vegetation Index values and unfavorable climatic conditions (precipitation deficit, elevated temperatures). In contrast, years with high NDVI and sufficient precipitation correspond to yield peaks, indicating a robust explanatory power of vegetation indices and climatic parameters for yield variations.

Accordingly, the analysis of descriptive statistics and correlations forms a solid empirical basis for further yield modeling. The normalized vegetation index and precipitation are the most significant determinants while temperature factor indicates the risk of thermal stress as a limiting element. The results confirm the need to integrate satellite and climate data into machine learning models as discussed in the next subsection.

5. 2. Model construction, performance and comparison

With identified patterns and interconnections between vegetation indices, climatic factors, and wheat yield, the next step was to apply machine learning methods to build predictive models. The main focus was on comparing the quality of different algorithm predictions and identifying the most significant factors determining yield variation.

Table 4 presents a comparative assessment of machine learning models' performance.

Table 4

Model performance comparison (CNN, RF, SVM)

Model	R^2	RMSE	MAE
CNN	–0.91	5.30	4.13
RF	0.85	1.50	1.03
SVM	0.85	1.47	0.54

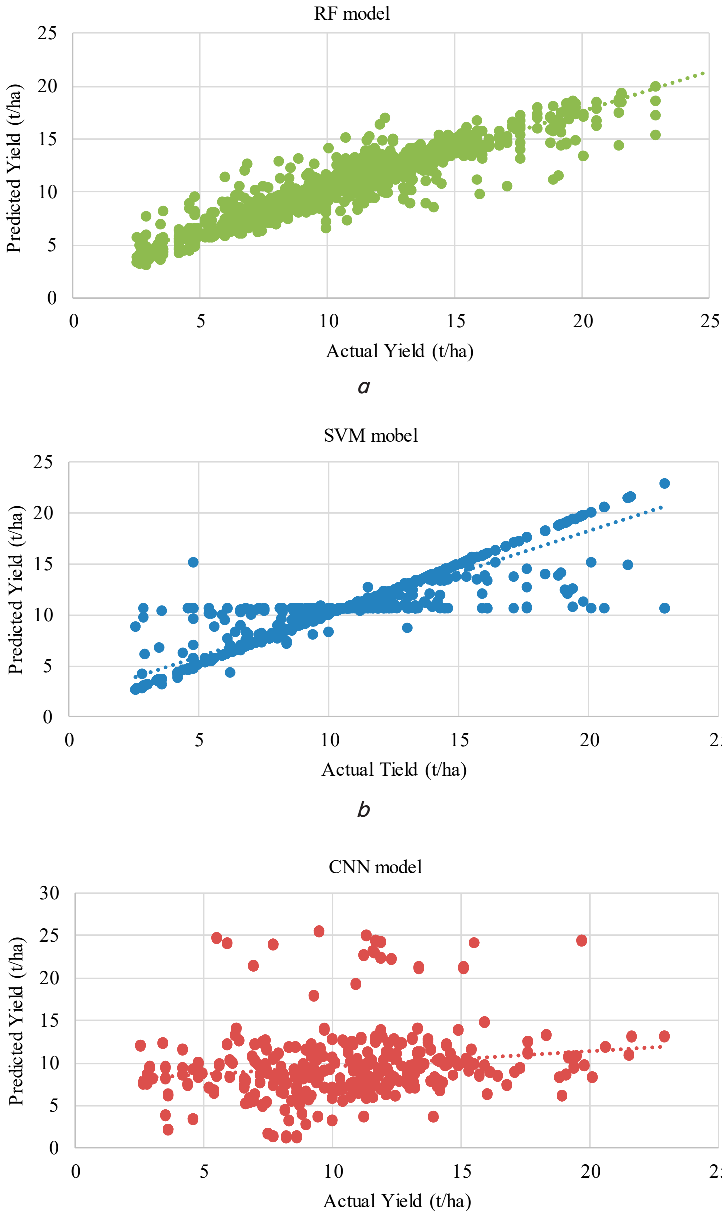
Random forest (RF) and support vector machine (SVM) performed the best with their determination coefficient of $R^2 = 0.85$. The error rates, on the other hand, were low: 1.50 and 1.47 RMSE, and 1.03 and 0.54 MAE, respectively.

The results indicate high forecast accuracy and good agreement between predicted and actual yield values. In contrast, the convolutional neural network (CNN) has shown a negative R^2 of -0.91 pointing towards the model's inability to adequately describe yield variations in this data set. High errors (RMSE = 5.30; MAE = 4.13) confirm significant deviations in CNN forecasts.

Fig. 2 below clearly demonstrates the differences in model accuracy through scatter plots of actual and predicted yield values.

For RF and SVM, points cluster along the equality line, which means closeness of predictions to the observed values. In CNN's case, significant scattering and systematic deviations are observed, confirming this architecture's predictive ability to be weak under the conditions of the used data.

Additional confirmation of the quality of models is provided by analyzing residuals, as presented in Fig. 3 below.



a

b

c

Fig. 2. Comparison of actual and predicted wheat yield: *a* — random forest model; *b* — support vector machine model; *c* — convolutional neural network model

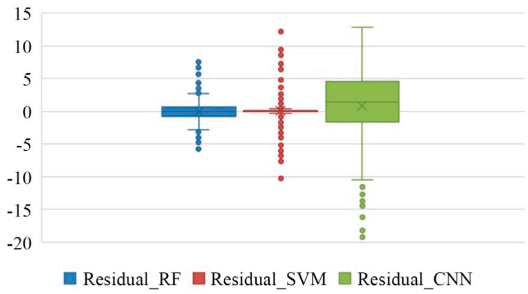


Fig. 3. Residual analysis (error distribution) for convolutional neural network, random forest, and support vector machine models

For RF and SVM, residuals are distributed uniformly around the zero line, their spread is limited and does not demonstrate pronounced systematic biases. At the same time, CNN is characterized by significant positive errors pointing towards yield overestimation and high model instability. Accordingly, RF and SVM can be considered as stable tools for crop forecasting while CNN requires further adaptation and fine tuning.

The results of assessing predictor significance using the RF method (Table 5) confirm the key role satellite indices and climatic factors play.

NDVI makes the largest contribution to predictions (0.345) reflecting the state of vegetation. The second most important factor is temperature (0.211) indicating yield's high sensitivity to thermal conditions. Sown area plays a significant role as well (0.160), which, in combination with biophysical parameters, forms production potential. The EVI (0.145) and precipitation (0.139) have a slightly smaller contribution but their influence is also statistically significant.

Comparison of actual and predicted yield values by year given in Table 6 allows for a more detailed assessment of the model stability.

RF and SVM demonstrate a high degree of consistency with the observed data in all years of the study, including those with unfavorable weather conditions (e.g., 2019 and 2021). CNN systematically deviates from the actual values, especially during years of extreme climatic conditions, once again confirming this model's low adequacy in the region under study.

Taken together, the analysis results show RF and SVM methods to have the greatest potential for forecasting wheat yield based on integration of climate and satellite data. NDVI is the main indicator of crop condition while climate parameters (temperature and precipitation) provide additional explanatory power of the models, further emphasizing importance of an integrated approach combining remote sensing and meteorological data for constructing reliable yield forecasts.

Table 5

Feature importance (random forest)

Rank	Variable	Type	Importance
1	NDVI	Satellite index	0.345
2	TEMP	Climate	0.211
3	AREA	Agrotechnical	0.160
4	EVI	Satellite index	0.145
5	PRECIP	Climate	0.139

Table 6

Comparison of actual and predicted wheat yield using CNN, RF, and SVM Models

Year	Actual	CNN	RF	SVM
2014	9.56	10.48	9.70	9.52
2015	11.25	9.95	11.41	10.95
2016	10.32	10.56	11.05	10.39
2017	11.33	10.26	11.36	11.20
2018	11.20	9.55	11.03	11.00
2019	7.49	9.01	8.02	7.87
2020	9.76	9.47	10.12	10.16
2021	7.10	9.97	7.43	7.16
2022	13.80	9.61	12.49	13.42
2023	10.62	10.22	10.75	10.38

In summary, integration of satellite indices and climate data in the RF and SVM models provides high forecast accuracy and confirms their practical applicability for the development of early warning systems and decision support in the agricultural sector.

6. Discussion of the results of assessing machine learning approaches for wheat yield prediction

The results of this study demonstrate the high efficiency of machine learning methods, particularly random forest and support vector machine, for forecasting wheat yields based on the integration of climate and satellite data. As shown in (Table 4), both models achieved an accuracy of $R^2 = 0.85$ with low RMSE and MAE values, confirming their robustness under conditions of climatic variability. This finding is further illustrated in (Fig. 2), where the scatter plots of actual and predicted yields for random forest and support vector machine cluster closely around the equality line, indicating stable prediction capacity. In contrast, the convolutional neural network demonstrated poor performance with a negative R^2 (−0.91) and high error values (RMSE = 5.30; MAE = 4.13), reflecting the model's inability to capture yield variability with the available dataset. This outcome is consistent with previous studies [25], which emphasize that deep learning approaches require large-scale and diverse datasets for stable performance.

The analysis of descriptive statistics and correlation patterns (Table 2, 3) confirmed that vegetation indices, particularly the normalized difference vegetation index (NDVI), play a decisive role in explaining yield variation. NDVI demonstrated the strongest correlation with yield ($r = 0.713$), followed by precipitation ($r = 0.517$), while temperature showed a negative correlation ($r = -0.348$), indicating the risk of thermal stress during the growing season. These findings align with the literature, where vegetation indices have been widely acknowledged as reliable predictors of crop growth [6, 7, 17] and where excessive temperatures were shown to negatively affect wheat development in Central Asia [20]. The long-term dynamics (Fig. 1) further support these results, since years with low NDVI and precipitation consistently coincided with yield decline.

The predictor importance analysis using random forest (Table 5) once again confirmed NDVI as the most influential variable (importance = 0.345), followed by temperature (0.211), sown area (0.160), and EVI (0.145). Precipitation, while somewhat less influential (0.139), remained statistically significant. Similar patterns were reported in other regional studies [18, 19], which also emphasized the decisive role of climatic and vegetation variables. In addition, our results provide empirical confirmation for Republic of Kazakhstan, a region insufficiently studied in the literature, thereby closing the gap highlighted in Section 2.

Comparison of actual and predicted yields by year (Table 6) showed that random forest and support vector machine successfully replicated yield dynamics even in unfavorable years such as 2019 and 2021, when productivity declined due to adverse weather. This result is consistent with international studies [21, 23] reporting the robustness of ensemble and support vector methods under variable climate conditions. By contrast, the convolutional neural network systematically deviated from observed values, especially in extreme years, which highlights the limitations of deep learning approaches when applied to relatively small datasets, as noted in [25].

The developed models can be directly applied in:

– governmental yield forecasting services (e.g., Ministry of Agriculture of Republic of Kazakhstan, regional agricultural departments) for early warnings and planning of grain balances;

– agricultural companies and large farms, which can use forecasts for optimizing the allocation of seeds, fertilizers, and water resources;

– international organizations such as FAO, which monitor food security in regions exposed to climate risks. Integration of these models into decision-support and early warning systems would reduce uncertainty, allow timely interventions, and strengthen resilience of agricultural production.

Originality of this work lies in the first systematic comparison of random forest, support vector machine, and convolutional neural network for wheat yield forecasting in Republic of Kazakhstan using integrated climatic and satellite data. Unlike previous international works [28–30], which often focused either on global datasets or on local case studies outside Central Asia, this study addresses the lack of regional research and demonstrates that ensemble and support vector approaches are particularly suitable for conditions of high climatic variability.

In summary, the study confirms that random forest and support vector machine represent reliable tools for forecasting wheat yields in Republic of Kazakhstan under high climatic variability. The integration of vegetation indices and climate indicators significantly improves forecast accuracy, closes existing research gaps on Central Asia, and provides a practical basis for early warning and decision-support systems aimed at enhancing food security.

At the same time, several limitations should be acknowledged. First, the study focused only on the Kostanay region, which may constrain generalizability to other agro-ecological zones. Second, MODIS data with a resolution of 250–500 m were used, which may be too coarse to capture field-level heterogeneity. Third, crop masks were not applied, meaning vegetation indices were calculated across the entire land cover rather than being restricted to wheat fields, potentially introducing noise. Fourth, the CNN approach showed instability under limited data conditions. Addressing these shortcomings – for example, by applying Sentinel-2 data with 10–20 m resolution, integrating crop classification maps, expanding the analysis to other regions of Republic of Kazakhstan and Central Asia, and testing recurrent neural networks (LSTM, GRU) better suited for time-series data – constitutes a promising direction for further research.

7. Conclusions

1. The analysis of descriptive statistics and correlations has confirmed the key role climatic and satellite factors play in forming wheat yield. The NDVI has shown the highest positive correlation with yield ($r = 0.713$), which represents its high information content in describing the condition of crops. Precipitation has a significant impact ($r = 0.517$) while temperature has a negative relationship ($r = -0.348$) indicating the risk of thermal stress during the growing season. These results are consistent with the world literature [6, 7, 20] and confirm the need for an integrated accounting of climatic and biophysical factors. Compared to previous studies in Central Asia [18, 19], our findings strengthen the evidence of vegetation indices as the dominant predictors of yield but provide new empirical results for Republic of Kazakhstan, a region that has been poorly represented in the literature.

2. A comparative analysis of machine learning models has shown the random forest and support vector machine algorithms to be the most effective for forecasting wheat yield. Their accuracy was $R^2 = 0.85$ with low RMSE (1.50 and 1.47) and MAE (1.03 and 0.54, respectively), which is comparable to or higher than similar results reported in other regional studies [21, 23]. The convolutional neural network (CNN) has demonstrated a negative coefficient of determination (–0.91) and high errors (RMSE = 5.30; MAE = 4.13). This can be explained by the relatively small dataset and the inability of CNN to capture temporal dependencies inherent in agricultural time series without larger and more detailed training data [25]. By contrast, the high accuracies of random forest and support vector machine are due to their robustness with limited samples and their ability to capture nonlinear relationships, confirming their advantages under conditions of high climate variability. The results of predictor significance analysis showed NDVI's leading role (0.345), followed by temperature (0.211), sown area (0.160), EVI (0.145) and precipitation (0.139), which confirms the need for a multi-component modeling approach. Practical significance of the study lies in the possibility of applying the developed models in early warning and decision support systems for the agricultural sector, improving agrotechnical planning, optimizing resource allocation and reducing food security risks in arid regions.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this study, whether financial, personal, authorship or otherwise, that could affect the study and its results presented in this paper.

Financing

This study has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP22686171).

Data availability

Data will be made available on reasonable request.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

References

1. Molotoks, A., Smith, P., Dawson, T. P. (2020). Impacts of land use, population, and climate change on global food security. Food and Energy Security, 10 (1). https://doi.org/10.1002/fes3.261
2. Yu, W., Yue, Y., Wang, F. (2022). The spatial-temporal coupling pattern of grain yield and fertilization in the North China plain. Agricultural Systems, 196, 103330. https://doi.org/10.1016/j.agsy.2021.103330
3. Lesk, C., Anderson, W., Rigden, A., Coast, O., Jägermeyr, J., McDermid, S. et al. (2022). Compound heat and moisture extreme impacts on global crop yields under climate change. Nature Reviews Earth & Environment, 3 (12), 872–889. https://doi.org/10.1038/s43017-022-00368-8
4. Malhi, G. S., Kaur, M., Kaushik, P. (2021). Impact of Climate Change on Agriculture and Its Mitigation Strategies: A Review. Sustainability, 13 (3), 1318. https://doi.org/10.3390/su13031318
5. Tamayo-Vera, D., Wang, X., Mesbah, M. (2024). A Review of Machine Learning Techniques in Agroclimatic Studies. Agriculture, 14 (3), 481. https://doi.org/10.3390/agriculture14030481
6. Shammi, S. A., Meng, Q. (2021). Use time series NDVI and EVI to develop dynamic crop growth metrics for yield modeling. Ecological Indicators, 121, 107124. https://doi.org/10.1016/j.ecolind.2020.107124
7. Swoish, M., Da Cunha Leme Filho, J. F., Reiter, M. S., Campbell, J. B., Thomason, W. E. (2022). Comparing satellites and vegetation indices for cover crop biomass estimation. Computers and Electronics in Agriculture, 196, 106900. https://doi.org/10.1016/j.compag.2022.106900
8. Fadl, M. E., AbdelRahman, M. A. E., El-Desoky, A. I., Sayed, Y. A. (2024). Assessing soil productivity potential in arid region using remote sensing vegetation indices. Journal of Arid Environments, 222, 105166. https://doi.org/10.1016/j.jaridenv.2024.105166
9. Jabed, Md. A., Azmi Murad, M. A. (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. Heliyon, 10 (24), e40836. https://doi.org/10.1016/j.heliyon.2024.e40836
10. Elbasi, E., Zaki, C., Topcu, A. E., Abdelbaki, W., Zreikat, A. I., Cina, E. et al. (2023). Crop Prediction Model Using Machine Learning Algorithms. Applied Sciences, 13 (16), 9288. https://doi.org/10.3390/app13169288
11. Asamoah, E., Heuvelink, G. B. M., Chairi, I., Bindraban, P. S., Logah, V. (2024). Random forest machine learning for maize yield and agronomic efficiency prediction in Ghana. Heliyon, 10 (17), e37065. https://doi.org/10.1016/j.heliyon.2024.e37065

12. Guo, Y. (2023). Integrating genetic algorithm with ARIMA and reinforced random forest models to improve agriculture economy and yield forecasting. Soft Computing, 28 (2), 1685–1706. https://doi.org/10.1007/s00500-023-09516-8
13. Ingole, V. S., Kshirsagar, U. A., Singh, V., Yadav, M. V., Krishna, B., Kumar, R. (2024). A Hybrid Model for Soybean Yield Prediction Integrating Convolutional Neural Networks, Recurrent Neural Networks, and Graph Convolutional Networks. Computation, 13 (1), 4. https://doi.org/10.3390/computation13010004
14. Uluocak, I., Bilgili, M. (2023). Daily air temperature forecasting using LSTM-CNN and GRU-CNN models. Acta Geophysica, 72 (3), 2107–2126. https://doi.org/10.1007/s11600-023-01241-y
15. Abdel-salam, M., Kumar, N., Mahajan, S. (2024). A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning. Neural Computing and Applications, 36 (33), 20723–20750. https://doi.org/10.1007/s00521-024-10226-x
16. Celis, J., Xiao, X., Wagle, P., Adler, P. R., White, P. (2024). A Review of Yield Forecasting Techniques and Their Impact on Sustainable Agriculture. Transformation Towards Circular Food Systems, 139–168. https://doi.org/10.1007/978-3-031-63793-3_8
17. Ashfaq, M., Khan, I., Alzahrani, A., Tariq, M. U., Khan, H., Ghani, A. (2024). Accurate Wheat Yield Prediction Using Machine Learning and Climate-NDVI Data Fusion. IEEE Access, 12, 40947–40961. https://doi.org/10.1109/access.2024.3376735
18. Jiang, P., Yuan, Y., Li, Q. (2024). Advanced precipitation enhances vegetation primary productivity in Central Asia. Ecological Indicators, 166, 112276. https://doi.org/10.1016/j.ecolind.2024.112276
19. Nurbekov, A., Kosimov, M., Islamov, S., Khaitov, B., Qodirova, D., Yuldasheva, Z. et al. (2024). No-till, crop residue management and winter wheat-based crop rotation strategies under rainfed environment. Frontiers in Agronomy, 6. https://doi.org/10.3389/fagro.2024.1453976
20. Su, F., Liu, Y., Chen, L., Orozbaev, R., Tan, L. (2023). Impact of climate change on food security in the Central Asian countries. Science China Earth Sciences, 67 (1), 268–280. https://doi.org/10.1007/s11430-022-1198-4
21. Sánchez, J. C. M., Mesa, H. G. A., Espinosa, A. T., Castilla, S. R., Lamont, F. G. (2025). Improving wheat yield prediction through variable selection using Support Vector Regression, Random Forest, and Extreme Gradient Boosting. Smart Agricultural Technology, 10, 100791. https://doi.org/10.1016/j.atech.2025.100791
22. Sonmez, M. E., Sabanci, K., Aydin, N. (2024). Convolutional neural network-support vector machine-based approach for identification of wheat hybrids. European Food Research and Technology, 250 (5), 1353–1362. https://doi.org/10.1007/s00217-024-04473-4
23. Ashfaq, M., Khan, I., Shah, D., Ali, S., Tahir, M. (2025). Predicting wheat yield using deep learning and multi-source environmental data. Scientific Reports, 15 (1). https://doi.org/10.1038/s41598-025-11780-7
24. Nigam, S., Jain, R., Singh, V. K., Marwaha, S., Arora, A., Jain, S. (2024). EfficientNet architecture and attention mechanism-based wheat disease identification model. Procedia Computer Science, 235, 383–393. https://doi.org/10.1016/j.procs.2024.04.038
25. Ma, J., Zhao, Y., Cui, B., Liu, L., Ding, Y., Chen, Y., Zhang, X. (2025). Prediction of Drought Thresholds Triggering Winter Wheat Yield Losses in the Future Based on the CNN-LSTM Model and Copula Theory: A Case Study of Henan Province. Agronomy, 15 (4), 954. https://doi.org/10.3390/agronomy15040954
26. Rouse, J. W., Haas, R. H., Schell, J. A., Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. NASA SP-351, 309–317. Available at: https://ntrs.nasa.gov/citations/19740022614
27. Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment, 83 (1-2), 195–213. https://doi.org/10.1016/s0034-4257(02)00096-2
28. Kussainov, T. A., Maitah, M., Kurmanov, N. A., Hájek, P., Tolysbaev, B. S., Baidakov, A. K. (2015). Economic Analysis of the Impact of Changing Production Conditions on Wheat Productivity Level. Review of European Studies, 7 (11). https://doi.org/10.5539/res.v7n11p125
29. Kurmanov, N., Bakirbekova, A., Adiyetova, E., Satbayeva, A., Rakhimbekova, A., Nabiyeva, M. (2025). ICTs' Impact on Energy Consumption and Economic Growth in the Countries of Central Asia: An Empirical Analysis. International Journal of Energy Economics and Policy, 15 (3), 8–16. https://doi.org/10.32479/ijeep.18779
30. Kurmanov, N., Kabdullina, G., Baidakov, A., Kabdolla, A. (2025). Renewable Energy, Green Economic Growth and Food Security in Central Asian Countries: An Empirical Analysis. International Journal of Energy Economics and Policy, 15 (2), 1–8. https://doi.org/10.32479/ijeep.17922